Genomic Evaluations with Many More Genotypes and Phenotypes

P.M. VanRaden[*]

Introduction

Reliabilities of genomic evaluations increase when more genotypes are matched to more phenotypes. Options to expand genomic evaluation include genotyping more animals, genotyping more markers, combining data sets across geographical and breed borders, using low cost, less dense marker subsets to obtain genotypes for a larger fraction of the population, phenotyping more animals and phenotyping more traits. Results and guidance are provided for each of those topics, including marker densities up to 500,000. New algorithms for haplotyping allow combining different marker sets and marker densities to obtain accurate genomic predictions for more animals with less expense. More options may require more research to optimize experimental designs and breeding plans.

More animals

Genotypes and phenotypes for thousands of animals are needed for successful genomic selection. Information from the reference population can be approximated as the sum of traditional reliabilities minus reliabilities of parent averages for the genotyped animals (VanRaden and Sullivan (2010)). Closest relatives provide the most information, but genotypes of previous generations also add accuracy. Numbers of progeny-tested dairy bulls that have already been genotyped are 12,142 in North America and about 20,000 in Europe. Genotypes and phenotypes for foreign animals can be easily exchanged and will add accuracy if genetic correlations across environments are high.

Currently 5,619 cow genotypes are used in the US reference population, but they add much less information than bull genotypes because of lower traditional reliability, potential biases from preferential treatment and selective genotyping. Genomic evaluations of progeny should not be biased by selective genotyping of parents, but evaluations of parents might be biased by genotyping only progeny with the best phenotypes. The Cooperative Dairy DNA Repository (US Department of Agriculture, Beltsville, MD) contains DNA that could be genotyped for an additional 10,000 Holstein bulls that have been progeny tested in North America and have a traditional reliability of >70% for milk yield. Breeders have great incentives to genotype candidates for selection but less incentive to share costs of genotyping or phenotyping additional animals to expand reference populations.

[*] Animal Improvement Programs Laboratory, ARS, USDA, 10300 Baltimore Ave., Beltsville, MD 20705, USA

More markers

More genetic markers can increase both reliability and cost of genomic selection. Genotypes for 50,000 markers now cost <US$250 per animal for cattle, pigs, chickens and sheep. Many more markers are expected to become available in the near future, and a few animals (such as the Holstein bull ToMar Blackstar USA1929410) already have been fully sequenced (3 billion DNA base pairs) by the US Department of Agriculture's Bovine Functional Genomics Laboratory (Beltsville, MD). Previously reliabilities of genomic predictions were compared for up to 50,000 actual or 100,000 simulated markers. Reliabilities for young bulls increased gradually as marker numbers increased from a few hundred up to 50,000 (Calus *et al.* (2008); VanRaden *et al.* (2009a); Weigel *et al.* (2009)), increased slightly when markers with low minor allele frequency were included (Wiggans *et al.* (2010)) and increased from 81 up to 83% as numbers of simulated markers increased from 50,000 to 100,000 using 40,000 predictor bulls (VanRaden *et al.* (2009b)).

Genotypes for 500,000 markers were simulated in this study for the 33,414 Holsteins with records in the North American database of actual genotypes as of January 2010. The population included 8,974 progeny-tested bulls, 14,061 young bulls, 4,348 cows with records and 6,031 heifers as well as 86,465 non-genotyped ancestors in the pedigrees. Haplotyping algorithms were tested using 1 simulated chromosome with a length of 1 Morgan, which is average for cattle, or using real genotypes for the same population. Gains in reliability were tested using 30 simulated chromosomes. The simulated percentages of missing genotypes and incorrect reads were 1.00 and 0.02%, respectively.

Simulating linkage. The simulation program of VanRaden (2008) was modified to generate linkage disequilibrium in the founding population (animals born before 1960). Previously Hardy-Weinberg equilibrium was generated for founding alleles, but adjacent markers become more correlated as marker densities increase. Most other studies (e.g. Meuwissen *et al.* (2001)) used thousands of generations of random mating to establish linkage. Methods to simulate linkage directly in the founding generation were derived and used in this study but might not provide the same linkage disequilibrium pattern as in actual genotypes.

