J. Dairy Sci. 93:2287–2292
doi:10.3168/jds.2009-2773

Selection and management of DNA markers for use in genomic evaluation

G. R. Wiggans,*[1] P. M. VanRaden,* L. R. Bacheller,* M. E. Tooker,* J. L. Hutchison,* T. A. Cooper,* and T. S. Sonstegard†
*Animal Improvement Programs Laboratory and
†Bovine Functional Genomics Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

To facilitate routine genomic evaluation, a database was constructed to store genotypes for 50,972 single nucleotide polymorphisms (SNP) from the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). Multiple samples per animal are allowed. All SNP genotypes for a sample are stored in a single row. An indicator specifies whether the genotype for a sample was selected for use in genomic evaluation. Samples with low call rates or pedigree conflicts are designated as unusable. Among multiple samples that qualify for use in genomic evaluation, the one with the highest call rate is designated as usable. When multiple samples are stored for an animal, a composite is formed during extraction by using SNP genotypes from other samples to replace missing genotypes. To increase the number of SNP available, scanner output for approximately 19,000 samples was reprocessed. Any SNP with a minor allele frequency of ≥1% for Holsteins, Jerseys, or Brown Swiss was selected, which was the primary reason that the number of SNP used for USDA genomic evaluations increased. Few parent–progeny conflicts (≤1%) and a high call rate (≥90%) were additional requirements that eliminated 2,378 SNP. Because monomorphic SNP did not degrade convergence during estimation of SNP effects, a single set of 43,385 SNP was adopted for all breeds. The use of a database for genotypes, detection of conflicts as genotypes are stored, online access for problem resolution, and use of a single set of SNP for genomic evaluations have simplified tracking of genotypes and genomic evaluation as a routine and official process.

Key words: genomic evaluation, genotyping, single nucleotide polymorphism

INTRODUCTION

Soller (1994) presented an overview of the potential for using marker-assisted selection for traits of economic importance. By 2004, the commercial use of genomic selection for livestock was growing but still small; for example, individual marker tests were available for diacylglycerol acyltransferase (DGAT) and growth hormone-releasing hormone for dairy animals (Dekkers, 2004). Since then, genomic evaluations based on genotypes from the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) have become the basis for selecting young bulls for use in AI and selling semen of bulls without a progeny test in the United States and Canada; they also influence selection of bull dams and service sires.

Numbers of animals with genomic evaluations calculated since the first unofficial release in April 2008 are given in Table 1. For Holsteins, the increase was almost 1,400 animals per month, which shows the high participation in the genomic evaluation program. Genomic evaluations became official for Holsteins and Jerseys in January 2009 and for Brown Swiss in August 2009.

Several issues must be considered in the selection of SNP for genomic evaluation so that evaluation accuracy is maximized (Wiggans et al., 2009). A SNP with frequent parent–progeny conflicts would reduce evaluation accuracy because it would be unreliable. Many missing genotype calls might indicate that a SNP was difficult to score and, therefore, also unreliable. Increasing the number of SNP used for genomic evaluation is expected to increase the accuracy of evaluations through better tracking of QTL (VanRaden et al., 2009b). As the number of genotyped animals increases, even SNP with a low minor allele frequency can contribute to evaluation accuracy.

In the United States, USDA initially maintained genotype data for dairy cattle by adding new genotypes to a master file at each genomic evaluation. Genotypes were checked for consistency with parent genotypes to maximize the probability that the genotype was from the indicated animal. Call rate requirements were im-

Received September 28, 2009.
Accepted January 18, 2010.
[1]Corresponding author: George.Wiggans@ars.usda.gov

Table 1. Numbers of genotyped animals with genomic evaluations by breed and evaluation date

Evaluation date	Holstein	Jersey	Brown Swiss
April 2008[1]	6,121	—	—
June 2008[1]	10,403	—	—
August 2008[1]	12,588	—	—
October 2008[1]	14,720	1,574	—
December 2008[2]	17,830	2,086	—
February 2009	19,596	2,380	—
April 2009	21,941	2,734	723[1]
June 2009	25,365	3,071	877[1]
August 2009	28,046	3,282	906
October 2009	30,618	3,482	949
December 2009	33,415	3,630	991

[1]Released as unofficial evaluations.

[2]Released as unofficial evaluations in December 2008 but considered to be official in January 2009.

posed because a low call rate usually indicates poor quality or small quantity of DNA. Parentage checking was quite effective in detecting switched samples and enabling correction of parentage errors. Because the final evaluation is a blend of genomic predictions and traditional evaluations that include mean genetic merit of parents, correct parent information contributes to evaluation accuracy.

