Genomic Data and Cooperation Result in Faster Progress

P.M. VanRaden[1], C.P. Van Tassell[2], G.R. Wiggans[1], T.S. Sonstegard[2], R.D. Schnabel[3], J.F. Taylor[3], and F. Schenkel[4]

[1]Animal Improvement Programs and [2]Bovine Functional Genomics Laboratories, USDA Agricultural Research Service, Beltsville, MD, [3]University of Missouri, Columbia, and [4]University of Guelph, ON, Canada.

Abstract

Genotypes for 38,416 markers of 5,335 Holstein bulls were combined with traditional evaluations to test predictive ability. Genomic evaluations were significantly (P < .0001) more accurate than official parent averages for all 27 traits tested. Squared correlations with future daughter deviations averaged 37% for young bulls as compared to 19% for parent averages. Correlations with future evaluations also increased for older, proven bulls. Genomic gains increased strongly with numbers of bulls genotyped and moderately with numbers of markers. Further increases in reliability are easily possible, for example by obtaining more genotypes for domestic or foreign proven bulls.

Introduction

Genomic selection uses thousands of markers to trace the inheritance of small chromosome segments. Each small DNA segment may account for only a very small fraction of the total genetic variance. Thus, very large numbers of genotyped animals may be required to estimate the many small effects that contribute to quantitative inheritance. Genetic markers act as a third source of data to increase the response to selection as compared to using only phenotypes and pedigrees.

Previous progress from marker assisted selection (MAS) was limited because markers or QTL with large effects explained only a small fraction of genetic variance for most traits (Schrooten et al., 2004). Progress increases when more markers are genotyped (Van der Beek, 2007). Recent MAS in France resulted in squared correlations with eventual daughter evaluations 5-19% higher for genomic predictions than for parent average (Boichard et al. 2006). Expected gains in reliability from simulation were slightly lower (Guillaume et al., 2008).

Breeding organizations and researchers in North America joined forces to develop genotyping methods and to obtain DNA from the very large families needed to produce accurate genomic predictions. This report summarizes initial results, examines potential increases in genetic progress, and outlines directions for future cooperation in this new field of genomic selection.

Genomic Data

Genotypes for 38,416 single nucleotide polymorphisms (SNP) of 6005 Holsteins were examined. The selected SNP were from the Illumina BovineSNP50 Genotyping BeadChip (Van Tassell et al., 2008) and had minor allele frequencies greater than 5% in Holsteins. Genotyping and DNA extraction was done at six locations: Bovine Functional Genomics Laboratory, University of Missouri, University of Alberta, Geneseek, GIVF, and Illumina. AI organizations in North America cooperated to contribute DNA and funding.

Genomic predictions were tested using historical data from August 2003 for 3,576 bulls born before 1999 to predict current data for 1,759 bulls born 1999-2002. Then, official April 2008 evaluations for 5,335 proven bulls and 75 cows with records were combined with genomic data to compute predictions for 623 young bulls and 29 heifers. Results were distributed to AI organizations and to animal owners in April 2008 to enable genomic selection.

Squared Correlations

Advantages of genomic selection were tested using weighted regressions of current daughter deviations on traditional and genomic evaluations computed from 2003 data. For young bulls, genomic predictions had squared correlations significantly ($P < .0001$) higher than those from parent average, with increases ranging from .38 to .02 for the 27 traits tested. Largest increase was for fat percentage and smallest for service sire calving ease. When averaged across traits, squared correlations were nearly twice as high (.37 vs. .19) for genomic predictions as compared to parent averages.

Squared correlations for proven bulls also increased significantly ($P < .001$) for 26 of 27 traits when genomic information was added. The only gain that was not significant was for service sire calving ease. Proven bulls included in these tests were those that added daughters and had at least 10% more reliability in 2008 than in 2003. Cows with records should have gains in reliability intermediate between those of young and proven bulls because traditional reliabilities for most cows are only somewhat higher than their parent average reliabilities.

