Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane

Xianming Wei, Phillip A. Jackson, Scott Hermann, Andrzej Kilian, Katarzyna Heller-Uszynska, and Emily Deomano

Abstract: Few association mapping studies have simultaneously accounted for population structure, genotype by environment interaction (GEI), and spatial variation. In this sugarcane association mapping study we tested models accounting for these factors and identified the impact that each model component had on the list of markers declared as being significantly associated with traits. About 480 genotypes were evaluated for cane yield and sugar content at three sites and scored with DArT markers. A mixed model was applied in analysis of the data to simultaneously account for the impacts of population structure, GEI, and spatial variation within a trial. Two forms of the DArT marker data were used in the analysis: the standard discrete data (0, 1) and a continuous DArT score, which is related to the marker dosage. A large number of markers were significantly associated with cane yield and sugar content. However, failure to account for population structure, GEI, and/or spatial variation produced both type I and type II errors, which on the one hand substantially inflated the number of significant markers identified (especially true for failing to account for GEI) and on the other hand resulted in failure to detect markers that could be associated with cane yield or sugar content (especially when failing to account for population structure). We concluded that association mapping based on trials from one site or analysis that failed to account for GEI would produce many trial-specific associated markers that would have low value in breeding programs.

Key words: association study, population structure, genotype × environment interaction, spatial variation, sugarcane.

Résumé: Peu d’analyses d’association ont simultanément tenu compte de la structure de la population, de l’interaction génotype × environnement (GEI) et de la variation spatiale. Dans cette analyse d’association réalisée chez la canne à sucre, les auteurs ont comparé des modèles tenant compte de ces facteurs et identifié l’impact de chaque composante du modèle sur la liste des marqueurs identifiés comme étant significativement associés aux caractères étudiés. Environ 480 génotypes ont été évalués pour le rendement en canne (TCH) et la teneur en sucre (CCS) sur trois sites tout en étant génotypés à l’aide de marqueurs DArT. Un modèle mixte a été employé pour l’analyse des données afin de tenir compte simultanément de l’impact de la structure de la population, du GEI et de la variation spatiale au sein d’un essai. Deux formes de données DArT ont été employées, soit des données discrètes de type standard (0, 1) et un score DArT continu reflétant la dose du marqueur. Un grand nombre de marqueurs était associé significativement avec le TCH et le CCS. Cependant, le fait de ne pas tenir compte de la structure de la population, du GEI ou de la variation spatiale a entraîné des erreurs de type I et de type II. Ces erreurs ont, d’une part, grandement augmenté le nombre de marqueurs significatifs (particulièrement dans le cas du GEI) et, d’autre part, empêché de détecter des marqueurs qui seraient associés au TCH et au CCS (particulièrement en ne tenant pas compte de la structure de la population). Les auteurs concluent que des analyses d’association fondées sur des essais réalisés sur un seul site ou des analyses qui ne tiendraient pas compte du GEI produiraient plusieurs marqueurs spécifiques d’un site d’essai et qui seraient peu utiles dans une programme d’amélioration génétique.
Introduction

Association mapping has become widely used to identify DNA markers associated with traits in many crops (Remington et al. 2001; Kraakman et al. 2004; Aranzana et al. 2005; Breezeello and Sorrells 2006; Agrame et al. 2007; Dhooe et al. 2008; Stich et al. 2008). Because of the often complex relationships between genotypes in breeding populations, it is necessary to separate marker–trait associations due to population structure from marker–trait associations due to linkage (e.g., Yu et al. 2006). Not surprisingly, approaches to overcome this problem have attracted much attention (Remington et al. 2001; Jannink and Walsh 2002; Zhao et al. 2007). Analytical methods that account for population structure include model-based clustering (Pritchard et al. 2000), principal component analysis (Price et al. 2006), kinship analysis (Yu et al. 2006), and genomic control (Devlin and Roeder 1999). Zhao et al. (2007) compared these methods using Arabidopsis populations and concluded that for a range of key criteria, kinship analysis worked best. They also reported that the model-based clustering method could not adequately define the structural complexity of the populations that were studied.

