Theor Appl Genet (2009) 119:555–565
DOI 10.1007/s00122-009-1065-7

ORIGINAL PAPER

Impact of depth of pedigree and inclusion of historical data on the estimation of additive variance and breeding values in a sugarcane breeding program

Felicity Claire Atkin · Mark J. Dieters · Joanne K. Stringer

Received: 29 September 2008 / Accepted: 9 May 2009 / Published online: 14 July 2009

Abstract Sugarcane breeders in Australia combine data across four selection programs to obtain estimates of breeding value for parents. When these data are combined with full pedigree information back to founding parents, computing limitations mean it is not possible to obtain information on all parents. Family data from one sugarcane selection program were analysed using two different genetic models to investigate how different depths of pedigree and amount of data affect the reliability of estimating breeding value of sugarcane parents. These were the parental and animal models. Additive variance components and breeding values estimated from different amounts of information were compared for both models. The accuracy of estimating additive variance components and breeding values improved as more pedigree information and historical data were included in analyses. However, adding years of data had a much larger effect on the estimation of variance components of the population, and breeding values of the parents. To accurately estimate breeding values of all sugarcane parents, a minimum of three generations of pedigree and 5 years of historical data were required, while more information (four generations of pedigree and 7 years of historical data) was required when identifying top parents to be selected for future cross pollination.

Communicated by F. van Eeuwijk.

F. C. Atkin · M. J. Dieters
School of Land, Crop and Food Sciences,
The University of Queensland,
St Lucia, QLD 4072, Australia
e-mail: m.dieters@uq.edu.au

F. C. Atkin (✉) · J. K. Stringer
BSES Limited, 50 Meiers Rd, PO Box 86,
Indooroopilly, QLD 4068, Australia
e-mail: fatkin@bses.org.au

J. K. Stringer
e-mail: jstringer@bses.org.au

Introduction

Parental selection is one of the most crucial steps to improve genetic gain in any breeding program. Breeding value (BV) is commonly used to measure the potential of an individual as a parent (White and Hodge 1989), and predicting it is one of the primary objectives of many breeding programs. The prediction of BV is affected directly by the estimation of variance components, especially the estimate of additive variance (Falconer and Mackay 1996). Thus, advances in genetic gain in a breeding program are functions of the accuracy of both genetic parameters and BVs.

Restricted Maximum Likelihood (REML) and Best Linear Unbiased Prediction (BLUP) methodology are regarded as one of the best tools to obtain reliable estimates of variance components and estimated BVs in animal and plant breeding where data are typically unbalanced (Meyer 1991). BLUP methodology is used routinely in animal breeding for estimating BV of individuals and, in recent years, has been more widely adopted in plant breeding (Davik and Honne 2005; Durel et al. 1998; Furlani et al. 2005; Oakey et al. 2006, 2007; Purba et al. 2001; Wei and Borralho 2000). Piepho et al. (2008) reviewed developments in the application of BLUP in plant breeding, including the use of pedigree information to exploit genetic correlations among relatives. They demonstrated that the use of BLUP, including pedigree information to estimate BV, has good predictive accuracy compared with other

procedures such as BLUP without pedigree information and the once-popular Additive Main Effect Multiplicative Interaction (AMMI). This was also demonstrated by Wei and Borralho (2000), Purba et al. (2001) and Furlani et al. (2005), where BLUP was superior to older methods of selection. This superiority is due to the ability of BLUP to use information from individuals by exploiting genetic correlations arising from the pedigree. The most common approach of including pedigree information involves the use of the numerator relationship matrix computed from the coefficient of coancestry (Henderson 1984).

Calculation of the coefficient of coancestry is based on several assumptions, including that: pedigree information of parents is detailed and accurate; the base populations of ancestors are unrelated; and effects of selection, mutation and genetic drift are negligible (Piepho et al. 2008). However, these assumptions do not hold for many animal and plant breeding programs. Often parents are genetically related, the knowledge of genetic relationships among parents is inaccurate, or the relationship among parents and families is unknown. Not including complete records back to the base population (Mehrabani-Yeganeh et al. 1999), errors in pedigree (Ericsson 1999; Long et al. 1990; Visscher et al. 2002), or ignoring pedigree information completely (Durel et al. 1998) can result in underestimated and biased estimates of additive variance and BVs, and hence, a reduction in potential genetic gain. However, the effects of different levels of pedigree structure on the estimation of BVs do not appear to have been investigated in plants.

Knowledge of the impact of pedigree structure on additive variance estimates and BV may be of interest to many plant breeders, as complete records are often not available, or inclusion of all records in models may become computationally demanding. Therefore, determination of the minimum level of pedigree information required for reliable genetic evaluation is of interest to many animal and plant breeders, and in particular to sugarcane breeders in Australia.

The BSES-CSIRO Plant Improvement program operates four regional selection programs to develop new sugarcane varieties for the Australian sugar industry. To assess the breeding potential of sugarcane parents, BSES intends to combine information from family trials from all four selection programs to more accurately predict BVs of parents in the breeding population. However, combining all available information from four selection programs, including pedigree information for over 3,200 parents, poses logistical issues due to computing limitations of estimating BVs of all parents in the pedigree. It is, therefore, important to consider what impact different amounts of data and pedigree information have on estimating BVs of sugarcane parents so that optimal methods can be chosen.

