Combining Different Marker Densities in Genomic Evaluation

P.M. VanRaden[1], J.R. O'Connell[2], G.R. Wiggans[1], and K.A. Weigel[3]
[1]Animal Improvement Programs Lab, USDA Agricultural Research Service, Beltsville, MD, USA
[2]University of Maryland School of Medicine, Baltimore, MD, 21201, USA
[3]University of Wisconsin, Madison, WI, 53706, USA

Abstract

Accurate genomic evaluations are less costly if many animals are genotyped at less than the highest density and their missing genotypes filled using haplotypes. Mixed density files for 45,870 animals were examined by reducing half of young animal or all animal genotypes from the observed 43,385 markers to a subset of 3,209 markers. For young Holsteins genotyped with 3,209 markers, the gain in net merit reliability was 79% of the gain from genotyping 43,385 markers. When half of the reference population had 3,209 markers, gain was 90% for young animals with 43,385 markers and 73% for young animals with 3,209 markers. Gain was 66% when all animals had only 3,209 markers. Simulated gain in reliability from increasing the number of markers to 500,000 was only 1.4%, but more than half of that gain could result from genotyping just 1,586 bulls at higher density. Reliability improved when more reference animals were genotyped at higher density. Individual reliabilities can be adjusted to account for number of markers and success of imputation.

Key words: haplotypes, imputation, genetic markers

Introduction

Genomic selection will be more efficient and affordable when breeders evaluate animals with different genetic marker sets available for different prices. Instead of genotyping all animals at the highest marker density, genotypes of different densities can be combined. The missing genotypes in the lower density sets can be filled (imputed) from genotypes or haplotypes of relatives or from matching allele patterns in the general population.

Lower density panels could be selected to include only the most significant markers from a larger set to maximize reliability for a particular trait, but reliability for other traits may be low if correlation to the selected trait is low. A second option is to include equally spaced, highly polymorphic markers and to impute the missing genotypes, giving increased reliability for all traits. Previous studies such as Weigel *et al.* (2009) have compared differing marker densities, but only a few recent studies have tested genomic evaluation using mixed density and imputation (Druet *et al.*, 2010; Habier *et al.*, 2009; Weigel *et al.*, 2010a).

Both lower and higher density marker panels have been designed for use in genomic evaluation. The imputation methods reported here have been tested on mixtures of simulated markers with a range of densities from 500 to 500,000. The current report investigates actual 346 and 3,209 marker subsets of the 43,385 marker genotypes in the North American database and also compares mixtures of 50,000 and 500,000 simulated markers for this same population.

Methods

Actual genotypes of 40,351 Holsteins, 4,064 Jerseys, and 1,455 Brown Swiss were used in comparing the full set of 43,385 markers (Wiggans *et al.*, 2010) to a subset of 3,209 evenly spaced markers selected for inclusion on an Illumina chip. The Holstein population included 24,306 males and 16,045 females, with 96% of the sires but only 31% of the dams genotyped. An earlier population of 25,365 Holsteins was used to test a subset of 346 selected markers with largest effects for net merit.

Two mixed density sets were constructed by 1) reducing half of all animals to low density (3,209 markers) or by 2) reducing half of only the young animals to 3,209 markers. Animals were assigned low density if the last digit of the identification number was even. A third analysis 3) used regressions on only the

3,209 markers for all animals to determine the loss from only using low density as compared to mixed density.

Haplotypes were formed and genotypes imputed using Fortran program findhap.f90. The program begins by dividing each chromosome into segments of about 250 markers and listing all haplotypes by matching each genotype to the list of haplotypes. This population haplotyping step is analogous to the fastPhase and IMPUTE methods tested by Weigel *et al.* (2010a) on the Jersey data file. The program ends with pedigree haplotyping steps to detect crossovers, fix noninheritance, and impute nongenotyped ancestors. Imputed genotypes are used in the evaluation only if at least 90% of the ancestor's haplotypes can be determined from progeny.

Genomic evaluations were computed with the iterative, nonlinear Fortran program densemap.f90 of VanRaden (2008). The model included a polygenic effect assigned 10% of genetic variance with 43,385 markers or 30% with 3,209 markers, and the remaining variance was modeled by the marker effects. For Holsteins, computing times using one processor were 2.2 hours to complete the haplotyping and 6.5 hours to complete 130 iterations for 5 traits evaluated together. Memory requirements were 0.7 gigabytes for haplotyping and 1.3 gigabytes to solve the genomic equations.