Updating genotype files for each evaluation did not allow easy recovery of genotypes for animals that had previously been excluded because of a conflict but could be recovered because of a pedigree correction. Only SNP currently being used for evaluation were maintained in the system. Increasing the number of SNP to be used required accessing the original data. To address those issues, USDA developed a database that supported storage of genotypes for all samples and all SNP.

For initial USDA genomic evaluation of dairy cattle, separate sets of SNP were used for the Holstein, Jersey, and Brown Swiss breeds (Wiggans et al., 2009). The development of a SNP genotype database was an opportunity to change the SNP selected for calculation of genomic evaluations. Thus, using the same set of SNP across breeds and increasing the number of SNP used for genomic evaluation could be investigated.

The objectives of this report are to 1) detail the structure of the genotype database developed by USDA for dairy cattle, 2) document how SNP genotypes in the database were selected for use in genomic evaluation, and 3) describe the procedure used to designate which sample genotype is used for each animal.

DATABASE TABLES

Using the IBM DB2 database management system (http://www-01.ibm.com/software/data/db2/), database tables were created to store genomic information. The tables are maintained with C programming language, and web interface is through ColdFusion (http://www.adobe.com/products/coldfusion). The application was patterned after USDA's strategy for yield information (Norman et al., 1994) and supports tight integration with USDA's pedigree data for dairy cattle. A web query was developed to allow an organization to record information for animals that it intends to genotype (i.e., nominate an animal) in a nomination database table. The query also allows the organization to confirm that the animal is included in the pedigree database as well as indicate that it is genotyping the animal; this procedure avoids submission of multiple samples for an animal and designates the recipient of the genomic evaluation. Genotyping laboratories also are provided access to the nomination table to enable verification of sample identification, thereby reducing the risk of incorrect assignment of genotypes.

Genotypes are submitted monthly from 4 commercial laboratories (DNA Landmarks, Saint-Jean-sur-Richelieu, Quebec, Canada; GeneSeek, Lincoln, NE; Genetic Visions, Middleton, WI; and Genetics and IVF Institute, Fairfax, VA). Conflict reports are provided to the laboratories and to genotype requesters shortly after the genotypes are received. The database currently contains BovineSNP50 BeadChip genotypes from commercial laboratories, those collected for the original research on USDA genomic evaluations (Wiggans et al., 2009), and those provided by the Swiss Brown Cattle Breeders Federation (Zug, Switzerland) for 523 Brown Swiss bulls.

The relational database table for genotypes has 1 row per sample and 18 columns for identification and descriptive information as well as a large character object (58,336 bytes) containing the genotypes for individual SNP (Table 2). A genotype for an individual SNP is the number of occurrences (0, 1, or 2) of the counted allele (arbitrarily chosen); if the allele indicator is missing, a genotype of 5 is assigned. The Illumina BovineSNP50 BeadChip barcode and position letter (A to L) uniquely identifies the sample, which allows multiple unique genotypes to be stored for each animal, which is identified by an internal sequence number (animal key). The inclusion of a breed evaluation code supports extraction by breed.

The usability indicator (Y = yes or N = no) identifies which sample (if any) from an animal will be used for genomic evaluation. When multiple genotypes are stored for an animal and they differ by <1,000 SNP, the genotypes are assumed to be from the same animal and the sample with the highest call rate is designated as usable. If ≥1,000 SNP differ, the most recent sample is retained and the conflicting genotype is stored with a negative animal key. This procedure allows the conflict-

Table 2. Column descriptions for genotype database table

Column no.	Description
1	Internal sequence number (animal key)
2	Illumina-assigned chip identifier[1] (barcode)
3	Sample location on the chip (A–L)
4	Date that row was added or identification was modified
5	Laboratory where chip was prepared and scanned
6	Evaluation breed to which genotype contributes
7	Sample identification when sent to the laboratory
8	AI organization or breed association requesting genotyping
9	Sample plate identifier
10	Amplification plate identifier
11	Sample location on sample plate
12	Date the chip was scanned
13	Number of SNP for which genotyping was successful (SNP read quantity)
14	Large character object that stores 58,336 SNP genotypes
15	Indicator to identify if genotype is usable (Y or N)
16	Number of parents with usable genotypes
17	Number of progeny with usable genotypes
18	Tissue from which DNA was extracted

[1]Illumina Inc., San Diego, CA.

ing genotype to be recovered if it is later determined to be correct. For an animal with a positive animal key and multiple genotypes, a composite genotype is formed during data extraction by using SNP genotypes from other samples to replace those missing from the sample designated as usable. The numbers of parents and progeny with usable genotypes also are stored and used in deciding which of 2 genotypes with a parent–progeny conflict is most likely to have been assigned to the correct animal.

