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 21201 USA
[2]USDA Animal Genomics and Improvement Lab
Beltsville, MD 20705 USA

Goals of project

- Examine the most important Quantitative Trait Loci (QTLs) in the SNP list
  - Genotyping laboratories began adding QTL gene tests in 2014
  - Some have large effects on traits we select or should select for
- Summarize imputed QTLs for quality and breed differences
  - Input vs. output genotypes, missing rates, Mendelian conflicts
- Examine marker density and SNP lists used in national evaluations
  - Summarize yield trait reliability gains for 79K vs 35K and 45K density

Gene tests examined

Gene test	Chr:Location	Gene function	Effects in cows or in humans
Polled	Casein Gene cluster	Yield Traits	Animal welfare, farm labor
ABCG2		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)

HO gene test frequency trends and missing genotypes

Gene test	Base Freq %	Current Freq %	Freq trend %	Input missing %
Polled	0.2	1.1	+0.9	85.0
ABCG2	94.1	99.1	+5.0	63.9
β-casein (a2)	46.9	32.5	-14.4	92.3
K-casein (1)	81.5	58.8	-22.7	74.7
K-casein (2)	90.7	86.4	-4.3	83.3
β-Lactoglobulin	44.6	61.4	+16.8	70.5
DGAT1	30.8	38.6	+7.8	85.7
BGHR	26.8	13.2	-13.6	72.7

From HO imputed genotypes. Missing % is from input data

Final allele frequency (%) comparison by breed

Breed	Polled	ABCG2	β-casein	K-casn1	K-casn2	β-Lact	DGAT1	BGHR
AY	0.6	99.9	52.0	84.8	65.3	33.2	8.6	22.1
BS	3.5	100.0	22.2	30.1	100.0	33.0	6.8	11.4
GU	1.1	99.7	7.2	65.1	99.7	16.0	60.6	17.9
JE	2.2	99.9	27.6	9.2	99.4	54.2	52.1	26.1
HO	1.0	97.4	39.1	72.5	89.8	51.6	30.1	19.7

Mendelian error rate (%) comparison by breed

Breed	Polled	ABCG2	β-casein	K-casn1	K-casn2	β-Lact	DGAT1	BGHR
AY	0.01	0.00	0.17	0.00	0.01	0.05	0.80	0.11
BS	0.18	0.00	0.10	0.12	0.00	0.12	0.51	0.03
GU	0.00	0.00	0.00	0.04	0.00	0.14	0.00	0.07
JE	0.50	0.00	0.17	0.13	0.00	0.03	0.09	0.08
HO	0.05	0.00	0.08	0.01	<0.01	0.02	0.67	0.10

Mendelian conflicts are after imputation.
Mendelian conflicts before imputation also examined from chip QC stats.

Frequencies of imputed gene tests: DGAT1

Breed	Tests (N)	Frequency	Imputed genotype codes (%)					
		(final, %)	AA	AB	BB	A?	B?	??
AY	15,110	8.6	88.11	8.84	0.07	2.85	0.06	0.07
BS	65,172	6.8	78.14	9.27	0.45	10.38	0.86	0.90
GU	7,620	60.6	14.00	43.24	33.23	3.49	5.66	0.38
JE	663,366	52.1	21.34	49.43	27.63	0.74	0.85	0.01
HO	5,669,157	30.1	46.10	42.70	9.60	1.12	0.48	0.00

Summary: Gene tests and imputation

- Gene tests are already imputed for all genotyped animals, all 5 breeds
  - Mendelian error rates low except for JE Polled and DGAT1 (most breeds)
- Gene content can be imputed for all non-genotyped animals
  - Extract the QTLs from the imputed genotypes
  - Use those as data to predict related animals, as in single-step GBLUP
- Since several laboratories sell the QTL gene tests, free imputed QTLs would be useful for customers

Gene tests should beat markers

- **The true QTL is expected to have a better genetic signal (allele substitution effect size or genetic SD) compared to other markers on the chip**
- **ABCG2 gene test** on chromosome 6, 36.6 Mb is the best signal and the **top ranked locus** for milk, protein % and net merit in HO

	Allele substitution effects				
	Milk	Protein	Fat %	Protein %	Net Merit
Rank	1	6	3	1	1

- **BGHR gene test** on chromosome 20, 31.8 Mb is the best signal and the **second ranked locus** for protein % in HO

DGAT1 is the exception

- **50K SNP** ARS-BFGL-NGS-4939 (***50K-4329***) on chromosome 14 at ~610 kb has the largest genetic SD genome-wide for the 5 HO yield traits: **milk, fat, protein, fat % and protein %**
- DGAT1 is 1,149 bp from 50K-4939 (611,019 bp) **has the 3rd largest effect**

		Trait Genetic SDs				
Marker	Freq	Milk	Fat	Protein	Fat %	Protein %
50K-4939	0.69	48.3	1.87	0.607	0.012	0.0063
DGAT1	0.39	39.4	1.51	0.509	0.010	0.0060
Delta %		22.5	23.9	19.1	23.0	4.7

Potential reasons DGAT1 is not the best

- Poor imputation quality?
  - Missing rate 86% for input genotype but <2% in imputation output
- Difficult to genotype?
  - Dinucleotide substitution AA->GC compared to SNP
  - Chips differ in quality of calls, DGAT1 is not usable on some chips
- Multiple assays on the same chip?
  - One chip reports 3 different DGAT1 tests, others report 2 tests
- Not the true QTL?

