REDUCED SNP PANELS – creation, realistic expectations, and use in different livestock industries

Written by Alison Van Eenennaam, UC Davis

Alison Van Eenennaam, PhD
Cooperative Extension Specialist
University of California
Department of Animal Science
One Shields Avenue Ph:(530) 752-7942
Davis, CA 95616 Fax:(530) 752-0175
Email: alvaneenennaam@ucdavis.edu
http://animalscience.ucdavis.edu/animalbiotech

High-density SNP chip assays (e.g. Illumina BovineSNP50,000 (**SNP50**), BovineHD ~770,000 SNP (**HD**)) are currently price prohibitive for many applications and species. There is considerable interest in developing low-density, low cost SNP assays for a variety of purposes including selection of breeding stock in species where individuals have a comparatively low value relative to the cost of high-density arrays, selection of replacement animals on commercial farms, parentage assignment, optimizing mate choice, and marker-assisted management. Two basic approaches can be used to develop low-density arrays. The first involves selecting SNPs that are the most highly associated with the trait of interest in the training data set. In the case of traits that are affected by very many genes with a small effect, as seems to be the case with most complex traits (Hayes *et al.* 2010), not all genes will be associated with a reduced set of markers. The other approach is to use a subset of SNP to "impute" high-density genotypes. Imputation is a method of dealing with missing genotypes by filling in values.

USE OF REDUCED SNP PANELS IN LIVESTOCK INDUSTRIES

There are a few published reports of high accuracy reduced SNP panels being used in company breeding lines of chicken and pigs (Table 1). In the swine industry a small number (96 or 196) of the most significant SNPs for each different trait were made into trait-specific SNP panels and in the poultry industry a 384 reduced SNP panel was used for high-density (41K) panel imputation of multiple traits,

Table 1. Company-reported accuracy estimates of commercial panels for livestock selection

Industry	Trait	# SNPs	Accuracy (r_g) estimate	Country	Breed	Company
Swine	Scrotal Hernia	96	0.30	US	Cross-bred	Genus/PIC[1]
Swine	Finisher mortality	96	0.30	US	Cross-bred	Genus/PIC
Swine	Total born	196	0.77	US	Cross-bred	Genus/PIC
Chicken	Body Weight	384 (being used for 41K imputation)	0.58	US	Broiler	Aviagen Ltd.[2]
Chicken	Hen house production		0.60	US	Broiler	Aviagen Ltd.

In the dairy industry a study compared the best makers selected from the 50K chip for 9 dairy traits, (Moser *et al.* 2010). Few were found to be in common between the different traits, and at least 1,000 of the highest ranked SNPs were required to get accurate predictions for each trait. The authors of this paper concluded that combining the highest ranked SNP for each trait onto a single chip was not a feasible approach to reducing genotyping costs. Weigel *et al.* (2009) compared dairy lifetime net merit correlations between molecular breeding values (**MBV**) based on SNP panels of varying size and progeny test data (Table 2).

[1] Deeb, N. *et al*. (2011) http://www.intl-pag.org/19/abstracts/P05n_PAGXIX_606.html
[2] Wang *et al*. (2011) http://www.intl-pag.org/19/abstracts/P05m_PAGXIX_580.html

Table 2. Correlations of April 2008 predicted transmitting abilities for lifetime net merit with August 2003 MBV for all SNP and selected or equally spaced reduced SNP sets in a testing set of 1,398 Holstein bulls (rPT_All), 1,195 bulls with genotyped sires (rPT_Sire), and 203 bulls without genotyped sires (rPT_NoSire) (Weigel *et al.*, 2009).

No. SNP	Method of SNP Selection	rPT_All	rPT_Sire	rPT_NoSire
300	Largest Effects	0.428	0.447	0.312
300	Equally Spaced	0.253	0.262	0.202
500	Largest Effects	0.485	0.503	0.369
500	Equally Spaced	0.333	0.348	0.245
750	Largest Effects	0.519	0.530	0.441
750	Equally Spaced	0.435	0.450	0.348
1,000	Largest Effects	0.537	0.549	0.460
1,000	Equally Spaced	0.422	0.438	0.321
1,250	Largest Effects	0.554	0.567	0.461
1,250	Equally Spaced	0.477	0.489	0.395
1,500	Largest Effects	0.559	0.576	0.445
1,500	Equally Spaced	0.518	0.534	0.412
2,000	Largest Effects	0.567	0.582	0.469
2,000	Equally Spaced	0.539	0.559	0.408
32,518	**All Available**	**0.612**	**0.627**	**0.511**