Actual R^2 is lower than expected reliability for several reasons: 1) daughter deviations contain error, 2) parents are selected, 3) genetic effects reside between rather than at the markers, and 4) the genotypes include a few errors. Also, gains in R^2 may have large standard errors because of the limited number of bulls predicted. A realized genomic reliability that partially accounted for these effects was obtained by dividing the R^2 values from PA and the genomic model by average reliability of the daughter deviations, and then the difference between the published and observed reliability of parent average was added to the adjusted genomic R^2. Expected reliabilities were obtained by inverting mixed model equations that included genomic relationships.

Population Size

More predictor bulls can increase reliability by providing more data to estimate each SNP effect. Large numbers of records are required to accurately estimate the small effects of individual genes. Numbers of bulls were compared using subsets of the bull genotypes as these became available. Squared correlations for net merit of younger bulls were compared using three progressively larger subsets that included 1402, 2391, and 3319 bulls. Methods were the same as for the full set of 5335 bulls, and results are in Table 1.

Table 1.
Squared correlations (x 100) for parent average and for genomic predictions of net merit from subsets of bulls.

Bulls		Squared correlation		
Older	Younger	PA	Genomic	Gain
1151	251	8	12	4
2130	261	8	17	9
2609	510	8	21	13
3576	1759	11	28	17

Gains in R^2 for net merit were nearly linear with increasing numbers of predictor bulls. Gains for most other individual traits followed this same pattern. While linear increases cannot continue indefinitely, the results suggest that the genotyping of additional predictor bulls will be profitable, and that genomic selection within small populations will not achieve the large gains obtained in this North American Holstein population. Genomic predictions were expected to have 67% reliability with 3576 older bulls included, whereas the realized reliability was 53% calculated from adjusted R^2. This compares to 30% for published PA of young bulls, for a reliability gain of 23% in net merit.

Numbers of genotyped animals could grow very quickly over the next few years. Computer programs were tested on a simulated data set larger than is currently available to determine what resources would be required and how reliability might further increase with population size. The largest data set used 15,197 older bulls to predict 5,987 younger bulls, for a total of 21,184. This largest test simulated not just North American bulls but instead included all bulls in the Interbull file born 1995-1997 with reliability in 2003 of >65% on U.S. scale for Net Merit.

Computer times for the largest data set were reasonable for both inversion and iteration. Calculation of genomic relationships required 20 hours, followed by inversion which required 33 hours. Traditional relationships among the 81,697 animals in the pedigree were calculated by tabular method using a series of animal subsets and required 7 Gigabytes of memory and 1.5 hours. The simulated number of markers was 40,000 and the simulated number of QTLs with effects was 10,000, but the distribution was heavy-tailed so that the largest QTL explained about 3% of genetic variation.

Nonlinear genomic predictions after 200 rounds of iteration and 13 hours of computing had .4% higher reliability than linear predictions from the same data. The genomic predictions are limited more by processing time than by memory. Average reliabilities for the younger bulls across 5 replicates were 74.3% and 74.7% for linear and nonlinear predictions, respectively, as compared to 80.3% expected by inverting the linear model equations that included genomic relationships among 15,197 older bulls. At equal population sizes, observed R^2 and reliability tend to be higher with simulated than with actual data.

Predictive ability should continue to increase with additional genotyping because many bulls are needed to separately estimate so many small genetic effects. Data sharing is very important to the success of genomic studies (National Human Genome Research Institute, 2003). In cattle breeding, new policies are needed to allow data sharing or access to predictions while ensuring that those who genotype historical populations can profit from their investment. Animal breeders should not expect free access to expensive data, but could benefit greatly from service for a fee. Faster progress should result from cooperation, negotiation, and trade than from isolated studies and total genomic secrecy.

Marker Densities

Many researchers are interested to know how increasing the numbers of available SNP will affect the accuracy of genomic selection. The edited set of 38,416 SNP with > 5% minor allele frequency in Holsteins (labeled 40K) was compared to subsets of exactly half or exactly one quarter of those SNP, resulting in 19,208 (labeled 20K) or 9,604 (labeled 10K) obtained by keeping every other or every fourth SNP sequentially across each chromosome.