To investigate this, family data from one sugarcane selection program was used to examine how different depths of pedigree affect estimates of additive variance components and BVs for one trait using two different genetic models. Knowledge of the magnitude of the effect of different depths of pedigree on estimates of BV, and how the selection of top parents differs, will be used to optimise the methods used in the future evaluation of sugarcane parents. This is particularly valuable if computational limitations mean that simpler models or reduced data sets must be used to allow a combined analysis of all programs. Here, we investigate the effect of using different amounts of historical data and pedigree structure on estimation of additive genetic variance and BV evaluation in a sugarcane breeding program.

Materials and methods

Experimental details

We used sugarcane family trials (i.e. Progeny Assessment Trials, PATs) from the Southern BSES selection program, established as part of the BSES-CSIRO Plant Improvement program in Queensland, Australia. Data were available from 10 years of PATs, from 1996 to 2005 at BSES Bundaberg, Australia (latitude 24°45′S). A total of 241,520 seedlings from 2,318 families were grown in 20 trials over this 10-year period (Table 1). The number of families grown each year varied from 233 to 411 full-sib families (Table 2) and was a mixture of proven (i.e. progeny had performed well in the past) and experimental (or untested) combinations. Two trials were planted each year containing mostly the same families, and hence, female and male parents. However, due to limitations in land availability, and for ease of management, some trials were comprised two or more sub-trials (Table 1). These sub-trials were connected by proven families and/or common parents.

The number of families in common across years varied greatly, 0–27% (Table 2), as a result of sparse and unreliable flowering and poor pollen fertility in sugarcane breeding (Berding and Skinner 1987). Nevertheless, there was considerably greater connectivity across years among the parents (Table 2).

Trials were laid out as a rectangular array of rows and columns using a randomised complete-block design with two replicates of each family. For each trial, families were planted as seedlings in a single-furrow plot 12.6 m long with a furrow spacing of 1.5 m, except for trials planted in 2002 where plots were 11.44 m long. Each family plot contained on average 20 individual seedlings with an intra-row spacing of 0.6 m. There were no guard rows planted between the family plots.

Table 1 Number of trials (and sub-trials) planted over 10 years of Southern PATs; the harvest month, the number of families, and female and male parents of the families, for each trial

Year	Trial number (no. sub-trials per trial)	Month of harvest	Families	Female parents	Male parents
2005	1 (3)	June	336	176	162
	2	October	240	122	123
2004	3 (3)	July	284	147	145
	4 (2)	November	241	132	130
2003	5 (2)	June	275	133	143
	6	October	223	120	122
2002	7 (2)	June	254	128	125
	8	September	197	106	108
2001	9 (3)	June	345	187	167
	10 (2)	August	281	158	148
2000	11 (3)	June	411	162	174
	12 (2)	September	364	148	160
1999	13 (4)	July	352	167	162
	14	September	142	91	94
1998	15 (2)	July	271	129	113
	16	October	217	110	101
1997	17 (2)	July	242	110	104
	18 (2)	October	244	111	104
1996	19	August	233	117	132
	20	October	233	117	132
Total	20 (39)		2,318	708	644

Table 2 Concurrence matrices of sugarcane families and parents of families planted each year from 1996 to 2005

Numbers in the lower off-diagonal are the numbers of families in common between pairs of years; and numbers in italic type face in the upper off-diagonal are the numbers of parents in common between pairs of years. Numbers in bold type face are the number of families/parents in each year

		2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	Year
		281	*108*	*86*	*69*	*83*	*70*	*65*	*50*	*45*	*45*	2005
2005	**336**		***253***	*118*	*87*	*80*	*73*	*76*	*47*	*41*	*39*	2004
2004	23	**284**		***248***	*95*	*81*	*80*	*76*	*57*	*47*	*39*	2003
2003	17	46	**276**		***220***	*87*	*80*	*75*	*57*	*44*	*49*	2002
2002	9	29	33	**259**		***309***	*158*	*145*	*109*	*106*	*91*	2001
2001	5	15	28	23	**345**		***295***	*160*	*117*	*108*	*98*	2000
2000	4	4	25	22	93	**411**		***299***	*113*	*106*	*98*	1999
1999	2	6	16	11	63	111	**352**		***204***	*114*	*102*	1998
1998	0	2	6	5	32	45	50	**271**		***183***	*108*	1997
1997	2	1	2	1	29	45	51	52	**244**		***211***	1996
1996	1	0	2	2	14	14	20	23	28	**233**		
Year	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996		

After approximately 12 months of growth, each plot was assessed for commercial cane sugar (CCS) expressed as a percentage of recoverable sucrose on a fresh weight basis, and cane yield (tonnes cane per hectare, TCH). TCH was assessed on a plot-mean basis by weighing all 20 seedlings per plot as a family unit. CCS was also assessed on a plot-mean basis, but only eight randomly chosen seedlings were measured per family plot. Here, we consider only data on CCS, because interplot competition is known to substantially affect the estimation of TCH in sugarcane (Stringer and Cullis 2002), and a method to model competition when combining data across trials has not been fully developed.

Trial data were combined to form ten different data sets, each with different years of data represented. The first data set contained only the 2005 trials (i.e. the most recent trial data). The second data set contained 2 years of data, i.e. 2005 and 2004. Each subsequent data set included the next year of trials, so that the last data set contained data from all 10 years of trials planted from 1996 to 2005.

