Proc. Assoc. Advmt. Anim. Breed. Genet. 18:129-132

VALIDATION AND ESTIMATION OF ADDITIVE GENETIC VARIATION ASSOCIATED WITH DNA TESTS FOR QUANTITATIVE BEEF CATTLE TRAITS

A. L. Van Eenennaam[1], R. M. Thallman[2], R. L. Quaas[3], K. Hanford[4], and E. J. Pollak[3]

[1] Department of Animal Science, University of California, Davis, CA, USA
[2] US Meat Animal Research Center, Clay Center, NE, USA
[3] Department of Animal Science, Cornell University, Ithaca, NY, USA
[4] Departments of Statistics and Animal Science, University of Nebraska, Lincoln, NE, USA

SUMMARY

The U.S. National Beef Cattle Evaluation Consortium (NBCEC) has been involved in the validation of commercial DNA tests for quantitative beef quality traits since their first appearance on the U.S. market in the early 2000s. This paper discusses pre-validation, analysis, and reporting issues based on our validation experiences. Estimates of DNA test performance (e.g. proportion of genetic variation accounted for by a DNA test panel) in representative populations will be required for incorporation of DNA data into the existing genetic evaluation infrastructure. Such incorporation is appealing as it presents results in an EBV format that is familiar to producers, and eliminates the choice that is implicit when EBVs and marker information are published in tandem.

INTRODUCTION

Prior to moving genetic markers from discovery populations to commercialization, it is important to validate their purported effects on the trait of interest in different breeds and environments, and assess them for correlated responses in associated traits (Barendse 2005). The biggest challenge to achieving this objective is the paucity of cattle populations with sufficient phenotypic data to assess the association between various traits and newly discovered genetic markers, and this makes it difficult and expensive to do large-scale field validations. Results from such validation studies to date have not been widely published (Burrow and Bindon 2005). The validation of panels of DNA markers that are proposed to be used commercially is not simply a repeat of the discovery process, but rather a critical activity to test the strength of support for the testing companies published claims based on independent data.

The NBCEC originally used the term "having validated" to mean finding a significant association "between genetic tests and traits as claimed by the commercial genotyping company based on phenotypes and genotypes derived from reference cattle populations" (Van Eenennaam *et al.* 2007). This process sometimes revealed that tests did not perform as expected, and in certain cases companies chose to withdraw their plans to market those tests.

During the past decade, the DNA testing industry matured from single gene tests to panels involving an ever-increasing number of markers with purported effects on multiple traits, and/or in specific cattle subpopulations. The NBCEC and DNA testing companies have struggled to find appropriately-phenotyped populations that were not involved in the discovery process for validation studies. Additionally, results from different validation populations genotyped with the same SNP panel were often inconsistent with respect to the significance of the association between the test and the trait(s), and sometimes even with respect to the direction of the association. This complicated the interpretation of validation results, and created confusion as to whether "validation" meant a test "worked" (i.e. was significantly associated with the trait) in one or more of the test populations, or had simply been tested by an independent third party. This exposed the process to marketing zeal and left producers somewhat stymied because the data generated did not help to inform decisions about the value proposition associated with investing in specific DNA tests. With the imminent commercialization of a multiplicity of products derived from high density

SNP assays, it seems an opportune time to address some of these concerns. We believe that the validation process needs to evolve from simply reporting the finding of a significant association between DNA test results and the trait of interest, towards an independent calibration approach that estimates the parameters that will be required to facilitate the incorporation of DNA test-based predictions of genetic merit into national genetic evaluation schema. Additionally, results have to be disseminated and interpreted in such a way as to provide industry with the information they require to make the best use of DNA information. Issues associated with the process of validation can be divided into three categories; pre-validation issues, analytical issues and issues associated with presenting results to scientific and industry audiences.

PRE-VALIDATION

DNA-test developers will typically use a "discovery" dataset(s) where a large number of SNP have been assayed on phenotyped animals to develop their test (Allan and Smith, 2008). Genotypes and phenotypes from these discovery populations will be used to develop "molecular breeding value" (MBV) prediction equations by summing the individual SNP additive effects of those loci that show the strongest association with the trait of interest. This DNA test is then the focus of a validation study where a representative sample of animals is genotyped for the markers included in the panel, and the resultant MBVs are compared to phenotypes to assess the accuracy of the test (Goddard and Hayes, 2007). It is essential that discovery populations not be used in validation studies (Barendse, 2005). Prediction equations will perform best in discovery populations in which the SNP associations were discovered and/or populations in which the SNP effects were estimated. Upon requesting a validation, it is therefore crucial that developers fully document which version of a SNP panel and prediction equation they plan to commercialize, and the populations that were used for discovery and training.

