A Comparison Of Various Methods For The Computation Of Genomic Breeding Values Of Dairy Bulls Using Software At Genomicselection.net

R. Mrode[*], M.P. Coffey [*] ,I. Strandén[¤], T. H.E. Meuwissen[±], , J. B.C.H.M. van Kaam[§], J.F. Kearney[†],D.P. Berry[‡],

Introduction

One obstacle to the implementation of new methodologies for genetic evaluation at a national level is the availability of reliable software that can handle large data sets. This implies that countries with human and technical resources available are usually the first to implement such new methodologies and often offer their software for use to other countries at a cost. The implication is a slow rate of adoption of new technologies, with the consequence that the international genetic evaluation centre (Interbull) are compelled to handle data from different models and methods. One of the possible hindrances for the implementation of recent advances in the use of single nucleotide polymorphisms (SNPs) data for the prediction of breeding values is relevant and reliable software. In an attempt to alleviate this problem, a number of interested scientists formulated a so called Clubware group for the development of relevant software for the computation of genomic breeding values. Such software, after due checking, will be made available at the website www.genomicselection.net. The purpose of this paper is to review the array of programmes at the website and report on their implementation for the estimation of direct genomic values (DGVs) using data from Ireland for milk, fat and protein yields, calving interval (CI) and survival (SUR). The computational time and accuracy of the DGVs from several methods are presented. In addition, the relative contributions of SNPs to DGVs at various gene frequencies are examined.

Material and methods

Genotypic and phenotypic data editing. Genotype data on 54,001 SNPs from the Illumina Bovine50 Beadchip on 1,096 Holstein-Friesian AI sires with daughters in Ireland were available. A total of 2,419 SNPs on the X-chromosome or without known position on the genome were discarded; a further 234 SNPs that did not follow Mendelian heritance patterns between sires and sons were also discarded. Genotypes of all 1,096 individuals were

[*] Scottish Agricutural College, Sir Stephen Watson Building, Bush, Penicuik, EH260PH
[†] Irish Cattle Breeding Federation, Highfield House, Bandon, Co. Cork, Ireland
[¤] MTT Agrifood Research, FIN-31600 Jokioinen, Finland
[±] Norwegian University of Life Sciences, N1432 As, Norway
[§] ANAFI- Italian Holstein Association, Via Bergamo 192, Cremona, Italy
[‡] Moorepark Dairy Production Research Center, Fermoy, Co. Cork, Ireland

compared between sire-son pairs, where available, for parentage verification. Comparison of expected average relationships and genomic relationships based on the genotype data were also undertaken as described later to validate relationships. Of the remaining SNPs, 3,776 were monomorphic and a further 3,423 had a minor allele frequency of <2%; these SNPs were discarded. SNPs were also discarded (in this order) if greater than 5% of calls were missing (n=1,566), SNPs were not segregating (n=15), SNP call clustering was poor (n=9), and SNP calls deviated ($P<0.1 \times 10^{-7}$) from Hardy-Weinberg equilibrium (n=322). Following all edits, 42,265 SNPs remained for inclusion in the analysis.

Predicted transmitting abilities (PTAs) with associated reliabilities and daughter yield deviations (DYDs) (expressed on the scale of PTA) for a range of performance traits evaluated by the Irish Cattle Breeding Federation in the January 2009 domestic genetic evaluations were available. DYDs for 305-day milk, fat and protein yield as well as geometric mean SCS ($\log_e$ somatic cell count) are estimated in Ireland using a repeatability animal model across the first five lactations. PTAs for calving interval and survival are estimated using a multi-trait animal model, including data from the first three lactations. PTAs for milk yield are used to adjust PTA for survival for differences in genetic merit of milk yield; hence, this survival trait is functional survival. Parental contribution to the reliability of each DYD or PTA was removed using the approach of Harris & Johnson (1998) and this adjusted reliability was used as a weight in the genomic evaluation where specified. PTAs were deregressed by dividing by their respective reliabilities.

Genomic evaluation methods. SNP effects and DGVs were estimated using linear models and Bayesian approaches. Two linear models were used: 1) BLUP with SNP effects assumed to be random and with equal variances (BLUP1), and 2) BLUP where the additive relationship matrix was replaced by a genomic relationship matrix (BLUP2).. The Bayesian methods were 1) BayesA (Meuwissen et al., 2001), 2) BayesB (Meuwissen et al., 2001), 3) fastBayesB (Meuwissen et al. 2009) and 4) BayesA modified to include a polygenic effect (BayesA-P). For the Bayesian methods, the MCMC chains were run for 80000 cycles with the first 24000 discarded as the burn-in period. In the case of BayesB, 20 Metropolis-Hastings cycles were evaluated within each MCMC chain. The proportion of SNP variances set to zero in BayesB was 0.66. A Sun workstation VG800 with 32 GB of Memory and eight 5GHZ processors was used for the analyses.

The relative contribution of each SNP to the DGV each animal for milk yield were computed for BLUP1 , BayesA and BayesB and averaged at three frequencies of the minor allele; high (0.83), medium (0.51) and low (0.16) to see how these differ among the methods.