Simulated genotypes and haplotypes can be more useful than real data for testing programs and hypotheses. Examples are analysis of larger data sets than are currently available or comparison of estimated haplotypes with true haplotypes, which are not observable in real data. Most simulations begin with all alleles in the founding generation in Hardy-Weinberg equilibrium and then introduce linkage using many non-overlapping generations of hypothetical pedigrees (Meuwissen *et al.* (2001)) or fewer generations of actual pedigree (VanRaden, 2008). Simulations can also include selection (Sargolzaei and Schenkel (2009)) or model divergent populations such as breeds (Toosi *et al.* (2009)). A goal of this study was to generate linkage directly in the founding population for a very large number of markers.

The steps used to generate initial linkage were to 1) simulate underlying, unobservable bi-allelic markers that each have an allele frequency of 0.5, 2) use the correlation of adjacent markers to introduce linkage among those in succession from one end of the chromosome to the other and 3) set minor allele frequencies for observed markers to <0.5 by randomly replacing a corresponding fraction of the underlying minor alleles by the major allele. The

benefit of the underlying markers and an autoregressive correlation structure is that a single linkage parameter can model the gradual decay of linkage as marker distances increase. The idea is similar to using underlying normal variables for categorical traits because the math is simpler on the underlying scale (Gianola and Fernando, 1986).

The underlying allele at the first locus on each chromosome is set to 1 or 2 if an initial uniform variable is above or below 0.5. Each subsequent allele in the founding population requires generating only two uniform random numbers: one to determine underlying linkage and a second to increase frequency of the major allele. Underlying alleles at subsequent loci are set to the same or the opposite allele as the previous locus if a uniform variable is less than or greater than 0.5 plus 0.5 times the correlation of the adjacent loci. Underlying alleles are converted to observed alleles using allele frequencies. If a second uniform variable is greater than twice the minor allele frequency, the underlying allele is overwritten with the major allele.

Correlations among adjacent underlying alleles were set to 0.965 with 15,000 markers per chromosome and 0.70 (equal to 0.965^{10}) with 1,500 markers per chromosome. Further testing may be needed to match actual and simulated linkage disequilibrium more closely.

Haplotyping. Unknown genotypes can be made known (imputed) from observed genotypes at the same or nearby loci of relatives using pedigree haplotyping or from matching allele patterns (regardless of pedigree) using population haplotyping. Haplotypes indicate which alleles are on each chromosome and can distinguish the maternal chromosome provided by the ovum from the paternal chromosome provided by the sperm. Genotypes indicate only how many copies of each allele an individual inherited from its two parents.

Many genotypes will be missing in the future when data from denser or less dense chips are merged with current genotypes from 50,000-marker chips or when two different 50,000-marker sets are merged as is being done in the EuroGenomics project using methods of Druet *et al.* (2008). Missing genotypes of descendants can be imputed accurately using low-density marker sets if ancestor haplotypes are available (Burdick *et al.* (2006); Habier *et al.* (2009)). At low marker densities, haplotypes provide higher accuracy than genotypes when included in genomic evaluation (Calus *et al.* (2008); Villumsen and Janss (2009)). Missing genotypes were not an immediate problem with data from a 50,000-marker set because 99% of genotypes were read correctly (Wiggans *et al.* (2009)).

Fortran program findhap.f90 was designed to combine population and pedigree haplotyping. Each chromosome was divided into segments of about 100 markers each. Each genotype was matched to the list of currently known haplotypes, which was sorted from most to least frequent for efficiency as haplotypes were found. If a match was found (no conflicting homozygote), any remaining unknown alleles in the found haplotype were imputed from homozygous genotypes. The individual's second haplotype was obtained by subtracting its first from its genotype, and the second was checked against remaining haplotypes. If no match was found, the new genotype (or haplotype) was added to the list. After completing population haplotyping, pedigrees were examined to resolve conflicts between parent and

progeny haplotypes, locate crossovers that created new haplotypes and impute haplotypes of non-genotyped ancestors from their genotyped descendants.

