A Single-Step Hybrid Marker Effects Model Using Random Regression for Stayability in Hereford Cattle

Theta Solutions, LLC 10615 San Marcos Rd, Atascadero, CA 93422 USA

1Theta Solutions, LLC, 10615 San Marcos Rd, Atascadero, CA 93422, USA
bruce@thetasolutionsllc.com (Corresponding Author)
2Agricultural Business Research Institute, University of New England, Armidale, NSW, 2351, Australia
3American Hereford Association, P. O. Box 014059, Kansas City, MO 64101, USA
4School of Agriculture, Massey University, Ruakura, NZ

Summary

Stayability in beef cattle can be one of the most important economically relevant traits and this is often reflected in maternal genetic indexes. However, as implemented in the U.S. beef industry, the accuracy of prediction remains relatively low for most bulls for most of their productive life (i.e., until sufficient numbers of calves are produced by daughters). A random regression approach to stayability has been proposed that can reduce the age at which bulls achieve high accuracy predictions by using data on all ages of cows and this method can incorporate genomic information. The objective of this study was to determine if implementation of stayability prediction using random regression in a single step hybrid marker effects model was feasible and effective for national cattle evaluation.

Calving records on 272,651 cows through twelve years of age were obtained from the American Hereford Association (AHA). Pedigree records included 1,158,750 animals of which 29,441 were genotyped and imputed to 54,947 SNP loci. Two Bayes C analyses were performed on a model that included random effects of marker and extra polygenic effects. The random regression was fit as a two-degree polynomial for each random effect in the model. Besides additive genetic direct effects which were partitioned into marker effects and an additional polygenic effect, permanent environmental effects and contemporary group effects were fitted as random. Two values of \( \pi(.95,.99) \), the prior probability a marker effect was 0, were evaluated.

Window variances were obtained for 1 megabase windows of markers. In the analysis when \( \pi = .95 \) 297 marker windows entered the model in at least 95% of the Gibb’s samples with 41 of the windows entering the model 100% of the time. When \( \pi = .95 \) the largest proportion of variance was described by a window on chromosome 3. That window accounted for 0.99% of the genetic variance accounted for by markers and was included in the model in 100% of the samples. When \( \pi = .99 \) the number of windows that entered the model 100% of the time was reduced but the proportion of additive genetic variance accounted for by the windows with largest effects was increased.

A random regression single step hybrid marker effects model implemented using a Bayes C sampling approach on large datasets was feasible. It resolved important quantitative trait loci and included extra polygenic effects. This approach should be particularly useful in improving the accuracy of prediction of stayability for young genotyped selection candidates.

Keywords: genetic prediction, hybrid marker effects models, single step, stayability, cattle
Introduction

Beef cows that have a greater chance of breeding each year improve herd profitability by reducing the cost of developing replacement heifers, reducing whole herd calving difficulty due to more older females, and having more calves sold at heavier weights (Garrick, 2006). Alternative genetic predictions have been proposed to improve mature cow productivity including longevity, survival, and stayability. Stayability (Snelling, et al., 1995) has been implemented in several U. S. beef cattle genetic evaluation programs and can be used to identify parents whose daughters would be more likely to remain productive.

Alternative statistical models can be applied to stayability data. Proportional hazards models can be used to predict the probability that a cow will survive to time \( t_m \) given that it has survived to time \( t_{m-1} \) (Ducrocq et al., 1988). However, implementation of that model has been limited to single trait sire models. Additionally, it would be challenging to include genomic information.

Most implementations of stayability in the U.S. use maximum a posteriori nonlinear mixed models. In those implementations, stayability is measured as a binary observation where 0 is assigned when a cow fails to breed and 1 is assigned to a cow that successfully calves at a given age. The limitations of this approach include the large numbers of iterations to solve the linear system in between rounds of Newton-Raphson iteration, and a requirement that contemporary groups have variation to be included in the analysis. Traditionally, stayability was expressed as the probability that a cow remains in the herd to a given age, typically 6 yr old, given that she had a calf as a 2 yr old. In this model, there are no phenotypic observations from daughters until they reach 6 years of age. Accordingly, Brigham et al. (2007) proposed a selection index approach to combine predictions of stayability to consecutive earlier ages and improve accuracy of bulls with daughters that have reached 3 to 6 years of age.