The tissue source identifies whether hair, semen, blood, or other tissue was the source of the DNA. That information is useful in understanding differences in call rate and, for blood, indicating the possibility of placental mixing of twin DNA.

Although conflicts for homozygous SNP genotypes can be detected if either parent has been genotyped (Wiggans et al., 2009), conflicts for heterozygous genotypes can be detected only if both parents have been genotyped. If $\geq$400 SNP (or $\geq$200 homozygote) genotypes conflict between a progeny and its parent, the animal genotypes are declared to be conflicting and, therefore, not usable.

Each sample genotype is compared with all other genotypes in the database to see whether a parent–progeny relationship exists that is not found in the pedigree. If such a relationship exists, the animal's genotype is not usable. Comparison with all other genotypes also allows detection of identical genotypes. A table that identifies animals with legitimate identical genotypes (identical twins, split embryos, and clones) is maintained, and their genotypes are usable. A parent–progeny relationship is accepted between the offspring of an animal and its identical twin or clone.

A call rate of $\geq$90% is required for autosomal SNP, and a call rate of $\geq$80% is required for X-specific SNP, which are used for validation of the sex of the genotyped animal. A table that identifies females with no heterozygous X-specific SNP as a result of inheriting both X chromosomes from the same ancestor is maintained.

Animal breed (Holstein, Jersey, or Brown Swiss) is validated through SNP that are almost monomorphic in 1 breed and have $\leq$30% of animals homozygous for that allele in another breed. A total of 622 SNP were selected with approximately equal numbers of monomorphic SNP for each of the 3 breeds. The number of SNP for which a genotype differs from the monomorphic genotype is counted separately for each breed, and the lowest breed count identifies the sample breed.

Most data conflicts are correctable. When a sire conflict is detected, the genomic sire usually is reported as a parent or progeny relationship not documented in the pedigree because most sires have been genotyped. Switches in sample identification are the most common reason for conflicts. Conflicts because of unreported identical genotypes usually can be eliminated by determining that the animals are identical twins or from a split embryo; that information then is added to the table of animals with identical genotypes. In some cases, 2 samples are erroneously collected from only 1 animal of a full-sib pair. Some errors remain because the animals are no longer of interest or are from research projects for which an accurate pedigree was not needed. Table 3 shows the number of animals for which the genotype was unusable for December 2009 genomic evaluations for various reasons even after record correction.

Table 3. Animals excluded from genomic evaluation as of December 2009 by reason for exclusion for 38,287 genotyped animals

Reason for exclusion	n
Autosomal call rate of <90%	30
Call rate for X-specific SNP of <80%	12
Count of >50 heterozygous X-specific SNP for a male	11
Count of ≤50 heterozygous X-specific SNP for a female	18
Sire conflict	127
Dam conflict	6
Progeny conflict	7
Identical genotype (not twin or clone)	1
Genomic parent–progeny relationship not found in pedigree data	40
All	251

Of the 38,306 animals with usable genotypes for December 2009 evaluations, 1,175 (3.1%) had >1 genotype and 82 (0.2%) had 3 genotypes. The additional genotypes resulted from identification errors and attempts to resolve conflicts. Both parents had been genotyped for 39% of the 14,571 genotyped calves born during the 24 mo preceding December 2009.

For animals with usable genotypes, SNP conflicts with parent genotypes are resolved by designating the genotype of the animal, or the parent, or both, as missing. Each animal is compared with all its genotyped parents and progeny, and the one with the smaller fraction of confirmations has its genotype for that SNP set to 5 (missing). If the fractions are equal (e.g., neither the animal nor its parent had any other genotyped parents or progeny), then the SNP genotype is set to 5 for both animals.

GENOTYPE CALLING

To obtain genotypes, Illumina BovineSNP50 BeadChips are scanned after processing with the Infinium HD Assay (Illumina Inc., San Diego, CA) to create color intensity files in which each allele is specified by red or green. Those files are input for GenomeStudio software (Illumina Inc., San Diego, CA), which assigns SNP marker genotypes by clustering assay results for each animal and SNP based on differences in signal and color intensity. Reprocessing of the intensity files was required to provide genotypes for SNP that had been skipped previously, primarily because of low minor allele frequency (**MAF**) for Holsteins.

