J. Dairy Sci. 92:3431–3436
doi:10.3168/jds.2008-1758

Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada

G. R. Wiggans,*[1] T. S. Sonstegard,* P. M. VanRaden,* L. K. Matukumalli,*† R. D. Schnabel,‡ J. F. Taylor,‡ F. S. Schenkel,§ and C. P. Van Tassell*

*Agricultural Research Service, USDA, Beltsville, MD 20705-2350
†Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110
‡Division of Animal Sciences, University of Missouri, Columbia 65211
§Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Ontario N1G 2W1, Canada

ABSTRACT

Nearly 57,000 single-nucleotide polymorphisms (SNP) genotyped with the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) were investigated to determine usefulness of the associated SNP for genomic prediction. Genotypes were obtained for 12,591 bulls and cows, and SNP were selected based on 5,503 bulls with genotypes from a larger set of SNP. The following SNP were deleted: 6,572 that were monomorphic, 3,213 with scoring problems (primarily because of poor definition of clusters and excess number of clusters), and 3,649 with a minor allele frequency of <2%. Number of SNP for each minor allele frequency class (≥2%) was fairly uniform (777 to 1,004). For 5 contiguous SNP assigned to chromosome 7, no bulls were heterozygous, which indicated that those SNP are actually on the nonpseudoautosomal portion of the X chromosome. Another 178 SNP that were not assigned to a chromosome but that had many fewer heterozygotes than expected were also assigned to the X chromosome. Existence of Hardy-Weinberg equilibrium was investigated by comparing observed with expected heterozygosity. For 11 SNP, the observed percentage of heterozygous individuals differed from the expected by >15%; therefore, those SNP were deleted. For 2,628 SNP, the genotype at another SNP was highly correlated (i.e., genotypes were identical for >99.5% of bulls), and those were deleted. After edits, 40,874 SNP remained. A parent–progeny conflict was declared when the genotypes were alternate homozygotes. Mean number of conflicts was 2.3 when pedigree was correct and 2,411 when it was incorrect. The sire was genotyped for >93% of animals. Maternal grandsire genotype was similarly checked; however, because alternate homozygotes could be valid, a conflict threshold of 16% was used to indicate a need for further investigation. Genotyping consistency was investigated for 21 bulls genotyped twice with differences primarily from SNP that were not scored in one of the genotypes. Concordance for readable SNP was extremely high (99.96–100%). Thousands of SNP that were polymorphic in Holsteins were monomorphic in Jerseys or Brown Swiss, which indicated that breed-specific SNP sets are required or that all breeds need to be considered in the SNP selection process. Genotypes from the Illumina BovineSNP50 BeadChip are of high accuracy and provide the basis for genomic evaluations in the United States and Canada.

Key words: genomic prediction, genotyping, single-nucleotide polymorphism

INTRODUCTION

The Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) provides a high-density assay of nearly 57,000 SNP (Matukumalli et al., 2009). As of October 2008, 14,720 Holsteins, 1,558 Jerseys, and 368 Brown Swiss had been genotyped with the Illumina BovineSNP50 BeadChip for use in genomic evaluations in the United States and Canada (VanRaden et al., 2009a). Before using those SNP for that purpose, the quality of the SNP and the accuracy of sample identification and pedigree information were investigated. Accuracy of genomic evaluations is reduced by mislabeled samples. Furthermore, SNP that are problematic to genotype may not be consistently scored and should not be used in genomic prediction.

A SNP that does not contribute to the accuracy of the evaluations can be eliminated to reduce computational effort and to improve stability of estimates of the effects of the remaining SNP. Hayes et al. (2009) imposed a minor allele frequency (**MAF**) of 2.5% using the Illumina BovineSNP50 BeadChip and a population of 798 Holstein bulls. They required a call rate of 90% and excluded SNP with a deviation of ≥600 from Hardy-Weinberg equilibrium. A MAF threshold of 5% is common in research on human genetics because SNP with lower frequencies require substantially greater

Received September 28, 2008.
Accepted February 17, 2009.
[1]Corresponding author: George.Wiggans@ars.usda.gov

sample sizes to detect small effects, which are typical for complex traits (Hirschhorn and Daly, 2005).

