Estimation Of The Proportion Of Variation Accounted For By DNA Tests. I. Genetic Variance

K. J. Hanford[*], R.M. Thallman[†], S. D. Kachman[*], L. A. Kuehn[†], R. L. Quaas[‡], R. J. Tempelman[§], R. L. Fernando[**], and E. J. Pollak[†].

Introduction

The proportion of additive genetic variation accounted for by a DNA test is a useful metric with which to quantitatively evaluate the merit of commercial DNA tests for marker-assisted selection (**MAS**) of seedstock. Until now, the estimation of this statistic has been considered a difficult problem. DNA test results can be presented in the form of molecular breeding values (**MBV**), which are continuous values, intended to predict the breeding values of animals based only on the DNA test results. They are typically expressed in units of the trait and assumed to be scaled equivalent to twice the EPD. However, in practice, they are often scaled differently. There should be an MBV for each trait that a DNA test is capable of predicting. Most current commercial DNA tests for quantitative traits in beef cattle are expressed as MBV (or closely related values), although some companies may use different names for them. In Thallman *et al.* (2009), a theoretically desirable estimator was derived that should be computationally feasible for data sets of the size likely to be useful for estimation. It is computed from a model that has desirable properties for the inclusion of MBV in the national cattle evaluation (**NCE**) system. This estimator of the proportion of additive genetic variation due to MBV (R_g^2) is equal to the square of the estimated genetic correlation (r_g) and is directly related to the variances and covariances that are required to incorporate MBV into NCE as described by Kachman (2008) and applied by MacNeil *et al*. (2009). Inclusion of MBV into NCE in lieu of genotypes is necessary because NCE does not have access to individual SNP genotypes associated with the commercialized tests. Analyses used for the independent validation of DNA tests have typically been conducted using a single trait sire model for the target trait in which the MBV is included as a covariate (Van Eenennaam *et al*. (2007)).

In this paper, we evaluate an estimator of the proportions of genetic variance from a two-trait animal model with the MBV and observed phenotype included as correlated traits. This methodology can readily be expanded to include multiple traits, so will be referred to as MT. We will compare this with the typical single trait sire model, where the estimator is the reduction in sire variance when compared to a reduced model in which the MBV covariate is dropped from the full model (RV). DNA tests also have potential for prediction of genetic merit of commercial cattle for application in marker-assisted management (**MAM)**, although

[*] University of Nebraska, Lincoln, NE, USA
† USDA-ARS U.S. Meat Animal Research Center, Clay Center, NE, USA
[‡] Cornell University, Ithaca, NY, USA
[§] Michigan State University, East Lansing, MI, USA
[**] Iowa State University, Ames, IA, USA

most tests currently being marketed are for use in MAS. Thallman *et al.* (2010) evaluated estimators for MAM.

Materials and Methods

For each of three levels of narrow sense heritability of y (h^2_{gy}) (0.1, 0.3, and 0.5) and four levels of R_g^2 (0.1, 0.2, 0.4, and 0.6), 500 independent replicates were simulated. Each replicate consisted of 1000 offspring. In NCE, the number of offspring per sire varies greatly; hence, a binomial random number generator was used to randomly determine the number of offspring for each sire with the proportion of success being defined as the inverse of the number of sires remaining. Therefore, the number of sires in each population ranged from 98 to 100, with the number of offspring per sire ranging from 1 to 27. Sires and dams (unknown) were considered unrelated. Each of the progeny was randomly assigned to one of 20 contemporary groups. The phenotypes for the replicates with only additive effects were generated as the sum of a contemporary group effect, a sire effect, an additive genetic component of the residual, and an environmental component of the residual. Sire effects and residual effects were generated separately as bivariate normal random variables with mean zero and scaled so that the phenotype had a phenotypic variance of one and the MBV had a phenotypic variance of $h^2_{gy} \times R_g^2$. For each of the same twelve combinations of h^2_{gy} and R_g^2, an additional 500 replicates with non-additive genetic effects were generated so that the narrow sense heritability of MBV (h^2_{gm}) was equal to 0.8, both the additive and non-additive genetic correlations were equal to $r_{g.}$, the phenotypic variance of the phenotype was one, and the phenotypic variance of the MBV was $h^2_{gy} \times R_g^2$. The single trait analyses were conducted using PROC Mixed of SAS[††] with an option that allows negative estimates of variances. Those estimates that fell within the parameter space were REML estimates and those that fell outside the parameter space were not REML. The two-trait analyses were conducted using ASReml[††]. Those parameters were within the parameter space with the exception of some numerical problems when the additive variance was on the boundary at zero in a preliminary analysis. The two-trait model was run for each replicate with the residual correlations estimated.