As a result of the global investment in SNP genotyping, it is likely that a large number of marker panels associated with various traits will be commercialized in the next couple of years. In genomic studies where many loci are tested against many traits, false positives are inevitable. It is important to delineate test claims and proposed target populations prior to the commencement of any validation studies. Otherwise there is an obvious temptation to use the validation process for discovery. Requiring the disclosure of results from all populations included in validation studies is also important to guard against the understandable temptation to go public with only favourable results. It will be a challenge to keep track of which SNPs are included in the various commercial offerings, especially as they mature and additional SNPs are added to previously tested panels. Therefore some system of nomenclature such as version numbers of tests is essential. While it might be assumed that adding additional SNPs to a panel will improve the accuracy of tests, it is not clear whether new panels should be tested in the same validation populations as the original panel to demonstrate the magnitude of this improvement, or in new populations. Equally unclear is what course of action should be taken if the new panel proves inferior to the old panel.

VALIDATION ANALYSIS

Initial validations performed by the NBCEC looked at individual marker effects to conclude whether the genotype at each individual locus, and the combined effect of all loci in the test, were significantly associated with trait phenotype(s) as claimed by the commercial genotyping companies (Van Eenennaam *et al.* 2007). A similar approach was taken by the "SmartGene for Beef" project in Australia, except in that case the amount of variation accounted for by each DNA marker in each cattle population was estimated using these gene frequencies and gene effects (http://agbu.une.edu.au/SmartGene%20Report11.pdf, accessed 10/6/09). As marker panels grew in size and intellectual property concerns regarding disclosure of the specific marker loci involved in a genetic test emerged, validation moved from testing the effect of individual loci towards testing a

single marker score, or molecular breeding value (MBV). The validation data analysis moved to a determination of whether the regression of phenotype on marker score for a single trait model (in which the marker score was a covariate) differed from zero. While reassuring to see this regression coefficient be non-zero, the significance of this result did not provide useful information in terms of decisions related to the value proposition of the test. A test that has a significant association with the trait of interest may nonetheless explain only a minor proportion of the genetic variance.

The proportion of variation accounted for by a DNA test seems to currently be the best metric available with which to quantitatively evaluate the merit of commercial products. The (co)variance estimates from which the proportion of additive genetic variation accounted for by a DNA test are computed will be requisite for the incorporation of DNA information into national beef cattle genetic evaluations as described by Kachman (2008b). This approach uses a two trait model with the marker score and phenotype included as correlated traits. A theoretically robust estimator of this statistic is the REML estimate of the genetic correlation squared (R_g^2) in a bivariate animal model for the target trait and the MBV, as the second trait (Thallman *et al.* 2009). This estimator has the advantage of producing estimates within the parameter space, and also should be computationally feasible given the size of typical validation data sets.

Simulation studies using this approach (Thallman *et al.* 2009) showed that this R_g^2 estimator tended to be closer to the simulated value of the proportion of variation accounted for by a DNA test than other statistical estimators. Australian researchers recently reported results from a marker panel (Pfizer Animal Genetics 56 SNP panel) evaluation run on four different populations using this approach. These data showed that the proportion of genetic variation accounted for by the molecular value predictions (i.e. MBVs) ranged from ~ 0 - 0.3 depending upon the population and the target trait. However, Kachman (2008b) warned that such estimates may be inaccurate in small data sets (< 1,000 records), and that this error will be exacerbated in traits with low heritability.

Unlike the dairy industry which has the advantage of large, single-breed phenotyped populations for marker discovery and validation (Van Raden *et al.* 2009), the data sets that have typically been used for validation in beef cattle are far from ideal for estimating additive variances and covariances. For example, in the U.S. the NCBA Carcass Merit Project used in the original validations, most breeds had between 400 and 600 progeny represented with records but these were progeny of relatively few sires (≤10). There is a clear need for large, well-organized, thoroughly-phenotyped populations for estimating genetic (co)variances. The development of such populations may require collaborative efforts, and the expenses involved are likely beyond the resources of any single company, or even a single country.