The purpose of this study was to 1) select SNP to be used in genomic evaluation of Holstein cattle by removing loci that were unscorable or monomorphic, had a MAF of <2% or a large deviation from expected heterozygosity, or were highly correlated with other SNP and 2) determine the accuracy of the genotypes by investigating parent–progeny conflicts and correspondence among genotypes produced by the repeated assay of an animal.

MATERIALS AND METHODS

Genotyped Animals

Holstein bulls (10,690) and cows (1,901) were genotyped for 56,947 SNP. Genotypes of those animals were included in the recently initiated USDA genomic evaluation program for the United States and Canada (VanRaden et al., 2009b). The main source of extracted DNA for bulls was semen from the Cooperative Dairy DNA Repository (Ashwell and Van Tassell, 1999) and from the National Center for Genetic Resources Preservation, ARS, USDA (Fort Collins, CO). The SNP were identified (Van Tassell et al., 2008) and selected (Matukumalli et al., 2009) to be polymorphic across a wide variety of breeds included in the International Bovine HapMap Project (International Bovine HapMap Consortium, 2006). Extraction of DNA and genotyping was conducted by the Bovine Functional Genomics Laboratory, ARS, USDA (Beltsville, MD); Division of Animal Sciences, University of Missouri (Columbia); Department of Agricultural, Food, and Nutritional Science, University of Alberta (Edmonton, Canada); GeneSeek (Lincoln, NE); Genetics and IVF Institute (Fairfax, VA); and Illumina. Scoring of marker genotypes was performed using Illumina BeadStudio software (v3.2.23).

Genotypes were represented by the number of the counted allele: 0, 1, and 2. A 5 indicated that the genotype could not be determined. Initially, only genotypes for polymorphic SNP with a MAF of ≥5% were included. As more genotypes became available, all polymorphic SNP were included. From the complete data set of genotyped bulls, the 5,579 bulls with genotypes reported for all polymorphic SNP were selected. Of those bulls, 1.4% had <90% of all SNP present and were excluded because they were considered to have a poor quality DNA sample. The final data set for determination of SNP to be used in genomic evaluation included genotypes from 5,503 bulls. By using only bulls, the presence of X-linked SNP could be determined. The entire set of 12,591 animals was used for the detection of parentage errors.

Quality of SNP Genotypes

Some SNP could not be easily scored because of an abnormal genotype clustering pattern. Clustering was then reassessed according to subsets of animals with similar genotyping protocols (e.g., same reagent date) to see if the likelihood of correctly identified SNP (call rate) could be improved through Illumina BeadStudio software so that the SNP could be retained. Other quality characteristics that were investigated included the portion of SNP with missing genotypes, parent–progeny conflicts, and SNP with an excess heterozygosity compared with expected heterozygosity from Hardy-Weinberg equilibrium. Genotypes from repeated genotyping of the same animal and genotypes of identical twins and split embryos were also investigated to determine their similarity.

X-Linked SNP

The Illumina BovineSNP50 BeadChip allows genotyping of 2,183 SNP that were not assigned to chromosomes in the Btau_4.0 bovine genome sequence assembly (Baylor College of Medicine, 2007). To test for the presence of X-linked SNP and other X-linked but incorrectly assigned SNP, genotypes from bulls were checked for heterozygosity (genotype of 1). Because bulls have only one X chromosome, they should have no heterozygous genotypes for nonpseudoautosomal loci. To allow for genotyping errors, SNP with ≤50 heterozygous (1%) bulls were considered to be on the X chromosome. The X-linked SNP were investigated for their usefulness in determining the sex of the genotyped animal. The presence of heterozygosity in cows was used to confirm that the lack of heterozygosity in bulls was not a genotyping problem.

Hardy-Weinberg Equilibrium

For autosomal SNP, the departure of the frequency of heterozygotes from that expected under Hardy-Weinberg equilibrium may indicate problems in accurately determining the genotype of that SNP or may suggest that the SNP is actually a duplicate rather than single genomic locus. If the difference between observed and expected genotype frequencies was >0.15, the SNP was excluded. That threshold was chosen so that only the most extreme outlier SNP would be excluded because selection could also have caused some departure from equilibrium.

