J. Dairy Sci. 102:2308–2318
https://doi.org/10.3168/jds.2018-15419

Modeling pedigree accuracy and uncertain parentage in single-step genomic evaluations of simulated and US Holstein datasets

H. L. Bradford,[1,2]* Y. Masuda,[3] J. B. Cole,[1] I. Misztal,[3] and P. M. VanRaden[1]
[1]Animal Genomics and Improvement Laboratory, Agriculture Research Service, USDA, Beltsville, MD 20705-2350
[2]Department of Animal and Poultry Science, Virginia Tech, Blacksburg 24061
[3]Department of Animal and Dairy Science, University of Georgia, Athens 30605

ABSTRACT

The objective of this study was to model differences in pedigree accuracy caused by selective genotyping. As genotypes are used to correct pedigree errors, some pedigree relationships are more accurate than others. These accuracy differences can be modeled with uncertain parentage models that distribute the paternal (maternal) contribution across multiple sires (dams). In our case, the parents were the parent on record and an unknown parent group to account for pedigree relationships that were not confirmed through genotypes. Pedigree accuracy was addressed through simulation and through North American Holstein data. Data were simulated to be representative of the dairy industry with heterogeneous pedigree depth, pedigree accuracy, and genotyping. Holstein data were obtained from the official evaluation for milk, fat, and protein. Two models were compared: the traditional approach, assuming accurate pedigrees, and uncertain parentage, assuming variable pedigree accuracy. The uncertain parentage model was used to add pedigree relationships for alternative parents when pedigree relationships were not certain. The uncertain parentage model included 2 possible sires (dams) when the sire (dam) could not be confirmed with genotypes. The 2 sires (dams) were the sire (dam) on record with probability 0.90 (0.95) and the unknown parent group for the birth year of the sire (dam) with probability 0.10 (0.05). An additional set of assumptions was tested in simulation to mimic an extensive dairy production system by using a sire probability of 0.75, a dam probability of 0.85, and the remainder attributed to the unknown parent groups. In the simulation, small bias differences occurred between models based on pedigree accuracy and genotype status. Rank correlations were strong between traditional and uncertain parentage models in simulation ($\geq$0.99) and in Holstein ($\geq$0.99). For Holsteins, the estimated breeding value differences between models were small for most animals. Thus, traditional models can continue to be used for dairy genomic prediction despite using genotypes to improve pedigree accuracy. Those genotypes can also be used to discover maternal parentage, specifically maternal grandsires and great grandsires when the dam is not known. More research is needed to understand how to use discovered maternal pedigrees in genetic prediction.

Key words: average relationship matrix, genotype, pedigree error, selective genotyping

INTRODUCTION

Pedigree errors occur for all species and can result from various types of recording and data entry errors, including the use of nonunique identification. These errors can be affected by management considerations such as grouped calving, use of natural service bulls, and use of multiple AI bulls. Pedigree accuracy can be verified with technology, enabling farmers and organizations to identify and to correct errors. Historically, breeding organizations discovered pedigree errors through blood groups, followed by microsatellites, and now SNP genotypes. Parentage testing was always required for US AI bulls, donor dams, and for 1 of every 3 Holstein (1 of 10 for Jersey) registered calves born by embryo transfer (**ET**), but was required for few non-ET registered females (1 of 500 US Jerseys) and for no-grade females in milk recording; other countries may have had similar requirements. Today, many more producers test parentage of their calves voluntarily as part of genomic prediction. Sire pedigree errors occurred for 10% of dairy cattle based on older technologies (Geldermann et al., 1986; Visscher et al., 2002; Weller et al., 2004). Genotypes have been used to confirm, to correct (Wiggans et al., 2012), and to discover (VanRaden et al., 2013) parentage in US Holstein. Parentage is confirmed when the genotypes of an animal and its parent are consistent

Received July 20, 2018.
Accepted November 14, 2018.
*Corresponding author: hbradford@vt.edu

with a parent-progeny relationship. As breeding organizations use genotypes to correct pedigrees, selective genotyping causes more accurate pedigree relationships between elite bulls and their daughters. Differences in daughter pedigree accuracy among bulls may create bias and prevent fair comparisons.

Pedigree errors affect genetic selection through biases in genetic parameters, EBV, and genetic trend. These errors caused heritability to be underestimated (Van Vleck, 1970; Geldermann et al., 1986) and reduced dispersion of EBV (Geldermann et al., 1986). Additionally, pedigree errors favored selection of young bulls instead of proven bulls (Israel and Weller, 2000), creating bias when selecting across generations. Bias also was created when selecting animals across countries, causing selection of domestic instead of foreign bulls (Banos et al., 2001). Pedigree errors decreased genetic gain by 4 to 17% (Geldermann et al., 1986; Israel and Weller, 2000) because of these issues. The lost genetic gain may be recovered by better modeling pedigree accuracy, enabling fair comparisons among animals.

The average numerator relationship matrix (Henderson, 1988) was developed for multiple-sire mating scenarios. Using this example, an animal had multiple possible sires and each sire had a probability of being the true sire. These probabilities could be equal or could vary as long as the probabilities sum to unity. These uncertain parentage models have been used in simulation (Perez-Enciso and Fernando, 1992; Cardoso and Tempelman, 2003) and in beef cattle (Cardoso and Tempelman, 2004; Shiotsuki et al., 2012) to model multiple-sire mating with unknown paternity in the offspring. Generally, uncertain parentage was better than having no pedigree or having unknown parent groups (**UPG**; Quaas, 1988) for sires, but worse than having the true pedigree. Uncertain parentage models could correct some of the problems caused by pedigree errors. To our knowledge, researchers have not applied these models to account for differences in pedigree accuracy caused by genotyping.

Dairy pedigrees are complex, with heterogeneous pedigree depth that is modeled with UPG. Heterogeneous pedigree accuracy occurs because selective genotyping causes differences in daughter pedigree accuracy across bulls. Our objective was to assess the performance of uncertain parentage models in simulation and in dairy cattle to better model differences in pedigree accuracy.

MATERIALS AND METHODS

Animal care and use committee approval was not needed, as data were obtained from existing databases.