BSES also has access to reliable pedigree information (Wei et al. 2006) for most modern sugarcane parents back to their founding populations (Roach 1989) when original crosses were made in the 1890s (Bremer 1961). This information was used to collate pedigrees for all 2,318

families from the Southern BSES family trials. Six pedigree files were generated for each of the ten data sets, each containing a different depth of pedigree. The first pedigree contained only the parents of each family grown; labelled two (2) generations (i.e. of the seedlings and their parents). This depth of pedigree treats the parents independently (i.e. their ancestors are unknown). The second pedigree generated for each data set included an additional generation of parents (i.e. the grandparents of each family). This pedigree level was labelled three (3) generations. For each depth of pedigree, an additional generation of parents was added, until there were seven generations included in the pedigree (i.e. ten data sets × six depths of pedigree). However, with each additional generation added to the pedigree for each data set, the number of new parents added progressively decreased (Table 3). Inclusions of pedigree information beyond seven generations back to the founding parents only resulted in the addition of two new parents. Hence, a pedigree of seven additional generations was considered a complete pedigree.

Statistical model

Consider a field experiment consisting of n plots arranged in a rectangular array of r rows and c columns ($n = r \times c$). If the data are sorted as rows within columns, then the mixed linear model for $\mathbf{y}$ is:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_u\mathbf{u} + \mathbf{e} \quad (1)$$

where $\mathbf{b}^{(b\times 1)}$ is a vector of fixed effects and includes an overall mean for each site, as well as site-specific modelling terms with associated design matrix $\mathbf{X}^{(n\times b)}$. As described by Gilmour et al. (1997), site-specific modelling terms include large-scale variation that is usually aligned with the rows and columns of a field trial (e.g. linear row and/or column effects) and/or variation arising from experimental procedures or management practices that have a recurrent pattern (e.g. serpentine harvesting). The vector $\mathbf{g}^{(mp\times 1)}$ contains the random genotypic effects of m genetic entities in each of p trials with associated design matrix $\mathbf{Z}_g^{(n\times mp)}$. The vector $\mathbf{u}^{(d\times 1)}$ contains the random non-genetic effects for modelling extraneous variation due to experimental procedures and blocking design factors specific to each trial, or sub-trial (in cases where a trial comprises of two or more sub-trials), with the associated design matrix $\mathbf{Z}_u^{(n\times d)}$. The vector of random residuals $\mathbf{e}^{(n\times 1)}$ is modelled using an autoregressive process of order 1 (AR1) in the row and column direction as described in Gilmour et al. (1997). In this paper, the vector $\mathbf{g}$ is partitioned in two ways. The first is called the parental or sire model (Henderson 1984, Mrode 2005), where $\mathbf{g}$ is the random genotypic effects of unique parents, as a sugarcane parent can be used as either a male or a female, or both. In the second model, the animal model (Henderson 1984; Mrode 2005), $\mathbf{g}$ is partitioned into random individual family effects. The vector $\mathbf{g}$ for both models can be further partitioned into additive and non-additive genetic effects as per Costa e Silva et al. (2004) for the parental model and Oakey et al. (2007) for the animal model. Full details of the ASReml code for both models can be found in the Appendix.

Statistical analysis

Each sub-trial was standardised as follows:

$$y_{ij(adj.)} = \left(y_{ij} - \bar{y}_i\right)/s_i \quad (2)$$

where $y_{ij(adj.)}$ was the jth CCS observation from sub-trial i, y_{ij} the jth original CCS from sub-trial i, $\bar{y}_i$ the CCS mean at sub-trial i and s_i was the estimated phenotypic standard deviation at sub-trial i. Therefore, the variance components from each analysis could be compared directly to assess the impact of depth of pedigree on genetic evaluation and the estimation of parental BVs using BLUP.

For each of the ten data sets, adjusted data for CCS were analysed using the animal and parental models. The animal model was similar to that used in Oakey et al. (2006, 2007). Here, each family and its parents were included in the

Table 3 Number of (grand-) parents included in pedigrees of varying generations used to analyse ten different datasets from 10 years of family trials from a sugarcane breeding program

	Data sets (years of data)									
	1	2	3	4	5	6	7	8	9	10
Depth of pedigree (number of generations)										
2	281	426	537	625	765	865	951	993	1,015	1,073
3	479 (*198*)[a]	686 (*260*)	833 (*296*)	943 (*318*)	1,100 (*335*)	1,220 (*355*)	1,320 (*369*)	1,369 (*376*)	1,394 (*379*)	1,454 (*381*)
4	608 (*129*)	845 (*159*)	1,005 (*172*)	1,118 (*175*)	1,262 (*162*)	1,383 (*163*)	1,485 (*165*)	1,531 (*162*)	1,556 (*162*)	1,619 (*165*)
5	671 (*63*)	908 (*63*)	1,067 (*62*)	1,180 (*62*)	1,321 (*59*)	1,442 (*59*)	1,545 (*60*)	1,591 (*60*)	1,617 (*61*)	1,679 (*60*)
6	695 (*24*)	933 (*25*)	1,090 (*23*)	1,204 (*24*)	1,341 (*20*)	1,460 (*18*)	1,562 (*17*)	1,608 (*17*)	1,632 (*15*)	1,693 (*14*)
7	702 (*7*)	936 (*3*)	1,094 (*4*)	1,207 (*3*)	1,344 (*3*)	1,463 (*3*)	1,565 (*3*)	1,611 (*3*)	1,634 (*2*)	1,696 (*3*)

[a] Italic type face in parentheses shows the number of additional parents added with each generation of pedigree for that given data set

pedigree, and every family plot was considered as an individual (even though each family replicate contained a different set of 20 seedlings per family). However, under the parental model only pedigree information from the parents of each family was used as family was not included in the pedigree.

