J. Dairy Sci. 99:1–8
http://dx.doi.org/10.3168/jds.2015-10038

Practical implications for genetic modeling in the genomics era[1]

P. M. VanRaden[2]
Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

Genetic models convert data into estimated breeding values and other information useful to breeders. The goal is to provide accurate and timely predictions of the future performance for each animal (or embryo). Modeling involves defining traits, editing raw data, removing environmental effects, including genetic by environmental interactions and correlations among traits, and accounting for nonadditive inheritance or nonnormal distributions. Data include phenotypes and pedigrees during the last century and genotypes within the last decade. The genomic data can include single nucleotide polymorphisms, quantitative trait loci, insertions, deletions, and haplotypes. Subsets must be selected to reduce computation because total numbers of variants that can be imputed have increased rapidly from thousands to millions. Current computation using 60,671 markers takes just a few days. Nonlinear models can account for the nonnormal distribution of genomic effects, but reliability is usually better than that of linear models only for traits influenced by major genes. Numbers of genotyped animals have also increased rapidly in the joint North American database from a few thousand in 2009 to over 1 million in 2015. Most are young females and will contribute to estimating allele effects in the future, but only about 150,000 have phenotypes so far. Genomic preselection can bias traditional animal models because Mendelian sampling of phenotyped progeny and mates is no longer expected to average zero; however, estimates of bias are small in current US data. Single-step models that combine pedigree and genomic relationships can account for preselection, but approximations are required for affordable computation. Traditional animal models may include all breeds and crossbreds, but most genomic evaluations are still computed within breed. Models that include inbreeding, heterosis, dominance, and interactions can improve predictions for individual matings. Multitrait genomic models may be preferred for traits with many missing records or when foreign records are included as pseudo-observations, but most countries use multitrait traditional evaluations followed by single-trait genomic evaluations. Genomic reliabilities are about 70% for the more heritable traits. Researchers must choose from many available models and explain how the models work so that breeders can more confidently apply the predictions in their selection programs.
Key words: genetic evaluation, genomic selection, mixed models, multitrait

INTRODUCTION

Genetic markers are central to dairy cattle selection programs, allowing accurate and affordable prediction of each animal's merit using tens of thousands of genotypes. Many new issues arise when models use genotypes along with phenotypes and pedigrees, but many previous principles of modeling remain true. Models must still separate genetic from environmental effects on traits, and genetic effects are now further separated and tracked across each chromosome by markers. Since 2008, rapidly growing genomic data sets and changes in selection programs require constant updating of evaluation systems. More available choices and algorithms allow researchers to improve accuracy and control bias in genetic rankings while adding more data.

Modeling involves defining traits, editing raw data, adjusting for environmental effects, including genetic by environmental interactions and correlations among traits, and accounting for nonadditive inheritance or nonnormal distributions. Major previous advances in US evaluation models were use of daughters' average production adjusted for dams' production in 1937, use of herd-year-season groups and heritability in 1962, use of sire and maternal grandsire pedigrees in 1974, use of all relatives in 1989, use of foreign data from multitrait, across-country evaluation (**MACE**) in 1995, and use of a multibreed model in 2007. Similar advances in modeling occurred in many other countries. The goal of modeling is to provide accurate and timely predictions of the future performance for each animal (or embryo).

Received June 30, 2015.
Accepted November 16, 2015.
[1]Presented as part of the Relevance of Modeling in the Genomics Era Symposium at the ADSA-ASAS Joint Annual Meeting, Orlando, Florida, July 2015.
[2]Corresponding author: Paul.VanRaden@ars.usda.gov

Accurate estimates of the environmental effects may also be helpful in management decisions and benchmarking to compare nongenetic factors, but this report will focus on genetics.

Goals of this report were to (1) summarize the choices of data available for genomic modeling, (2) compare methods to test different models for predictive ability, and (3) describe how changes in breeding programs and available data may require changes in models.

GENETIC MODELING

Data

Genotypes can include SNP, QTL, insertions, deletions, and haplotypes; total numbers of known variants have increased rapidly since 2000 from hundreds to thousands to millions (Table 1). About 30 million total variants have been identified in cattle (Daetwyler et al., 2014) and could, in theory, be imputed for each genotyped animal. Thus, imputation can generate far more data than any routine analysis could affordably include. Research, therefore, focuses on estimating effects of many variants from progeny-tested bulls or phenotyped cows and then including the variants with larger estimated effects on future chips and in routine evaluations (Hayes et al., 2014; Wiggans et al., 2014; Brøndum et al., 2015).

The variants included are increasingly chosen from previous estimates or bioinformatics instead of additional random markers. Until 2013, patents prevented some variants from being used, but patents on naturally occurring variants are no longer valid in the United States (US Supreme Court, 2013). As costs of genotyping by sequencing decline, new variants and mutations could be detected with each new animal. Whole-genome sequencing can read all DNA at various depths of coverage, and exome sequencing can read just the sections of DNA that code for expressed genes (McClure et al., 2014).

Pedigrees were the basis of selection for many decades. Accuracy was excellent for bulls in AI service and for embryo transfer (**ET**) calves because of careful checking but was not so good for commercial cows. Genomic testing can discover or correct ancestry because nearly all AI sires and maternal grandsires are now genotyped for dairy cattle in many countries (Tooker et al., 2015). Evaluation models have not accounted for uncertain paternity and thus have underestimated the true genetic differences among such sires (Foulley et al., 1987), but paternity is certain when daughters are genotyped. Correct pedigrees are useful in imputation, avoiding inbreeding, and in quality control of the genotyping process, such as to avoid switched samples, but are no longer as important in modeling because pedigree relationships are being replaced by genomic relationships.

Phenotypes can be defined in many ways. A main goal is consistent definition across time and countries so that larger data sets can be used to estimate small effects of individual genes. Foreign data can improve reliability if traits have high correlations across countries (Lund et al., 2011), and this has led to widespread international genotype exchanges. Historical phenotypes can increase reliability if the recent reference population is small (Cooper et al., 2015) but could decrease reliability if generations of recombination have changed the linkage patterns or allele frequencies or if the traits have changed over time (Lourenco et al., 2014). Selection previously emphasized traits with high heritability, but large reference populations now allow progress from genomic selection even for traits with low heritability.

Environmental factors such as herd-year-season, age, parity, days in milk, and milking frequency were included in models or used to preadjust data for many decades. Data edits for phenotypes are important but are not affected by the addition of genotypes, and so previous methods should apply. Edits for genotypes are becoming more complex as new chips, new variants, and sequence data are added. Each of the 17 chips now included has different patterns of missing data and often different names for the same variant, requiring complex merges and edits (Nicolazzi et al., 2014) because error rates can also differ by chip or source of data. Imputation is now a key step in the US evaluation because only 25% of all animal genotypes are measured and 75% are missing and imputed. Imputation uses statistical methods and pedigrees to phase observed higher-

Table 1. Growth in number of variants and animals available to estimate genetic effects

Year	Variants	Reference animals[1]	Example reference
1991–2004	367	1,415	Ashwell et al., 2004
2007–2010	42,503	16,646	Wiggans et al., 2011
2011–2013	636,967	15,842	VanRaden et al., 2013
2014–future	28,300,000	234 + imputed	Daetwyler et al., 2014

[1]Holstein reference animals in April 2015 included 27,464 progeny-tested bulls and 136,184 phenotyped cows.