One processor took 2 hours to find haplotypes for 43,385 actual markers of 33,414 Holsteins. For the same population, time increased only to 2.5 hours with 500,000 simulated markers but with 500 markers per segment. Computing time increased much less than linearly because most haplotypes were excluded as not matching after just the first few markers. Genotype storage required 13 gigabytes for 500,000 markers, but haplotype storage required only 2.5 gigabytes. Shared haplotypes were stored just once, and only index numbers were stored for individuals instead of full haplotypes. Paternal alleles were determined correctly for 95% of heterozygous markers, and linkage was determined correctly for 98% of adjacent pairs of heterozygous markers in simulated data. Ninety-five percent of missing high-density marker genotypes were imputed correctly with population haplotyping. Pedigree haplotyping can be used to impute missing genotypes efficiently for non-genotyped ancestors or progeny with lower marker density.

Simulated genotypes for 1,479 Jerseys and 713 Brown Swiss were also used in testing the haplotyping programs. True haplotypes from the simulation allow checking proportions of correctly called linkages and paternal allele origins. Correct calls were summarized for each animal to determine how successful the algorithm was for different members of the pedigree. Estimates of genotype or haplotype accuracy will be needed with real data because true values are not available for comparison. Pedigree files included 86,465 Holstein, 16,306 Jersey and 3,969 Brown Swiss non-genotyped ancestors. Genotypes, linkages and haplotypes were estimated for those animals and compared with their true genotypes and haplotypes from simulation. For each heterozygous marker, paternity was considered to be correctly called if the allele presumed to be from the sire was actually from the sire. Linkage was considered to be correctly called if estimated phase matched true phase for each adjacent pair of heterozygous markers.

Table 1 shows results from 50,000 markers for the three breeds. For Holsteins, correctly called genotypes improved from 99% for raw data to 99.97% after haplotyping. Many non-genotyped ancestors had sufficiently accurate imputed data to meet the 90% call rate required for genotyped animals. Thus, 1,308 Holstein ancestors could have their imputed genotypes included in genomic evaluation. Nearly all of those animals were dams because most sires have already been genotyped, whereas only about 30% of dams have been genotyped. About 95% of paternal alleles were determined correctly because nearly all sires were genotyped. The most popular sires and dams had 100% correctly called linkages and paternal alleles, whereas animals with fewer close relatives had somewhat fewer correct calls.

Table 1: Frequency of correctly called genotypes, linkages, and paternity by breed and animal group

			Correct calls (%)		
Breed	Number	Animal group	Genotype	Linkage	Paternity
Holstein	1,308	Imputed	97.37	98.8	92.0
	5,369	Progeny tested	99.97	99.3	95.2
	11,646	Young	99.97	99.3	96.2
Jersey	141	Imputed	97.87	98.9	90.8
	1,361	Progeny tested	99.93	99.1	94.9
	706	Young	99.94	99.3	94.5
Brown Swiss	56	Imputed	96.63	97.8	87.0
	506	Progeny tested	99.90	98.6	94.6
	207	Young	99.90	99.5	95.2

The more precise information from haplotypes is useful both in understanding biology and in modeling the genome. Current single-nucleotide polymorphism (SNP) chips detect only genotypes, whereas new sequencing tools can directly detect haplotypes by reading DNA base pairs on one strand at a time. Currently the sequence segments are too short to reconstruct whole chromosome haplotypes easily.

Reliability. Reliability for young bulls averaged 84.0% with 500,000 simulated markers for all genotyped animals as compared with 82.6% using a 50,000-marker subset. Observed reliabilities from actual genotypes may be lower than those from simulation (VanRaden *et al.* (2009a)) and are affected by distribution of quantitative trait loci, linkage among markers and selection within the population. A heavy-tailed distribution was used in simulation of effects of quantitative trait loci and in nonlinear (Bayes A) evaluation. With 500,000 markers, one processor required 2.5 days to complete 150 iterations for the 5 replicates. Convergence was poor for the highly correlated marker effects but was acceptable for the breeding value estimates.