As indicated by Zhao et al. (2007), some methods to account for population structure might be ineffective because of inadequacy in capturing the complexity of population interrelationships. This may well be true in the genetically complex polyploid sugarcane. In a previous study of sugarcane (Wei et al. 2006), attempts were made to account for population structure by assigning genotypes to subpopulations as identified using the program STRUCTURE (Pritchard et al. 2000). However, this division was somewhat arbitrary, as there were no clear discontinuities in the population to unambiguously define subpopulations.

In addition to population structure, acquisition of accurate phenotypic trial data used in identifying marker–trait associations is essential to the success of association mapping. Data for association mapping experiments are generally collected from field trials and thus genetic effects derived from these data could be affected by genotype by environment interaction (GEI) and spatial variation at any given site (Jackson 1992; Jackson and Hogarth 1992; Smith 1999; Cooper et al. 2001; Stringer and Cullis 2002). Relative to the research on the impact of population structure, there appears to be less research on the impact of environmental effects in association mapping studies (Malosetti et al. 2007), despite a number of published studies on biparental mapping populations (Piepho 2000; Boer et al. 2007; Emrich et al. 2008). Failure to account for spatial variation and environmental effects will certainly reduce the power of detecting quantitative trait loci (QTL) in mapping populations and could inflate QTL main effects (e.g., Emrich et al. 2008). In the Australian sugarcane breeding program, GEI has been reported to be significant in a number of studies (Jackson and Hogarth 1992; Gilbert et al. 2006), and spatial variation (Gilmour et al. 1997) within a trial appears common and can be accounted for in routine data analysis (e.g., Stringer and Cullis 2002; Smith et al. 2007; Wei et al. 2007). Analysis of phenotypic data from multiple sites needs to include GEI and spatial variation for detecting the association between markers and traits over a wide range of environments.

In this paper we report the results of a study undertaken to detect associations between genetic markers and two important sugarcane traits and examine the impact that population structure, GEI, and spatial variation from three field trials have on the results.

Materials and Methods

Field Trials and Phenotypic Data

Three field trials were planted in two major sugar production areas in Australia in 2006: two in the Burdekin region (Kalamia Mill area or KAL and Pioneer Mill area or PNR) and one in the Herbert region (Victoria Mill area or VIC). Originally, 480 sugarcane clones were propagated for each site, of which half were selected parental clones from the Australian sugarcane breeding program and commercial cultivars, and the other half were clones randomly taken from 30 unselected families (eight clones per family). Owing to germination failures in some clones during the propagation process, slightly different numbers of clones were planted in each site. However, most clones were planted across all three sites (Table 1). Within each trial, a randomized complete block design with two replicates was used, with each clone planted in one plot per block. The individual plot size was 1 row \( \times \) 10 m, with an inter-row spacing of 1.5 m in each trial.

During the harvesting season in 2007, cane yield (t/ha) of each plot was measured directly in the field. Three stalks were taken at random from each plot and commercially extractable sucrose content (expressed as commercial cane sugar, CCS) was estimated using procedures commonly used in sugarcane breeding programs. This involved extracting juice using either a small mill (VIC and PNR trials) or a hydraulic press (KAL trial). Brix and pol were measured in the extracted juice, from which CCS was determined using commonly applied procedures (BSES 1984).

Marker Data

All clones in the field trials were genotyped using DArT markers (Heller-Uszynska et al. 2010). For each clone, two types of polymorphic DArT markers were obtained: 1998 markers were classed as “discrete” (i.e., able to be scored clearly as “present” (1) or “absent” (0)), and 15,360 markers were classed as “continuous” (i.e., score is related to the number of copies of each marker allele per genotype, which could be valuable in a highly polyploid species like sugarcane). Only 1,531 discrete markers were used in the analyses after markers with frequency greater than 0.95 or
Table 1. Number of clones tested in each of the three trials; numbers on the diagonal are the number of clones in each site, and numbers in the off-diagonal are the number of clones common to each combination of sites.

