

Genetic marker technology and its impact on the dairy industry

June 2009

Bennet Cassell, Extension Dairy Scientist, Genetics and Management, Virginia Tech

What's all the fuss?

Major changes in genetic evaluations of dairy cows don't come along every day. Producers notice the genetic base changes every five years. We have recently added traits to the indexes like Net Merit because it made sense to do so. These changes, though, use systems that have been in place for some time. Over the past year, though, a fundamental change has been under way. A very new kind of information about genetic merit of dairy cows has become available and is being integrated into genetic evaluations.

The information comes from a technology called "dense SNP arrays". The device that measures all these SNPs is called the Illumina Bovine SNP50 beadchip. What this technology does is reveal the genetic makeup of individual animals at about 50,000 locations in their DNA. That's 50,000 sites out of the about 3,000,000,000 possible sites in the DNA of dairy cattle, so "dense" certainly doesn't mean that we could reproduce DNA code from this technology. This kind of information is called SNP data or genomic data. SNP, by the way, stands for "single nucleotide polymorphism" or, literally, differences between individuals in a population in the nucleotide structure of DNA at a single point.

About 38,000 of these 50,000 pieces of information have proven to be useful, meaning that they differ enough from dairy animal to dairy animal to help track transmission of genetic material from generation to generation. Animals that are related share some DNA through descent from a common ancestor(s) and thus show more agreement in genomic data than non related animals. Greater agreement exists between close relatives, especially if they also happen to be inbred. Genomic data does not change across an animal's lifetime, and can be measured very early in life, which is one of the major advantages of the information. Many important phenotypes, or characteristics we directly observe such as milk records on daughters of a bull, are only available when the animal is older. Phenotypic information becomes more useful in predicting genetic merit of an animal over time, since more records accumulate either on the maturing animal or on their progeny.

Dense SNP array technology uses blood or tissue or hair or semen samples. Only a single sample is needed in the life of an animal, but the test costs about \$250 per animal. This isn't a test that most dairy farmers would be willing to use on a whole herd, at least at this stage of the game. It has been used on AI bulls of historic importance as well as currently active AI sires and young bulls in AI sampling programs. Dairy males can only be genotyped if sponsored or nominated by a major AI organization because of proprietary control of the technology. This exclusivity for bulls will be in effect until 2013. However, dairy females may be genotyped by individual owners. The tests have been performed on a growing number of potentially elite females in the Holstein, Jersey, and Brown Swiss breeds. Some high end purebred sales have begun to feature genomic tested females and sales prices are beginning to reflect the results of those tests. A February 2009 summary of genotyping showed that 22,344 dairy animals had been processed using the SNP50 chip. Many of these animals, of course, were Holstein, and the four birth years with largest number of animals genotyped were 2005-2008.

A little basic biology

It is hard to discuss genomic data without some understanding of the structure of the DNA molecule. The shape of the DNA molecule is something like a twisted ladder. It is made up of two strands or sequences of four nucleic acids, adenine (A), thymine (T), cytosine (C), and guanine (G). There are about three billion of these sites, spread across 30 chromosomes in the bovine, and bonded together in pairs to form the "rungs" of the ladder. Adenine - A - always pairs with T, and C always pairs with G. DNA is a very complex and very long molecule.

Many living organisms - humans and cows among them - are "diploid" - meaning that two chromosomes of DNA carry the genetic code necessary for gene expression. These pairs of DNA strands are by no means identical, with the variation between the two copies serving important roles in the evolution of species. The 50,000 SNPs (nucleic acid sites) occur in a linear sequence in two strands (legs on the ladder) on those pairs of ladders on each of 30 chromosomes. Thus, there are actually four strands or sequences of nucleic acids, which pair up with the rungs described above. The two strands in a ladder exist in a unique relationship where one sequence is a very predictable complement of the other. A nucleotide sequence of AGCT on one strand would be associated with a paired strand of TCGA.