The database allows for 58,336 SNP, the total number of SNP assays that was initially placed on the BovineSNP50 BeadChip. Of those, 56,947 were included in the research set used by Wiggans et al. (2009); the other SNP were not considered functional after manufacturing by Illumina. The number of included SNP was further reduced to 50,972 after removing SNP found not to be useful because they were monomorphic across all cattle breeds, contained other polymorphisms (flanking SNP or insertion/deletions) that affect either oligonucleotide hybridization or SNP signal intensity and color detection (i.e., SNP clustering), were null alleles, or had patterns that could not be scored as the result of detected misinheritances or genome duplications.

Genotypes for additional SNP were collected because 1) the large increase in the number of genotyped animals increased the usefulness of SNP with an MAF of <5%, 2) many monomorphic Holstein SNP were useful for Jerseys or Brown Swiss, and 3) some parentage SNP had been excluded (Heaton et al., 2007; Matukumalli et al., 2009). Because only 40,874 SNP had been called for prior genomic evaluations (Wiggans et al., 2009), approximately 19,000 genotypes called before March 2009 were redone for Holsteins, Jerseys, and Brown Swiss.

Groups of samples were assembled based on manufacturing date and reagent lot number for clustering by GenomeStudio. A group size of approximately 300 was used because it could be processed in reasonable time and provided enough observations to justify cluster adjustments. Use of the previous cluster file enabled GenomeStudio to call most SNP automatically. However, when the call rate was <93%, clustering was checked to see whether it could be manually improved by adjusting the boundaries between the 3 SNP genotypes. Processed genotypes were then checked against genotypes currently included in evaluations to detect identification corrections that had been made to the evaluation data set but not to the identifications associated with the intensity files. After genotypes were determined, cluster files were provided to the 4 commercial laboratories to standardize the set of called SNP and provide an initial clustering to maximize consistency across laboratories.

SNP SELECTION

A total of 29,548 sample genotypes (25,594 Holstein, 3,083 Jersey, and 871 Brown Swiss) were extracted.

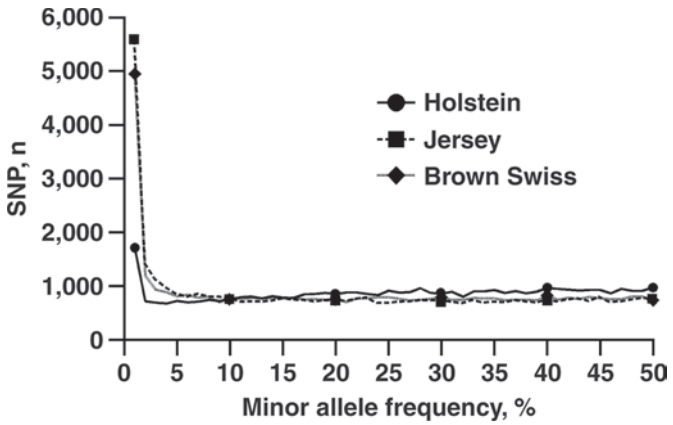

Figure 1. Minor allele frequency by breed (Holstein, Jersey, and Brown Swiss) for 43,385 SNP selected for calculating USDA national genomic evaluations.

From those genotypes, MAF were determined, and SNP with an MAF of ≥1% for at least 1 breed were retained. Next, call rate was checked. For a SNP with an MAF of 50%, a call rate of ≥90% was required. That requirement increased to 100% as MAF declined to 0. Similarly, the percentage of parent–progeny conflicts allowed was decreased from 1 to 0% as MAF decreased from 50%. Remaining SNP were checked against all other SNP with similar MAF, and redundant SNP were eliminated. Genomic evaluations were calculated using the same SNP set across breeds to determine whether the presence of many monomorphic SNP affected the rate of convergence.

Numbers of SNP by reason for exclusion from use in genomic evaluation are given in Table 4. Of the 14,951 SNP excluded, 2,378 were removed because of the new checks for low call rate or high rate of parent–progeny conflict.

Of the 43,385 SNP that could be used for genomic evaluation, Jerseys and Brown Swiss both had >11%, with an MAF of ≤1% compared with <4% for Holsteins (Figure 1). For all breeds, numbers of SNP were fairly uniform (2% of total SNP) for MAF from 2 to 50%. Inclusion of SNP with MAF of <1% did not affect convergence of solutions for SNP effects.

