AN EFFICIENT BAYESR THAT HAS COMPARABLE SPEED TO GBLUP AND ENABLES LARGE MULTI-TRAIT ANALYSES

E.J. Breen[1], I.M. MacLeod[1], P.N. Ho[1], J.E. Pryce[1,2], H.D. Daetwyler[1,2] and M.E. Goddard[1,3]

[1] Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia; [2] School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia; [3] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC, 3052 Australia

SUMMARY

Bayesian methods of genomic prediction are often more accurate but much slower than GBLUP. Here we present a new method (BayesR3) for implementing BayesR which is 18 times faster than previous versions of BayesR and is comparable in speed to GBLUP.

INTRODUCTION

As Bayesian MCMC based inference procedures for genomic analysis are considered slow the aim of this study was to produce a faster implementation of uni- and multi-variate BayesR (Erbe et al. 2012; Kemper et al. 2018). BayesR's mixture model a prior assumes each SNP is drawn from a mixture of 4 zero mean normal distributions of SNP effects:

$$p(g_j|\pi,\sigma^2_\Sigma) = \pi_1 \times 1_1 \times N(0, 0 \times \sigma^2_\Sigma) + \pi_2 \times 1_2 \times N(0, 10^{-4} \times \sigma^2_\Sigma) + \pi_3 \times 1_3 \times N(0, 10^{-3} \times \sigma^2_\Sigma) + \pi_4 \times 1_4 \times N(0, 10^{-2} \times \sigma^2_\Sigma)$$

Where $1_k = \begin{cases} 1 & \text{with probability } \pi_k \\ 0 & \text{with probability } 1-\pi_k \end{cases}$; $p(g_j|\pi,\sigma^2_\Sigma)$ gives the prior probability that SNP j has an effect, given π and σ^2_Σ; $\sum_k \pi_k = 1$; σ^2_Σ is the sum of all the genetic variances.

A simple blocked Gibbs sampling approach to estimate SNP effects, is given by:

$$y = V_1 g_1 + V_2 g_2 \ldots + V_B g_B + e$$

Where y is a vector of phenotypes, V_i the genotypes in block i, g_i the SNP effects and e the residual error vector. In block i, the right-hand sides $r = V_i' e_i$ are calculated. Each SNP's least squares estimate at the kth iteration, k in the set {1, 2, ..., b}, is given by:

$$\hat{g}_j^k = \frac{r + V_{ij}' V_{ij} \hat{g}_j^{k-1}}{V_{ij}' V_{ij}}$$

After which, $\hat{g}_j^k$ is sampled as specified by (Erbe et al. 2012). $V_{ij}' V_{ij}$ is the diagonal element of block i. After each SNP has been sampled the right-hand sides, r, are updated for SNP $j+1$ by adding to it: $(V_i' V_i)_j (\hat{g}_j^{k-1} - \hat{g}_j^k)$, where $(V_i' V_i)_j$ is the jth column of $V_i' V_i$. After all block SNPs have been processed sequentially b-times, e is then updated for block i+1:

$$e_{i+1} = e_i - V_i (g_i^b - g_i^1)$$

RESULTS

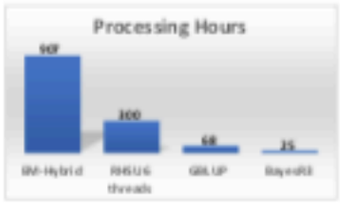

Figure 1: Comparison of actual processing times in hours between BM-Hybrid, R4SU 6 threads and GBLUP and BayesR3. Data set given in Figure 2 and chain length 100,000. Colus' Right Hand Side Updating (R4SU Colus 2014) was run using Gerhard Moser's (2015) fortran 90 bayesR2.f90 program available at https://github.com/gmoser/bayesR. This program contains several runtime options, so we applied the recommended default settings, with block size of 6.

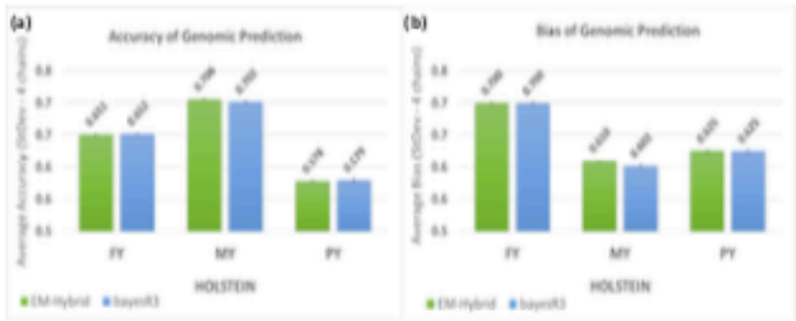

Figure 2: (a) Accuracy = correlation (phenotype, Genomic Breeding Value) for the Bayes R (BM-Hybrid (green) and BayesR3 (blue) using 633, 374 SNPs and 97,634 dairy cows and bulls. (b) Bias given by the slope of the regression. Chain length 100,000, Block size 75. FY for yield, MY milk yield and PY protein yield. The data set was composed of Holstein (68%), Jersey (14%) and crossbreds (17%). Validation set: 1251 Holstein Bull's (all validation animals had no sires in the reference population)

Figure 3: Manhattan plot of posterior probabilities (PP > 0.05) for Multi-trait BayesR3 analysis of Mid-Infrared (MIR) spectra from 5131 individual cow milk samples, analysed as 20 PCA components. Square symbols indicate SNP from the MY single trait analysis that had a PP > 0.8 and overlapped with the multi-trait analysis. Several of these SNP are in or very close to annotated genes known to influence milk composition. MCMC chain length 60,000 with burn in 30,000. For the high density array the processing time was 14 hours. Milk samples from Holstein (80%), Jersey (6%), and cross-breed cows (8%) collected in Spring 2017 from 21 commercial herds were analysed for milk composition by an infrared spectrometer. Each spectrum included 899 data points, representing the absorption of infrared light through the milk sample at wavelengths from 649 to 3,999 cm-1 regions. The spectrum were filtered to produce 537 wavenumbers per sample, and from a PCA of these wavenumbers the 20 components explaining 95% of the variance formed the 20 PCA traits. The fixed effects in the model were the mean, herd ID (34 level's), breed (70 levels) and age at calving (covariate in months).

DISCUSSION

Here we have presented a much faster method for implementing BayesR. The increase in speed is achieved by sampling the SNP effects in blocks and cycling through the SNPs within a block a number of times before moving to the next block. That is, there are inner cycles within a block and outer cycles among the blocks. This allows us to take a given number of samples the SNP effects, using Gibbs sampling, much faster than previously possible.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114-29.
Colus MR. Right hand-side updating for fast computing of genomic breeding values. Genet Sel Evol. 2014;46:24.
Kemper KE, Bowman PJ, Hayes BJ, Visscher PM, Goddard ME. A multi-trait Bayesian method for mapping QTL and genomic prediction. Genet Sel Evol. 2018;50(1):10.
Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. 2015;11(4):e1004969