As some parents act as both male and female in sugarcane, a unified estimate for each parent was obtained. Each of the ten data sets was analysed using the six different depths of pedigree (as shown in Table 3) for each of these two models, with a total of 120 analyses performed. All analyses were performed using ASReml (Gilmour et al. 2006).

Adjusted CCS data were then summarised by additive and additive-by-environment variance components. Even though we are interested in the effects of pedigree structure on the estimation of additive variance and BV, ignoring non-additive effects often results in bias of additive genetic variance (Henderson 1985; Quinton and Smith 1997; Wei and Van der Werf 1993). Hence, we partitioned both additive and non-additive genetic effects.

For each data set, the goodness-of-fit test for each analysis, using different levels of pedigree information, was assessed using the Akaike Information Criterion (AIC), where models with smaller AIC values are superior in terms of goodness of fit (Akaike 1974). Additive genetic variance estimates, and additive-by-environment variance estimates (or additive-by-year variance estimates as each year trials are planted at the same location), from each analysis were also compared to aid in determining the minimum depth of pedigree needed to reliably assess parental BV.

BVs estimated from each analysis were collated only for the 281 parents of the families grown in the 2005 series of PATs. These parents were common to all pedigree files used. They are also the most current parents in the Southern BSES breeding population, and therefore, are of direct interest for future breeding. The top 70 parents, i.e. the top 25% of 281 parents, from each data set were identified from the analyses using seven generations of pedigree information as these are the parents of most interest to the breeders for further cross-pollination. Estimates of BVs of these 70 parents using six different pedigree levels were compared using a Pearson correlation within each data set.

Estimates of BV of the 281 common parents were also compared using a Pearson correlation across all data sets and pedigree levels. As the true BV of each parent is unknown, our best estimates of the 'true' BVs were obtained using 10 years of data and seven generations of pedigree, that being the most comprehensive data and pedigree information available. Correlations were calculated between the 'true' BVs of parents obtained from this analysis and the estimated BVs of parents obtained from all other analyses using varying degrees of data and pedigree levels to investigate the interaction between years of data and depth of pedigree. Again, parents in the top 25% for 'true' BV were identified. Estimates of BV of these top parents using different amounts of data and pedigree levels were also compared using a Pearson correlation. Spearman rank correlations were also calculated between the rankings of the top 70 parents for 'true' BV and rankings of parents obtained from analysing varying amounts of data. Correlations were used to ascertain how differently the top parents were being ranked relative to the assumed true BV.

Results

Depth of pedigree marginally affected the modelled spatial variation for each trial (or sub-trial). Some spatial effects (e.g. linear row and linear column effects) applied were no longer significant as additional generations were included to pedigree structures in the analysis of each data set.

Even though non-additive effects were estimated in analyses for CCS, they were not considered in the interpretation of the results, as we are only interested in additive genetic effects in estimating BVs. In addition, variance components and BVs estimated for both the parental and animal models using varying depths of pedigree and years of data were very similar. This was reflected in a correlation coefficient of 0.97 when the BV estimates from the animal and parental models, using the most comprehensive data set and pedigree information, were compared using a Pearson correlation. Therefore, only the results from one model, the parental model, are presented.

Some differences in additive and additive-by-environment genetic variance estimates were observed for each data set when the six different pedigree structures were applied (Figs. 1, 2). Initial estimates of additive variance for CCS were low, but increased as additional years of data were included in analyses (Fig. 1). Additive variance estimates were very similar when 3 or more years of data were included in analyses.

Additive-by-environment variance estimates showed a reverse trend to the estimates of additive variance (Fig. 2), in that variance estimates decreased as the years of data increased. However, both additive and additive-by-environment estimates for each data set stabilised with the inclusion of five or more generations of pedigree information.

Standard errors of additive and additive-by-environment variance estimates are not presented in this paper for brevity. Standard errors of additive variance estimates within each data set remained steady or increased marginally as additional pedigree information was included. In contrast, the standard errors decreased as additional years of data were included. For example, standard errors of additive variances ranged from 0.035 for the analysis of 1 year of data and seven generations of pedigree, down to 0.015 for the analysis of 10 years of data and seven generations of pedigree.