Uncertain Parentage Model

Differences in pedigree accuracy were modeled with the average relationship matrix (Henderson, 1988). The EBV was written as

$$u_i = 0.5\sum_{j=1}^{n\,sire} p_{ij}u_j + 0.5\sum_{k=1}^{n\,dam} p_{ik}u_k + m_i,$$

where u_i was the EBV for animal i, p_{ij} was the probability that j was the parent of i, n *sire* (n *dam*) was the number of possible sires (dams) for i, and m_i was the Mendelian sampling term. With this derivation, the probabilities summed to unity for all possible sires, and the same was true for dams. At most, 2 possible sire (dam) contributions were considered: the sire (dam) on record or the UPG corresponding to that sire's (dam's) birth year. Previously, rules were developed to create $\mathbf{A}^{-1}$ from uncertain parentage (Famula, 1992; Perez-Enciso and Fernando, 1992) as

$$\begin{array}{c} \\ \text{animal} \\ \text{sire} \\ \text{sire UPG} \\ \text{dam} \\ \text{dam UPG} \end{array}\begin{bmatrix} 1 & -0.5p_{is} & -0.5(1-p_{is}) & -0.5p_{id} & -0.5(1-p_{id}) \\ & 0.25p_{is}^2 & 0.25p_{is}(1-p_{is}) & 0.25p_{is}p_{id} & 0.25p_{is}(1-p_{id}) \\ & & 0.25(1-p_{is}) & 0.25(1-p_{is})p_{id} & 0.25(1-p_{is})(1-p_{id}) \\ & \textit{symmetric} & & 0.25p_{id}^2 & 0.25p_{id}(1-p_{id}) \\ & & & & 0.25(1-p_{id})^2 \end{bmatrix},$$

where s was the sire and d was the dam on record. The contributions to $\mathbf{A}^{-1}$ were multiplied by the inverse of the variance of Mendelian sampling as

$$\left[\left(1-0.25\sum_{j=1}^{n\,sire}\sum_{j'=1}^{n\,sire} p_{ij}p_{ij'}a_{jj'} - 0.25\sum_{k=1}^{n\,dam}\sum_{k'=1}^{n\,dam} p_{ik}p_{ik'}a_{kk'}\right)\sigma_u^2\right]^{-1},$$

where $a_{jj'}$ is the additive relationship between j and j' and σ_u^2 is the additive variance. The Mendelian sampling variance reduced to Henderson (1976) when parents were accurate.

We made the following assumptions for pedigree certainty. All combinations existed for pedigree certainty: 2 accurate parents, 1 accurate and 1 uncertain parent, and 2 uncertain parents. If the parent-progeny relationship was confirmed through genotyping, then we used the parent on record and assumed accurate parentage. If the parent-progeny relationship was not confirmed, then the uncertain parentage was modeled, and 2 possible assumptions were considered: intensive and extensive. The intensive assumption modeled expected pedigree errors similar to those tracked in the Council for Dairy Cattle Breeding (**CDCB**; Bowie, MD; https://

www.uscdcb.com/) database. In those data, genotyped animals had 9% sire errors and 3% dam errors when the parent was genotyped. These values were based on pedigree corrections where initial parentage before genotyping was saved as a reference. These values were likely underestimated, because, over time, producers were no longer required to submit initial parentage information before genotyping, and the genotype results were used to discover pedigree and to provide expected parentage. Hence, probabilities were rounded up to 10 and 5% errors, respectively, to better match reality; this approach resulted in probabilities of 0.90 for the recorded sire and 0.95 for the recorded dam. The remaining contributions were attributed to the appropriate UPG. These assumptions mean that the maternal grandsire (**MGS**) had a $0.95 \times 0.90 = 0.86$ probability of being correct and that the maternal great-grandsire (**MGGS**) had a $0.95 \times 0.95 \times 0.90 = 0.81$ probability of being correct. The extensive assumptions modeled expected pedigree errors similar to those reported in pasture-based dairy production systems (Stephen et al., 2018). This approach resulted in probabilities of 0.75 for the recorded sire and 0.85 for the recorded dam. These assumptions mean that the MGS had a $0.85 \times 0.75 = 0.64$ probability of being correct and that the MGGS had a $0.85 \times 0.85 \times 0.75 = 0.54$ probability of being correct.

All analyses used single-step genomic BLUP (Aguilar et al., 2010; Christensen and Lund, 2010) with UPG based on $\mathbf{H}^{-1}$ (the unified pedigree and genomic relationship matrix; Misztal et al., 2013). In all cases, **G** (the genomic relationship matrix) was constructed by blending $0.95\mathbf{G}$ with $0.05\mathbf{A}_{22}$, the numerator relationship matrix for genotyped animals, and **G** was scaled to have the same mean diagonal and off-diagonal elements as $\mathbf{A}_{22}$. No other adjustments were made to $\mathbf{H}^{-1}$. Modifications were made to the BLUPF90 family of programs (Misztal et al., 2018) to account for parentage uncertainty in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$.

Simulation

Data. Data were simulated using QMSim v1 (Sargolzaei and Schenkel, 2009), and 3 simulations for sex-limited traits were performed based on 3 heritabilities (0.1, 0.3, and 0.5). The simulations were replicated 20 times with the mean and standard deviation (**SD**) of replicates reported. The historical population started with 5,000 individuals, steadily reduced to 250 at historical generation 1,000, and steadily increased to 30,000 at historical generation 1,100. The last historical generation had 350 males and 29,650 females. From the last historical generation, 50 males and 14,950 females were selected to be founders of the current population that maintained a constant population size. These individuals were mated randomly with 1 offspring per female. Individuals were selected based on BLUP EBV for 10 overlapping generations, with 30% of young animals selected to replace the oldest individuals of each sex. All females had phenotypes except the most recent generation, resulting in 82,224 (178) phenotypes (SD).