Combinations of marker densities can improve reliability at lower cost. Transition to higher density chips will require including multiple marker sets in one analysis because breeders will not re-genotype most animals. To determine the number of higher density genotypes needed, three data sets were simulated to include genotypes from both 500,000- and 50,000-marker chips, and the missing genotypes were imputed using findhap.f90. Table 2 shows results from analysis of the three mixed densities as well as those from 50,000 or 500,000 density alone using the same 5 data replicates. Increased reliability will require genotyping more than 3,726 of the 33,414 animals at higher density. Initially 80% of genotypes were missing, but only 6% of genotypes were missing after haplotyping.

Fewer markers for more animals

Fewer markers can be used to trace chromosome segments within a population once identified by high-density haplotyping. Without haplotyping, regressions could simply be computed for available SNP and the rest disregarded. Reduced SNP subsets were examined

Table 2: Missing genotypes before and after haplotyping and reliabilities for genomic evaluations from simulated data by marker density for genotyping and number of animals genotyped with 500,000 markers (n)

Genotype missing rates and genomic reliability	Single density: 50,000; n = 0	Mixed density: 50,000 and 500,000			Single density: 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	11	6	4	0.05
Genomic reliability (%)	82.6	81.5	82.5	83.1	84.0

using every 10th, 100th or 1000th of the original 500,000 markers. Polygenic variance was assumed to be 70, 30, 10 and 0% of genetic variance with 500, 5,000, 50,000 and 500,000 markers, respectively. Respective reliabilities obtained as squared correlations of estimated and true breeding values averaged across 5 replicates were 39, 70, 83 and 84% for 14,061 young bull predictions. With haplotyping, effects of both observed and unobserved SNP can be included.

Two simulated mixed-density data sets had 50,000 markers for cows and progeny-tested bulls but only 5,000 or 500 markers for young animals. Low- and high-density evaluations were compared for progeny that had both parents genotyped at high density. Reliabilities averaged 80% for young animals if 5,000 markers were genotyped and the other 45,000 imputed as compared with 70% from 5,000-marker regression. At 500-marker density, inheritance probabilities were computed for each marker instead of simply assigning either parental haplotype. The Fortran program sparsehap.f90 was developed to compute inheritance probabilities and genomic evaluations of progeny using parents' high-density haplotypes. Results (table 3) from this approach agree with those of Habier *et al.* (2009). Reliabilities averaged 70% when young animals were genotyped for 500 markers and both parents were genotyped for 50,000 as compared with 39% from 500-marker regression. Reliabilities averaged 77% with 5,000 markers for young animals (somewhat less than the 80% with mixed-density haplotyping) because the inheritance probability approach did not use low-density genotypes of young animals to help assign haplotypes for reference animals.

Table 3: Reliability (R^2) of simulated genomic evaluations from tracing true or estimated parent haplotypes using 500 or 5,000 markers as compared with genotyping progeny for 50,000 markers

Markers (n)	Correlation with 50,000-marker genomic estimated breeding value	R^2 (%)				R^2 gain due to use of estimated haplotypes (%)[α]
		Parent average	Haplotypes used		50,000 markers	
			True	Estimated		
500	0.92	44	71	70	83	67
5,000	0.96	44	77	77	83	85

[α]Computed as (low density R^2 − parent average R^2)/(high density R^2 − parent average R^2).

The haplotyping success rates from findhap.f90 are sufficient to implement the low-density genotyping methods of Habier *et al.* (2009). Correlations were similar whether true or estimated haplotypes were used.

More breeds

Genotypes from other breeds can be viewed as additional data, but marker effects from one breed do not accurately predict genetic merit for other breeds. Correlations should increase with higher marker density, but high reliability may still require many genotyped animals in each breed. Breeds with population sizes smaller than for Holsteins have less reliable genomic predictions because fewer phenotypes are observed for each chromosome segment. For genomic evaluation of crossbreds, data from purebreds should be combined so that effects of chromosome segments from both populations will be well estimated. Unless prevented by breed association rules, introgression of favorable DNA from other breeds can happen automatically with genomic selection.

