Genomic Selection and its Effects on Dairy Cattle Breeding Programs

K.A. Weigel[*], G. de los Campos[*,†], A.I. Vazquez[*,†], C.P. Van Tassell[‡], G.J.M. Rosa[*], D. Gianola[*], J.R. O'Connell[§], P.M. VanRaden[‡], and G.R. Wiggans[‡]

Introduction

The availability of high-throughput assays for genotyping single nucleotide polymorphisms (SNP) has led to the genotyping of thousands of dairy cattle, mostly progeny tested bulls in artificial insemination (AI) programs or young bulls that are candidates for such programs, using the BovineSNP50 BeadChip (Illumina, Inc., San Diego, CA) or similar platforms. As a result, various methods for estimating SNP effects and predicting direct genomic values (DGV) of selection candidates have been developed. In the United States (US), genomic data have been used to enhance predicted transmitting abilities (PTA) for production, conformation, and fitness of dairy cattle since January, 2009. Changes in reliability (REL) due to inclusion of genomic data are shown in Table 1 (VanRaden et al., 2009).

Table 1. Changes in reliability due to the inclusion of genomic data in national genetic evaluations in the United States (VanRaden et al., 2009).

Trait	Holstein	Jersey	Brown Swiss
Net Merit	+24%	+8%	+9%
Milk Yield	+26%	+6%	+17%
Fat Yield	+32%	+11%	+10%
Protein Yield	+24%	+2%	+14%
Fat Percentage	+50%	+36%	+8%
Protein Percentage	+38%	+29%	+10%
Productive Life	+32%	+7%	+12%
Somatic Cell Score	+23%	+3%	+17%
Daughter Pregnancy Rate	+28%	+7%	+18%

Despite the aforementioned gains in REL of young selection candidates, the price of high-density genotyping assays may limit their application to males and elite females. The development of inexpensive, low-density genotyping platforms with, e.g., 300 to 3,000 SNP, could stimulate widespread implementation of genomics on commercial farms.

[*] University of Wisconsin - Madison, 1675 Observatory Drive, Madison, WI, 53706, USA

[†] University of Alabama - Birmingham, 1665 University Boulevard, Birmingham, AL, 35294, USA

[‡] Agricultural Research Service, United States Department of Agriculture, 10300 Baltimore Avenue, Beltsville, MD, 20705, USA

[§] University of Maryland, 655 West Baltimore Street, Baltimore, MD, 21201, USA

This paper summarizes four recent studies regarding: 1) the relationship between published, genome-enhanced PTA (GPTA) and subsequent daughter performance; 2) the predictive ability of DGV resulting from selection of low-density SNP based on magnitude of estimated effects; 3) the accuracy of imputation of high-density genotypes from equally spaced low-density SNP, and 4) the precision of DGV derived from imputed high-density genotypes.

Materials and methods

Relationship between published predictions and subsequent daughter performance. To assess how well genomic evaluations are working in practice, we evaluated data from US Holstein bulls that had only genomic data in January 2009 and ≥ 50 milking daughters in August 2009. In routine genetic evaluations, genomic data are combined with pedigree data when computing the official, published PTA that are released to the industry. Therefore, instead of using the official August 2009 GPTA, we computed correlations of January 2009 predictions with unofficial, "traditional" August 2009 daughter yield deviations (DYD) that contained no genomic information.

Selection of low-density SNP based on magnitude of estimated effects. The training set consisted of high-density SNP genotypes (from the BovineSNP50 BeadChip) and August 2003 progeny test PTA for lifetime net merit (LNM) of 3,305 Holstein bulls born from 1952 to 1998. The predictive ability of models fitted in this study was evaluated in a testing set consisting of high-density SNP genotypes and April 2008 progeny test PTA for LNM of 1,398 Holstein bulls born from 1999 to 2002. Among bulls in the testing set, 85.5% had genotyped sires and 70.9% had genotyped maternal grandsires in the training set. Genotypes for 32,518 loci, after removal of SNP with minor allele frequency (MAF) < 0.05 or complete linkage disequilibrium (LD) with adjacent SNP, were coded as 0 (homozygous for allele B), 1 (heterozygous), 2 (homozygous for allele A), or missing.