Compare DGAT1 and 50K-4939 discordance by chip

- December 2022 Holstein genotype calls
- 9 chips have both loci genotyped
- 762,754 animals have genotype calls at both markers
- 46,051 (6%) animals have discordant calls

CHIP INFO				ANIMAL INFO MARKER PAIRS		
NUM	ABBR	MARKERS	NAME	NUM	DISCORDANT	PERCENT
12	ELD	9072	EuroG10K	410	0	0
17	GH2	139914	GGP_Bovine_150K	36406	813	2.23
18	G7K	7083	GGP_Bovine_7K	34480	239	0.69
19	GP4	30113	GeneSeek_Genomic_Profiler_LD_Version_4	112135	327	0.29
25	G9K	8984	GGP_Bovine_9K	452687	37417	8.27
29	ID3	53450	Irish_Beef_and_Dairy_Chip_Version_3	121	1	0.83
31	EL7	10706	EuroG10K_Version_7	582	0	0
42	GM3	94121	GGP_Bovine_100K	30606	1676	5.48
43	G65	65320	GGP_Bovine_65K	95327	5578	5.85

DGAT1 regression results by chip

- **6,830 of the 46,051 animals with discordant genotypes have phenotypes (daughters' or own) for yield**
- **Single variant regression adjusting for genetic relationship matrix**
- **50K-4939 beats DGAT1 on GGP 9K with N=6281 and 8.27% discordance**

	GGP 9K, N=6281, Discord=8.27%			
	Marker P-value		Abs Marker effect	
	50K-4939	DGAT1	50K-4939	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

SNP “heritability” of discordant genotypes

- **Fit linear mixed model with SNP as the outcome to estimate SNP h^2 (Gengler)**
  - **SNP ~ *mean + g + e*, *g* ~N(0,s^2G)**
- **Low h^2 is indicates potential genotyping error**
- **GGP 9K with N=25,000 DGAT1 has low h^2**

	GGP 9K, N=25,000, Discord=8.27%	
	50K-4939	DGAT1
Marker h^2	0.976	0.160

Summary: DGAT1 gene test

- Analysis of discordant genotype calls
  - Regression ruled out imputation quality
  - SNP heritability indicates genotyping quality
- The value of a gene test depends being a true QTL and genotype quality
  - An incorrect or poor-quality test may reduce imputation and prediction accuracy
  - DGAT1 currently adds no additional value over the 50K-4939
- The difference between 50K-4939 and DGAT1 was replicated in JE
- *To further investigate DGAT1 assay quality is needed and planned*

USA SNP list history

Year	Reference	Breeds	Added information	Markers (1000s)		HO 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

Benefits from selected markers and QTLs

- Predict 2022 from 2019 PTAs
- 6,899 young Holstein bulls now proven
- Compare 3 SNP densities
  - 79K current SNP list
  - 45K with ¼ of the HD SNPs
  - 35K subset of 50K chip
- ~2% higher R^2 (79K vs. 45K)
- Regressions very similar

Trait	35K	45K	79K
	Squared correlations (%)		
Milk	81.3	82.0	84.5
Fat	84.3	84.7	86.2
Protein	82.9	83.2	84.9
	Regressions (B_1)		
Milk	1.09	1.09	1.09
Fat	1.08	1.08	1.08
Protein	1.05	1.05	1.05

Summary: marker test

- **While most countries still use 45-50K markers the US industry continues to update SNP list with additional markers and QTLs**
- **There was 1-2% reliability gain with 79K vs 45K**
- **Worth > $10 million / year nationally**

Conclusions:

- **Long-read whole genome sequencing is required to accurately identify QTLs to maximize reliability**
  - **Multiple QTLs between markers**
    - **Gene clusters such as casein**
  - **Non-SNP variation such as DGAT1**
  - **The number of QTLs across the dozens of current plus future traits is greater than 50K or 79K markers**

Acknowledgments

- CDCB staff and industry cooperators for data
- AGIL staff and USDA funding of project 8042-31000-113-000-D, “Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals”
- The NIH National Heart, Lung, and Blood Institute (NHLBI) funding of grant U01 HL137181