To make the pedigree more realistic, we simulated incomplete and inaccurate pedigrees. We created incomplete pedigrees by randomly removing the sire with 0.08 probability and the dam with 0.18 probability. These probabilities were chosen to create more missing dams than sires while still being able to simulate errors for the remaining pedigrees. Inaccurate pedigrees consisted of differences in daughter pedigree accuracy for 2 groups of sires. In each generation, 15 new males were selected, and the 5 best males by BLUP EBV had daughters with accurate pedigrees; the remaining 10 males had daughters with possible pedigree errors. The replacement parent was selected from other parents of the progeny's generation. We simulated 2 assumptions for inaccurate pedigrees based on intensive and extensive dairy production systems. In the intensive system, these daughters had a 0.09 probability of a wrong sire and a 0.03 probability of a wrong dam; again, these probabilities aligned with pedigree errors in the CDCB data. In the extensive system, these daughters had a 0.25 probability of a wrong sire and a 0.15 probability of a wrong dam. The sire error rate was similar to that in New Zealand (Stephen et al., 2018) and the dam error rate was similar to beef cattle (Pollak, 2005; Carolino et al., 2009), as reports of dam errors outside the United States could not be found at the time of publication. The extensive system was meant to demonstrate a reasonable worst-case scenario for dairy pedigree errors.

We simulated a dairy cattle genome that consisted of 29 chromosomes with a total length of 2,319 cm, 50,000 biallelic SNP, and 500 biallelic QTL. The SNP and QTL had 0.5 allele frequencies to begin the historical population, 2.5×10^{-5} mutations per meiosis per loci, and 1 crossover per meter per meiosis. The SNP were equally spaced throughout the genome and had a mean (SD) pooled squared correlation coefficient among all SNP combinations per chromosome of 0.30 (0.02) based on internal calculations in QMSim. The QTL were randomly placed in the genome with effects from a Gamma distribution (shape = 0.4, scaled internally to match the heritability).

The most recent 5 generations were selectively genotyped. Within each generation, males and females were ranked in the top or bottom half based on BLUP EBV. The top half of males had a 0.35 probability of being

Table 1. Numbers of animals (SD) with missing or inaccurate pedigrees for simulated data for the trait with 0.3 heritability

Pedigree relationship	Missing pedigree		Inaccurate pedigree	
	All	Genotyped	All	Genotyped
Intensive system[1]				
Sire	11,658 (99)	1,454 (45)	9,461 (148)	748 (33)
Dam	27,197 (106)	3,384 (39)	3,133 (66)	247 (63)
Extensive system[2]				
Sire	11,658 (99)	1,464 (45)	26,465 (395)	2,092 (52)
Dam	27,197 (106)	3,384 (39)	15,825 (281)	1,252 (41)

[1]Simulated 9% sire errors and 3% dam errors.
[2]Simulated 25% sire errors and 15% dam errors.

genotyped, with the constraint that all sires were genotyped, and the bottom half of males had a 0.10 probability of being genotyped. The top half of females had a 0.35 probability of being genotyped and the bottom half of females had a 0.20 probability of being genotyped. This structure gave preference to better genetic merit animals and to females, resulting in 18,686 (8) genotyped animals (SD). Descriptive statistics for missing and inaccurate parents based on genotype status were presented in Table 1 for the 0.3 heritability trait. The other traits were similar and differed by less than 1 SD.

Model. Genotyped individuals had accurate parentage with $\mathbf{A}^{-1}$ created as usual (Henderson, 1976). The remaining individuals had uncertain parentage for their sire and dam with assumptions as described previously. Pedigree errors were simulated for intensive and extensive systems, and each simulation was analyzed separately with each set of uncertain parentage assumptions (2 data simulations × 2 model assumptions). Statistics were provided in Table 2 for numbers of uncertain pedigrees for the 0.3 heritability trait. Counts were similar for the other 2 traits and differed less than 1 SD. The UPG were defined based on generation as founders plus generations 1 to 4, 5 to 7, and 8 to 10. Sex was not used to define UPG because the UPG would have too few animals represented. Phenotypes were truncated at generation 9 and validations were performed for animals born in generation 10. Accuracy and dispersion were assessed for genotyped animals from generation 10. Accuracy was the correlation between true breeding value (**TBV**) and EBV. Dispersion was b_1 from TBV $= b_0 + b_1$EBV. Bias was $(\text{TBV} - \text{EBV})/\sigma_u$. For validation of the intensive simulation, there were 3,563 (52) genotyped individuals (SD) with correct parents, 1,489 (36) genotyped individuals with incorrect parents, 6,459 (65) not genotyped individuals with correct parents, and 3,438 (57) not genotyped individuals with incorrect parents. For validation of the extensive simulation, there were 3,128 (62) genotyped individuals (SD) with correct parents, 1,924 (46) genotyped individuals with incorrect parents, 4,421 (112) not genotyped individuals with correct parents, and 5,477 (114) not genotyped individuals with incorrect parents.

Holstein

Data. Data were obtained from the CDCB for the December 2017 Holstein evaluation. Phenotypes were edited to include 34 million records since 2000. Pedigrees were edited to remove instances where a progeny was older than the parent. Figure 1 presents the year of birth distribution for animals with at least 1 missing parent. The complete pedigree had 66 million animals, and a 3-generation pedigree for all phenotyped and genotyped animals had 21 million animals. We analyzed the data with the complete and with the reduced pedigree. The 1.8 million genotypes were edited to include all genotyped bulls (n = 231,396) and cows with at least 1 phenotype (n = 442,258), resulting in 673,654 genotyped animals with 60,671 SNP.

Model. We made the following assumptions about pedigree accuracy. Holsteins had accurate parentage if the parent and offspring were both genotyped. Parentage was also accurate for both parents of AI sires and dams of AI sires because SNP, microsatellites, or

Table 2. Numbers of animals (SD) with uncertain pedigrees for a simulated trait with 0.3 heritability

Pedigree relationship	Uncertain parentage assumptions	
	Intensive[1]	Extensive[2]
Sire	12,408 (93)	12,408 (93)
Dam	5,465 (58)	5,465 (58)
Sire and dam	50,127 (181)	49,530 (190)
Neither	96,501 (160)	97,098 (169)

[1]Assume the sire is correct with probability 0.90 and the dam is correct with probability 0.95.
[2]Assume the sire is correct with probability 0.75 and the dam is correct with probability 0.85.