<table>
<thead>
<tr>
<th></th>
<th>Victoria</th>
<th>Kalamia</th>
<th>Pioneer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victoria</td>
<td>392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kalamia</td>
<td>376</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td>Pioneer</td>
<td>377</td>
<td>449</td>
<td>453</td>
</tr>
</tbody>
</table>

less than 0.05 were excluded. All continuous markers were used in the association study.

Marker data were used to assess the level of linkage disequilibrium (LD) in the population studied. Fisher’s exact probability was used to test for associations between DArT markers that were common to both the association mapping population and the Q1650 genetic map (Ai etc. 2005; K.S. Aitken et al., unpublished data). For each pair of markers a contingency table (presence versus absence) was established and the Fisher probability (phi) was computed using the vcd package in R (R Development Core Team 2006).

Statistical analysis

Phenotypic data were first analyzed independently for each trial using a mixed model that best accounted for spatial variation effects and other natural and extraneous effects (Gilmour et al. 1997). Optimal spatial variation models for each site were then combined in a single multi-environment analysis. This combined model, along with the additive relationship coefficient from the pedigree of all the clones in the trials and their ancestors, was used to detect marker–trait associations. This model is similar to the model developed by Yu et al. (2006). Our model included multiple environment trials and their ancestors, was used to detect marker–trait associations. This model is similar to the model developed by Yu et al. (2006). Our model included multiple environment trials, so genetic correlations between sites were included, and residual effects at each plot were assumed to be correlated with others. However, our model did not include the Q matrix from STRUCTURE (Pritchard et al. 2000). Analysis of each marker was conducted independently. The model could be written as follows:

\[ y = X_b + X_m Z_u + Z_a + R \]

where

- **y** = a vector of measured value for cane yield or CCS;
- **b** = a vector of fixed effects related to field trials, e.g., location of trials;
- **m** = a vector of fixed effects of a marker and its plate number in the laboratory;
- **u** = a vector of random effects related to field trials, e.g., replicate, row, or column within a trial;
- **a** = a vector of random polygenic effects for clones in trials and their ancestors with \( N(0, G) \), where

\[
G = A \otimes \begin{pmatrix}
\sigma_2^{2_{\text{VIC}}} & 0 & 0 \\
0 & \sigma_2^{2_{\text{KAL}}} & 0 \\
0 & 0 & \sigma_2^{2_{\text{PNR}}}
\end{pmatrix}
\]

and **A** is the numerator relationship matrix derived from pedigree (Henderson 1976), \( \sigma_2^{2_{\text{I}}} \) is genetic variance at the \( i \)th trial, \( \sigma_2^{2_{\text{Q}}} \) is genetic covariance between the \( i \)th and \( j \)th trials, and \( \otimes \) represents the Kronecker product operator between two matrices (see Smith et al. 2007 for a simple example of Kronecker products); **R** = residual effects from environmental factors and errors, modelled as diag(**R**), where \( \mathbf{R}_i = \mathbf{I}_i \otimes \Sigma_r \) for the \( i \)th trial, \( \sigma_r^{2} \) is a scale parameter, and \( \Sigma_c \), and \( \Sigma_r \) are the \( c_i \times c_i \) and \( r_j \times r_j \) correlation matrices corresponding to the column and row dimensions of the \( i \)th trial, respectively (Smith et al. 2007); and \( X_1, X_2, Z_1, \) and \( Z_2 \) are design or incidence matrices.

Therefore, this model simultaneously accounted for the effects of population structure (via pedigree), GEI (via **G** structure), and spatial variations within a trial (via **R** structure) (see Butler et al. 2006 for details about **G** and **R** structures).

