Whole Genome Analysis of Heifer Puberty in Brahman Cattle

Y. D. Zhang¹, B. Tier¹ and R. J. Hawken²

Introduction

Whole genome association studies have been increasingly used for QTL discovery. Analyses can performed using single marker (for example, single-nucleotide polymorphism, SNP) regression, but have also evolved to multiple marker (SNPs) regression to facilitate genomic selection. A major area of importance to the Northern Australian beef industry is female reproductive rate; specifically the number of calves produced over the lifetime of breeding females. An important component of female reproductive rate is age at puberty of heifers. Reduction of age at puberty of heifers can increase calving rates (Taylor and Rudder 1986). This study compared the two types of methods for estimating SNP effects for age at puberty.

Materials and Method

Data. Brahman cows from the CRC for Beef Genetic Technologies (Beef CRC) were used in this study. Age at puberty (AP) is defined as the age when the first corpus luteum (AGECL) was observed following frequent ovarian scanning. This population consisted of 1007 animals representing 56 half-sib families (Johnston et al. 2009). The heritability of AGECL was estimated at 0.56, with a mean of 750 days. Eight hundred and thirty seven Brahman females were genotyped using the Illumina Bovine SNP50 array (Illumina Inc, Hayward, California). Genotypes were subjected to quality control (GC score, consistency between duplicate samples) and exclusion of monomorphic SNPs, resulting in 50691 SNPs used for whole genome association analysis.

Statistical Models. Three models were used. A base animal model was used on all 1007 records to estimate fixed and additive genetic effects (estimated breeding value, EBV),

\[ y = \mu + Xb + Za + e, \]

with \( \text{Var}(a) = A\sigma_a^2 \) and \( \text{Var}(e) = I\sigma_e^2 \), where \( y, b, a \) and \( e \) were vectors of phenotypic observations, fixed (eg cohort, month of birth, herd of origin etc), additive genetic and residual effects, \( X \) and \( Z \) were incidence matrices relating \( b \) and \( a \) to the observations, respectively. Residuals from this model were calculated as \( y' = y - Xb - Za \) to form a second dataset (APRes).

A second model was used to estimate single SNP effects:

\[ y = \mu + Xb + Za + mg + e, \]

where \( m \) is the SNP genotype, as the number of copies of an allele (0, 1 or 2) and \( g \) is its allele effect, \( y \) is a vector of observations (AP or APRes). Individual SNP effects were estimated for AP and APRes data from genotyped animals only, using the restricted maximum-likelihood

Cooperative Research Centre for Beef Genetic Technologies, Australia.

¹Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351
²CSIRO Livestock Industries, Queensland Bioscience Precinct, Brisbane, Australia
with ASReml software (Gilmour et al. 2006). Furthermore, SNPs with minor allele frequency (MAF) < 0.05 were excluded from post analysis.

The model for jointly estimating SNP effects from multiple regression was:

\[ y = \mu + Xb + Za + \sum_i \mathbf{m}_i \mathbf{g}_i + e, \]

where \( y \) was a vector of the observations of individuals (AP or APRes); \( X, Z, b, a, \) and \( e \) are as in the first model; \( \mathbf{m}_i \) is a vector containing the genotypes (0, 1 or 2) of each individual at locus \( i \), \( \mathbf{g}_i \) is the random allele substitution effect for marker \( i \), and \( e \) is a random vector of residual with \( e \sim N(0, \delta_{ei}^2) \). SNP effect \( g_i \) was solved using a modified BayesB (BayesBFast) method (Fernando, 2008). The prior distribution for the variance of marker effects, \( \delta_{gi}^2 \), was an admixture of distributions with probability \( \pi \) at \( \delta_{gi}^2 = 0 \) and an inverted chi-square distribution with probability \( (1-\pi) \) at \( \delta_{gi}^2 > 0 \). Analysis was implemented using a Markov chain Monte-Carlo of 500,000 cycles with burn-in period of 10,000. The probability \( \pi \) was set to 0.9, i.e. 10% of 50691 SNP were fitted in model jointly for each iteration. Results were captured on every 10 cycles for posterior analysis. SNP effects were calculated as the average posterior mean of effects for SNPs fitted in model.

