Quality and Value of Imputing Gene Tests for All Animals

Jeffrey R. O'Connell[1] and Paul M. VanRaden[2]

[1] University of Maryland School of Medicine, Baltimore, MD, USA

[2] USDA-ARS Animal Genomics and Improvement Laboratory, Beltsville, MD, USA

Abstract

Genomic selection is driven by genotyping arrays designed for uniform coverage of the genome because most quantitative trait loci (QTLs) underlying the heritability of the trait are unknown. Laboratories have improved the arrays since 2014 with custom content by adding selected QTLs discovered from whole-genome sequencing (WGS) and high-effect markers from higher-density arrays. Breed differences, missing data rates, and error rates were investigated for eight QTL gene tests currently imputed for all genotyped animals of 5 breeds plus crossbreds. Gene content for each gene test was predicted for non-genotyped relatives using mixed model methods like those used in single-step genomic evaluations, allowing potential direct selection across all animals. For the 8 QTL studied, Mendel error rates were low except for polled in Jerseys and *DGAT1* in most breeds. Allele effects for *DGAT1* were smaller than two nearby flanking single nucleotide polymorphism (SNPs) because *DGAT1* was genotype quality was poor on several arrays. For yield traits, 79K predictions including selected markers and QTLs had 1-2% higher reliability than 45K or 35K predictions excluding those SNPs.

Key words: Gene tests, imputation, marker selection, dairy cattle

Introduction

Genotyping laboratories began adding QTL gene tests in 2014 following the US Supreme Court decision that natural genetic variants should not be patented. Accuracy of imputing QTL genotypes for other animals can be affected by which arrays include the QTLs. Each year, new QTLs may be discovered and included. The SNP list used in US evaluations was updated frequently to include selected markers and QTLs from more breeds and higher density chips or from sequence (Al-Khudhair et al., 2021; Olson et al., 2012; VanRaden et al., 2009, 2017; Wiggans et al., 2016), with gains in reliability across traits expected to total about 3% (Table 1).

Some QTLs have effects larger than markers on traits we select or should select for. Goals of the project were to examine the most important QTLs currently used, summarize quality and breed differences of raw and imputed genotypes, estimate gene content for non-genotyped animals, and estimate gains in reliability of prediction from including or excluding the selected markers and gene tests.

Materials and Methods

Genotypes were examined from December 2022 official evaluations of the Council on Dairy Cattle Breeding (CDCB) for 5,669,157 Holstein, 663,366 Jersey, 65,172 Brown Swiss, 15,110 Ayrshire, and 7,620 Guernsey to summarize allele frequencies by breed (Table 2), Mendelian conflicts (Table 3) for eight important QTLs, and missing rates before and after imputation with *DGAT1* as an example (Table 4). Gene content was estimated for all non-genotyped relatives by predicting their genotypes from relatives using Gengler (2007) method. To potentially include such QTLs in a selection index, non-genotyped candidates for selection also need estimates of their unknown QTLs.

For the QTLs studied (Table 5), some have economic merit not yet included in national selection indexes such as 1) polled mutations near 1:2578598 (chromosome: position on ARS-UCD1 map) that suppress horn growth, improve animal welfare, and reduce farm labor, 2) β-casein allele (a2) at 6:84451299 in a milk protein gene that may improve

digestibility, and 3) two κ-casein alleles near 6:84451299 that affect cheese yield. The three casein QTLs are in a 200kb gene duplication region. Other QTLs mainly affect traits already in selection such as 4) *diacylglycerol O-acyltransferase 1 (DGAT1)* at 14:611019 affecting fatty acid metabolism, percentages, and yields of fat and protein, 5) *Bovine growth hormone receptor (BGHR)* at 20:31888449 affecting protein percentage, 6) *β-lactoglobulin (BLG)* at 11:103259232 with large effects on yield especially in Brown Swiss, and 7) *ATP binding cassette subfamily G member 2 (ABCG2)* at 6:36599640 with the largest effect for milk, fat %, protein %, and net merit in Holsteins, but the favorable allele is now nearly fixed at 2.5%, while fixed in other breeds (Table 2). Many other QTLs have recessive lethal effects and carrier status is reported, but those were not part of this study.

Genomic predictions using three SNP densities from 2019 yield trait data for 6,899 young Holstein bulls now proven allowed estimating the value of including selected markers and QTLs. The current 79K official list was compared to the 35K subset of only markers from the original 50K array and two 45K chips constructed by augmenting the 35K chip with independent sets of ¼ of the high density (HD) SNPs, respectively.