To examine the impact of the effects from population structure, GEI, and (or) spatial variations, one or more of the terms in eq. 1 were modified. This was done as follows:

1. \( \mathbf{A} = \mathbf{I} \), an identity matrix, for excluding population structure effect or assuming all clones were genetically unrelated;

\[
G = A \otimes \begin{pmatrix}
\sigma_2^{2_{\text{VIC}}} & 0 & 0 \\
0 & \sigma_2^{2_{\text{KAL}}} & 0 \\
0 & 0 & \sigma_2^{2_{\text{PNR}}}
\end{pmatrix}
\]

for failing to account for GEI, i.e., assuming that all trials were genetically independent, even though almost all of the clones were identical in all trials; and

2. \( \mathbf{R} = \sigma_r^{2} \mathbf{I} \) for excluding spatial variation effects or assuming all plots were independent from their neighbour plots.

As a result, the base or optimal model was modified to form seven different models, summarized in Table 2. Comparison between the base model (M0) and model M1 was used to determine the impact of population structure; M2 to M4, the combined impact of population structure and GEI and (or) spatial variation; and M5 to M7, the impact of GEI and (or) spatial variation only. It should be noted that GEI can be interpreted as correlation between two environments (Burdon 1977). As implied in the modification of the **G** matrix, the impact of GEI was used to test the impact of analyzing multiple environment tests independently in the association study.

All analyses were conducted by ASReml-R (Butler et al. 2006) under the **R** platform (R Development Core Team 2008).

Results

Performance and genetic control of cane yield and CCS in field trials

In general, means observed for cane yield and CCS in all three trials were within the range of what would be regarded as typical for performance of sugarcane in the Herbert and Burdekin regions in Australia (Table 3). As expected, cane
yield was higher in the two Burdekin trials (>140 t/ha) than in the trial in Herbert (77 t/ha).

Broad-sense heritability of each trial and the genetic correlations between pairs of the three trials were estimated based on a combined analysis following eq. 1 (but excluding $X^2_m$ and pedigree for $Z^2_a$). The results are presented in Table 3. Both cane yield and CCS were under strong genetic control, as estimates of broad-sense heritability were high, ranging from 0.56 to 0.80 for cane yield and from 0.52 to 0.76 for CCS. The high estimates of heritability indicated that the field trials produced good-quality data for the association study. Genetic coefficients of variation (i.e., ratio of genetic standard deviation to grand mean) indicated that variation in cane yield would have a larger impact than variation in CCS on genetic variation in sugar yield (Table 3). Genetic correlations between trials were high for both traits, ranging from 0.63 to 0.94 for cane yield and from 0.67 to 0.78 for CCS. These genetic correlations suggest that GEI was not important relative to overall genetic effects and that estimation of an overall marker effect across all trials was more important than identifying marker effects specific to individual sites.

**Linkage disequilibrium present in the population**

The Fisher probability (phi) was plotted against the distance (cM) between linked markers on the Q165q genetic map (310 markers with 1044 associations) (Fig. 1). Phi values from unlinked marker pairs (46 851 associations) were also plotted to allow visual comparison of phi between linked and unlinked markers. The population showed good evidence of LD decay in relation to genetic distance (Fig. 1). As in Raboin et al. (2008), the strongest LD appeared in the first 5 cM and a clear decay occurred over distances of 30 cM. The high level of LD found in this population supported attempts to identify markers linked to QTL.

**Number of markers declared as being significantly associated with traits**

More markers were observed as being significantly associated with both cane yield and CCS than would be expected under the null hypothesis of no association (Table 4), based on the optimal model (M0). For example, at $P \leq 0.01$, there were 47 discrete and 352 continuous markers associated with cane yield and 42 discrete and 377 continuous markers associated with CCS. Simulation showed that type II errors were consistent with what was expected (results not shown) and thus about 15 and 154 discrete and continuous markers would be expected to be declared at this $P$ value by random chance. Of these markers declared at $P \leq 0.01$, 28 and 15 markers for cane yield and CCS, respectively, were declared as significant when expressed in both discreet (1–0 classification) and continuous (normalized intensity score) forms of measurement.