Table 1. Counts of Holstein SNP selected to be excluded from use in genomic evaluation by reason for exclusion

Category/reason for exclusion	n
SNP genotyped	56,947
Monomorphic	6,572
Scoring problems	3,213
Minor allele frequency of <2%	3,649
Heterozygosity deviation of >0.15 from Hardy-Weinberg equilibrium expectation	11
High correlation with other SNP[1]	2,628
SNP selected for use in genomic evaluation	40,874

[1]Genotypes were all the same (0–0, 1–1, and 2–2) or all opposite (0–2, 1–1, and 2–0).

Detection of Highly Correlated SNP

The MAF were calculated, and each SNP was compared with every other SNP that had a MAF within 2.5 percentage units to determine if the 2 SNP were highly correlated. Two SNP were declared highly correlated if the genotypes were all the same (0–0, 1–1, and 2–2) or all opposite (0–2, 1–1, and 2–0). To allow for genotyping errors, up to 27 bulls with other genotype combinations were allowed (0.5% of genotyped animals).

Parentage Verification and Determination

The SNP genotypes for each animal were compared with the SNP genotypes of each of its parents to determine how many times homozygous SNP were discrepant for a parent–progeny pair. Based on preliminary analysis, an occurrence of >200 discrepancies was considered to be a parent–progeny conflict. If a conflict was found or if neither parent had been genotyped, the animal's SNP genotypes were compared with those of every other animal to determine if there was a parent–progeny relationship with another animal. A duplicate genotype would not generate conflicts and, therefore, would also appear to be consistent with a parent or progeny relationship. Such duplicates could be identical twins, split embryos, clones, or sample labeling errors. An animal of the appropriate sex with only a few SNP genotype conflicts (not identical) and a suitable interval between its birth and that of the subject animal was designated as the putative parent. For an animal with an ungenotyped parent, the animal was rejected for genomic prediction if a putative parent was detected.

Because relatively few cows were genotyped, the parentage checking system was extended to include maternal grandsires (**MGS**). Because 2 meioses separate an animal from its MGS, alternate homozygous SNP genotypes are possible. Those should, however, be less frequent than for unrelated animals. Consideration of gene frequency could improve the precision of this test. The frequency of conflicts was investigated to determine a threshold for which a true MGS would be unlikely to have more than that number of conflicts. If an animal's MGS was considered likely to have been misidentified, the animal's genotype was flagged as a potential problem rather than excluded.

Jersey and Brown Swiss

A similar SNP selection procedure was applied for genotypes of Jersey and Brown Swiss bulls. To avoid revising the scoring procedures used to determine Holstein genotypes, only the 43,547 polymorphic SNP that could be scored reliably for Holsteins were considered. After excluding bulls with <90% of scored Holstein SNP, the data sets for determination of SNP to be used in genomic evaluation included genotypes from 941 Jersey and 344 Brown Swiss bulls.

RESULTS AND DISCUSSION

Table 1 shows the number of Holstein SNP selected to be excluded from use in genomic evaluation for various reasons. Deletions included 6,572 monomorphic SNP, 3,213 SNP with scoring problems, and 3,649 SNP with a MAF of <2%. Scoring problems primarily included genotype clustering abnormalities (poor definition of clusters and too many or too few clusters) and flanking polymorphisms that affected assay performance. Because SNP with a low MAF (<2%) were expected to have little impact on genomic evaluation, they were excluded. This MAF minimum also provided a means to exclude monomorphic SNP that had been incorrectly genotyped as polymorphic. VanRaden et al. (2009b) used a minimum SNP MAF of 5% for genomic predictions but suggested that SNP with lower MAF could be included accurately as sample size increased. Because genotypes are now available for many more animals, the MAF threshold was lowered to 2%. The number of SNP for each MAF class that was ≥2% was fairly uniform (777–1,004).