Only the most recent data from April 2010 was used in this study instead of using truncated data. In most previous studies, differing models or data subsets were compared by predicting recent data from historical data. Squared correlations among subset GEBV, full set GEBV, and parent average (PA) were used to obtain the gain in reliability above PA reliability from using a marker subset as a percentage of the gain from the full set:

$$\% \text{ of gain} = 100\ [\text{corr}^2(\text{subset, full set}) - \text{corr}^2(\text{PA, full set})] / [1 - \text{corr}^2(\text{PA, full set})]$$

Simulated 500,000 marker genotypes were used to estimate the numbers of animals needed for higher density genomic selection. Three simulated data sets included 1,586, 3,726, or 7,398 bulls genotyped with 500,000 markers and the remainder of the 33,414 Holsteins genotyped with 50,000 markers. Using one processor, haplotyping required 3.1 hours and 2.8 gigabytes of memory, and evaluation of 5 traits (replicates) required 150 iterations, 2.5 days, and 7.9 gigabytes of memory.

Maximum genomic reliability that can be obtained in practice (REL_{max}) is limited by the maximum marker density and by the size of the reference population. As the reference population becomes infinitely large, reliability should approach 1 minus the fraction of polygenic variance (poly). Total daughter equivalents (DE_{max}) from the reference population can be obtained by summing traditional reliabilities (REL_{trad}) minus the reliabilities of parent average (REL_{pa}), multiplying by the ratio of error to sire variance (k), and dividing by the equivalent reference size (n) needed to achieve 50% genomic REL (VanRaden and Sullivan, 2010):

$$DE_{max} = \sum(REL_{trad} - REL_{pa})\ k / n.$$

Conversion of DE_{max} to genomic REL should account for genotyped SNP not perfectly tracking all QTL in the genome because full sequences are not available. Multiplication by 1 - poly prevents reliability from reaching 100%. If all reference animals are genotyped at the highest chip density, expected genomic REL for young animals without pedigree information can be calculated as:

$$REL_{max} = (1 - \text{poly})\ DE_{max} / (DE_{max} + k).$$

Genomic reliabilities for individual animals can account for their traditional reliabilities, numbers of markers genotyped, quality of imputation, and relationship to the reference population. Each animal's traditional REL is converted to daughter equivalents (DE_{trad}), and these are added to DE_{max} adjusted for any additional error introduced by genotyping at lower SNP density. The reduced daughter equivalents from genomics (DE_{gen}) can be calculated from the squared correlation between estimated and true genotypes averaged across loci (REL_{snp}) for each animal as:

$$DE_{gen} = k\ REL_{max}\ REL_{snp} / (1 - REL_{max}\ REL_{snp})$$

Animals less related to the reference population may have lower DE_{gen} (Liu *et al.*, 2010). The animal's total reliability REL_{tot} is computed from the sum of the daughter equivalents as:

$$REL_{tot} = (DE_{trad} + DE_{gen}) / (DE_{trad} + DE_{gen} + k)$$

Results

Numbers of non-genotyped dams that had at least 90% of haplotypes imputed from progeny were 2254 Holsteins, 184 Jerseys, and 68 Brown Swiss. Squared correlations of their genomic with traditional evaluations for net merit were 0.76 for Holsteins, 0.81 for Jerseys, and 0.91 for Brown Swiss. Larger correlations are expected with smaller reference populations. Squared correlations of genomic evaluations for imputed dams obtained when half of their progeny had 3,209 markers genotyped or when all progeny had 43,385 markers were 0.87 for Holsteins, 0.94 for Jerseys, and 0.93 for Brown Swiss. Thus, inclusion of some progeny with only 3,209 markers resulted in less accurate imputation of their dams and less gain in reliability.

When reference animals all had 43,385 SNPs, squared correlations were 0.90 or higher between young animal GEBV computed using full data or imputed using 3,209 SNPs (Table 1). Gains in reliability from 3,209 SNPs were 79-88% of the gain from 43,385 SNP if haplotyping was used, but were only 61-63% if regressions on the 3,209 SNPs were used in analysis 3.