Despite a substantial difference in genetic coefficients of variation between cane yield and CCS (Table 3), the number of significant markers was similar at all $P$ values (Table 4). As expected from the low between-trait genetic correlation (data not shown), the number of common markers between cane yield and CCS was small (Table 4). For example, there were only two discrete and five continuous markers that were significant for both cane yield and CCS at $P \leq 0.01$, and virtually none when $P$ values were lower. This is consistent with the reported low genetic correlation between cane yield and CCS in previous studies (e.g., Jackson 2005).

### Table 2. Terms included in models for examining the impact of population structure, genotype × environment interaction, and (or) spatial variation on detecting association between markers and sugarcane yield and sugar content.

<table>
<thead>
<tr>
<th>Model</th>
<th>Population structure via pedigree</th>
<th>Genotype × environment interaction</th>
<th>Spatial variation within a trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>M1</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>M2</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>M3</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>M4</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>M5</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>M6</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>M7</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Table 3. Trait means and standard errors, genetic coefficients of variation, broad-sense heritability (diagonal) at each trial, and genetic correlations (above diagonal) between trials for cane yield (TCH, t/ha) and sugar content (CCS, %).

<table>
<thead>
<tr>
<th>Trait</th>
<th>Site</th>
<th>Mean ± SE</th>
<th>Genetic coefficient of variation</th>
<th>Victoria</th>
<th>Kalamia</th>
<th>Pioneer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCH</td>
<td>Victoria</td>
<td>76.6±1.5</td>
<td>23.3</td>
<td>0.80±0.02</td>
<td>0.72±0.05</td>
<td>0.63±0.06</td>
</tr>
<tr>
<td></td>
<td>Kalamia</td>
<td>142.7±2.4</td>
<td>19.2</td>
<td>0.56±0.04</td>
<td>0.94±0.05</td>
<td>0.59±0.04</td>
</tr>
<tr>
<td></td>
<td>Pioneer</td>
<td>147.3±3.6</td>
<td>21.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCS</td>
<td>Victoria</td>
<td>16.2±0.1</td>
<td>4.1</td>
<td>0.76±0.02</td>
<td>0.78±0.04</td>
<td>0.73±0.06</td>
</tr>
<tr>
<td></td>
<td>Kalamia</td>
<td>15.8±0.1</td>
<td>6.1</td>
<td>0.74±0.02</td>
<td>0.67±0.06</td>
<td>0.52±0.04</td>
</tr>
<tr>
<td></td>
<td>Pioneer</td>
<td>13.8±0.2</td>
<td>7.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The impact of excluding population structure, GEI, and (or) spatial variations within a trial from analyses was substantial (Fig. 2). In general, models that failed to account for one or more of these factors inflated the number of significant markers, regardless of marker type (discrete or continuous) or trait (cane yield or CCS). This was most evident for models that failed to account for GEI (M3, M4, M6, and M7); for example, 2.0 (discrete) and 1.5 (continuous) times more markers were detected to be significant with M6 than with M0 for cane yield, and corresponding figures were 2.6 and 2.4 times for CCS. This was exacerbated when both GEI and population structure were excluded from analyses (M3); there were 3.9 (discrete) and 2.5 times more markers with M3 than with M0 for cane yield, and 6.6 and 4.2 times more for CCS. Spatial variations within a trial seemed to have much less impact than population structure and GEI.

Table 4. Number of DArT markers associated with cane yield and sugar content at different P values with model M0 applied.