More recently Jamrozik et al. (2013) proposed a random regression (RR) approach to predict stayability to consecutive calvings. That RR model is appealing because it can easily accommodate observations at multiple cow ages, while handling missing values. A prediction to a specific desired age can be made on the trajectory predicted for each animal. The RR model can easily accommodate multiple traits and additional random effects, including marker effects models (MEM; Fernando, et al., 2016), and data from contemporary groups with no variation contribute to the analysis.

Previous work (Spiedel et al., 2017) has demonstrated the potential for identifying QTL for stayability. Defining a RR as a marker effects model (MEM; Fernando, et al., 2016) is straightforward.

Our objective was to determine if implementation of stayability using RR in a MEM in national cattle evaluation was feasible and effective.

Materials and methods

Data

Calving records on 272,651 cows through twelve years of age were obtained from the American Hereford Association (AHA). Because stayability is defined as sustained cow fertility given the cow had a first calf as a two-year-old, the 896,374 repeated cow calving records did not include calving observations from females when they were 2 year old first calf heifers. Pedigree records included 1,158,750 animals. Only data used since the
implementation of the AHA’s Whole Herd Total Performance Recording (TPR) program were used.

Genotypes on 29,441 AHA registered animals from several different densities of bead array SNP chips were imputed to a common density of 54,947 SNP loci using FImpute (Sargolzaei, et al., 2014) and Bolt (Golden and Garrick, 2017). Only loci assigned to autosomes and those with minor allele frequency >0.10 were subsequently used. The final genomic data used in the analysis included 42,635 SNP loci.

Model

Two BayesC analyses were performed with $\pi = .95$ or $\pi = .99$ using the Bolt software package (http://www.thetaSolutionsLLC.com). The two analyses were performed to determine if similar marker effects were resolved with a more rigid prior probability of marker effects being 0. The MEM fit was,

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = Xb + C^0c^0 + C^1c^1 + P^0p^0 + P^1p^1 + D^0d^0 + D^1d^1$$

$$+ \begin{bmatrix} Z_n^0 \\ Z_g^0M_g \end{bmatrix} \begin{bmatrix} u_n^0 \\ \alpha_n^0 \end{bmatrix} + \begin{bmatrix} Z_n^1 \\ Z_g^1M_g \end{bmatrix} \begin{bmatrix} u_n^1 \\ \alpha_n^1 \end{bmatrix} + e$$

(1)

where $y_n$ ($y_g$) was a vector of annual 0 or 1 stayability observations on non-genotyped (genotyped) individuals; $X$ was a matrix relating fixed effects of year of record and covariate of age of cow ($b$) to observations in $y$; $C^0$ ($C^1$) was an incidence (covariable) matrix relating the intercept (slope) of random contemporary group effects, $c^0$ ($c^1$), to observations in $y$; $P^0$ ($P^1$) was an incidence (covariable) matrix relating the intercept (slope) of random permanent environment effects, $p^0$ ($p^1$), to observations in $y$; $D^0$ ($D^1$) was an incidence (covariable) matrix relating the intercept (slope) of independent additive random extra polygenic effects, $d^0$ ($d^1$), to observations in $y$; $Z_n^0$ ($Z_n^1$) was an incidence (covariable) matrix relating intercept (slope) animal additive direct genetic effects accounted for by the imputed SNP markers, $u_n^0$ ($u_n^1$) for non-genotyped animals; $Z_g^0$ ($Z_g^1$) was an incidence (covariable) matrix relating intercept (slope) marker effects, $\alpha^0$ ($\alpha^1$), to observations in $y_g$; and $M_g$ was the matrix of marker genotype values of -1, 0 or 1 at each locus indicating homozygous, heterozygous or opposite homozygous genotypes.

The extra polygenic additive genetic effect terms, $d$, were fit assuming .5 genetic variance was not captured by markers, and these effects were uncorrelated to the other random genetic effects. The contemporary groups were constructed according to Jamrozik et al. (2013). Six additive genetic groups were included using the method of Westell et al. (1988).

Jamrozik, et al. (2013) found that a 4-degree polynomial had the best fit, but in previous work with this dataset we determined that the model was poorly conditioned. Therefore, we obtained variance components for the 2-degree polynomial from J. Jamrozik (personal communication) to apply to this study.

The MEM mixed model equations (MME) constructed were those from the hybrid model of Fernando et al. (2016) but with the addition of extra polygenic effects not explained by markers. The MME were first solved using a preconditioned conjugate gradient (PCG)
method and then four parallel BayesC Gibb’s single site samplers were seeded with the PCG solutions. A total of 40,000 samples were obtained, after seeding with different random numbers, and discarding 1,000 samples (from each parallel chain) for burn in. The method of Golden et al. (2015) was used to perform the Gibb's sampling on a workstation with 4 Nvidia Titan X (P) graphics processing units on an ASRock X99 Extreme11 motherboard, 128G of memory and an Intel Xeon E5-2643 processor.