Table 3. Percent (unless noted) of Holsteins with missing or uncertain parent(s) for all and genotyped animals

	Missing pedigree		Uncertain pedigree	
Pedigree relationship	All	Genotyped	All	Genotyped
3-generation pedigree				
Sire	3.5	0.01	0.01	0.2
Dam	9.6	7.0	1.5	46.6
Sire and dam	11.1	0.2	96.7	1.4
Neither	75.8	92.7	1.8	51.7
Total (n)	20,917,044	673,654	20,917,044	673,654
Complete pedigree				
Sire	4.9	0.01	0.01	0.2
Dam	9.0	7.0	1.2	46.6
Sire and dam	15.5	0.2	97.3	1.4
Neither	70.6	92.7	1.5	51.7
Total (n)	65,927,043	673,654	65,927,043	673,654

blood groups confirmed these pedigrees. For remaining animals, we modeled uncertain parentage as previously described for the intensive production system. Statistics were provided in Table 3 for numbers of missing and uncertain pedigrees in Holstein. We compared EBV from models assuming known parentage and those assuming uncertain parentage. The EBV were compared to be consistent with the simulation results, even though PTA were standard in the industry. Based on the EBV comparisons, the models were not validated further.

Single-step genomic BLUP with the algorithm for proven and young was used with 10,000 random core animals (Misztal et al., 2014). Preliminary testing showed 10,000 random core animals was not different from 15,000 or 20,000. Milk, fat, and protein were analyzed in single-trait, repeatability animal models as

$$y_{ijklmn} = herd_i + parity_j + \beta_1 inb_k + herdsire_l + u_m + pe_m + e_{ijklmn},$$

where y is the milk, fat, or protein adjusted 305-d yield; *herd* is the herd-management group; *parity* is the age-parity group; β_1 is the regression coefficient for inbreeding, *inb*; *herdsire* is the random herd-sire interaction; u is the random additive genetic effect; *pe* is the random permanent environmental effect; and e is the random residual. Inbreeding from the official evaluation was used for all analyses (VanRaden, 1992). All random effects were assumed to be normally distributed, with variances

$$\text{var}\begin{bmatrix} herdsire \\ u \\ pe \\ e \end{bmatrix} = \begin{bmatrix} \mathbf{I}\sigma^2_{hs} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}\sigma^2_u & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_{pe} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}\sigma^2_e \end{bmatrix},$$

where $\mathbf{I}$ was the identity matrix and σ^2_{hs} was the variance for the herd-sire interaction. Residuals were weighted by a diagonal matrix, $\mathbf{D}$, with weights calculated based on lactation length and deviations from constant heritability (VanRaden et al., 1991; Wiggans and VanRaden, 1991). In the national evaluation, UPG were originally defined based on breed, country of origin, and selection path, but were redefined for the subset of data used in this analysis. Unknown parent groups were categorized based on year of birth as 0 to 1999, 2000 to 2001, 2002 to 2003, 2004 to 2005, 2006 to

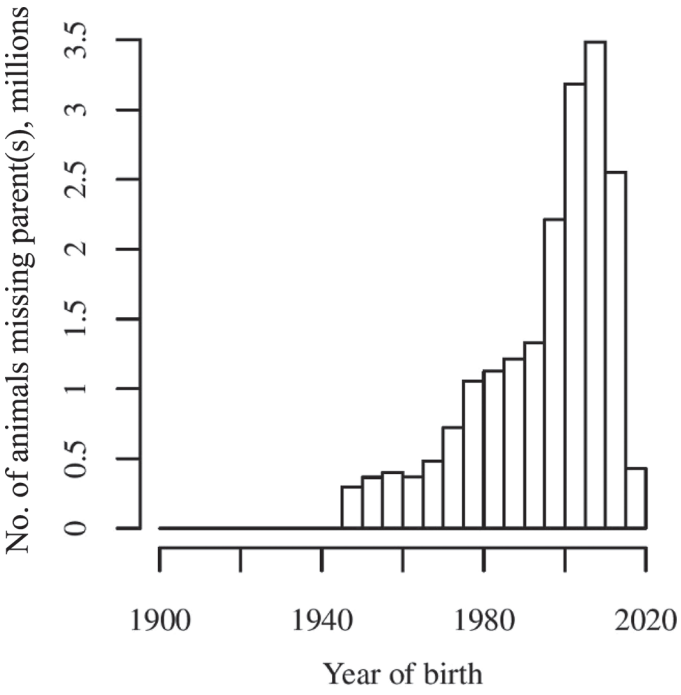

Figure 1. Histogram of the birth year for animals with at least 1 missing parent in the full US Holstein pedigree.

2007, 2008 to 2009, 2010 to 2011, and 2012 to 2017; everything before 2000 was grouped together because the data were cut at 2000. Maintaining a constant interval would make some older UPG difficult or impossible to estimate because of poor ties to phenotypes. Everything after 2012 was grouped together because females born in the last 2 yr have no phenotypes and do not contribute to estimating the last UPG. Selection path was not used to define UPG, as complicated UPG definitions have led to unreasonable EBV predictions in single-step genomic BLUP (**ssGBLUP**; Y. Masuda, unpublished data).

RESULTS AND DISCUSSION

Figure 2 presents the percent of daughters genotyped for bulls born at the start of genomic selection in 2009 and born in 2013, making them the youngest bulls that could have daughters lactating in 2017. These bulls had at least 1 phenotyped daughter and 1 genotyped daughter. Initially, most bulls had very few genotyped daughters, but, as more females were genotyped, some bulls had a greater proportion of daughters genotyped. Hence, the genotyping structure changed over time and will continue to change. The same distribution was presented in Figure 3 for the top 10% of net merit AI bulls from the CDCB official evaluation. Figure 4 showed the net merit and percent of daughters genotyped for these AI bulls. Although most bulls had a small proportion of their daughters genotyped, more variation existed in daughter genotyping for top net merit bulls than the entire bull population. The US dairy industry is changing rapidly, with greater genetic control by AI companies, and the variation in daughter genotyping could reflect a shift toward genotyping more elite individuals at an earlier age. Additionally, the more extreme net merit bulls had a greater proportion of their daughters genotyped, and this trend likely will continue as more females are genotyped. The genotyped daughters had parentage confirmed and the performance of those females was attributed to the correct sire. For daughters without genotypes, some had pedigree errors and were attributed to the wrong sire. This variation in pedigree accuracy was a concern for appropriately ranking sires with large differences in daughter pedigree accuracy, as pedigree errors caused EBV to regress toward the mean.