Only 11 SNP were excluded because of a large (>0.15) deviation between frequencies of observed and expected heterozygotes. Some SNP that would have had a large

deviation had already been excluded as unscorable. Five SNP from chromosome 7 and 178 SNP that were previously unassigned were assigned to the X chromosome because they had <50 (1%) heterozygous genotypes, which is considerably fewer than the 216 expected with a minimum MAF of 2%. The 5 contiguous SNP from chromosome 7 did not have other evidence of genotyping problems to explain the lack of heterozygous bulls.

An additional 2,628 SNP were excluded because of a high correlation with other SNP. In most cases, the highly correlated SNP were adjacent. However, 44 unassigned SNP were highly correlated with a SNP that was assigned to a chromosome, and 45 unassigned SNP were highly correlated with each other. In 9 cases, SNP from different chromosomes were highly correlated, which suggested that at least one of the SNP was not correctly assigned; however, with many more SNP than bulls, such similarity could happen by chance. The largest set of highly correlated SNP contained 12 SNP. The SNP retained from each set of highly correlated SNP (in preference order) was 1) a parentage SNP identified by Heaton et al. (2002), 2) a SNP validated by Illumina for the BovineSNP50 BeadChip, or 3) the SNP with the best genotype scoring characteristics (call rate). The set of selected SNP will be reevaluated if calling procedures change or if sufficient additional genotyping samples become available to allow exclusion or inclusion of additional SNP according to the established SNP quality criteria.

A total of 40,874 SNP were selected for use in genomic evaluation (Table 1). Hayes et al. (2009) qualified 38,259 SNP using the Illumina BovineSNP50 BeadChip and a MAF of ≥2.5%.

Of the 5,503 bulls used for SNP selection, 96% had <1% missing genotypes. Of the 40,874 selected SNP, 90% had <0.1% missing genotypes; the maximum missing was 3%. For the 10,690 genotyped bulls, most had no heterozygous SNP in the nonpseudoautosomal region of the X chromosome (mean of 0.12). The 3 bulls with >20 heterozygous SNP in that region also had an unusually high number of missing genotypes, which indicated a possible scoring problem for those samples. Cows had a mean of 224 heterozygous X-linked SNP. A few cows had low numbers of heterozygous X-linked SNP. One cow with only 8 heterozygous SNP was investigated further and found to have both X chromosomes identical by descent from one bull. The number of heterozygous X-linked SNP is a useful tool for detecting mislabeled genotypes and errors in the sex recorded with pedigree information.

Conflicts between parent and progeny SNP genotypes provide a powerful tool for detecting sample mislabeling and laboratory errors. With the large number of genotyped animals, most animals had a genotyped sire (93%). When an animal's parent was accepted as matching based on 37,811 autosomal SNP, the number of SNP conflicts averaged 2.3 (range of 0–89). When the reported parent was considered to be incorrect, the number of SNP conflicts averaged 2,411 (range of 754–3,507).

Because only 16% of dams had been genotyped, agreement between MGS (91% genotyped) and grandprogeny SNP genotypes was investigated. Inheritance of SNP from an MGS is affected by Mendelian sampling (2 meioses separate the animals); therefore, MGS–grandprogeny SNP conflicts were expected. Figure 1 shows the distribution of SNP conflicts for both correctly and incorrectly reported MGS. A threshold for alternate homozygous conflicts of 16% was adopted to identify putatively incorrect MGS that merited additional investigation.

Genotypes were next checked for the presence of duplicates. To minimize computational requirements, only every 10th SNP was considered. Sixteen sets of identical genotypes were found: 1 set of 3 animals from nuclear transfer, 3 pairs of split embryos, 5 pairs from embryo transfer, and the remaining 7 pairs from naturally occurring identical twins. The genotypes of those animals were harmonized. First, the sample with the fewest missing SNP genotypes was determined. When possible, the missing SNP values were then filled using the genotype(s) of the other animal(s), and that genotype was used for the other animal(s). The most complete genotypes had a mean of 35 missing SNP genotypes (range of 2–167), which were filled. For the other animal(s), the number of SNP genotypes that differed averaged 229 (range of 5–2,345). Nearly all of those differences were the result of filling a missing SNP genotype. The cow with 2,345 SNP differences had only 5 that were conflicts, and her number of missing SNP was within the 90% threshold for call rate. The number of SNP conflicts for duplicate genotypes averaged 2.6 (range of 0–25). Harmonization of duplicate genotypes ensures that genetically identical animals receive the same genomic evaluation.