REPORTING

Despite analyses and validation work in both Australia and the US, it is not clear if the data that is currently being reported is providing users with the type of information they need to make informed decisions about the use of a particular test. Validation teams have objectively presented their findings on websites with the optimistic view that "decisions very much depend on the individual business' attitude to risk and can only be made effectively by the individual business." While this is undoubtedly true, most producers are unlikely to have had sufficient training in quantitative genetics to correctly interpret the results and get to the point of making such decisions based on risk considerations. Many producers are already wary of this historically-oversold technology based on past experience (Barendse *et al.* 2005; Casas *et al.* 2005). It is not evident how reporting findings that a test explains a proportion ranging from 0 to 0.15 of the additive genetic variation associated with the target trait, has a regression coefficient of 0.26 (±0.3), and a *p* value of 0.001 provides information that helps in the decision-making process. The accuracy of a DNA test at predicting the true genetic merit of an animal is primarily driven by the amount of additive genetic variation accounted for by the DNA test. Current estimates for this correlation

and the proportion of the variation accounted for by existing tests (the square of that correlation) are generally low with the exception of DNA tests for tenderness where available estimates include 0.25 (Allan and Smith, 2008) and 0.016-0.299 from an Australian evaluation of the Pfizer Animal Genetics 56 SNP panel, (http://www.beefcrc.com.au/Aus-Beef-DNA-results, accessed 10/6/09). Over time it is envisioned that genetic tests will have markers associated with the majority of important genes influencing a trait and the marker effects will have been assessed in large enough training populations to provide for accurate SNP effect estimates meaning genotyping results will be highly predictive of the true breeding value (BV) of an animal.

In the interim however, from the user's perspective, perhaps the most useful information that could be provided is how much a DNA test improves the accuracy of expected breeding values (EBV). That is, how much improvement in the accuracy associated with an EBV can be expected from adding DNA test information. Publishing traditional EBVs and marker information separately, as is currently the case, is confusing and can lead to incorrect selection decisions when marker scores predict only a small proportion of the genetic variance (Crews *et al.* 2008). Developing an approach to calculate marker-assisted EBVs would seem to be the next logical step.

Selection index methodology has been applied to combine marker scores and polygenic EBVs using a linear index weighted on the accuracy of traditional BV and the proportion of genetic variance attributable to the marker score (Amer 2007; Crews *et al.* 2008). Kachman (2008a) simulated the effect of DNA marker scores on the accuracy of tenderness expected progeny difference (EPD). Assuming a heritability of 0.4 and a 0.45 genetic correlation between the marker score and shear force, he found that the beef improvement association accuracy of the sire tenderness EPD increased from 0.27 to 0.31 when DNA marker information was combined with information from 10-phenotyped progeny. Expressing the effect of marker information on accuracy is appealing because it makes use of the existing genetic evaluation infrastructure, and presents a metric that is familiar to producers.

CONCLUSIONS

The validation process and analyses have evolved as the DNA testing industry has matured from single gene tests to panels involving an ever-increasing number of markers. With products derived from high density SNP assays on the horizon, it seems an opportune time to reassess the process and consider how it can best be used to provide the estimates of the genetic (co)variance parameters that will be required to facilitate the incorporation of information into national genetic evaluation systems. Additionally, careful consideration must be given to industry dissemination of independent validation studies.

REFERENCES

Allan, M.F., and Smith, T.P.L. (2008) *Meat Sci*. **80**:79.
Amer, P.R. (2007) *Proc. Assoc. Advmt. Anim. Breed. Genet*. **17**:479.
Barendse, W. (2005) *Aust. J. Exp. Agric.* **45**:831.
Barendse, W., Bunch, R.J., and Harrison, B.E. (2005) *Anim. Genet*. **36**:86.
Burrow, H.M., and Bindon, B.M. (2005) *Aust. J. Exp. Agric*. **45**:941.
Casas, E., White, S. N., Riley, D. G. , et al.. (2005*) J. Anim. Sci.* **83**:13.
Crews, D. H., Jr., Moore, S. S., and Enns, R. M. (2008) *Proc. 40th Beef Improv. Fed.* **40**:44.
Goddard, M. E. and Hayes, B. J. (2007) *J. Anim. Breed. Genet*. **124**:323
Thallman, R.M., Hanford, K.J., Quaas, R.L., et al. (2009) *Proc. 41st Beef Improv. Fed.* **41**:184.
Kachman, S. D. (2008a). *40th Beef Improv. Fed.* http://www.bifconference.com; accessed 10/6/09
Kachman, S. D. (2008b). *CD-ROM 9th Genetic Prediction Workshop, Beef Improv. Fed.*
Van Eenennaam, A.L, J. Li, J., Thallman, R.M., Quaas, R.L., et al. (2007) *J. Anim. Sci.* **85**:891.
VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., et al. (2009) *J. Dairy Sci*. **92**:16-24.