For the simulation, traditional and uncertain parentage models yielded similar results for accuracy and dispersion. Results follow for the traditional model with validations for young genotyped animals. All uncertain parentage results (either intensive or extensive assumptions) were within 1 SD of the certain parentage. For the intensive simulation, accuracies (SD) were 0.52

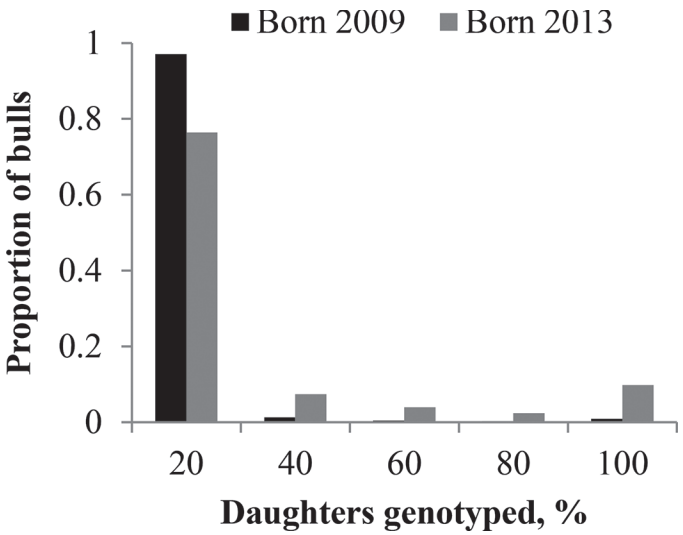

Figure 2. Percent of daughters genotyped for Holstein bulls born in 2009 and 2013.

(0.04) for the 0.1 heritability trait, 0.65 (0.03) for the 0.3 heritability trait, and 0.71 (0.02) for the 0.5 heritability trait. Dispersions (SD) were 0.76 (0.07) for the 0.1 heritability trait, 0.83 (0.06) for the 0.3 heritability trait, and 0.85 (0.04) for the 0.5 heritability trait. For the extensive simulation, accuracies (SD) were 0.50 (0.04) for the 0.1 heritability trait, 0.63 (0.04) for the 0.3 heritability trait, and 0.69 (0.02) for the 0.5 heritability trait. Dispersions (SD) were 0.71 (0.08) for the 0.1 heritability trait, 0.79 (0.06) for the 0.3 heritability trait, and 0.81 (0.04) for the 0.5 heritability trait. In all cases, the intensive uncertain parentage assump-

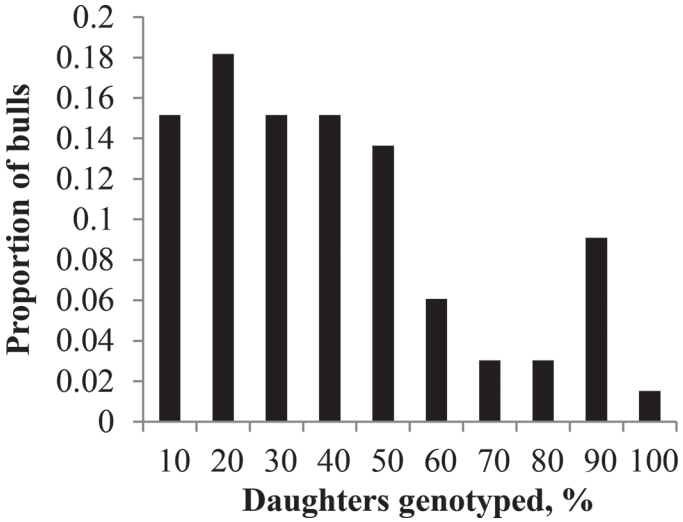

Figure 3. Percent of daughters genotyped for the top 10% of net merit Holstein AI bulls.

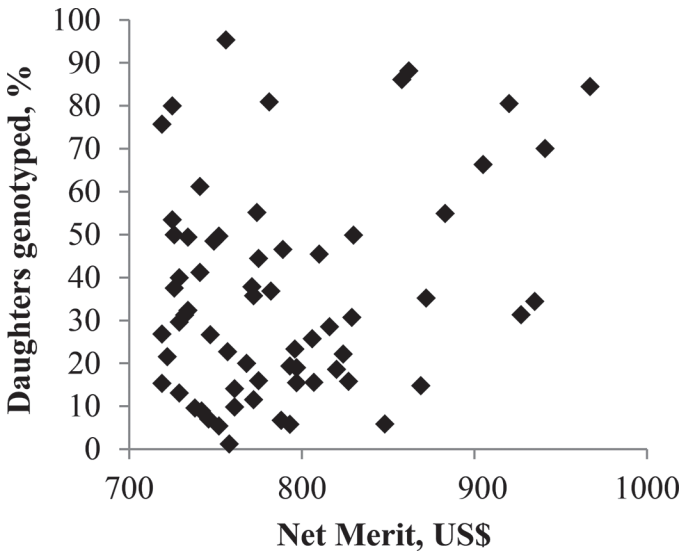

Figure 4. Net merit and percent of daughters genotyped for 65 Holstein AI bulls in the top 10% for net merit.

tions were numerically more accurate and less overdispersed, although the difference was typically 0.01. In all cases, the over dispersion could be corrected by using omega less than 1 in ssGBLUP. All models had similar numbers of rounds to convergence. Hence, modeling parentage certainty did not change the accuracy or dispersion of EBV for the most recent generation. Variation of EBV was similar between models, with slightly more variation with the uncertain parentage model. Previously, pedigree errors reduced the variance of EBV and increased regression coefficients (Israel and Weller, 2000; Banos et al., 2001). Hence, accounting for pedigree accuracy helped compensate for the reduction in dispersion caused by pedigree errors.

Scaled bias was presented in Table 4 for the intensive pedigree error simulation and in Table 5 for the extensive pedigree error simulation. For genotyped young animals with accurate pedigrees, those animals had less bias with the uncertain parentage model, with the extensive assumptions being least biased even for the simulation with fewer pedigree errors (Table 4). In some cases, the bias was not different from 0 when animals had accurate pedigrees. The opposite happened for genotyped young animals with inaccurate pedigrees, where the extensive assumptions were the most biased, even for the extensive simulation that had pedigree errors simulated at the same proportion as the extensive uncertain parentage (Table 5). These results could be caused by selective genotyping, where better animals were more likely to be genotyped and also more likely to have accurate pedigrees. When animals with pedigree errors were genotyped, a small UPG contribution did not help the bias that already existed because of the pedigree error. Using uncertain parentage increased bias for young animals without genotypes. Many of those animals likely had correct pedigrees, resulting in a slight bias from assuming pedigrees were not accurate. The uncertain parentage model did not reduce bias for animals with pedigree errors but was expected to be less biased for those animals. The bias did not change predictably with different heritabilities.