The dense SNP array test (actually 50,000 tests) measures a small segment of the sequence of nucleotides, using a DNA-like chemical construction called a "primer". A

sequence is useful for genomic mapping if a single nucleotide in the sequence is variable or "polymorphic". A recognizable sequence for the primer used to interrogate one SNP could, for instance, be GGGAGGG. If such a sequence existed on one (of four) DNA strands, the primer would "anneal" to it and the test would record a positive result - a YES answer. The parallel strand of DNA would always read CCCTCCC to compliment such a sequence, but the primer would not recognize that sequence. The other chromosome of the pair might also include the sequence GGGAGGG, which would produce a second "YES", but it **could** contain GGGTGGG or GGGGGGG or GGGCGGG, each of which would produce a NO. The primers on the dense SNP chip will record a YES or NO answer for both pairs of chromosomes. In one method of recording such answers, 0 means "two NO answers", 1 means "one YES and one NO" and 2 means "two YES answers". There are other methods to record the results, so a challenge when using genomic data is to know how the SNP test results are stored. Regardless, a list of 50,000 zeros, ones and twos makes for very boring reading.

How the SNP data help predict genetic merit

The sites that are measured are distributed more or less evenly throughout the DNA of dairy cattle, which is an important feature of the SNP50 chip. Since only 50,000 out of three billion possible sites are measured, about 60,000 nucleic acid sites are skipped for each one that is recorded. Thus, almost all of the functional units of an animal's DNA are NOT measured. However, SNP sites that are recorded bracket or are imbedded in or bracket chunks of functional DNA. The sequence of nucleic acids, measured by SNP results, at specific sites allows scientists to track the movement of those significant chunks of DNA from generation to generation.

This is a simplistic explanation, but is the basis of the relationship between SNP array data and genetic merit as we use the term in animal breeding. Statistical procedures associate the 0, 1 or 2 data at each site with phenotypes of the animal herself or (more importantly with today's technology) with phenotypes of progeny. Thus, dairy cows with phenotypes but no SNP data are extremely important in revealing what the SNP sequences mean in AI bulls or elite cows. When SNP sequences in animals without records or progeny are similar to those of ancestors with progeny data, the genetic predictions assume that relationships of SNP and phenotype from earlier generations still hold. Thus, genomic predictions on an AI young sire are based on his dense SNP array results and predictions from the relationship of genomic and phenotypic information in related, older animals. Since the SNP sequences for a young animal reflect the genes actually inherited from sire and dam, genomic predictions for that animal reflect genes actually inherited rather than anticipating an average sample of genes from each parent. Genomic predictions of genetic

merit gain accuracy over pedigree evaluations because they are not restrained by the expectation that full sibs, for instance, inherited an average sample of genes from each parent. With traditional pedigree evaluations, each offspring is assumed to have inherited an average sample of genes from his or her sire and dam. Full sibs have equal "parent average" of PA, but are expected to share only half of their genes as copies of the same genes in their parents. Genomic data depart from the expectation of "average sample from each parent" by actually measuring the genetic material each individual received from each parent. Some full sibs are more closely related than the expected 50% of genes in common, while other full sib pairs are less closely related. The degree of relationship between relatives is not necessarily the same for all traits on which selection may be practiced. Genomic data can tell the sire analysts at bull studs which full sib bull calf to sample from an ET flush. There is no need to guess which youngster is best and no need to sample two or three full brothers unless more than one appears to have inherited favorable gene samples. Genomic predictions of full sibs are NOT the same. "Pick of flush" has taken on a new meaning with genomic evaluations.

Published genomic predictions combine traditional PA with genomic data. The two evaluations are combined using selection index theory which weights each piece according to the amount of information it contains. Published genomic evaluations, gPTA's, then, are a blend of information long familiar to dairy cattle breeders and a unique, new information predicted from the genome of the animal evaluated. Research shows that the combination is more useful than either kind of information alone. Yet some of the problems with traditional PA estimates continue in gPTAs. Results from AI sampling programs have shown that many PA's are inflated for many AI young sires. Most of this upward bias comes from very high evaluations on elite bull dams. There is no one simple answer to how bull mother proofs get inflated, but daughters of the young sires don't face the same careful management that benefited their paternal grandmother. Genomic information corrects much, but not all of this problem, as traditional PA's are part of the gPTA.