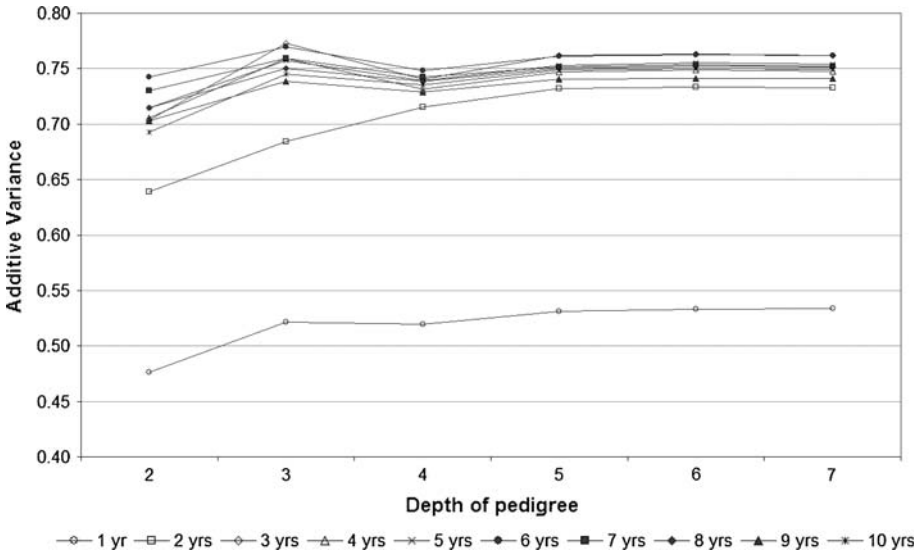

Fig. 1 Additive genetic variance estimates for each data set using different depths of pedigree for CCS for 10 years of PATs using the parental model

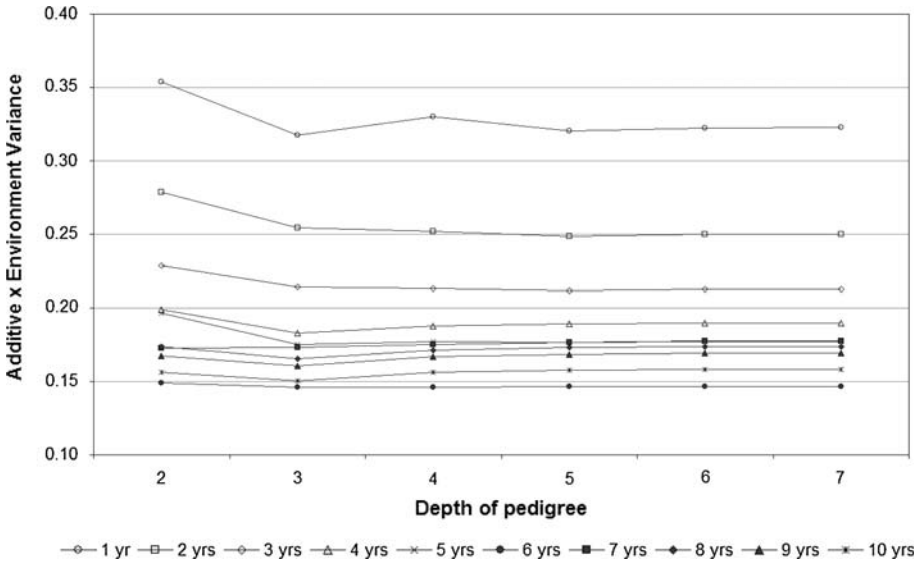

Fig. 2 Additive-by-environment variance estimates for each data set using different depths of pedigree for CCS for 10 years of PATs using the parental model

The same trend was observed for standard errors of additive-by-environment variance estimates.

The AIC was used to help determine the optimum number of generations needed in pedigrees. AIC values for each data set decreased steadily as additional generations were included in pedigree information, and then plateaued when five or more generations of pedigree were included with the analyses. For example, in the analysis of 4 years of data the AIC value for two generations was 2,673, three generation 2,613, four generations 2,567, five generations 2,560, six generations 2,559, and seven generations was 2,559. This trend in the decline of AIC values was typical of that observed in all data sets analysed.

Initial correlation coefficients indicated that BVs of parents in the top 25% using only two generations of pedigree information for each data set were the most different compared to full pedigree information (seven generations); however, rankings of parents did not change considerably, if at all, when four or more generations of pedigree were included (Table 4). Although, in most cases all coefficients were quite high.

Then we assumed that 'true' BV estimates for all 281 parents in common were obtained from the analysis of seven generations of pedigree and 10 years of data. There was a greatest difference between initial BV estimates using limited information (i.e. 1 year and two generations) compared with the 'true' BV (Table 5), but as years of data and/or pedigree information were added, BV estimates improved for both models when compared with 'true' BVs. There were very few differences in estimated BVs using at least 5 years of data and three generations of pedigree information (Table 5).

Additionally, when parents in the top 25% for BV from each analysis were identified and compared with the top parents for 'true' BV, the differences were greatest when

Table 4 Correlation coefficients of BVs of parents in common in the top 25% (i.e. the top 70 parents) by comparing BV estimates using seven generations with fewer generations within each data set for the parental model

	Data sets (years of data)									
	1	2	3	4	5	6	7	8	9	10
Depth of pedigree (number of generations)										
2	0.87	0.80	0.81	0.84	0.82	0.83	0.81	0.84	0.84	0.87
3	0.95	0.94	0.65	0.97	0.98	0.98	0.98	0.98	0.98	0.99
4	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	–	–	–	–	–	–	–	–	–	–

Table 5 Correlation coefficients of BVs for all 281 parents in common, comparing BV estimates using varying years of data and pedigree information with the 'true' estimate from seven generations in the pedigree and 10 years of data for CCS for the parental model