<table>
<thead>
<tr>
<th>P value</th>
<th>Discrete</th>
<th>Continuous</th>
<th>Discrete</th>
<th>Continuous</th>
<th>Common</th>
<th>Discrete</th>
<th>Continuous</th>
<th>Common</th>
<th>Discrete</th>
<th>Continuous</th>
<th>Common</th>
<th>Discrete</th>
<th>Continuous</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>77</td>
<td>768</td>
<td>144</td>
<td>1228</td>
<td>72</td>
<td>136</td>
<td>1380</td>
<td>56</td>
<td>14</td>
<td>136</td>
<td>14</td>
<td>14</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>15</td>
<td>154</td>
<td>47</td>
<td>352</td>
<td>28</td>
<td>42</td>
<td>377</td>
<td>17</td>
<td>2</td>
<td>42</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>2</td>
<td>15</td>
<td>15</td>
<td>64</td>
<td>10</td>
<td>7</td>
<td>55</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Significant markers from models that failed to account for population structure, GEI, and (or) spatial variations generally did not overlap with those detected with the base model (M0). The most dissimilar models were those without population structure (M1, M2, M3, and M4), where generally fewer than half of the markers identified by M0 were detected. For example, considering cane yield at a significance threshold of $P \leq 0.01$ when population structure was excluded, only 21 of 47 discrete and 138 of 352 continuous markers were in common with markers declared significant using the optimal model (M0). For CCS at the same threshold, 18 of 42 discrete and 160 of 377 continuous markers were in common with those detected with the base model (M0). Additional failure to account for GEI and (or) spatial effects appeared not to decrease these ratios further. Although there were some differences, models that did not
account for GEI and (or) spatial variations (M5, M6, and M7) detected most of the markers detected by M0. For discrete markers, when both GEI and spatial variation were excluded, the number of overlapping markers was 46 (47 by M0) for cane yield and 42 (42 by M0) for CCS. There was slightly less agreement for continuous markers, but it was still very high: 303 (352 by M0) for cane yield and 358 (377 by M0) for CCS.

**Discussion**

**Significant markers**

This study detected a number of markers associated with cane yield or CCS. These associations appear to be robust in terms of different types of markers, with a high degree of overlap between discrete and continuous markers. The proportion of significant markers in this study was consistent with a previous study of disease resistance in sugarcane (Wei et al. 2006). Relative to the number of significant markers expected by random chance, about 2 times more markers were detected to be associated with cane yield or CCS in this study (at $P \leq 0.01$, Table 4). The corresponding figure for four different diseases was 1 to 3 times more AFLP markers (Wei et al. 2006).

Unfortunately, we were not able to find a reliable method to determine the effect of the size of a marker under the current mixed model framework. However, it was reported that with a set of markers, a reasonable amount of phenotypic variation in sugarcane could be explained. For example, Wei et al. (2006) found that 11 AFLP markers could account for 59% of smut rating variation. For a biparental population, Alwala et al. (2009) reported that 35.5%–48.4% of phenotypic variation for Brix or 22.1%–36.2% of phenotypic variation for pol could be explained by five to seven markers.

**Population structure**

A number of researchers have used both the kinship rela-
tionship and the Q matrix from STRUCTURE (Pritchard et al. 2000) to account for the effect of population structure and to separate identification of associations due to linkage and those due to population structure (e.g., Yu et al. 2006; Crossa et al. 2007; Iwata et al. 2007). In a previous study, Wei et al. (2006) grouped a sample of parents and cultivars in the BSES–CSIRO breeding program into eight subpopulations based on criteria set by the program STRUCTURE (Pritchard et al. 2000). However, the same program failed to conclusively group the current population. This may suggest very complex interrelationships for which simple classification may not be possible, and thus formation of genetic population groups may not capture the complexity of the population (Zhao et al. 2007). A simulation study (data not shown) showed that by incorporating pedigree into a model, as in this study, two markers that were simulated to be linked with QTL could be correctly distinguished from two markers that were not linked but were associated because of population structure. Without pedigree in the analyses, all four markers were claimed to be significantly associated with QTL.