The high density chip (32,518 SNP) provided a correlation of 0.612 for all bulls, with a significant advantage for bulls with genotyped sires. By comparison, correlations between progeny test data and MBVs from 300 to 2,000 selected SNP panels ranged from 0.428 to 0.567, and correlations between progeny test data and MBVs from 300 to 2,000 equally spaced SNP ranged from 0.253 to 0.539. In every case, the predictive ability of MBV from selected SNP was greater than for equally spaced SNP.

However because low-density assays composed of selected SNP will be breed and trait-specific, the authors concluded it would be more efficient to use equally spaced SNP that would facilitate imputation of high-density genotypes, rather than to focus on prediction of MBVs from smaller panels that contain only a few hundred selected SNP with large estimated effects for a single trait. The preferred option is to use evenly spaced SNP to infer or impute the sequence of missing SNPs based on the high density genotype of key ancestors. The dairy industry is currently using the Illumina BovineLD **(LD)** Genotyping chip which is comprised of 6,909 SNP for imputation. As of February 2012, the USDA national genotype database for dairy cattle included LD genotypes for 19,515 animals.

A hybrid of these two approaches involves selecting a subset of highly ranked SNP that are themselves within evenly-spaced segments of approximately equal size for imputation (Habier *et al.* 2009; Moser *et al.* 2010). The GeneSeek Genomic Profiler 80K (**GGP-80K**) is an example of such a product. It was developed with around 80,000 SNPs. Although 80K may not seem to be a low-density SNP assay, it is relative to the 770,000 HD! The GGP-80K includes SNP50 and HD SNP with the largest effects on net merit index. Consideration also was given to spacing as well as maintaining around 30,000 SNP50 SNP for imputation to HD. The GGP-80K genotypes are expected to improve the accuracy of imputation and genomic evaluation in the dairy industry because of these additional SNP (Wiggans *et al.*, 2012).

USE OF REDUCED SNP PANELS IN THE BEEF INDUSTRY

Until relatively recently, commercialized DNA tests for marker-assisted selection in beef cattle targeted only a handful of traits, specifically marbling, tenderness and feed efficiency (Van Eenennaam *et al.* 2007). Recent tests on the U.S. market target more than 10 traits including growth, maternal, and carcass traits. One of these tests is a product based on the Illumina SNP50 (**Pfizer 50K**, Pfizer Animal Health, Kalamazoo, MI) and the other is a 384 SNP panel (**Igenity 384**, Duluth, GA). Both products have been trained for Angus cattle. The accuracies (genetic correlation (r_g)) between MBV and phenotypic trait of interest in American Angus Association data are in the range of 0.24-0.65 (Table 3).

Table 3. Genetic correlation between genomic results (MBV) and phenotypic trait of interest (American Angus Association data) by genomics company[3].

Trait	Genetic Correlation (r)/(r^2%)	
	Igenity 384	Pfizer 50K
Calving Ease Direct	.47 (22%)	.33 (11%)
Birth Weight	.57 (32%)	.51 (26%)
Weaning Weight	.45 (20%)	.52 (27%)
Yearling Weight	.34 (12%)	.64 (41%)
Dry Matter Intake (component of RADG)	.45 (20%)	.65 (42%)
Yearling Height	.38 (14%)	.63 (40%)
Yearling Scrotal	.35 (12%)	.65 (42%)
Docility	.29 (.08%)	.60 (36%)
Milk	.24 (06%)	.32 (10%)
Mature Weight	.53 (28%)	.58 (34%)
Mature Height	.56 (31%)	.56 (31%)
Carcass Weight	.54 (29%)	.48 (23%)
Carcass Marbling	.65 (42%)	.57 (32%)
Carcass Rib	.58 (34%)	.60 (36%)
Carcass Fat	.50 (25%)	.56 (31%)

Such high accuracies for multiple traits when using a single 384 SNP panel contrasts from findings with reduced panels in other animal industries. Likewise the accuracy estimates associated with the 50K Pfizer product are higher than would have been predicted by deterministic modeling based on the number of phenotypic records used in the training populations. These high accuracies might be explained if there are relationships between animals in the population that was used for training (i.e. high accuracy Angus AI bulls), and the evaluation population (i.e. registered Angus cattle). This is undoubtedly the case, and would likely be the case for most breeds where the training population involves widely-used (i.e. high-accuracy) sires. Markers can predict family relationships between animals, independently of linkage disequilibrium between the markers and genes (Habier *et al.* 2007). If animals in the training and target populations share DNA segments from a small number of ancestors and are only a few generations apart, a relatively small number of markers will be able to track segments shared between related animals (Moser *et al.* 2010).