Rank correlations (SD) were 0.99 (<0.01) between traditional and uncertain parentage models for all animals and males in the simulations. Rank correlations (SD) were at least 0.98 (<0.01) between intensive and extensive uncertain parentage assumptions for all ani-

Table 4. Scaled bias [(true breeding value − EBV)/σ_u] and SD for models with certain or uncertain parentage based on genotype status and pedigree accuracy in a simulation with few pedigree errors[1]

	Genotyped		Not genotyped	
Pedigree model	Accurate	Inaccurate	Accurate	Inaccurate
Heritability = 0.1				
Certain[2]	0.25 (0.06)	−0.27 (0.08)	0.14 (0.05)	0.55 (0.07)
Intensive[3]	0.13 (0.07)	−0.39 (0.09)	0.15 (0.06)	0.59 (0.07)
Extensive[4]	0.03 (0.07)	−0.45 (0.09)	0.26 (0.06)	0.64 (0.07)
Heritability = 0.3				
Certain	0.25 (0.04)	−0.29 (0.05)	0.12 (0.03)	0.64 (0.06)
Intensive	0.17 (0.04)	−0.37 (0.05)	0.19 (0.03)	0.68 (0.06)
Extensive	0.09 (0.04)	−0.42 (0.05)	0.33 (0.04)	0.73 (0.06)
Heritability = 0.5				
Certain	0.19 (0.04)	−0.30 (0.04)	0.10 (0.03)	0.67 (0.06)
Intensive	0.12 (0.04)	−0.35 (0.04)	0.18 (0.03)	0.71 (0.06)
Extensive	0.06 (0.04)	−0.38 (0.04)	0.33 (0.03)	0.76 (0.06)

[1]Simulated 9% sire errors and 3% dam errors.

[2]Assume all pedigrees are correct.

[3]Assume the sire is correct with probability 0.90 and the dam is correct with probability 0.95.

[4]Assume the sire is correct with probability 0.75 and the dam is correct with probability 0.85.

Table 5. Scaled bias [(true breeding value − EBV)/σ_u] and SD for models with certain or uncertain parentage based on genotype status and pedigree accuracy in a simulation with many pedigree errors[1]

	Genotyped		Not genotyped	
Pedigree model	Accurate	Inaccurate	Accurate	Inaccurate
Heritability = 0.1				
Certain[2]	0.31 (0.06)	−0.16 (0.08)	0.18 (0.05)	0.42 (0.06)
Intensive[3]	0.20 (0.07)	−0.26 (0.08)	0.20 (0.05)	0.44 (0.06)
Extensive[4]	0.11 (0.07)	−0.34 (0.08)	0.30 (0.06)	0.50 (0.06)
Heritability = 0.3				
Certain	0.29 (0.04)	−0.20 (0.04)	0.15 (0.04)	0.47 (0.04)
Intensive	0.22 (0.03)	−0.28 (0.04)	0.22 (0.04)	0.50 (0.05)
Extensive	0.15 (0.04)	−0.35 (0.04)	0.35 (0.04)	0.57 (0.05)
Heritability = 0.5				
Certain	0.24 (0.04)	−0.21 (0.04)	0.13 (0.03)	0.48 (0.05)
Intensive	0.18 (0.04)	−0.27 (0.04)	0.21 (0.03)	0.52 (0.05)
Extensive	0.12 (0.04)	−0.32 (0.04)	0.36 (0.04)	0.60 (0.05)

[1]Simulated 25% sire errors and 15% dam errors.

[2]Assume all pedigrees are correct.

[3]Assume the sire is correct with probability 0.90 and the dam is correct with probability 0.95.

[4]Assume the sire is correct with probability 0.75 and the dam is correct with probability 0.85.

mals, males, and young animals. Both models ranked animals similarly, resulting in the same selection decisions. The EBV differences were evaluated as traditional minus uncertain parentage and scaled by σ_u; the minimum difference (SD) was −0.42 (0.08) for the intensive assumptions and −0.98 (0.11) for the extensive assumptions. The maximum difference (SD) was 0.29 (0.04) for the intensive assumptions and 0.73 (0.09) for the extensive assumptions. The 2 uncertain parentage assumptions were compared as intensive minus extensive and scaled by σ_u; the minimum difference (SD) was −0.67 (0.08), and the maximum difference was 0.48 (0.06). No large differences (>1 SD) occurred across the heritabilities or the 2 simulated pedigree errors. Although rank correlations were strong, few animals had moderate changes in EBV that could affect selection decisions. The EBV differences increased with a greater probability for uncertain parentage as expected. Limited comparisons previously existed for traditional and uncertain parentage predictions, especially in accounting for pedigree accuracy.

For all 3 traits, rank correlations were 0.99 for the top 100 Holsteins, AI bulls (n = 581), and genomic young bulls (n = 2,215), indicating little re-ranking between models. Table 6 shows EBV differences between traditional and uncertain parentage models after base adjusting. We adjusted bases so EBV for 2010-born animals had a 0 mean, as in the official evaluation. Some animals had EBV that differed by 0.75 genetic SD, but AI bulls differed less. Pedigree depth had minimal effect on the EBV differences. The magnitude of the difference between models was similar to those in the simulation.