The PTA of sires in the training set were regressed on SNP genotypes using the Bayesian LASSO (BL) method of Park and Casella (2008), which was implemented via the Gibb's sampler using the R software (R Foundation for Statistical Computing, 2008). As a reference, estimated effects of all 32,518 SNP were used to compute DGV of bulls in the testing set, and these were compared with August 2008 progeny test PTA for LNM. In addition, subsets of selected SNP were created by sorting the original 32,518 SNP by the absolute value of posterior means of their estimated effects and choosing the top 300, 500, 750, 1,000, 1,250, 1,500, or 2,000 SNP. Subsets of 300, 500, 750, 1,000, 1,250, 1,500, or 2,000 equally spaced SNP were also created. Subsequently, SNP effects were recomputed using the aforementioned BL model to regress August 2003 PTA of bulls in the training set on genotypes for each of the fourteen subsets of selected or equally spaced SNP. Lastly, DGV of bulls in the testing set for LNM were computed using estimated SNP effects from the training set and corresponding genotypes of bulls in the testing set, and correlations between the resulting DGV and August 2008 progeny test PTA for LNM were computed.

Imputation of high-density genotypes using equally spaced SNP. Genotypes of 3,146 Jersey dairy cattle (2,656 males and 490 females) for 43,385 SNP, which represented the subset of SNP on the BovineSNP50 BeadChip with a call rate of > 90%, < 1% parent-progeny conflicts, incomplete LD with adjacent SNP, and MAF > 1% in the Holstein, Jersey, or Brown Swiss breed, were coded as described previously. The population was divided into a reference panel, consisting of animals that were genotyped for all 43,385 SNP, and a study sample, in which genotypes were masked for a randomly chosen 20, 60, 80, 90, 95, 98, or 99% of loci. A subset of 2,542 animals born from 1953 to 2006 was used as the reference panel, and a subset of 604 animals born from 2007 to 2009 was used as the study sample. Three chromosomes were considered, BTA1, BTA15, and BTA28, but results are shown only for BTA15, which contained the median number of SNP (1,399). After removal of SNP with unknown physical position on the chromosome based on the UMD2 assembly of *B. taurus* (Zimin et al., 2009), 1,377 SNP remained. After masking 20 to 99% of SNP on BTA15 for animals in the study sample, the number of available SNP from which to impute masked genotypes ranged from 14 to 1,102, depending on the masking rate.

Many algorithms have been developed for constructing haplotypes and imputing genotypes in humans (e.g., Scheet and Stephens, 2006; Kong et al., 2008; Howie et al., 2009). In this study, masked genotypes were imputed using the haplotype clustering algorithm of Scheet and Stephens (2006), implemented via the fastPHASE version 1.2 software (University of Washington TechTransfer Digital Ventures Program, Seattle, WA), and the hidden Markov model algorithm of Howie et al. (2009), implemented via the IMPUTE version 2.0 software (Department of Statistics, University of Oxford, United Kingdom). For computational reasons, the number of haplotype clusters in fastPHASE 1.2 was set to 32, whereas the number of conditioning states used in IMPUTE 2.0 was set to 40. In addition, with fastPHASE 1.2, BTA15 was analyzed in its entirety, whereas with IMPUTE 2.0, computational feasibility was enhanced by breaking BTA15 into two pieces of equal size.

Prediction of genomic breeding values using imputed high-density genotypes. Genotypes of 1,762 Jersey sires for 42,552 SNP, which represented the subset of 43,385 SNP that are presently used for routine genomic evaluation of US dairy cattle after removal of SNP with unknown physical position based on the UMD2 assembly of *B. taurus* (Zimin et al., 2009), were coded as described previously. Phenotypes, which represented the result of progeny testing, consisted of PTA for milk yield, protein percentage, and daughter pregnancy rate (DPR). The training set contained 1,446 sires with ≥ 10 milking daughters in May 2006, whereas the testing set contained 316 sires that had no milking daughters in May 2006 and ≥ 10 milking daughters in April 2009. April 2009 PTA values were computed using a traditional animal model and contained no genomic information.

Beginning with list of SNP from the BovineSNP50 Beadchip that were considered as potential candidates for a low-density chip, based on high MAF across a broad range of *Bos taurus* and *Bos indicus* breeds, we created equally spaced subsets in which 93.1% (all but 2,942), 96.6% (all but 1,468), 98.3% (all but 741), or 99.1% (all but 366) of the original 42,552 loci were masked. The number of unmasked SNP per chromosome ranged from 48 to 175 for a masking rate of 93.1%, 24 to 89 for a masking rate of 96.6%, 12 to 44 for a masking rate of 98.3%, and 6 to 20 for a masking rate of 99.1%. Next, masked genotypes were imputed using the aforementioned algorithm of Howie et al. (2009), implemented via the IMPUTE 2.0 software, using 40 conditioning states. For computational reasons, BTA1 to BTA11 were broken into three pieces of equal size, BTA12 to BTA24 were broken into two pieces, and BTA25 to X were analyzed in their entirety. Genotype probabilities from IMPUTE 2.0 (which range from 0 to 1 for possible genotypes of 0, 1, and 2 at a given locus) were used to compute the expected dosage value at each locus.