Results and discussion

Comparisons of the two alternative estimators are presented in Table 1. Those data simulated with low h^2_{gy} had more problems with convergence and unreasonable estimates. At h^2_{gy} =.1, the combinations with higher R_g^2 had more problems with convergence, due to the estimates of $r_{g.}$ converging to 1, resulting in a singular genetic covariance matrix. Therefore, any MT model analysis that didn't converge was not included in the summarizations. At h^2_{gy} = 0.1 and R_g^2 = 0.6, the number of non-converging replicates was close to one third (146 out of

[††]Reference herein to any specific commercial products by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

500). For the RV model, any replicate where the percent reduction in sire variance was greater than 5000% was not included in the summary and any negative percent reduction was set to zero. When simulated under an additive effects-only model, across the range of parameter sets considered, the mean of the estimated R_g^2 with the MT model tended to be closer to the true values than the mean of the estimated R_g^2 with the RV model, although both estimators performed reasonably for most parameter sets. However, the MT estimator tended to be biased up when the true value of R_g^2 was close to zero, relative to the mean standard error. This was particularly noticeable at low heritability, whether non-additive effects were simulated or not. At h^2_{gy} = 0.1 and R_g^2 = 0.6, the MT estimator was seriously biased down (regardless of whether non-additive effects were simulated (0.522) or not (0.524). This problem is likely to become more pronounced in smaller data sets. However, the RV estimators can produce negative estimates, as well as estimates > 1 of the proportion of variance explained. Provided REML is used to estimate the (co)variance parameters, the MT estimator has the statistical properties of REML estimators, including the advantage of producing estimates within the parameter space. The root mean squared errors (**RMSE**) of MT estimator were uniformly smaller than RV estimator. An additional advantage of the MT estimator is that its standard error can be readily computed as twice the absolute value of the genetic correlation times the standard error of the genetic correlation. This is not feasible for the RV estimator because it is computed from different analyses of the same data. The problem with the additive variance of the observed trait being nearly zero obviously occurs much more frequently when the heritability is low and/or the standard error is too large because of population size and structure. Considerably larger populations will be required for traits with low heritability.

Conclusions

The squared genetic correlation between the observed trait and the MBV from the MT model appears to be a better estimator of the proportion of genetic variation accounted for by a DNA test (in the form of an MBV), than that from the RV model. The estimation and reporting of proportion of genetic variation accounted for by DNA tests will enhance the incorporation of DNA testing into the national cattle evaluation system, and allow the beef industry to utilize this technology much more effectively and extensively than it is currently being utilized. The proportion of variation accounted for by DNA tests should be a very useful tool for cattle producers to use in determining the value of DNA tests in their breeding programs and production systems. The first example of the MT estimator in independent validation is at http://www.beefcrc.com.au/Aus-Beef-DNA-results (accessed 2/25/10).

References

Kachman, S. (2008). *Proc. of the 9th Genetic Prediction Workshop, BIF, Kansas City, MO, Dec. 8-10*: 92-98.

MacNeil, M.D., Nkrumah, J.D., and Northcutt, S.L. (2009). *J. Anim. Sci*., 88:517-522.

Thallman, R., Hanford,. K., Quaas, R., *et al*. (2009). *Proc. of the BIF 41st Ann Res Symposium, Sacramento, CA, Apr. 30-May 3*: 184-209.