**Results and discussion**

The entire time to setup and solve this problem using Bolt (Golden and Garrick, 2017) was just over four hours, with approximately three hours required to perform the Gibb’s sampling.

Marker samples for stayability to age 9 were calculated from each intercept and slope sample’ sample of marker effects. Window variances were obtained for 1 megabase windows of markers (Sun, et al., 2011) and are shown expressed as a proportion of genetic variance in figure 1 for \( \pi = .95 \) and figure 2 for \( \pi = .99 \) . Figures 3 and 4 show the posterior probabilities of inclusion in the model for all the marker effects.

In the analysis when \( \pi = .95 \) there were 297 marker windows that entered the model in at least 95% of the Gibb’s samples with 41 windows entering the model 100% of the time. Chromosome 2 had the greatest number (5) entering the model in 100% of the samples. Four of those 5 windows were clustered towards the beginning of the chromosome. The windows with 100% PPI were at 5, 10, 15, 19 and 30 Mb. Twenty three of the autosomes had at least 1 locus included in the model in 100% of the Gibb’s samples.

When \( \pi = .95 \) the largest proportion of variance was described by a window on chromosome 3 that began at 89.7 Mb and ended at 90.7 Mb. That window accounted for 0.99% of the total variance accounted for by the markers and was included in the model in 100% of the samples. Several coding regions in this window of the bovine genome have been identified (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=43351630#map) and include 2 genes, C8A and C8B involved in immune response and PRKAA2, which has been associated with Polycystic Ovary Syndrome in women but Sproul et al., (2009) showed no such association although the protein product associated with this gene is expressed in mammalian ovaries.

The second largest signal when \( \pi = .95 \) was on chromosome 1 between base 89.1 Mb and 90.1 Mb, and accounted for 0.81% of the variation due to markers and entered the model for 97% of the samples. Other nearby windows also showed signal and were likely a result of the same effect. This window has virtually no known identified coding genes, however the adjacent prior window, which also had strong signal, has at least 10 identified genes.

When \( \pi = .99 \) the number of windows that entered the model in at least 95% of the samples was reduced to 141. Only 38 of the loci entered the model in 100% of the Gibbs samples. Chromosome 1 had the greatest number of windows (6) entering the model 100% of the time, with chromosomes 2 and 3 having 4 windows entering the model in 100% of the samples. Greater proportions of variance accounted for by windows was achieved with \( \pi = .99 \) . The number of autosomes that had windows entering the model at least 100% of the time was reduced to 18 with 38 total windows entering the model 100% of the time. Virtually the same regions showed important associations with additive genetic effects on stayability with \( \pi = .99 \) versus \( \pi = .95 \).
Conclusions

The hybrid MEM of Fernando et al. (2016) implemented as a random regression model and using a Bayes C sampling approach on large stayability datasets is feasible. It appears to resolve important quantitative trait loci and can include extra polygenic effects. This approach should be useful in improving the accuracy of prediction of stayability for young genotyped selection candidates.

This study demonstrates that Gibbs sampling for large analyses is feasible using a relatively low cost workstation with high performance graphics processing units. An additional advantage of using sampling is that it provides direct assessment of the prediction error variances and covariances. Estimates from current methods for approximating prediction error variance can exhibit bias and cannot be readily used to approximate prediction error covariances.

Future studies should determine the assumed $\pi$ that maximizes accuracy of prediction. We propose that in production analyses only those loci from windows that enter the model at high frequency be fit, along with an extra polygenic effect. Accuracy nearly equal to the Bayes C analysis can be achieved with less computational effort.

Figure 1. Proportion of total marker effect variance due to 1 megabase windows for the 29 autosomes with $\pi=0.95$. Chromosomes are consecutive and indicated by the alternating red/blue colour. Windows are consecutive within chromosome.
Figure 2. Proportion of total marker effect variance due to 1 megabase windows for the 29 autosomes with π=.99. Chromosomes are consecutive and indicated by the alternating red/blue colour. Windows are consecutive within chromosome.

Figure 3. The posterior probability of inclusion of the marker effects windows for π=.95.
Figure 4. The posterior probability of inclusion of the marker effects windows for $\pi=.99$. 
List of References