After imputation of masked genotypes, May 2006 PTA for milk yield, protein percentage, and DPR of sires in the training set were regressed on SNP genotypes using the BL model described previously. First, the model was fitted to the training data of 1,446 Jersey sires using all 42,552 SNP as covariates. This provided estimated SNP effects, which were used to compute reference DGV for sires in the testing set. Next, estimated SNP effects from the BL model with 42,552 covariates were used to compute DGV using imputed dosage values for each of the four masking rates described previously. To evaluate the difference in predictive ability between imputing high-density genotypes versus the alternative of leaving them as missing, we also fit four reduced BL models containing 2,942, 1,468, 741, or 366 SNP that were unmasked in the corresponding testing sets. Estimated SNP effects from these reduced models were used to compute DGV of sires in the training set, as described earlier. Lastly, to evaluate the potential loss in accuracy that might occur in future generations if some animals whose phenotypes contribute to the estimation of SNP effects lack high-density genotypes, we considered a scenario in which a random 50% of sires in the training set also had only low-density SNP genotypes. In the case, only 723 sires provided reference haplotypes for imputation of high-density genotypes in the remaining 1,039 sires. Thus, the 42,552 high-density genotypes used for estimation of SNP effects in the full BL model represented a mixture of actual and imputed genotypes. As previously, four full BL models were fitted with 1,446 phenotypes and 42,552 regression coefficients, but with varying proportions of actual genotypes and imputed genotype dosage values, as well as four reduced BL models with 2,942, 1,468, 741, or 366 SNP covariates. Estimated SNP effects were multiplied by corresponding genotype dosage values and summed across markers to obtain DGV for 316 sires in the testing set, and resulting DGV were correlated with April 2009 PTA from progeny testing.

Results and discussion

Relationship between published predictions and subsequent daughter performance. As shown in Table 2, 238 Holstein bulls had official PTA in January 2009 based only on genomic data, as well as DYD in August 2009 based on ≥ 50 milking daughters. Note that only 60 bulls had ≥ 50 daughters for daughter pregnancy rate. Average January 2009 REL based on parent average (PA) information was 42% for yield traits, 39% for somatic cell score (SCS), and 26% for DPR, whereas REL of the GPTA, which included pedigree and genomic data, averaged 72%, 67%, and 62%, respectively. Data from an average of 71 daughters per bull (62 for DPR) provided an average REL of 84% for yield traits, 67% for SCS, and 62% for DPR in August 2009. The average REL of January 2009 GPTA for SCS was equivalent to that of the August 2009 PTA based on 71 daughters, whereas average REL of the August 2009 PTA for DPR was still lower than REL of the corresponding GPTA in January 2009. August 2009 DYD should be considered as preliminary, as changes may occur when additional progeny records become available. Correlations between August 2009 DYD from progeny testing and January 2009 PA and GPTA are also shown. In every case, correlations were greater with GPTA than PA.

Table 2. Comparison of January 2009 parent averages (PA) and genome-enhanced predicted transmitting abilities (GPTA) for milk, fat, protein, somatic cell score (SCS), and daughter pregnancy rate (DPR) with August 2009 daughter yield deviations (DYD) for US Holstein bulls whose first-crop daughters calved between January and August.

	Milk	Fat	Prot	SCS	DPR
No. Bulls	238	238	238	237	60
Reliability (Jan '09 PA)	42%	42%	42%	39%	36%
Reliability (Jan '09 GPTA)	72%	72%	72%	67%	62%
No. Daughters (Aug '09 DYD)	71	71	71	71	62
Reliability (Aug '09 DYD)	84%	84%	84%	67%	52%
Correlation (Jan '09 PA, Aug '09 DYD)	0.444	0.540	0.476	0.376	0.213
Correlation (Jan '09 GPTA, Aug '09 DYD)	0.624	0.695	0.632	0.531	0.341

Selection of low-density SNP based on magnitude of estimated effects. Correlations between DGV of 1,398 Holstein bulls in the testing set and corresponding April 2008 PTA for LNM from progeny testing are shown in Table 3. The reference model with 32,518 SNP provided a correlation of 0.612 for all bulls, with a significant advantage for bulls with genotyped sires. By comparison, correlations between progeny test PTA and DGV from 300 to 2,000 selected SNP ranged from 0.428 to 0.567, and correlations between progeny test PTA and DGV from 300 to 2,000 equally spaced SNP ranged from 0.253 to 0.539.