Table 7 had the SD of EBV for both models. Accounting for uncertain parentage slightly increased variability of EBV and recovered some of the lost dispersion that was caused by the pedigree errors. Analyses with complete pedigrees had slightly more variation in EBV

Table 6. Summary statistics for EBV differences from certain and uncertain parentage models in Holstein

	3-generation pedigree (n = 20,917,044)			Complete pedigree (n = 65,927,043)		
Trait (kg)	Minimum	Mean	Maximum	Minimum	Mean	Maximum
All animals						
Milk	−1,092	3	904	−1,088	13	1,033
Fat	−41	0	30	−41	0	26
Protein	−43	0	26	−44	0	34
AI bulls, n = 581						
Milk	−542	−182	267	−536	−180	266
Fat	−21	−8	5	−23	−8	5
Protein	−16	−6	3	−16	−6	3

Table 7. Standard deviation for Holstein EBV from models assuming certain or uncertain parentage

	3-generation pedigree (n = 20,917,044)		Complete pedigree (n = 65,927,043)	
Trait (kg)	Certain	Uncertain	Certain	Uncertain
All animals				
Milk	1,353	1,363	1,405	1414
Fat	50.6	50.9	52.3	52.2
Protein	41.2	41.4	44.3	44.5
AI bulls, n = 581				
Milk	1,590	1,673	1,607	1,687
Fat	62.3	65.3	62.9	65.8
Protein	43.5	45.6	43.9	45.9

than analyses with 3-generation pedigrees, although the increase was generally less than the increase from modeling uncertain parentage. We did not validate the models, as all results were expected to be equivalent. Although single-trait analyses were used, multiple-trait models were not expected to create larger differences between models. Larger differences could occur for traits with greater heritability, as phenotypes would contribute more information to EBV. Accounting for differences in pedigree accuracy did not change the ranking of selection candidates for production traits.

Within-country pedigree errors previously affected international predictions. Pedigree errors reduced the genetic correlations between countries (Banos et al., 2001), resulting in greater genotype-by-environment interactions between countries than appropriate. Pedigree errors in one country caused slight changes in rank on the international scale, but the greatest rank changes were within the country with the pedigree errors (Banos et al., 2001). As all countries were expected to have pedigree errors at different rates, more re-ranking likely occurs internationally, causing bias toward selection of domestic bulls. Current results indicate pedigree errors have limited effect within country. Although the errors would compound across countries, the consequences of inaccurate pedigrees for international predictions were likely limited.

These results depended on the assumptions for the uncertain parentage model. For the simulation, accuracy and dispersion for young genotyped animals did not meaningfully differ between different uncertain parentage assumptions, and predictions were not necessarily less biased when using the correct uncertain parentage assumptions. For the Holstein data, the only probabilities tested were 0.90 for the sire and 0.95 for the dam and were based on pedigree errors previously identified in these data. Not all populations would have the data to approximate sire and dam probabilities. Probabilities would need to be tested to determine the best assumptions, although the simulations were reasonably robust to using the wrong uncertain parentage assumptions. Alternatively, the probability of correct parentage could be treated as unknown and predicted from the data using a Bayesian approach (Cardoso and Tempelman, 2003), but frequentist and Bayesian approaches did not differ in EBV prediction in beef cattle (Shiotsuki et al., 2012). The Bayesian approach could enable better discrimination among animals with uncertain parentage to identify those that are more or less likely to have pedigree errors based on all other information. This approach was not used here because this was a preliminary investigation of the topic. Given how similar our results were, we do not expect a Bayesian approach to be markedly different.

Widespread genotyping enabled Holstein pedigree discovery when pedigrees were not available. New genotypes were compared with the database to confirm or to identify parents, grandparents, and so on. In Holsteins, the accuracies of pedigree discovery were 100% for sires, 97% for MGS, and 92% for MGGS if male ancestors were genotyped (VanRaden et al., 2013). When parentage was not known, pedigree sires were identified but dams were identified less frequently because 48% of dams were not genotyped. Additionally, dam pedigree was missing for 14% of genotyped animals.

In many cases when the dam was not known nor genotyped, the animal's MGS, and potentially even MGGS, can be identified. Figure 5 shows an example pedigree when the dam was not known but the paternal lineage was discovered. Hence, the pedigree was discovered for the paternal lineage on the maternal side but was not used for prediction because, for example, no dam was available to link the animal to its discovered MGS. The discovered pedigree can be used by creating virtual or placeholder dam identification numbers to fill in pedigree gaps. These dams would have discovered sires and unknown dams. If the MGGS was discovered, a virtual maternal granddam would be created with unknown dam and discovered sire. Each virtual female would have a unique identification number and would

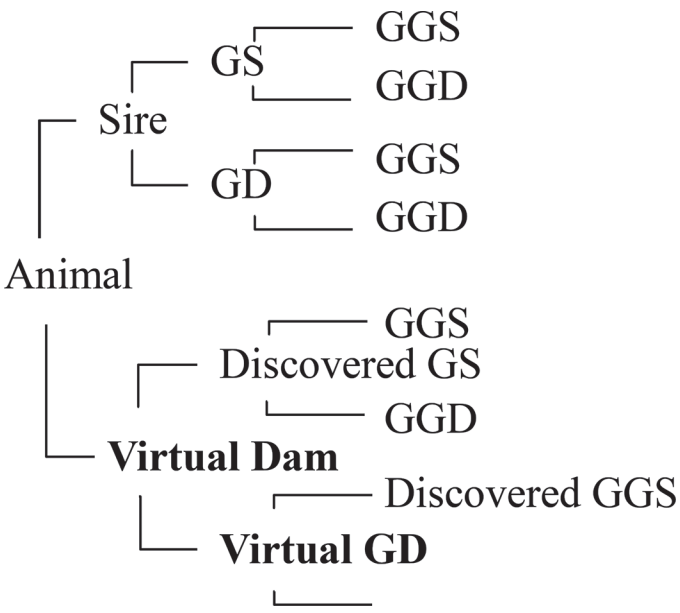

Figure 5. Example 3-generation pedigree with grandsire/granddam (GS/GD) and great-grandsire/great-granddam (GGS/GGD) when maternal GS and GGS were discovered by genotypes and dam and maternal GD were not known.

have relationships through the offspring and the discovered sire.

If this approach was implemented in 2018, 205,200 virtual dams and 194,429 virtual granddams would be created to fill in discovered MGS and MGGS based on 2,012,868 genotyped Holsteins (G. Wiggans, Council for Dairy Cattle Breeding, Bowie, MD, personal communication). Additionally, cows could be matched to calves based on the herd, cow fresh date, calf birth date, and the discovered pedigree from genotyping. In this way, 18,000 cows were identified as probable dams of genotyped animals. These approaches provided maternal pedigree information for 78% of the genotyped animals that currently have no maternal pedigree. Implementing this approach could reduce the effect of poor behavior of unknown parent groups in ssGBLUP by reducing the amount of missing pedigree for selection candidates.