	Data sets (years of data)									
	1	2	3	4	5	6	7	8	9	10
Depth of pedigree (number of generations)										
2	0.73	0.80	0.83	0.84	0.89	0.90	0.91	0.92	0.93	0.93
3	0.77	0.84	0.90	0.91	0.96	0.97	0.98	0.99	0.99	0.99
4	0.79	0.87	0.91	0.93	0.97	0.98	0.99	0.99	1.00	1.00
5	0.80	0.88	0.91	0.93	0.97	0.98	0.99	0.99	1.00	1.00
6	0.80	0.87	0.91	0.93	0.97	0.98	0.99	0.99	1.00	1.00
7	0.80	0.87	0.91	0.93	0.97	0.98	0.99	0.99	1.00	–

the least amount of information was included in analyses (Table 6). As more data and pedigree information were included in analyses, the correlation coefficients for BV and number of parents in common in the top 25% increased. If we only increased the number of years included, and not the depth of pedigree, the precision of estimating the BV of parents was 0.87 with a maximum of 59 of the parents in common (84%) with the top 70 parents for 'true' BV (Table 6). In addition, increasing only the depth of pedigree, and not the years of data, resulted in a lower correlation coefficient and fewer parents in common compared with 'true' BV (Table 6). However, when the depth of pedigree and years of data were both increased, the number of parents in common also increased considerably. When 7 or more years of data and four generations were included, at least 90% of the top parents were being identified with a correlation coefficient of 0.96 for BVs (Table 6).

Correlation coefficients from comparing BVs and rankings of parents in the top 25% were very similar and showed the same trend; therefore, we have only presented the results from the Pearson correlation when BVs of parents were compared.

Discussion

Piepho et al. (2008) suggested that not including complete pedigree records often resulted in biased estimates of additive variance and BV. This has been demonstrated by Mehrabani-Yeganeh et al. (1999) and Durel et al. (1998), where improvements in estimates of BLUPs and heritabilities using a complete pedigree structure were shown to give significantly greater gains than not using pedigree information at all. However, we know of no published research into the effects of different depths of pedigree on estimates of BV. In our study, adding pedigree information back to the base population improved the estimates of both additive variance of the population and BVs for each parent. We also showed that minimal pedigree information yielded biased estimates of additive variance and BV, and selection accuracy of parents is low compared with using all pedigree information. Estimates of both additive variance and BV improved as pedigree information was added.

Even though not including complete pedigree information in analyses can lead to biased estimates of additive variance and BVs (Piepho et al. 2008), in our study additive genetic variances were estimated with the same precision when five or more generations were included. This was also reflected in the AIC values plateauing after five or more generations were included in analyses, and BVs were estimated with the same accuracy within each data set even when identifying the top-performing parents (Table 4). However, this may reflect our data structure, where relatively few parents were added to each pedigree after the inclusion of five generations (Table 3). In addition, approximately 10% of the parents in the pedigree were produced from polycrosses (where more than one male is used to fertilise a female), or produced prior to the 1950s when the Australian sugarcane breeding program started controlled pollination (Skinner 1959). This has resulted in some missing data in the pedigree where some parents are unknown. Therefore, as older generations are added, they appear to contribute relatively little additional pedigree information for BLUP methodology to exploit, and consequently, the accuracy of estimates using five or more generations should be similar. This was also demonstrated when BVs of the top 70 parents were compared within each data set using varying levels of pedigree information, where BVs estimated from at least four generations of pedigree were estimated with the same accuracy.

Unfortunately, there is no formal test, like the AIC, to aid in determining how much data is needed to reliably estimate BV. Therefore, we can only rely on comparing variance estimates and their standard errors, and correlation coefficients obtained from analyses. Regardless of how much pedigree information is included, additive genetic variance is underestimated when only 1 year of data was analysed,

Table 6 Correlation coefficients of BVs for parents in the top 25% (i.e. the top 70 parents), comparing BV estimates using varying years of data and pedigree information with the 'true' estimate from seven generations in the pedigree and 10 years of data for CCS for the parental model

	Data sets (years of data)									
	1	2	3	4	5	6	7	8	9	10
Depth of pedigree (number of generations)										
2	0.58 (*44*)[a]	0.68 (*49*)	0.72 (*51*)	0.73 (*52*)	0.72 (*54*)	0.77 (*55*)	0.79 (*58*)	0.82 (*59*)	0.84 (*58*)	0.87 (*59*)
3	0.59 (*46*)	0.71 (*53*)	0.79 (*53*)	0.80 (*55*)	0.84 (*58*)	0.91 (*61*)	0.93 (*62*)	0.95 (*62*)	0.95 (*64*)	0.99 (*64*)
4	0.59 (*48*)	0.74 (*55*)	0.78 (*55*)	0.81 (*56*)	0.84 (*61*)	0.93 (*62*)	0.96 (*65*)	0.97 (*65*)	0.98 (*67*)	1.00 (*68*)
5	0.60 (*49*)	0.74 (*55*)	0.78 (*56*)	0.81 (*57*)	0.84 (*60*)	0.93 (*62*)	0.97 (*65*)	0.97 (*65*)	0.98 (*68*)	1.00 (*70*)
6	0.59 (*48*)	0.74 (*56*)	0.78 (*56*)	0.81 (*57*)	0.84 (*60*)	0.93 (*63*)	0.97 (*65*)	0.97 (*66*)	0.98 (*68*)	1.00 (*70*)
7	0.59 (*48*)	0.74 (*56*)	0.78 (*56*)	0.81 (*57*)	0.84 (*60*)	0.93 (*63*)	0.97 (*65*)	0.97 (*66*)	0.98 (*68*)	–

[a] Italic type face in parentheses shows the number of parents in common in the top 25% compared with 'true' BV estimates using 10 years of data and seven generations of pedigree

while additive genetic variances estimated from 3 or more years of data were very similar. Additionally, standard errors of additive variance estimates decreased as additional years of data were included in analyses, in contrast to pedigree information, indicating that including additional years of data improves the accuracy of estimates of additive genetic variance more than pedigree information. However, just comparing additive variance components does not consider what impact varying amounts of data have on estimates of parental BVs, or the interaction of varying amounts of data and pedigree information have on BVs.