More phenotypes

Progeny testing has generated the valuable phenotypes that genomic selection now uses, and a continuing source of new phenotypes is needed. Young bulls will sire much larger fractions of the population than in the past, and numbers of progeny per young bull will increase. This may generate a larger number of useful phenotypes even if formal programs and incentives for participation in data recording decline. Huge numbers of daughters for a few bulls may no longer occur because of more rapid turnover of generations. Instead phenotypes will be distributed more evenly across more bulls and thereby provide more information.

Phenotypes can be more directly matched to genotypes in national genomic evaluations using a 1-step instead of multi-step model (Aguilar *et al.* (2010)). The main advantage of this approach is to account properly for selection on genotypes. However, other biases may occur because phenotypes for many animals are matched to genotype probabilities instead of to observed genotypes. Use of probable or imputed genotypes is helpful only if probabilities are very high (as demonstrated in table 1). Estimates of gene content will not be precise for most non-genotyped animals in national evaluations. New methods may be needed to account for selection while using only animals with observed or well-imputed genotypes to estimate SNP effects.

More traits

Genomic predictions are most accurate for traits with long histories of recording because more phenotypes are available in the reference population. New traits can be added, but predictions may be poor until many records accumulate. Often new traits may have moderate to high correlations with previously recorded traits or combinations of traits. Therefore, multi-trait genomic models may be needed to combine recent data for new traits with historical data for correlated traits. Multi-trait models may reasonably use the same genetic correlations for SNP as for breeding values because breeding values are the sum of SNP effects.

Conclusions

Genotypes and genomic computations are rapidly expanding the data and tools available to breeders. With many new technologies and options, experimental design is becoming a more important part of animal breeding to balance the speed, reliability and cost of selection. Breeders and breeding companies need accurate advice on the potential of each investment to yield returns. Very high marker density increases reliability slightly (1.4%) in simulation, whereas lower densities could allow breeders to apply cost-effective genomic selection to many more animals. New methods for combining information from multiple data sets can improve gains with less cost. New computer programs that combine population haplotyping with pedigree haplotyping performed well with mixed-density genotypes for 500 to 500,000 markers simulated for 33,414 animals. Breeders also need to continue collecting phenotypes, especially for traits with lower heritabilities or without long histories of data recording.

Acknowledgments

The author thanks Jeff O'Connell and George Wiggans for many helpful discussions, Mel Tooker for assistance with computing, and Suzanne Hubbard for technical editing.

References

Aguilar, I., Misztal, I., Johnson, D.L. *et al.* (2010). *J. Dairy Sci.*, 93:743–752.

Burdick, J.T., Chen, W.M., Abecasis, G.R. *et al.* (2006). *Nature Genet.*, 38:1002–1004.

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W. *et al.* (2008). *Genetics*, 178:553–561.

Druet, T., Fritz, S., Boussaha, M. *et al.* (2008). *Genetics*, 178:2227–2235.

Gianola, D. and Fernando, R.L. (1986). *J. Anim. Sci.*, 63:217–244.

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

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). *Genetics*, 157:1819–1829.

Sargolzaei, M. and Schenkel, F.S. (2009). *Bioinformatics*, 25:680–681.

Toosi, A., Fernando R.L., and Dekkers J.C.M. (2009). *J. Anim. Sci.*, 88:32–46.

VanRaden, P.M. (2008). *J. Dairy Sci.*, 91:4414–4423.

VanRaden, P.M. and Sullivan, P.G. (2010). *Genet. Sel. Evol.*, 42:in press.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R. *et al.* (2009a). *J. Dairy Sci.*, 92:16–24.

VanRaden, P.M., Wiggans, G.R., Van Tassell, C.P. *et al.* (2009b). *Interbull Bull.*, 39:67–72.

Villumsen, T.M. and L. Janss (2009). Bayesian genomic selection: the effect of haplotype length and priors. *BMC Proc.*, 3(Suppl. 1):S11.

Weigel, K.A., Vazquez, A., de los Campos, G. *et al.* (2009). *J. Dairy Sci.*, 92:5248–5257.

Wiggans, G.R., VanRaden, P.M., Bacheller, L.R. *et al.* (2010). *J. Dairy Sci.*, 93:in press.