CONCLUSIONS

With the assumptions in this study, variation in pedigree accuracy had only a minor effect on Holstein EBV for production traits. The EBV ranking did not change for AI bulls or for elite animals when pedigree accuracy was accounted for with an uncertain parentage model. Further research is needed to establish if these results can be verified for different traits and are sensitive to the model assumptions. Genotypes were used to correct pedigree relationships without biasing predictions because of heterogeneous pedigree accuracy for bulls' daughters. Genotypes can also be used to discover pedigrees and further research is needed to evaluate the effect of using discovered maternal pedigrees.

ACKNOWLEDGMENTS

The authors thank the Council on Dairy Cattle Breeding (Bowie, MD) for providing data under USDA Nonfunded Cooperative Agreement 58-1245-3-228N. We also thank L. R. Bacheller and M. E. Tooker (Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD) for providing technical assistance. This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE; Oak Ridge, TN) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA); ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-SC0014664. All opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of USDA, ARS, DOE, or ORAU/ORISE. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The authors appreciate the helpful comments of 2 anonymous reviewers.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. https://doi.org/10.3168/jds.2009-2730.

Banos, G., G. Wiggans, and R. Powell. 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. J. Dairy Sci. 84:2523–2529. https://doi.org/10.3168/jds.S0022-0302(01)74703-0.

Cardoso, F. F., and R. J. Tempelman. 2003. Bayesian inference on genetic merit under uncertain paternity. Genet. Sel. Evol. 35:469. https://doi.org/10.1051/gse:2003035.

Cardoso, F. F., and R. J. Tempelman. 2004. Genetic evaluation of beef cattle accounting for uncertain paternity. Livest. Sci. 89:109–120. https://doi.org/10.1016/j.livprodsci.2004.02.006.

Carolino, I., C. O. Sousa, S. Ferreira, N. Carolino, F. S. Silva, and L. T. Gama. 2009. Implementation of a parentage control system in Portuguese beef-cattle with a panel of microsatellite markers. Genet. Mol. Biol. 32:306–311. https://doi.org/10.1590/S1415-47572009005000026.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2. https://doi.org/10.1186/1297-9686-42-2.

Famula, T. R. 1992. Simple and rapid inversion of additive relationship matrices incorporating parental uncertainty. J. Anim. Sci. 70:1045–1048. https://doi.org/10.2527/1992.7041045x.

Geldermann, H., U. Pieper, and W. Weber. 1986. Effect of misidentification on the estimation of breeding value and heritability in cattle. J. Anim. Sci. 63:1759–1768. https://doi.org/10.2527/jas1986.6361759x.

Henderson, C. 1988. Use of an average numerator relationship matrix for multiple-sire joining. J. Anim. Sci. 66:1614–1621. https://doi.org/10.2527/jas1988.6671614x.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

Israel, C., and J. Weller. 2000. Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. J. Dairy Sci. 83:181–187. https://doi.org/10.3168/jds.S0022-0302(00)74869-7.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943–3952. https://doi.org/10.3168/jds.2013-7752.

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2018. Manual for BLUPF90 family of programs. Vol. 2018. Accessed Dec. 17, 2018. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf.

Misztal, I., Z.-G. Vitezica, A. Legarra, I. Aguilar, and A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130:252–258. https://doi.org/10.1111/jbg.12025.

Perez-Enciso, M., and R. Fernando. 1992. Genetic evaluation with uncertain parentage: A comparison of methods. Theor. Appl. Genet. 84:173–179. https://doi.org/10.1007/BF00223997.

Pollak, E. 2005. Application and impact of new genetic technologies on beef cattle breeding: A 'real world' perspective. Aust. J. Exp. Agric. 45:739–748. https://doi.org/10.1071/EA05047.

Quaas, R. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338–1345. https://doi.org/10.3168/jds.S0022-0302(88)79691-5.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: A large-scale genome simulator for livestock. Bioinformatics 25:680–681. https://doi.org/10.1093/bioinformatics/btp045.

Shiotsuki, L., F. Cardoso, J. I. V. Silva, G. Rosa, and L. G. d. Albuquerque. 2012. Evaluation of an average numerator relationship matrix model and a Bayesian hierarchical model for growth traits in Nellore cattle with uncertain paternity. Livest. Sci. 144:89–95. https://doi.org/10.1016/j.livsci.2011.11.002.

Stephen, M., J. Bryant, M. Camara, and D. Meadows. 2018. Developing metrics to rank individual herds according to data quality. In 2018 Interbull Meeting. Auckland, New Zealand. Accessed Dec. 19, 2018. http://interbull.org/static/web/1615_MelissaStephen.pdf.

Van Vleck, L. D. 1970. Misidentification in estimating the paternal sib correlation. J. Dairy Sci. 53:1469–1474. https://doi.org/10.3168/jds.S0022-0302(70)86416-5.

VanRaden, P. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. J. Dairy Sci. 75:3136–3144. https://doi.org/10.3168/jds.S0022-0302(92)78077-1.

VanRaden, P. M., T. Cooper, G. Wiggans, J. O'Connell, and L. Bacheller. 2013. Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. J. Dairy Sci. 96:1874–1879. https://doi.org/10.3168/jds.2012-6176.

VanRaden, P. M., G. Wiggans, and C. Ernst. 1991. Expansion of projected lactation yield to stabilize genetic variance. J. Dairy Sci. 74:4344–4349. https://doi.org/10.3168/jds.S0022-0302(91)78630-X.

Visscher, P. M., J. Woolliams, D. Smith, and J. Williams. 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J. Dairy Sci. 85:2368–2375. https://doi.org/10.3168/jds.S0022-0302(02)74317-8.

Weller, J. I., E. Feldmesser, M. Golik, I. Tager-Cohen, R. Domochovsky, O. Alus, E. Ezra, and M. Ron. 2004. Factors affecting incorrect paternity assignment in the Israeli Holstein population. J. Dairy Sci. 87:2627–2640. https://doi.org/10.3168/jds.S0022-0302(04)73389-5.

Wiggans, G. R., T. Cooper, P. VanRaden, K. Olson, and M. Tooker. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95:1552–1558. https://doi.org/10.3168/jds.2011-4985.

Wiggans, G. R., and P. VanRaden. 1991. Method and effect of adjustment for heterogeneous variance. J. Dairy Sci. 74:4350–4357. https://doi.org/10.3168/jds.S0022-0302(91)78631-1.