Results & Discussion

A true QTL is expected to have a better genetic signal (effect size or genetic SD) compared to nearby markers on the chip and that was true for most QTLs. For Holsteins, the *ABCG2* gene test had the best signal and the top ranked locus for milk, fat %, protein % and net merit. The *BGHR* gene test had the best signal and the second ranked locus for protein %. But the *DGAT1* gene test had a smaller effect than two nearby markers, and so attention was focused on *DGAT1*.

A locus from the 50K chip (ARS-BFGL-NGS-4939) on chromosome 14 at 609,870 bp had the largest genetic standard deviation (SD) genome-wide for the five Holstein yield traits: milk, fat, protein, fat % and protein %. That locus is 1,149 bp away from *DGAT1*, and another locus from the high-density chip (BovineHD1400000216) also had larger effects than *DGAT1*. Poor imputation quality was ruled out by comparing SNP regressions using only cows with direct calls for *DGAT1* and the 50K SNP. Genotypes from nine of the 52 chips and 1,377,604 Holsteins had both loci, 46,051 (6%) had discordant calls (gene test vs. marker), of which 6,830 had phenotypes. Six GeneSeek chips accounted for most of the data and had varying discordant rates (Table 6). The GeneSeek Genomic Profiler (GGP) 9K had the most genotyped animals (452,687), highest discordant rate (8.27%), and 92% (6281) of the phenotyped animals. GGP 9K regression effect sizes were greater and p-values smaller for the 50K SNP (Table 7). Genotype quality of GGP 9K was then assessed using SNP heritability (Gengler 2007) for 25,000 animals with discordant calls on that chip. The 50K SNP had heritability 0.98 and *DGAT1* only 0.16, indicating poor genotype quality as the likely source. Discordant calls for *DGAT1* on other chips also had low heritability although sample size was much smaller.

Because some valuable gene tests are sold by laboratories rather than delivered with array genotypes, freely imputed QTLs could benefit breeders and progress. Decreasing costs of whole genome sequence data will increase power of QTL discovery, and more QTL genotypes should increase imputation accuracy, prediction accuracy, and economic gain. Regressions averaged 1.07 and were nearly equal across the 3 densities. Reliabilities of yield traits for 79K averaged 1.2% higher than 45K and 2.0% higher than 35K, worth potentially > $10 million every year nationally. Eventually, more QTLs should be included to further improve predictions.

Conclusions

Gene tests were already imputed for all genotyped animals of all five breeds. Mendelian error rates were low for QTLs except for Polled in Jerseys and *DGAT1* in most breeds. Imputed *DGAT1* tests were statistically less significant for all yield traits compared to two nearby chip SNPs (one HD and one 50K), direct *DGAT1* gene tests also had smaller effects than the best markers, and SNP heritability indicated that *DGAT1* genotyping quality was the cause of later imputation errors, though the GGP 7K and linkage disequilibrium (LD) V4 had low discordance rates. Further investigation of problematic chips is warranted. Gene content was imputed for all non-genotyped animals by extracting QTLs from the imputed genotypes and using those as data to predict related animals. Accumulated gains in reliability for yield from adding selected markers and QTLs were 1-2%, a little less than previous studies indicated. Most gains were from larger reference populations.

Acknowledgments

The authors thank CDCB staff and industry cooperators for contributing data to the National Cooperator Database, and Emmanuella Ogwo for computing and providing gene test summary statistics.

References

Al-Khudhair, A., VanRaden, P.M., Null, D.J., and Li, B. 2021. Marker selection and genomic prediction of economically important traits using imputed high-density genotypes for 5 breeds of dairy cattle. *J. Dairy Sci*. 104(4):4478–4485

Gengler, N., Mayeres, P. and Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal* 1:21-28.

Olson, K.M., VanRaden, P.M. and Tooker, M.E. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci*. 95(9):5378–5383.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., and Schenkel, F.S. 2009. *Invited review:* Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92(1):16–24.

VanRaden, P.M., Tooker, M.E., O'Connell, J.R., Cole, J.B., and Bickhart, D.M. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* 49:32.

Wiggans, G.R., Cooper, T.A., VanRaden, P.M., Van Tassell, C.P., Bickhart, D.M., and Sonstegard, T.S. 2016. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. *J. Dairy Sci.* 99(6):4504–4511.