Previous USDA genomic evaluations had been based on 38,416 SNP for Holsteins, 31,658 SNP for Jerseys, and 34,593 SNP for Brown Swiss (VanRaden et al., 2009a). Based on the methods of VanRaden et al. (2009b) for assessing improvement in reliability of genetic evaluations from inclusion of genomic predictions, the added SNP increased reliability across traits (not shown) by 0.4 percentage units for Holsteins (range of −0.5 to 0.9) and 0.3 percentage units for Jerseys (range of −0.5 to 0.7); reliability across traits for Brown Swiss (not shown) decreased by 0.2 percentage units (range of −1.0 to 1.2), most likely because few Brown Swiss genotypes were available. Therefore, a single set of 43,385 SNP was adopted for all breeds for calculation of genomic evaluations in August 2009.

CONCLUSIONS

Methods for managing genomic data have been developed to allow storage in a database, immediate determination and reporting of conflicts, convenient access to genotypes, and documentation of reasons for exclusion. A web query enables requesters to nominate animals, which provides a check on animal identification and indicates who should receive the evaluation. Another query shows identification information for all genotypes for an animal and, if none are usable, the reasons why. A single set of 43,385 SNP for use across breeds was implemented in August 2009 to simplify data management and allow research on across-breed evaluation.

The evolution of SNP assays in marker density (both higher and lower) and cost will generate a need to be able to manage genotypes across assay platforms. The database developed can be adapted by 1) adding a platform code to indicate the set of SNP contained in the genotype segment, 2) standardizing for the highest density platform and then either assigning a genotype of 5 (missing) or imputing missing SNP genotypes for data from lower density platforms, or 3) creating sister tables for genotypes from each platform.

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 (Washington, DC) and by the National Association of Animal Breeders (Columbia, MO), Holstein Association USA (Brattleboro, VT), American Jersey Cattle Association (Reynoldsburg, OH), and Brown Swiss Cattle Breeders' Association (Beloit, WI). Genotypes were contributed by Semex Alliance (Guelph, Ontario, Canada) and the Center for Genetic Improvement of Livestock (University of Guelph, Ontario, Canada). Frank Ross (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD) developed the web nomination system. The authors thank Kent Weigel (University of Wisconsin–Madison) and Suzanne Hubbard (Animal Improvement Programs Laboratory, ARS, USDA) for manuscript review and suggestions and Alicia Beavers (Bovine Functional Genomics Laboratory, ARS, USDA, Beltsville, MD) for technical assistance in refining genotypes and confirming sample exchanges.

Table 4. Total number of SNP by reason for exclusion from use in genomic evaluation

Reason for exclusion	n
Maximum number of SNP for Illumina Bovine SNP50 BeadChip[1]	58,336
Insufficient number of oligo-bound beads (3 microns) to assay SNP accurately	1,389
Unscorable SNP	5,975
Minor allele frequency of <1%	3,488
Call rate of <90%	2,017
Parent–progeny conflict rate of >1%	361
Highly correlated	1,721
Used for genomic prediction	43,385

[1]Illumina Inc., San Diego, CA.

REFERENCES

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82(E-Suppl.):E313–E328.

Heaton, M. P., W. M. Snelling, T. P. Smith, J. W. Keele, G. P. Harhay, R. T. Wiedmann, G. L. Bennett, B. A. Freking, C. P. Van Tassell, T. S. Sonstegard, L. C. Gasbarre, S. S. Moore, B. Murdoch, S. D. Mckay, T. Kalbfleisch, and W. W. Laegreid. 2007. A marker set for parentage-based DNA traceback in beef and dairy cattle. Abstract P516 in Proc. XV Plant Anim. Genome Conf., San Diego, CA. http://www.intl-pag.org/15/abstracts/PAG15_P05k_516.html Accessed Nov. 16, 2009.

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 4:e5350.

Norman, H. D., L. G. Waite, G. R. Wiggans, and L. M. Walton. 1994. Improving accuracy of the United States genetics database with a new editing system for dairy records. J. Dairy Sci. 77:3198–3208.

Soller, M. 1994. Marker assisted selection–An overview. Anim. Biotechnol. 5:193–207.

VanRaden, P., G. Wiggans, M. Tooker, J. Hutchison, and L. Bacheller. 2009a. Revised marker set used in genomic predictions. Changes to evaluation system (August 2009). http://aipl.arsusda.gov/reference/changes/eval0908.html Accessed Sept. 24, 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.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.