Our comparison of the accuracy to estimate BVs of sugarcane parents, and identify top-performing parents, from different amounts of data and pedigree information, to the 'true' BV of each parent, shows that greater amounts of information are required. If sugarcane breeders are only interested in the BVs of all the current parents, and not selecting the top parents for cross pollination, then a minimum of 5 years of data and three generations of pedigree appears to be adequate to ensure the precision of estimating BVs is very high. If the same amount of data is used to select the top parents for cross-pollination (i.e. the top 70 parents), then the precision of identifying these top parents is reduced, with only 58 of the 70 top parents in common. This means that 17% of the top parents would be overlooked for cross-pollination. To ensure that no more than 10% of the top parents are identified incorrectly, a minimum of 7 years of data and four generations of pedigree information is required.

As BSES intends to combine information from family trials from all four selection programs to more accurately predict BVs of parents in the breeding population, it is important to consider the issue of possible computing limitations in estimating BVs of over 3,200 sugarcane parents. Even though analyses are much easier and faster to process when minimal information is used, this leads to a reduction of the accuracy of estimating both additive genetic variance and BVs of parents. While including all available information (10 years of data and seven generations of pedigree) to estimate BVs for 1,600 parents in only one of the four BSES selection programs is computationally achievable, less family data and/or pedigree information will be necessary when the four selection programs are combined. We show that adding years of data clearly has a larger effect on the estimation of both additive variance and BVs of parents, than adding pedigree information. However, the number of years of data included in analyses should be no less than 5 years when assessing the BVs of all parents, or less than 7 years when selecting the top parents for future breeding, while the pedigree information can be reduced to three or four generations.

When estimating BV of sugarcane parents, either the animal or the parental model can be used, as there was very little difference in variance components and BVs estimated. Both models require the same minimum number of years and pedigree information to estimate BV. However, in sugarcane, families may be replicated within and across trials, but individual seedlings are not, as each family plot is represented by 20 different full-sibs and each measurement is recorded on a plot (family)-mean basis. It also appears that this data structure is unique to the sugarcane family trials in the BSES-CSIRO Plant Improvement program, and neither the animal nor the parental model has been used on this type of data before where only family means are available. It is unclear that what effects analysing family-mean data have on estimating additive and non-additive variance components. This may pose a problem for sugarcane breeders, as within-family variance cannot be properly estimated, and is only based on a family mean.

Even though non-additive genetic variances are not considered when estimating BVs of parents, they play an important role in selecting sugarcane families and individual clones to advance through to the next stage of selection. Estimated non-additive genetic variances were considerably different between the animal and parental models. Non-additive genetic variance was between one-third and one-half the size of the additive genetic variance estimated using the parental model, while no non-additive genetic variance was detected using the animal model. Whether the animal or

parental model can be used for both crossing and selecting purposes should be investigated further using simulated data. Although both models provide similar results, the parental model appears more suitable to the sugarcane family data structure, as family plots are not really individuals (as the animal model assumes), but rather comprise 20 individual (unreplicated) seedlings. Therefore, we recommend that the parental model is used by BSES breeders to estimate BV of sugarcane parents.

Acknowledgments Thanks to the Sugar Research and Development Corporation (SRDC) who partially fund Felicity Atkin's research through a Sugar Industry Research Scholarship. We thank the BSES Limited Plant Improvement technicians who contributed to the production, collection and collation of the data used in this study. We would also like to thank the reviewers for their comments which have greatly improved this manuscript.

Appendix: ASReml code

The code to fit the parental and animal models presented in this paper is given below.

ASReml code to run the parental model:

```
1     Combined anlaysis of 1996 to 2005 PATs S P - ccs
2      SubTrial 39 !A
3      Trial 20 !A
4      Family 2584 !A
5      Female 1696 !P
6      Male 1696 !P
7      Nrow 44 !I
8      Plot 22 !I
9      ccs
10     ccsa
11    S_pat9605_pedigree.csv !skip 1 !ALPHA !MAKE
12    S_pat9605.csv !skip 1 !nodisplay !maxit 50 !ddf-1
13    ccsa ~ mu SubTrial,
14          at(SubTrial,10,13,21,27,28,32,34,35,38,39).lin(Plot),
15          at(SubTrial,14,27,32,34).lin(Nrow),
16          !r at(SubTrial,37).Nrow,
17          at(SubTrial,20,24,30,31,38).Plot,
18          Female and(Male) Trial.Female -Trial.Male and(Trial.Male)
            Family Trial.Family,
19          !f mv
20    39 2 0
21    12 Plot AR .05 #MQN05-11E(SubTrial,1)
22    41 Nrow AR .2
23    12 Plot AR .05 #MQN05-11L(SubTrial,2)
24    41 Nrow AR .2
25    .
26    .
27    .
28    16 Plot AR .1 #MQN96-11L(SubTrial,39)
29    31 Nrow AR .38
```