When half of the reference animals had only 3,209 SNP genotyped, gains in reliability for net merit for progeny genotyped with 43,385 SNP were 90% of gains from the full data for Holstein, 82% for Jersey, and 84% for Brown Swiss (Table 2). Respective gains in reliability for young animals genotyped with 3,209 SNP decreased to 73%, 56%, and 72% of the full set gains when half of the reference animals also had only 3,209 SNP genotyped. Jersey and Brown Swiss PA results differed in the two subsets due to small numbers.

Table 1. Squared correlations and percentage of reliability gain using 3,209 to impute 43,385 markers for Holstein young animals.

	Squared correlations		3K gain
Trait	3K, 43K	PA, 43K	% of 43K
Net merit	0.90	0.52	79
Milk	0.92	0.52	83
Fat	0.92	0.52	83
Protein	0.92	0.56	82
Fat %	0.92	0.34	88
Protein %	0.92	0.47	85
Productive life	0.93	0.50	87
Somatic cells	0.91	0.42	85
Pregnancy rate	0.94	0.54	86

Table 2. Percentage of net merit reliability gain for young animals when half of all animals had 3,209 or 43,385 markers (part data).

Markers for	Squared correlations		% of
young animals	Part, Full	PA, Full	Full gain
Holstein			
43,385	0.95	0.51	90
3,209	0.87	0.51	73
Jersey			
43,385	0.91	0.51	82
3,209	0.81	0.57	56
Brown Swiss			
43,385	0.94	0.63	84
3,209	0.93	0.73	72

The lower density panel of 346 markers selected for net merit gave gains that were smaller and more variable across traits (14-55%) as compared to the 43,385 gain when evaluated using 346 regressions (Table 3).

Table 3. Squared correlations and percentage of reliability gain with 346 selected vs. 43,385 markers.

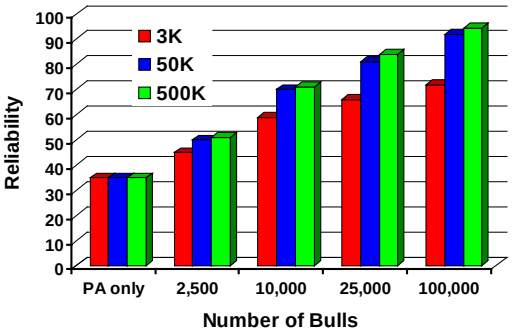
	Squared correlations		
Trait	346, 43K	PA, 43K	% of gain
Net merit	0.73	0.60	33
Milk	0.62	0.51	22
Fat	0.69	0.52	35
Protein	0.63	0.54	20
Fat %	0.69	0.31	55
Protein %	0.60	0.46	26
Productive life	0.65	0.55	22
Somatic cells	0.50	0.42	14
Pregnancy rate	0.63	0.56	16

Gains increased to about 80% when evaluated using methods of Habier *et al.* (2009) for animals with 346 SNP and both parents genotyped for 43,385 SNP, but remained small if parents were not genotyped. Gains were 90% for progeny genotyped with 3,209 SNP and both parents with 43,385 SNP. Gains were above 70% if parents were not genotyped. A primary advantage of using more markers in young animal selection is more precise evaluation of those without genotyped parents. Results in Table 1 are similar to those obtained by Weigel *et al.* (2010b) from Jersey genotypes, but results in Table 2 are more favorable, probably because of the use of pedigree information in the haplotyping algorithm.

With 500,000 simulated markers for all genotyped animals, reliability for young bulls averaged 84.0% as compared with 82.6% using a 50,000-marker subset (Table 4). Reliabilities for three mixed densities were intermediate, ranging from 83.4% to 83.7%. Percentage of missing alleles that could not be determined from haplotypes ranged from 5.3% with 1586 bulls to 1.5% with 7,398 bulls. Recent refinements to the haplotyping algorithm have improved the call rates and reliabilities compared to earlier tests on the same data.