Multiple genotypes for 21 bulls were investigated to determine the degree of variation among repeated genotypes. The number of instances in which a SNP genotype was missing for one sample but present for another of the same bull averaged 198 (range of 20–616). Concordance for scored SNP was extremely high (range of 99.96–100%). The genotype with the fewest missing SNP genotypes was used for animals with >1 set of genotypes.

Only SNP that were polymorphic in Holsteins were used for SNP selection for other breeds because geno-

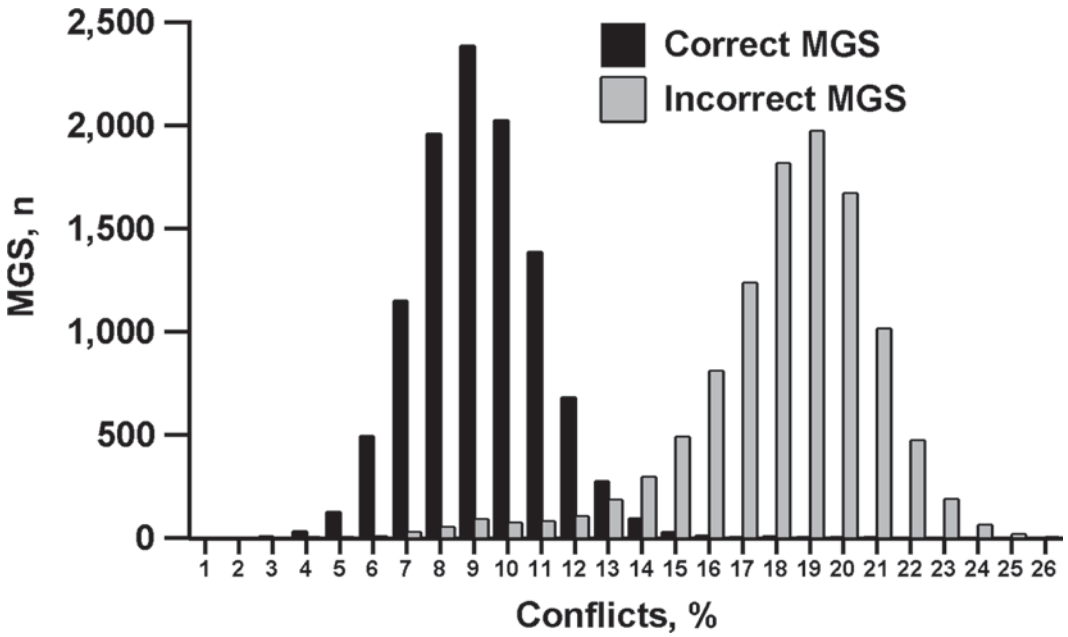

Figure 1. Distribution of conflicting SNP between grandprogeny and correctly or incorrectly reported maternal grandsires (MGS).

typing Holsteins was the top priority and considerable time and effort could be saved by restricting SNP scoring to potentially useful SNP. Of the original 43,547 polymorphic Holstein SNP, 2,845 were monomorphic in Jerseys, and 31,628 polymorphic SNP remained after imposing a MAF of ≥3% and removing highly correlated SNP. The corresponding numbers for Brown Swiss were 2,347 monomorphic SNP and 34,593 remaining after requiring a MAF of ≥3% and removing highly correlated SNP. The small number of genotyped Jersey and Brown Swiss bulls with available genotypes makes the data preliminary, but some indication of breed differences in monomorphic SNP was apparent. Selection of SNP for genomic evaluation of Jerseys and Brown Swiss will be repeated when genotypes from more bulls become available. Monomorphic Holstein SNP will be considered, and a minimum MAF of 2% will be used. With more bulls, the number of SNP that are monomorphic or have low MAF may decrease.