Table 3. Correlations of April 2008 predicted transmitting abilities for lifetime net merit with August 2003 direct genomic values for all SNP and selected or equally spaced SNP in a testing set of 1,398 Holstein bulls (r_{PT_All}), 1,195 bulls with genotyped sires (r_{PT_Sire}), and 203 bulls without genotyped sires (r_{PT_NoSire}) (Weigel et al., 2009).

No. SNP	Method of SNP Selection	r_{PT_All}	r_{PT_Sire}	r_{PT_NoSire}
300	Largest Effects	0.428	0.447	0.312
300	Equally Spaced	0.253	0.262	0.202
500	Largest Effects	0.485	0.503	0.369
500	Equally Spaced	0.333	0.348	0.245
750	Largest Effects	0.519	0.530	0.441
750	Equally Spaced	0.435	0.450	0.348
1,000	Largest Effects	0.537	0.549	0.460
1,000	Equally Spaced	0.422	0.438	0.321
1,250	Largest Effects	0.554	0.567	0.461
1,250	Equally Spaced	0.477	0.489	0.395
1,500	Largest Effects	0.559	0.576	0.445
1,500	Equally Spaced	0.518	0.534	0.412
2,000	Largest Effects	0.567	0.582	0.469
2,000	Equally Spaced	0.539	0.559	0.408
32,518	All Available	0.612	0.627	0.511

In every case, the predictive ability of DGV from selected SNP was greater than for equally spaced SNP. In a related study, Vazquez et al. (2009) noted that a low-density platform containing SNP with largest estimated effects for lifetime net merit in US Holsteins provided correlations of 0.40 to 0.55 with subsequent progeny test PTA for individual production and fitness traits, but correlations tended to be greater for production than fitness. Furthermore, low-density assays composed of selected SNP will be breed and trait-specific. Because of these challenges, it may be more efficient to genotype a slightly larger set of equally spaced SNP that would facilitate imputation of high-density genotypes, rather than to focus on prediction of DGV from smaller regression models that contain only a few hundred selected SNP with large estimated effects, and the feasibility of this alternative is discussed below.

Imputation of high-density genotypes using equally spaced SNP. The mean, minimum, and maximum proportion of masked SNP genotypes on BTA15 that were imputed correctly in the study sample is shown in Table 4. With fastPHASE 1.2, the mean proportion imputed correctly ranged from 0.66 to 0.73 when only 1% or 2% of genotypes were unmasked in the study sample, versus 0.75 to 0.89 when 5 to 10% of genotypes were unmasked, as would be the case for a medium-density panel with 2,000 to 4,000 SNP. As the percentage of unmasked genotypes increased to 20, 40, or 80%, the proportion of correct genotypes ranged from 0.90 to 0.99. When the proportion of masked genotypes was very high, e.g., 98 or 99%, IMPUTE 2.0 was slightly more accurate than fastPHASE 1.2, and when 90 or 95% of genotypes were masked in the study sample IMPUTE 2.0 was significantly more accurate. On the other hand, when 20, 40, or 80% of SNP genotypes were unmasked, fastPHASE 1.2 was more accurate, because accuracy of IMPUTE 2.0 peaked at approximately 0.90 to 0.95.

Table 4. Mean (minimum, maximum) proportion of masked SNP genotypes on chromosome 15 (1,377 total SNP) that were imputed correctly in a future study sample composed of 604 US Jersey cattle, using a reference panel composed of 2,542 animals of the same breed (Weigel et al., 2010b).

Proportion of SNP Unmasked in the Study Sample	fastPHASE 1.2 (32 haplotype clusters)	IMPUTE 2.0 (40 conditioning states)
0.01	0.701 (0.574, 0.766)	0.730 (0.533, 0.925)
0.02	0.726 (0.596, 0.797)	0.780 (0.568, 0.981)
0.05	0.780 (0.644, 0.856)	0.890 (0.682, 0.999)
0.10	0.874 (0.732, 0.960)	0.924 (0.762, 0.998)
0.20	0.951 (0.841, 0.993)	0.932 (0.778, 1.000)
0.40	0.984 (0.890, 1.000)	0.935 (0.772, 1.000)
0.80	0.992 (0.946, 1.000)	0.930 (0.663, 1.000)