Testing of accuracy of genomic evaluations. The 1,096 AI sires were divided into a training and validation dataset. The number of sires used per trait differed depending on the sire's number of daughters per trait in Ireland. Only sires with an adjusted reliability of >40% were used as the training population. The validation dataset for milk production traits were bulls born post-1996 and which had at least 40 Irish daughters. Validation data for calving interval and survival were bulls with at least 40 Irish daughters with a reliability for calving interval and survival of at least 65% and born after 1995 and 1994, respectively.

Statistics used to determine the accuracy of the genomic evaluation were based on the validation dataset and included 1) the correlation between the DGVs and the traditional EBVs (tEBVs), 2) the regression of the tEBVs on DGVs, as well as the 3) mean, and 4) standard deviation of the difference between the DGVs and tEBVs.

Results and discussion

Table 1: Correlation and regression of DGV on tEBV as well as the mean and standard deviation (SD) of the bias

Trait	Method					
	BLUP1	BLUP2	BayesA	BayesA-P	FastBayesB	BayesB
Milk yield						
Correlation	0.68	0.69	0.70	0.72	0.67	0.65
Regression	1.31	0.78	0.99	1.49	1.30	0.98
Mean bias (kg)	-31.2	18.1	-4.6	-165	4.4	-6.6
SD of bias	147.8	146.8	141.4	144.6	148.7	149.6
Fat yield						
Correlation	0.65	0.67	0.68	0.62	0.65	0.58
Regression	1.32	0.8	0.89	1.09	1.35	0.58
Mean bias (kg)	-3.82	-0.48	-1.01	-318	-3.80	-0.50
SD of bias	4.89	4.74	4.53	4.92	4.88	5.73
Protein yield						
Correlation	0.67	0.69	0.67	0.68	0.67	0.53
Regression	1.31	0.78	0.93	1.36	1.34	0.37
Mean bias (kg)	-2.21	-1.75	-1.02	5.09	-2.37	-0.74
SD of bias	4.05	4.01	3.96	4.06	4.06	6.59
Calving interval						
Correlation	0.70	0.71	0.66	0.69	0.70	0.49
Regression	1.18	0.75	0.77	1.29	1.19	0.3
Mean bias (d)	4.02	0.28	0.11	-41.5	3.93	0.23
SD of bias (d)	2.35	2.43	2.55	2.42	2.36	4.71
Survival						
Correlation	0.58	0.59	0.56	0.57	-	0.57
Regression	1.22	0.76	0.93	1.48	-	1.10
Mean bias (%)	-1.6	-0.45	-0.45	25.3	-	-0.61
SD of bias (%)	1.46	1.47	1.47	1.49	-	1.70

Most of the software at http://www.genomicselection.net was initially developed using the small example in Ben Hayes lecture notes (Hayes, 2007) and this study involves their first usage in a real and large data set. The computation time for BLUP1 was about 5 minutes but the time required for BayesA, BayesA-P and BayesB was 14.6, 22.9 and 65 hours, respectively. However fastBayesB took only 49 minutes for milk but needed significantly less time of about 5 minutes for all other traits. With the exception of BayesB, these times should be acceptable for national genomic evaluations although the time required is expected to increase with the number of animals genotyped. Statistics on the accuracy of the different methodologies are summarised in Table 1. The correlation between DGVs and tEBVs for the production traits varied from 0.53 for 0.72 and were lower for BayesB. The correlations between the DGVs and tEBVs were slightly lower for CI and SUR, ranging from 0.49 to 0.70. The lower correlations may be attributable to the smaller training population size and lower reliabilities of the tEBVs. BLUP1 and BayesA-P, on average, over predicted the tEBVs, with the regressions of DGV on tEBVs larger than 1 for most traits but fastBayesB and BayesB were lowest for most traits. The mean and SD of the bias between DGV and tEBVs were consistent across methods except the unexplainable large mean difference of BayesA-P for milk and fat. An error was found in the set up of BayesB for survival at a late hour, hence no results. The mean relative contribution (RC) of each SNP to DGV (standard deviation (SD) in brackets) for milk yield for BLUP1, BayesA and BayesB were 8.6 (158), 13.2 (78) and 7.8 (58) kg for SNPs at high frequency respectively. The corresponding values for the medium frequency were -0.84 (96), 1.15 (70) and -0.84 (56) and -0.88 (35), -1.70 (21) and -1.64(17) for the low frequency. The RC increased with allele frequencies as expected given the coding of genotype in this study and were in general similar for three methods but the lower SD of RC for BayesA and BayesB might be due to use of different variances for each SNP and more so for the later with small SNP effects set to zero.

Conclusion

The results from the initial use of software programmes in www.genomicselection.net indicate that they handle large datasets adequately and the site has the potential for providing dependable and working software for calculating DGVs. The accuracies from the various linear and non-linear methods utilised have been similar although there are some differences in predictive ability. The mean relative contribution of SNPs increased with allele frequency but was more varied for the BLUP method, may be due to the assumption of same variance for all SNPs.

Acknowledgements

The contribution of other scientists in Club-ware: R. Thompson, J. Woolliams, E. Mantysaari, H. Jorjani and E. Bongcam-Rudloff and Ben Hayes for his original programs.

References

Hayes. B. J. (2007) Short course notes, Iowa University, USA

Harris B and Johnson D. (1989) Interbull *bulletin* 17:31-36

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). *Genetics*. 157:1819-1829.

Meuwissen, T.H.E., Solberg, T.R., Shepherd, R. and Woolliams, J.A. (2009) *Genet. Sel. Evol*. 41:2.