Thallman, R., Hanford,. K., Kachman, S., Kuehn, L., Quaas, R., Tempelman, R., Fernando, R., and Pollak, E. (2010). In *Proc 9th WCGALP* (submitted).

Van Eenennaam, A.L., Li, J., Thallman, R.M., *et al*. (2007). *J. Anim. Sci.*, 85:891-900.

Table 1: Proportion of Additive Variance Explained by MBV

Simulation Parameters			Two Trait Model: Genetic Correlation Squared				Full & Reduced ST: Reduction in Variance	
h^2_{gy}	R_g^2	N	Mean Estimate ± Standard Error of Mean	Root Mean Square Error	Mean Std Error of Est	Std Dev of Est	Mean Estimate ± Standard Error of Mean	Root Mean Square Error
Data Simulated from Additive Model Only								
0.1	0.1	448	0.179 ± 0.009	0.20	0.27	0.18	0.296 ± 0.059	1.36
0.1	0.2	437	0.234 ± 0.010	0.21	0.26	0.21	0.457 ± 0.106	2.40
0.1	0.4	410	0.382 ± 0.012	0.24	0.33	0.24	[a]0.603 ± 0.075	1.74
0.1	0.6	354	0.522 ± 0.012	0.24	0.35	0.23	0.972 ± 0.117	2.69
0.3	0.1	500	0.124 ± 0.005	0.10	0.08	0.10	0.114 ± 0.004	0.14
0.3	0.2	492	0.218 ± 0.006	0.13	0.13	0.13	0.216 ± 0.007	0.24
0.3	0.4	498	0.393 ± 0.007	0.16	0.13	0.16	0.386 ± 0.008	0.29
0.3	0.6	490	0.588 ± 0.007	0.16	0.14	0.16	0.592 ± 0.009	0.31
0.5	0.1	498	0.106 ± 0.004	0.09	0.06	0.09	0.102 ± 0.004	0.13
0.5	0.2	500	0.202 ± 0.004	0.10	0.08	0.10	0.199 ± 0.005	0.19
0.5	0.4	500	0.388 ± 0.005	0.12	0.10	0.12	0.387 ± 0.006	0.26
0.5	0.6	500	0.600 ± 0.005	0.11	0.09	0.11	0.603 ± 0.005	0.27
Data Simulated from Additive & Non-Additive Model								
0.1	0.1	442	0.171 ± 0.009	0.20	0.29	0.19	0.177 ± 0.015	0.38
0.1	0.2	430	0.248 ± 0.011	0.23	0.30	0.23	[a]0.380 ± 0.073	1.67
0.1	0.4	400	0.398 ± 0.012	0.25	0.37	0.25	[a]0.567 ± 0.063	1.48
0.1	0.6	357	0.524 ± 0.012	0.24	0.37	0.23	0.924 ± 0.115	2.63
0.3	0.1	500	0.134 ± 0.005	0.12	0.11	0.12	0.116 ± 0.005	0.15
0.3	0.2	500	0.224 ± 0.006	0.15	0.13	0.14	0.199 ± 0.006	0.21
0.3	0.4	500	0.416 ± 0.008	0.17	0.16	0.17	0.394 ± 0.008	0.29
0.3	0.6	487	0.631 ± 0.007	0.16	0.17	0.16	0.616 ± 0.008	0.31
0.5	0.1	500	0.118 ± 0.004	0.09	0.08	0.09	0.105 ± 0.004	0.13
0.5	0.2	500	0.214 ± 0.005	0.12	0.11	0.12	0.197 ± 0.005	0.19
0.5	0.4	500	0.416 ± 0.006	0.13	0.12	0.13	0.400 ± 0.006	0.28
0.5	0.6	499	0.604 ± 0.005	0.12	0.11	0.12	0.589 ± 0.006	0.26

[a]One replicate of 500 had percent reduction in sire greater than 5000% and was not included in the summary