Results and Discussion

Single SNP. After exclusion of SNPs with MAF<0.05, 2138 of remaining 31539 SNPs were found to be significant (P<0.05) for AP. Although the same model was fitted for AP and APRes, more significant SNPs were detected for AP than for APRes (Table 1). Partial confounding of alleles with fixed effects is most likely the cause of this difference, (eg cohort). The correlation of all significant SNP effects between AP and APRes datasets was high (0.99). Among SNPs showing significant effects (P<0.05) for APRes, 90% of them were identified as also significant for AP. The variance explained by SNP effects (P<0.05) for AP (\( \sum 2pq\alpha^2 \)) was 222198 days\(^2\), 30 times bigger than the REML estimate of additive genetic variance. The corresponding figure was 93379 days\(^2\) for SNPs at a significance of P<0.01. Significant SNPs (P<0.05) for AP were found to be distributed across genome with a number of obvious clusters (Figure 1a).

Table 1. Counts of Significant SNPs for AP and APRes, from single SNP association analyses.

<table>
<thead>
<tr>
<th></th>
<th>N*</th>
<th>P&lt;0.05</th>
<th>P&lt;0.01</th>
<th>P&lt;0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>31539</td>
<td>2138</td>
<td>546</td>
<td>130</td>
</tr>
<tr>
<td>APRes</td>
<td>31539</td>
<td>1307</td>
<td>295</td>
<td>74</td>
</tr>
<tr>
<td>Common</td>
<td>1153</td>
<td>272</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>

*SNPs with MAF < 0.05 were excluded in the post analysis.

Multiple SNP. SNP effects were calculated as posterior means for AP and APRes. The posterior mean of variance explained by SNP effects for AP was 3564 days\(^2\), being 27% of phenotypic (13048 days\(^2\)) or 48% of genetic variance of this trait. The estimated polygenic variance was 3401 days\(^2\). The estimated SNP, polygenic and residual variances for APRes were 3203, 84 and 745 days\(^2\), respectively. The SNP effects from the AP or APRes are
consistent, with a correlation between them of 0.92. The distribution of SNP effects across the genome is shown in Figure 1b. The correlation between EBV and TBV from multiple analyses is high for both AP and APRes, 0.95 vs 0.85, respectively. The correlation between AP and TBV was 0.69, close to that between AP and EBV. Absolute value of SNP effects for AP ranged from 0 to 10.6 days with a mean close to zero, a corresponding range from 0 to 6.0 days for APRes.

Figure 1. Distribution of significant SNPs effects (P<0.05) for AP from single analysis (a, red point represents SNP with P<0.01). The corresponding SNP effects from BayesB analyses are in b.

SNP effects between single and multiple analysis. Comparison between single and multiple analyses is limited to SNP sets either being significant (P<0.05) from the single SNP analyses or in the highest ranked subsets from multiple analyses and with MAF > 0.05. For significant SNPs (P<0.05) in single SNP analysis, Pearson’s correlation coefficients of estimated effects
were 0.83 and 0.86 between single and multiple analyses for AP and APRes, respectively. Most of SNPs in the highest effect set from multiple analyses were also found to be significant from single analyses (P<0.05), as shown in Table 3. Of 546 highly significant SNPs (P<0.01) from single analyses, only 231 SNPs were in the set derived from multiple analyses. Furthermore, the single SNP analysis highly overestimated the effects. For example, average SNP effects for AP in the top 2500 from multiple analyses were 16.0 days for single analyses and 1.1 days for the multiple, and 20.2 vs 1.7 days in the top 500 set.

Table 3. Number of SNPs in the highest ranked subsets from multiple analyses and significant in single SNP analyses.

<table>
<thead>
<tr>
<th>TOP*</th>
<th>AP#</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>P&lt;0.05</td>
<td>P&lt;0.01</td>
<td>P&lt;0.05</td>
<td>P&lt;0.01</td>
</tr>
<tr>
<td>2500</td>
<td>1526</td>
<td>520</td>
<td>869</td>
<td>272</td>
</tr>
<tr>
<td>1000</td>
<td>857</td>
<td>412</td>
<td>500</td>
<td>196</td>
</tr>
<tr>
<td>500</td>
<td>479</td>
<td>317</td>
<td>346</td>
<td>148</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>99</td>
<td>93</td>
<td>57</td>
</tr>
</tbody>
</table>

* N SNPs with the greatest effects in multiple analyses; # number of SNPs at 5% or 1% significance from single analyses represented in the TOP subsets.

Conclusion

Single SNP association analyses overestimated SNP effects and is not appropriate for estimating genomic breeding values. Variance explained from single SNP effects is overestimated, while the estimate from multiple SNP effects was at a similar scale of the REML estimate. Use of phenotype values could produce better SNP effect estimates than use of residuals. This study demonstrated advantages of multiple SNP analysis over the single SNP approach, particularly for whole genome selection, because accurately estimated SNP effect is an important component for efficient genomic breeding value prediction.

References