Genomic data and genetic evaluations

Studies of the effectiveness of genomic predictions have been conducted by the Animal Improvement Programs Laboratory, ARS, Beltsville, MD starting in Fall of 2007. Scientists have devised methods to predict genetic merit from SNP data and have tested the genomic predictions against the genetic evaluations that have been routinely available to the industry.

The most recent test performed in 2009 to establish the validity of genomic evaluations used SNP scans, pedigree relationships, and progeny records from a "predictor" group of 4,422 Holstein bulls heavily used in AI and born before 2000. Another group of 2,035 younger bulls with progeny records available and born after 2000 served as the test group of animals to be predicted. Genomic evaluations and parent averages from the older group of bulls used performance records terminated at the end of 2004. None of the test group bulls contributed progeny to those records. The tests involved a wide variety of traits as shown in the Table below.

Two other breeds were involved in these studies. In the Jersey breed, the prediction equations came from 1149 bulls born before 2000 along with 212 cows with data. The test group consisted of 388 bulls born after 2000. In Brown Swiss, the predictions were from 472 bulls born before 2000, and the test group was 150 bulls born after 2000. There were many fewer "predictor" bulls in the Jersey and Brown Swiss breeds. Consequences of the smaller numbers are evident in the results.

Table 1. Increase in Reliability of genetic evaluations of young bulls when genomic predictions are compared to traditional Parent Average information.			
	Gain (%) in Reliability from genomic data		
Trait	Holstein	Jersey	Brown Swiss
Net Merit	24	8	9
Milk	26	6	17
Fat	32	11	10
Protein	24	2	14
Fat %	50	36	8
Protein %	38	29	10
Productive life	32	7	12
Somatic cell score	23	3	17
Daughter pregnancy rate	28	7	18
Final score	20	2	5
Udder depth	37	20	8
Foot angle	25	11	-1

Table 1, based on data available in April 2009, shows the increase in Reliability from using genomic evaluations of young sires compared to using Parent Average. Reliability increased by using genomic data for all traits in all three breeds except for foot angle in Brown Swiss. . The increases were always greatest for Holsteins, and sometimes by quite a large amount. This result shows that genome data must be available on large numbers of animals

each with accurately recorded phenotypes to build predictions of genetic merit for animals not yet old enough to have such phenotypic data. Consequently, the impact of genomic evaluations in the Holstein breed exceeds that for Jerseys. Increases in accuracy for Jerseys do not always exceed Brown Swiss despite the larger number of genotyped animals. Accuracy of predictions is affected by how genes operate, and the mechanisms are not necessarily the same in different breeds.

The increases are greater for some traits than for others for reasons which we do not always understand - yet. This is very new science. The greatest increase due to SNPs was for milk components expressed as a percentage, especially for fat. We have known for some time of a major gene called DGAT1 that exists at a reasonable frequency in two forms in dairy cows. One form of this allele causes higher fat percentages than the other. SNP chip data detect this allele, improving the accuracy of prediction of fat % over traditional pedigree methods where the status of DGAT1 is unknown. Table 1 shows major improvement in accuracy for udder depth, but an explanation for that result is not readily available. We will learn more, and probably much more, about genetic control of individual traits over time.

In a second test, Holstein proofs from the 2004 data were used to make selection decisions among AI young sires and AI proven bulls. The top 20 young sires were chosen based on traditional PA and on genomic evaluations. There may have been some overlap in the two groups. Two sets of top 20 proven bulls were then chosen by the traditional as well as the genomic PTA methods, again with the possibility of overlapping. For proven bulls, addition of genomic information raises Reliability by about 3 percentage units for bulls with first crop daughters, so the gain in accuracy is less than for young sires.