ASReml code for the animal model:

```
4      Family 2584 !P
5      Female 1696 !A
6      Male 1696 !A
.
.
.
18          Family Trial.Family ide(Family) Trial.ide(Family),
.
.
```

```
##############################

Explanation of Asreml code for each line number for the parental
model:

1 - heading line
2 - number of sub-trials
3 - number of trials
4 - number of families - alphanumeric variable
5 - number of female parents - pedigree applied
6 - number of male parents - pedigree applied
7 - number of rows
8 - number of plots
9 - sugar content
10 - sugar content - standardised
11 - pedigree file name, first line of pedigree file contains headings
12 - data file name, first line of data file contains headings
13 - ccsa is the variable to be analysed, intercept (mu) is fitted,
     SubTrial is fitted as a fixed effect
14 - linear Plot is fitted as a fixed effect at the SubTrial level
15 - linear Row is fitted as a fixed effect at the SubTrial level
16 - Row is fitted as a random effect at the SubTrial level
17 - Plot is fitted as a random effect at the SubTrial level
18 - uses a biparental model where var(Female and(Male)) = ¼ additive
     variance and var(Family) = ¼ dominance variance
     (in Female and(Male) overlays the design matrices for males and
     females so only one prediction for each parent is given and
     represents GCA; Family represents SCA)
19 - missing values are fitted as fixed in sparse set of terms
20 - variance header line, 39 R structures (for each SubTrial)
     are the direct product of 2 variances, no G structure
21 - 12 levels for Plot, AR(1) fitted in the column direction
22 - 41 levels for Row, AR(1) fitted in the row direction
23-29 - as above for the 39 sub-trials


Explanation of Asreml code for each line number for the animal model:

4 - number of families - pedigree applied
5 - number of female parents - alphanumeric variable
6 - number of male parents - alphanumeric variable
.
.
.
18 - uses individual (or animal) model where var(Family) = additive
     variance and var(ide(Family)) = dominance + epistatsis
.
.
```

References

Akaike H (1974) A new look at statistical model identification. IEEE Trans Autom Control 19:716–722

Berding N, Skinner JC (1987) Traditional breeding methods. Copersucar International Sugarcane Workshop, Copersucar, Piracicaba, Brazil

Bremer G (1961) Problems in breeding and cytology of sugar cane. Euphytica 10:59–78

Costa e Silva J, Borralho NMG, Potts BM (2004) Additive and non-additive genetic parameters from clonally replicated and seedling progenies of *Eucalyptus globulus*. Theoret Appl Genet 108:1113–1119

Davik J, Honne BI (2005) Genetic variance and breeding values for resistance to a wind-borne disease [*Sphaerotheca macularis* (Wallr. ex Fr.)] in strawberry (*Fragaria* × *ananassa* Duch.) estimated by exploring mixed and spatial models and pedigree information. Theoret Appl Genet 111:256–264

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

Ericsson T (1999) The effect of pedigree error by misidentification of individual trees on genetic evaluation of a full-sib experiment. Silvae Genet 48:239–242

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

Furlani RCM, de Moraes MLT, de Resende MDV, Furlani E Jr, Goncalves P, Filho WVV, de Paiva JR (2005) Estimation of variance components and prediction of breeding values in rubber tree breeding using the REML/BLUP procedure. Genet Mol Biol 28:271–276

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

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2006) ASReml user guide, 2.0th edn. VSN International Ltd, Hemel Hempstead

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph

Henderson CR (1985) Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. J Anim Sci 60:111–117

Long TE, Johnson RK, Keele JW (1990) Effects of errors in pedigree on three methods of estimating breeding value for litter size, backfat and average daily gain in swine. J Anim Sci 68:4069–4078

Mehrabani-Yeganeh H, Gibson JP, Schaeffer LR (1999) Using recent versus complete pedigree data in genetic evaluation of a closed nucleus broiler line. Poult Sci 78:937–941

Meyer K (1991) Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. Genet Select Evol 23:67–83

Mrode RA (2005) Linear models for the prediction of animal breeding values, 2nd edn. CABI Publishing, Wallingford

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. Theoret Appl Genet 113:809–819

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

Piepho HP, Mohring J, Melchinger AE, Buchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228

Purba AR, Flori A, Baudouin L, Hamon S (2001) Prediction of oil palm (*Elaeis guineensis,* Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). Theoret Appl Genet 102:787–792

Quinton M, Smith C (1997) An empirical check on best linear unbiased prediction genetic evaluation using pig field recording data. Can J Anim Sci 77:211–216

Roach BT (1989) Origin and improvement of the genetic base of sugarcane. Proc Aust Soc Sugar Cane Technol 11:34–48

Skinner JC (1959) Controlled pollination of sugar cane. Bureau Sugar Exp Stations Tech Commun 1:1–19

Stringer JK, Cullis BR (2002) Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. Aust J Agric Res 53:911–918

Visscher PM, Woolliams JA, Smith D, Williams JL (2002) Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J Dairy Sci 85:2368–2375

Wei M, Van der Werf JHJ (1993) Animal model estimation of additive and dominance variances in egg production traits of poultry. J Anim Sci 71:57–65

Wei X, Borralho NMG (2000) Genetic gains and levels of relatedness from best linear unbiased prediction of *Eucalyptus urophylla* for pulp production in southeastern China. Can J For Res 30: 1601–1607

Wei X, Jackson PA, McIntyre CL, Aitken KS, Croft B (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population structure. Theoret Appl Genet 114:155–164

White TL, Hodge GR (1989) Predicting breeding values with applications in forest tree improvement. Kluwer Academic Publishers, Dordrecht