Individual haplotypes would provide extremely useful information. Haplotype blocks could be used in place of individual SNP as the basis for genomic prediction and for imputation of missing SNP genotypes of genotyped animals as well as entire missing genotypes for ungenotyped relatives of genotyped animals. Errors in mapping SNP would affect the accuracy of haplotyping. This approach is being investigated, but the large number of SNP and large population with missing genotypes for most dams makes haplotyping a difficult and potentially time-consuming process.

CONCLUSIONS

The Illumina BovineSNP50 BeadChip provides an accurate and reliable set of SNP for use in genomic evaluation. For Holsteins, 40,874 SNP had a MAF of >2%, were not highly correlated with other SNP, and did not have scoring problems. Relative importance of the different SNP selection criteria was not identified because retention of most of the eliminated SNP, which added little information, would not have affected overall accuracy and might have made estimates of individual SNP effects less stable. The selected SNP had few parent–progeny conflicts when the pedigree relationship was correct but usually had >1,000 conflicts when it was not. The X-chromosome SNP in the nonpseudoautosomal region provided a good check on the sex of the animal. Only rarely did a cow have a low number of X-linked heterozygous SNP. The similar genotypes produced for identical twins, split embryos, and clones and repeated genotypes on the same animal are further evidence of the consistency of the assay.

Selection of SNP is breed specific. Sets of monomorphic SNP differ among breeds. A common set of SNP across breeds would simplify data management and allow investigation of the value of data from one breed in predicting genetic merit for another. Some SNP quality issues could cause a SNP to be useful for one breed but not scorable for another.

The value of additional SNP increases with the number of genotyped animals. In this study, the MAF re-

quirement was dropped from the 5% used by VanRaden et al. (2009b) to 2% because with >12,000 genotyped animals, even SNP with a MAF of 2 to 5% have enough animals with the less frequent allele to obtain a useful estimate. Elimination of highly correlated SNP reduced the computational burden without loss of accuracy and may aid in convergence of solutions.

The selection of SNP is likely to evolve as additional crossing-over events are included in the data and gene frequencies change, either because of increased sample size or changes in selection pressure on the alleles themselves. An area of further study would be to determine if the accuracy of estimated SNP effects was improved by removing additional highly correlated SNP. The extensive parentage testing possible because of the high percentage of genotyped sires and the high reliability of the genotyping allows great confidence in the genomic prediction system.

ACKNOWLEDGMENTS

This project was supported by National Research Initiative Grants 2006-35205-16888 and 2006-35205-16701 from the USDA Cooperative State Research, Education, and Extension Service and by the National Association of Animal Breeders, Holstein Association USA (Brattleboro, VT), and American Jersey Cattle Association (Reynoldsburg, OH).

REFERENCES

Ashwell, M. S., and C. P. Van Tassell. 1999. The Cooperative Dairy DNA Repository—A new resource for quantitative trait loci detection and verification. J. Dairy Sci. 82(Suppl. 1):54. (Abstr.)

Baylor College of Medicine. 2007. Bovine Genome Project. http://www.hgsc.bcm.tmc.edu/projects/bovine/index.html Accessed Dec. 11, 2008.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, and M. E. Goddard. 2009. Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

Heaton, M. P., G. P. Harhay, G. L. Bennett, R. T. Stone, W. M. Grosse, E. Casas, J. W. Keele, T. P. L. Smith, C. G. Chitko-McKown, and W. W. Laegreid. 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mamm. Genome 13:272–281.

Hirschhorn, J. N., and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6:95–108.

International Bovine HapMap Consortium. 2006. An overview of the Bovine HapMap Project. Page 60 in Proc. 30th Inter. Conf. Anim. Genet., ISAG 2006. Colégio Brasileiro de Reprodução Animal, Belo Horizonte, Brazil. (Abstr.)

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE (accepted).

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. Taylor Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252.

VanRaden, P., G. Wiggans, T. Sonstegard, and L. Walton. 2009a. Genomic evaluations become official. Changes to Evaluation System (January 2009). http://aipl.arsusda.gov/reference/changes/eval0901.html Accessed April 3, 2009.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009b. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.