Table 2. Selection based on combinations of genomic and progeny data on AI young sires and proven bulls in AI service		
Top 20 bulls (2004 data)*	Average NM\$, 2009	Difference from NM\$ in 2004
Young bulls, traditional PA	\$395	-\$278
Young bulls, gPTA	\$516	-\$130
Proven bulls, traditional PTA	\$381	-\$96
Proven bulls, gPTA	\$463	-\$30

*Five years' information is added between the two sets of proofs. Young bulls add progeny test results while older bulls add many second crop daughters.

The best group of bulls based on the 2009 NM\$ evaluations were young bulls chosen five years earlier on genomic evaluations and no progeny. The average Net Merit for this group was \$53 higher than the second best group of bulls, proven bulls in AI with genomic data included in their proofs. So, genomic information on young bulls allowed selection to identify a very good group of bulls. An important caveat is that the evaluations on genomically evaluated young bulls declined by an average of \$130 over the five year period. Some of these 20 bulls declined a lot more than \$130, while some change less. The message is that genomic proofs cannot yet identify the very best individual bulls before progeny data are available, but genomic data does reduce the slippage from PA to PTA by over 50%. The group of young bulls selected on traditional PA declined by \$278. A second observation also jumps out. When genomic data was added to first crop progeny data, the average proof change from first crop daughters to lots of second crop daughters was reduced from a drop of \$96 to only \$30. First crop proofs with genomic information were very stable. This will allow breeders to identify individual bulls for high-value matings more accurately than previously possible and should increase the rate of genetic progress.

Plans for genetic evaluations using genomic data

The Animal Improvement Programs Laboratory at Beltsville, MD published the first genomic PTA's in January 2009. Some breeders and industry personnel had experience with genomic proofs earlier in 2008, however. Preliminary proofs were released to owners of individual animals during the time that the procedures for using genomic data were under development. As the story below relates, selection decisions have been made based on these initial proofs.

The first genomic PTA that I saw was on a young bull, the result of an ET flush, which was shared with me by a Virginia dairy farmer in the summer of 2008. This flush produced three bull calves, full brothers. One went to stud on the basis of a favorable genomic evaluation. One of the other two left on the farm (the one I was shown) had a genomic PTA of \$448 compared to an official Parent Average which did not include genomic information of \$507. This young bull did not go to stud because he did not get a particularly favorable sample of genes from his sire and dam. However, the genomic PTA of \$448 is quite respectable. He would be a well above average contributor to a natural service genetic improvement program. The third brother was not so fortunate. He might be used as a breeding bull, but he is not as likely to be a genetic improver as his two brothers. The Reliability of the genomic PTA for Net Merit was 63% for all three bulls, compared to 34% for traditional PA. The brothers differed in genomic PTA, but not in

Reliability, as the same amount of information was available on each. This story has been re-told many times with different breeders in recent months.

The role of young sires in the semen market place has changed dramatically since genomic proofs were first released in January 2009. Many of the high ranking young bulls based on gPTAs are available to dairy farmers. And the predicted genetic merit of those bulls is outstanding. An April '09 proof sheet from the bull stud GENEX shows 20 young bulls with no progeny among the best 25 bulls for NM\$ at that stud. GENEX is by no means the only stud offering outstanding gPTA bulls to dairy farmers on a regular basis. The semen price says that they are good bulls. This new genetic resource is priced high to reflect its expected true value to producers.

Dairy farmers should give this new offering a serious look. The best bulls in any stud have ALWAYS been the young bulls in waiting, but we didn't know which ones were the good ones. Genomic evaluations increase Reliability of NM\$ on young bulls in sampling programs from about 36% to about 66% based on evaluations released in January 2009. An accuracy of 66% is excellent for finding the top group of young bulls, but I'm not ready to bet the farm on finding the best individual two or three bulls on such information, at least based on what we know about genomic proofs today. Use groups of such bulls. Use bulls with combinations of genomic data and progeny records more heavily, depending on their genetic merit. The tradeoff between accuracy and rank for NM\$ is always a key element in deciding how much to use individual bulls.