Results for 5 yield traits, 3 fitness traits, and Net Merit were obtained using the nonlinear model. Table 2 compares increases in R^2 above parent average for the three SNP densities, and indicates that 20K density would provide about 90% of the gain provided by 40K density, and 10K density would provide about 80% of the gain.

Table 2.
Squared correlations for parent average and for predictions with differing marker densities.

		Marker density		
Trait	PA	10K	20K	40K
Net Merit	11	25	26	28
Milk	28	45	47	49
Fat	15	41	43	44
Protein	27	45	46	47
Fat %	25	59	61	63
Protein %	28	48	53	58
Longevity	17	24	25	27
Somatic cell	23	34	36	38
Days open	20	27	28	29

As more bulls are genotyped, more phenotypes are available to estimate each SNP effect and more crossover events are observed between adjacent loci. This increases the value of having more SNP. Affordable SNP chips with higher density will likely become available in the future. However, exchange and combination of data is much easier if all populations choose to genotype a common set of SNP.

Progress

Genetic progress results from choosing the best males and females of one generation to be parents of males and females of the next via the four paths of traditional selection (Rendel and Robertson, 1950). With marker assisted selection, breeders have the opportunity to select on genotype and phenotype, only genotype, only phenotype, or to not select within each of the four paths. Optimum strategies depend on the accuracies, generation intervals, and costs for each type of selection within the four paths of selection. Equations similar to Schrooten et al. (2005) were used to compare strategies and determine progress.

In the near future, males should all be genotyped before selection for breeding. If predictions at one year of age have reliability of 60% for total merit, optimum use of young bulls vs. progeny tested bulls could increase to 90% from 20% currently. About 1% of females might be profitably genotyped with the current chip, and many more with a less expensive SNP subset. The path of females to produce sons could include 80% heifers with genotype only and 20% cows with genotype and phenotype. Females to produce daughters will still be a limiting pathway unless embryo transfer or sexed semen increase in popularity. Genetic progress should increase by about 50% when breeders are convinced that predictions are accurate.

Summary

Genomic predictions were much more accurate than PA as verified by use of historical data from 2003 to predict 2008 daughter deviations. R^2 improved when more bulls and more markers were used in predictions. Genomic evaluations of proven bulls were also significantly more reliable than traditional evaluations from the animal model.

Reliabilities increased further and computing times were reasonable when the same methods were applied to simulated genotypes for 21,184 bulls from all countries in the Interbull file instead of just North American bulls. Genomic selection has immediate benefits that will grow as more genotypes become available from larger populations. Rapid progress is possible with access to very large genomic data sets. More cooperation is required for genomic selection than was required for progeny testing.

Acknowledgments

This project was supported by National Research Initiative Grant nos. 2006-35205-16888 and 2006-35205-16701 from the USDA Cooperative State Research, Education, and Extension Service and by the National Association of Animal Breeders. Many of the results were computed by Mel Tooker and Leigh Walton.

References

Boichard, D., S. Fritz, M.N. Rossignol, F. Guillaume, J.J. Colleau, and T.Druet. 2006. Implementation of marker-assisted selection: practical lessons from dairy cattle. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Communication 22-11.

Guillaume, F., S. Fritz, D. Boichard, and T. Druet. 2008. Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40:91–102.

National Human Genome Research Institute. 2003. Reaffirmation and extension of NHGRI rapid data release policies: large-scale sequencing and other community resource projects. Available at http://www.genome.gov/page.cfm?pageID=10506537

Rendel, J. M. and A. Robertson. 1950. Estimation of genetic gain in a closed herd of dairy cattle. J. Genetics 50:1.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223.

Schrooten, C., H. Bovenhuis, J.A.M. Van Arendonk, and P. Bijma. 2005. Genetic progress in multistage dairy cattle breeding schemes using genetic markers. J. Dairy Sci. 88:1569-1581.

Van der Beek, S. 2007. Effect of genomic selection on national and international genetic evaluations. Interbull Bull. 37:111–114.

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. Interbull Bull. 37:33–36.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:(Submitted).

Van Tassell C.P., T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren and T.S. Sonstegard, 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5:247–252.