Reliabilities expected with larger reference populations and larger marker densities are in Figure 1. Expectations in the graph are for net merit using a single density, but combined densities instead allow genotypes to be imputed, bringing reliabilities much closer to those possible when all animals are genotyped at highest density. The graph reflects the 1.4% increase in reliability observed at highest density rather than the 10% polygenic variance assumed in U.S. evaluations. Methods to estimate proportions of correctly called genotypes or squared correlations of estimated and true genotypes are needed for individual animals so that REL_{snp} can be included in the published REL.

Figure 1. Expected reliabilities by reference population size using only 3K, 50K, or 500K SNP.



Conclusions

Mixed marker sets can give good reliabilities for all animals at less cost. Animals genotyped at lower density can have their missing genotypes imputed from higher density haplotypes of relatives or from other members of the population. Average gains in reliability with 3,209 SNP for young animals were 79-88% of those with 43,385 SNP if imputing was used but only 61-63% without imputation. A smaller set of 346 markers selected for net merit provided 80% of the gain in reliability if both parents were genotyped at high density, but gain was much lower if parents were not genotyped and only 33% if regression instead of imputation was used.

The reference population can also include animals with lower density genotypes after imputing these to the higher density. When half of the reference population was genotyped with 3,209 SNP, gains in reliability were 90% of those from the full Holstein data set for progeny genotyped with 43,385 SNP and 73% for progeny genotyped with 3,209 SNP.

When higher density panels are introduced, mixed density datasets may be the only option because breeders will not regenotype all reference animals. With 500,000 simulated markers, reliability increased by 1.4%. Most of that gain could be achieved using only a few thousand animals genotyped at higher density, and only 2-6% of the missing genotypes could not be determined for the animals with 50,000 markers observed.

Differing marker sets for large populations can be combined with just a few hours of computation. Further improvements to imputation algorithms may allow smaller fractions of animals to be genotyped at highest density. For animals genotyped at lower density, reliabilities are lower if reliabilities of imputed genotypes are less than 1. More precise estimates of reliability will allow breeders to properly balance benefits vs. costs of using different marker sets.

Acknowledgements

Curt Van Tassell selected the list of 3,209 markers, Bob Schnabel helped to revise marker locations, and Mel Tooker assisted with computation.

References

Druet, T. 2010. Integrating data from different marker panels in human genetics. *Interbull Bulletin 41,* 43-48.

Habier, D., Fernando, R.L. & Dekkers, J.C.M. 2009. Genomic selection using low-density marker panels. *Genetics 182,* 343–353.

Liu, Z., Seefried, F., Reinhardt, F. & Reents, R. 2010. Approximating reliabilities of estimated direct genomic values. *Interbull Bulletin 41,* 29-32.

VanRaden, P.M. 2008. Efficient methods to compute genomic evaluations. *J. Dairy Sci. 91,* 4414-4423.

VanRaden, P.M. & Sullivan, P.G. 2010. International genomic evaluation methods for dairy cattle. *Genet. Sel. Evol. 42,* 7.

Weigel, K. A., de los Campos, G., González-Recio, O., Naya, 452 H., Wu, X. L., Long, N., Rosa, G.J.M. & Gianola, D. 2009. Predictive ability of direct genomic values for lifetime net merit using of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J. Dairy Sci. 92,* 5248-5257.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M. & Wiggans, G.R. 2010a. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J. Dairy Sci. 93,* 2229-2238.

Weigel, K.A., de los Campos, G., Vazquez, A.I., Rosa, G.J.M., Gianola, D. & Van Tassell, C.P. 2010b. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *J. Dairy Sci.* (accepted).

Wiggans, G.R., VanRaden, P.M., Bacheller, L.R., Tooker, M.E., Hutchison, J.L., Cooper, T.A. & Sonstegard, T.S. 2010. Selection and management of DNA markers for use in genomic evaluation. *J. Dairy Sci. 93,* 2287-2292.

Table 4. Missing genotypes before and after haplotyping and reliabilities by marker density and by number of animals genotyped with 500,000 markers (n).

	Single density:	Mixed density:			Single density:
Genotype missing rates and genomic reliability	50,000; n = 0	50,000 and 500,000			500,000; n = 33,414
		n = 1,586	n = 3,726	n = 7,398	
Missing before (%)	1	88	80	70	1
Missing after (%)	0.05	5.3	2.3	1.5	0.05
Genomic reliability (%)	82.6	83.4	83.6	83.7	84.0