Table 1. History of US SNP list revisions and reported gains in reliability of Holstein predictions

Year	Reference	Breeds	Added information	Markers (1000s)		HOL Reliability	
				Added	Total	Gain (%)	Total
<2008		All	Parent average		0		27
2009	VanRaden	HO	Chip genotypes (50K)	38	38	23	50
2012	Olson	3	More breeds (JE, BS)	5	43	0	50
2013	Wiggans	HO	Add HD markers (GHD)	18	61	0.5	67
2016	Wiggans	HO	Add HD markers (GH2)	16	77	1.5	68
2019	VanRaden	HO	Add sequence SNPs	2	79	1.2	69
2020	Al-Khudhair	5	Add HD, other breeds	+5, -5	79	0	69

Table 2. Final allele frequencies for the eight QTLs including gene content for all animals of each breed

Breed	Polled	ABCG2	β-casein	κ-casein1	κ-casein2	β-Lact	*DGAT1*	BGHR
RDC	0.6	99.9	52.0	84.8	65.3	33.2	8.6	22.1
BSW	3.5	100.0	22.2	30.1	100.0	33.0	6.8	11.4
GUE	1.1	99.7	7.2	65.1	99.7	16.0	60.6	17.9
JER	2.2	99.9	27.6	9.2	99.4	54.2	52.1	26.1
HOL	1.0	97.4	39.1	72.5	89.8	51.6	30.1	19.7

Table 3. Mendelian error rates by breed for imputed genotypes of eight QTLs

Breed	Polled	ABCG2	β-casein	κ-casein1	κ-casein2	β-Lact	*DGAT1*	BGHR
RDC	0.01	0	0.17	0.00	0.01	0.05	0.80	0.11
BSW	0.18	0	0.10	0.12	0.00	0.12	0.51	0.03
GUE	0.00	0	0.00	0.04	0.00	0.14	0.00	0.07
JER	0.50	0	0.17	0.13	0.00	0.03	0.09	0.08
HOL	0.05	0	0.08	0.01	<0.01	0.02	0.67	0.10

Table 4. *DGAT1* imputed allele and genotype frequencies and genotypes missing in input

Breed	Tests (N)	Frequency (%)							
		Allele	Imputed genotype codes						Genotypes
		A	AA	AB	AB	A?	B?	Missing	Missing
RDC	15,110	8.6	88.11	8.84	0.07	2.85	0.06	0.07	71.6
BSW	65,172	6.8	78.14	9.27	0.45	10.38	0.86	0.90	91.0
GUE	7,620	60.6	14.00	43.24	33.23	3.49	5.66	0.38	89.0
JER	663,366	52.1	21.34	49.43	27.63	0.74	0.85	0.01	74.2
HOL	5,669,157	30.1	46.10	42.70	9.60	1.12	0.48	0.00	85.7

Table 5. Locations and effects of eight QTLs examined

Gene test	Chr:Location	Gene function	Effects in cows or in humans
Polled	1:2578598	Grow horns	Animal welfare, farm labor
ABCG2	6:36599640	Membrane transport	Yield and NM$ (biggest effect)
β-casein (a2)	6:84451299	Milk protein	More digestible? (JE protein%)
K-casein (1)	6:85656772	Milk protein	Increased cheese yield
K-casein (2)	6:85656792	Milk protein	Increased cheese yield
β-Lactoglobulin	11:103259232	Milk fat	Human allergies (BS yield & %)
DGAT1	14:611019	Fat and protein %	Fatty acid metabolism, obesity
BGHR	20:31888449	Growth hormone	Protein% (2nd biggest effect)

Table 6. Descriptive statistics for six GeneSeek chips tested for *DGAT1* calling

Chip info		Animal info		
Name	Markers	Genotyped	Discordant (N)	Discordant (%)
GGP 7K	7083	34480	239	0.69
GGP 9K	8984	452687	37417	8.27
GGP LD V4	30113	112135	327	0.29
GGP 65K	65320	95327	5578	5.85
GGP 100K	94121	30606	1676	5.48
GGP 150K	139914	36406	813	2.23

Table 7. Regression results for GGP 9K chip for *DGAT1* vs. nearby 50K SNP using 6,281 genotyped animals

	Marker P-value		Abs (marker effect)	
	50K	*DGAT1*	50K	*DGAT1*
Milk	8.9E-45	2.6E-02	70.932	11.887
Fat	1.4E-19	3.1E-02	2.062	0.546
Protein	4.9E-13	9.8E-01	0.967	0.004
Fat %	2.2E-94	3.8E-04	0.016	0.003
Protein %	4.1E-42	1.2E-04	0.003	0.001