[3] Northcutt. S.L. (2011) http://www.angus.org/AGI/GenomicChoice11102011.pdf (updated 11/18/2011)

The practical implication of markers picking up family relationships is that the accuracy of marker-based selection will decay over generations within breed. This was demonstrated in German Holstein cattle where the additive-genetic relationships between training and validation animals were found to be a good indicator of accuracy (Habier *et al.* 2010). Effectively this means that the accuracy of prediction equations will decrease as the relationship between the training population and the evaluation population becomes more distant. From the perspective of seedstock breeders, this might not be an issue as elite seedstock typically provide the next generation of selection candidates and so selection candidates will most likely be closely related to the training population. However, such tests are likely to be less accurate across lines of Angus cattle that have few close relatives in the training data set. Practically this means that SNP effects will have to be re-estimated frequently to include data from each generation of selection candidates, although this may create logistical complications for genetic evaluation entities, especially if they do not have access to both the phenotypes and the genotypes or if additional costly phenotyping is required.

THE FUTURE

At the current time the costs of genomic testing tend to exceed the value that is returned to any single sector of the beef industry. Seedstock producers are sometimes sending DNA to different labs for pedigree verification, genetic defect testing, and genomic enhanced EPDs, and combined genotyping costs have sometimes been in excess of $200 per animal. This is cost-prohibitive, and the inefficient practice of extracting DNA multiple times from the same animal will soon become a relic of the past.

As genotyping costs continue to decline, it is likely that assays involving thousands of SNP will become inexpensive and perhaps "the norm", and there will be no need to select a few hundred of the "best" SNP to develop a low-cost, reduced-SNP panel. Genomic technology providers are already starting to develop cheaper multipurpose SNP panels with thousands of SNP markers. For example, GeneSeek has developed a product called the Genomic Profiler (GGP) using the "add-on" capability to add custom SNP to the existing Illumina LD chip. This 8,655 SNP product includes the 6,909 LD evenly-spaced SNP for imputation and additional SNP for proprietary single-gene tests for recessive conditions including genetic defects, detection of haplotypes that affect fertility in dairy cattle, imputation of microsatellite alleles to facilitate parentage validation, and improved imputation by including SNP from the now-obsolete Illumina GoldenGate Bovine3K Genotyping BeadChip. As SNP for an increasing number of uses are included on a single chip, the benefit derived from these multiple applications will increase the value derived from genotyping. It may be that low-cost SNP panels and the combined value derived from using DNA information for multiple purposes across the beef-cattle supply chain will push the economics of genomics over the tipping point towards more widespread industry adoption (Van Eenennaam and Drake, 2012)

REFERENCES

Habier D., Fernando R.L. and Dekkers J.C.M. (2007) Genetics 177: 2389.

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

Habier D., Tetens J., Seefried F.-R., Lichtner P. and Thaller G. (2010) Genet. Sel. Evol. 42: 5.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J. and Goddard M.E. (2010) Plos Genet 6:1.

Moser G., Khatkar M.S., Hayes B.J. and Raadsma H.W. (2010) Genet. Sel. Evol. 42: 37.

Van Eenennaam A.L., Li J., Thallman R.M., Quaas R.L., Dikeman M.E., Gill C.A., Franke D.E. and Thomas A.G. (2007) J. Anim. Sci. 85: 891.

Van Eenennaam A.L. and Drake D.J. (2012). Anim. Prod. Sci. 52: 185.

Weigel K.A., de los Campos G., Gonzalez-Recio O., Naya H., Wu X.L., Long N., Rosa G.J.M. and Gianola D. (2009) J. Dairy Sci. 92: 5248.

Wiggans G.R., VanRaden P.M., Cooper T.A., VanTassell C.P., Sonstegard T., and Simpson B. (2012) J. Dairy Sci. *In press*.