As in previous reports (Remington et al. 2001; Thornsberry et al. 2001; Crossa et al. 2007; Zhao et al. 2007), we found evidence for spurious associations (Fig. 2). False positives, identified by the exclusion of pedigree in the model, appeared to be trait dependent and were more severe for CCS than for cane yield. We also found evidence for many type I errors, which were not detected if population structure was ignored. This impact of population structure on the model appeared to have a greater effect on the markers identified than the effects of GEI and (or) spatial variation (Fig. 2).

As mentioned above, we were not able to find a reliable method to determine the effect of size of a marker. Thus, we cannot comment on whether the impact of population structure (that is the same for the effects of GEI and (or) spatial variation) also extends to the effect of size. Biased estimates of the effect of size could lead to an inefficient implementation of markers in a breeding program.

GEI

It was surprising to observe a substantial increase in the number of significant markers in our study when GEI was not accounted for, that is, when the three sites were assumed to be independent (Fig. 2). Such a discrepancy (when trials were assumed to be independent vs. correlated) was not expected given the sites selected in the study. These sites (two in the same region) were highly genetically correlated (genetic correlations between 0.63 and 0.94 for cane yield and between 0.67 and 0.78 for CCS; see Table 3), such that they could be served by one selection program. Therefore, data from more trials would not add much more information, which could be confirmed by the high correlation between genetic values predicted from a model without GEI (or assumed independent sites) and a model with GEI (or assumed correlated sites): $r_e = 0.90$ for cane yield and 0.83 for CCS. The substantially higher number of significant markers from models without GEI suggests that many markers were site specific for both cane yield and CCS. Therefore, it is likely that association studies that are based on phenotypic data collected in only one site or that fail to account for GEI would detect a large number of markers specific only to the given site. Such markers are not expected to be effective in a breeding program because they are specific to a site that could not be easily defined in terms of easily measured factors such as rainfall, temperature, and soil attributes. Therefore, markers identified for that site could not be confidently applied at other sites.

It should be noted that we are not seeking to apply markers in all environments over the whole sugarcane-growing area in Australia. We expect that markers detected in this association study are stable enough for regions that are highly correlated genetically, based on phenotypic data, and also to some extent for regions for which limited resources will not allow separate breeding programs. We believe that it is essential to have a relatively stable set of markers that are associated with sugarcane traits over a defined area. To achieve that, more trials must be planted even for a region that is generally regarded as uniform and GEI must be effectively accounted for in association studies.

Spatial variation

The impact of spatial variation within a trial on detecting significant markers was minor. However, correlations between rows and between columns in all three trials were relatively small. In this study correlations for both traits were generally less than 0.3 and mostly close to zero, whereas in other sugarcane studies correlations could be as high as 0.6 (Stringer and Cullis 2002). Therefore, the impact of spatial variation could become larger in trials where neighbouring plots share more similar microenvironments than plots that are far apart.

Future research

The overall objective of this work was to determine if and how DNA markers may be applied to sugarcane breeding programs. While encouraging results have been reported here in terms of providing evidence for marker–trait associations for the commercially important traits cane yield and CCS, one of our major concerns is the repeatability of the marker–trait associations found to date and the probability that marker–QTL recombination will limit marker application in breeding. Further research is planned to investigate these issues. In addition, further studies will also include modelling the expected gains from marker-assisted breeding in recurrent selection for breeding value with multiple markers. Implementation of a marker-assisted breeding program is also envisaged to help quickly validate expected gains and (or) identify unexpected complications in applying DNA markers to sugarcane breeding.

Acknowledgements

This work was funded through the Cooperative Research Centre for Sugarcane Industry Innovation through Biotechnology, BSES Limited, and CSIRO. We thank numerous people who have helped with this work, including people at Diversity Arrays Technology Pty Ltd for provision of the marker data; John Foreman, Michael Hewitt, Hu Fengduo, and others for management of field trials; Karen Aitken for supplying Q165® map information prior to publication; and Andrew George and Emma Huang for helpful discussions about the data analysis.
References