Prediction of genomic breeding values using imputed high-density genotypes. The mean proportion of masked SNP genotypes for which the most likely genotype provided by IMPUTE 2.0 matched the original BovineSNP50 genotype call was averaged across animals, loci, and chromosomes for 316 Jersey sires in the testing set (study sample). At a masking rate of 93.1%, the proportion of genotypes imputed correctly was 0.912, whereas at masking rates of 96.6, 98.3, or 99.1%, corresponding means were 0.875, 0.789, or 0.735, respectively. When masking was applied to 316 sires in the testing set plus 723 sires in the training set (i.e., random 50%), means were 0.896, 0.860, 0.782, or 0.733 for masking rates of 93.1, 96.6, 98.3, or 99.1%, respectively, indicating that halving the size of the training set (reference panel) led to a small reduction in imputation accuracy. Table 5 shows correlations between DGV derived using estimated SNP effects from the May 2006 training set in conjunction with actual or imputed genotypes (or genotype dosage values) of sires in the testing set and the actual April 2009 progeny test PTA of sires in the testing set for milk yield, protein percentage, and DPR. Reference values correspond to the full BL model in which none of the 42,552 loci were masked in the training or testing set.

Table 5. Correlations between predicted direct genomic values and corresponding April 2009 predicted transmitting abilities from for milk yield, protein percentage, and daughter pregnancy rate from progeny testing using full or reduced models with 42,552 or 366, 741, 1,468, or 2,942 SNP covariates, respectively, with or without imputation of masked genotypes for bulls in the testing set or bulls in the testing set and a random 50% of bulls in the training set (Weigel et al., 2010a).

Masking Rate (number of unmasked SNP)	Reduced Model with Masked SNP Deleted	Full Model with Imputing in Testing Set	Full Model with Imputing in Testing + 50% of Training Set	Reference Model with No Masking
		Milk Yield		
93.1% (2,942)	0.617	0.673	0.640	0.673
96.6% (1,468)	0.515	0.649	0.628	0.673
98.3% (741)	0.532	0.525	0.537	0.673
99.1% (366)	0.492	0.367	0.472	0.673
		Protein Percentage		
93.1% (2,942)	0.687	0.740	0.690	0.770
96.6% (1,468)	0.614	0.676	0.658	0.770
98.3% (741)	0.539	0.546	0.534	0.770
99.1% (366)	0.504	0.468	0.506	0.770
		Daughter Pregnancy Rate		
93.1% (2,942)	0.608	0.642	0.641	0.674
96.6% (1,468)	0.585	0.619	0.610	0.674
98.3% (741)	0.544	0.572	0.546	0.674
99.1% (366)	0.518	0.470	0.506	0.674

Conclusion

Results to date indicate that genomic selection using high-density SNP genotypes will greatly enhance genetic progress in dairy cattle. However, at current prices genotyping may be limited to males and elite females. The development of low-density assays containing selected SNP with large estimated effects or, more likely, low-density assays containing equally spaced SNP that will facilitate imputation of high-density genotypes, could lead to widespread adoption of genomics on commercial farms. Potential applications include selection of replacement heifers on farms that use gender-enhanced semen, preliminary genomic screening of young bulls or potential bull dams, parentage discovery, genome-enhanced mate selection, and genome-guided management protocols.

References

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2009). *Genetics* 182:343-353.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). *PLoS Genetics* 5:e1000529.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). *Nature Genet.* 40:1068-1075.

Park, T., and Casella, G. (2008). *J. Am. Stat. Assn.* 103:681-686.

Scheet, P., and Stephens, M. (2006). *Am. J. Human. Genet.* 78:629-644.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., and Schenkel, F. (2009) *J. Dairy Sci.* 92:16-24.

Vazquez, A. I., Rosa, G. J. M., Weigel, K. A., de los Campos, G., and Gianola, D. (2009). *J. Dairy Sci.* 92(Suppl. 1):125.

Weigel, K. A., de los Campos, G., Vazquez, A. I., Rosa, G. J. M., Gianola, D., and Van Tassell, C. P. (2010a). *J. Dairy Sci.* (submitted).

Weigel, K. A., Van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., and Wiggans, G. R. (2010b). *J. Dairy Sci.* (in press).

Weigel, K. A., de los Campos, G., González-Recio, O., Naya, H., Wu, X. L., Long, N., Rosa, G. J. M., and Gianola, D. (2009). *J. Dairy Sci.* 92:5248-5257.

Zimin, A. V. , Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassel, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A., and Salzberg, S. L. (2009). *Genome Biol.* 10:R42.