One place where gPTAs may have an especially important impact is in the genetic merit of bulls for which sexed semen is available. In the recent past, studs were understandably reluctant to offer their top bulls via the sexed semen option. The sexing process is extremely wasteful of semen, and studs were reluctant to accept that waste on high demand proven bulls. Genomic proofs on youngsters make a whole new group of bulls available for marketing through sexed semen. Genetic merit of bulls available in the sexed semen format seems to be improving. Of course, demand for sexed semen is a function of prosperity among dairy farmers. Demand for sexed semen softened in 2009 with the decline in milk prices.

Prospects for using SNP chips in commercial herds

A test that costs \$250 per heifer won't attract much interest from commercial producers. However, a new kind of genetic improvement product is being developed to take advantage of SNP chip technology. The generic name that seems to be sticking is "low density SNP chips". These SNP chips screen for 300-500 SNPs instead of 50,000 on

the dense SNP chip. A company called Igenity currently offers these low density SNP chips at around \$35 per sample. I saw an advertisement recently for skilled employees to develop this technology for Pfizer. Other segments of the dairy industry are at work on low density chips as well, including researchers at USDA who developed the dense SNP chip. This area of dairy herd management technology is attracting considerable venture capital at this time.

The SNPs to be included in low density chips will be chosen from among the most effective of the SNPs on the dense 50K SNP panel. The purpose, just as for the dense SNP chip, will be to predict genetic merit for a variety of traits. Some traits such as milk, productive life, somatic cell score and pregnancy rate, are polygenic and thousands of genes affect animal performance. Other traits are favorable or unfavorable recessives like BLAD, CVM, DUMPS, or red coat color. Single-gene traits can be evaluated with a great deal of accuracy by low density SNP chips if those chips include the right SNPs. The polygenic traits are tougher and it remains to be seen how useful low density SNP chips will be for such traits. Under some circumstances, such as purchasing replacement heifers for which no pedigree data are available, low density SNP chips would provide some information for selection decisions. In those cases, a \$35/head investment could be recovered even with relatively low accuracy. Some progressive producers may wish to combine low density SNP screening with extensive use of sexed semen to select the genetically better heifers from a larger than normal group of potential replacements.

A specially selected group of about 100 SNPs which will find their way into most low density offerings have proven to be very useful for verifying or perhaps even establishing parentage. These SNPs would be especially useful to improve the accuracy of progeny testing programs. Large herds, which have been increasingly important to progeny testing programs, will be an ideal place to use this technology. If (when) this technology is widely applied to progeny test programs, the industry can look forward to more accurate first proofs, more stable proofs, and greater customer satisfaction from sire selection. Accurate identification of first crop progeny will increase differences in genetic evaluations of bulls as well.

The Future

We stand on the doorstep of a new era in genetic improvement of dairy cattle. The potential of genomic evaluations is enormous. We are just beginning to make use of that potential, yet a number of benefits are already having an effect. Accuracy of female selection, long a weak link in genetic progress, can be increased dramatically by this new

technology. The limitation is the cost of the test itself, which means that many females will not be evaluated. A new step has been added to the selection process prior to progeny testing. In the past, young sires destined for AI sampling programs had to have outstanding parents. Today, an outstanding genomic prediction in the young bull is required - good parents AND a good sample of genes. This change not only increases odds of success, it opens the door to a more varied group of pedigrees for AI sampling programs.

One study showed that use of genome information to select sires of replacement females in dairy herds could increase the rate of genetic progress by at least 50%. Accuracy, even with genomic data, is less for young bulls than for progeny tested bulls, but the reduction in generation interval from at least six to two or three years more than offset the loss in accuracy. That study is based on what we know today, which isn't much, really. The very first actual genome sequences only appeared in September 2007. We certainly don't have a great deal of experience in how to make best use of the information. And the technology on which genomic predictions are based will only improve. I look forward to these sweeping changes, and to the opportunities and the challenges they present,