Understanding Genomics: An Introduction to the Terminology

Jay Shannon

Holstein Association of Canada, P.O. Box 610, Brantford ON N3T 5R4
Email: jshannon@holstein.ca

Take Home Messages

- The bovine genome was the first ruminant species to be sequenced. The comprehensiveness of dairy cattle data has enabled the dairy industry to quickly take advantage of these extensive opportunities for improvement.
- Despite very small effective population sizes of cattle, nucleotide diversity is relatively high in the Holstein breed to allow genomic selection to make further progress in production, longevity, conformation, health and fertility traits.
- All cells in the body contain DNA comprised of 3 billion sub-units called nucleotides arranged over the 30 chromosome pairs. The nucleotides are coded as being one of four types: A, T, G or C.
- A SNP (pronounced snip) marks a nucleotide location where variation exists in the population. There are only 2 to 4 million SNPs on the bovine genome, over two million of which have been discovered to date. The development of high-throughput, high-density SNP chips, most notably the Illumina BovineSNP50 Beadchip, has significantly advanced genomic testing and predictive capabilities.
- By genomic testing a large number of historical bulls using the 50K SNP chip and establishing a predictive relationship between the SNPs and sire proofs for all traits, genomics provides significant gains in accuracy for young animals.
- As more bulls are tested and larger SNP chips become available, the accuracy of genomic predictions for young bulls and heifers will lead to significant change in how the dairy industry is modeled in terms of genetic improvement.
- With the higher accuracies and a shortened generation interval, greater progress will be achieved for all economically important traits especially

WCDS Advances in Dairy Technology (2010) Volume 22: 327-338

for lower heritable traits. Hard to measure traits such as feed efficiency, health and nutritional components may become possible using genomic selection.

- Genomics will be the focus of livestock research and genetic selection in the dairy cattle industry for the next two decades.

Introduction

A 13-year global initiative to fully sequence the human genome has led to many associated projects to sequence the genome of numerous other species including the bovine. For over six years, three hundred international researchers from 25 countries worked together on a $53M US project to fully sequence the genome of an inbred 8-year-old Hereford cow from Montana; making the bovine the first mammalian livestock species to be sequenced. Her high resolution sequence was complemented by lower level sequences from six other breeds including Holstein, Angus, Jersey, Limousin, Norwegian Red and Brahman. Primary funding ($25M) for this research was provided by the National Human Genome Research Institute, as well as many other contributors including $5M from Genome Canada through Genome BC.

Interest in sequencing the bovine genome related to the similarity of the bovine genome with that of humans. Dr. Harris Lewis from the University of Illinois said “cattle and other ruminants diverged from a common ancestor that gave rise to humans about 95 million years ago, yet cattle and humans still share a high degree of conservation of their chromosome architecture” (Woolcott E. 2009). The bovine genome provides valuable insight into how the human genome works with concern to health and disease since humans share roughly 80% of their genes with bovine (National Human Genome Research Institute 2009). The key factor in the decision to sequence the cattle genome was related to her ability to efficiently convert low energy forages unfit for human consumption into high energy milk and associated components. By comparing the bovine genome with humans and mice, bovine researchers discovered significant reassembling of genes involved in digestion, immunity, reproduction and lactation (Tellam R.L. 2009).

Effective population size is a derivation of the number of unique animals in a population that actually contribute genes to subsequent generations. Cattle have witnessed a rapid decline in effective population size due to domestication some 11 million years ago, subsequent breed formation and intensive selection practices over the last half century (Figure 1). By comparison, the human effective population size has exponentially increased over a similar time period (The Bovine HapMap Consortium 2009). Nucleotide diversity is another measure of genetic diversity. The good news is bovine HapMap researchers have discovered close to 40% more nucleotide variation in the Angus and Holstein breeds than is found in the human

population (The Bovine HapMap Consortium 2009). So despite the small effective population size, bovine researchers suggest ample variation still exists to enable the use of high-density genotyping techniques to make continued improvement in production, longevity, conformation, health and reproduction.

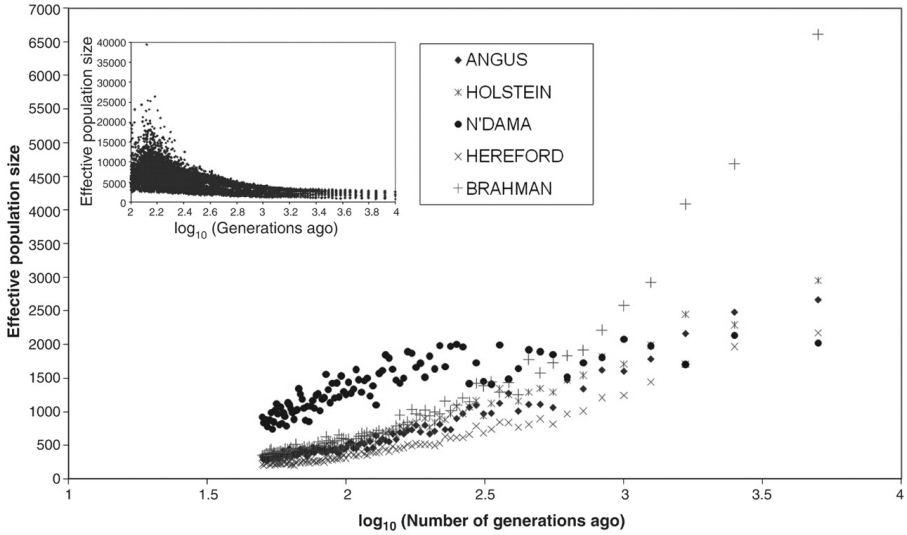

Figure 1: Rapid Decline in Effective Population Size (The Bovine HapMap Consortium 2009)

■ Understanding the Biology

All cells in the body (with the exception of mature red blood cells) contain the complete chemical DNA (deoxyribonucleic acid) which comprises all instructions needed for cell functions. The genome is the complete set of DNA. The genome is arranged into distinct chromosomes, of which there are 23 for humans and 30 for bovine. The cows' genome contains only 22,000 protein-coding genes and roughly 500 microRNA genes (The Bovine Genome Sequencing and Analysis Consortium, et al 2009). These genes contain instructions for making proteins, which are fundamental in building and maintaining cellular structure and function.

The DNA for both humans and cows contains approximately 3 billion long chain subunits called nucleotides, which are the building blocks of DNA. Each of the 3 billion nucleotides can be coded as being one of four different types: A (adenine), T (thymine), G (guanine) or C (cytosine). The exact sequence or ordering of the nucleotides (i.e. CCTAGGCGTTACAA) along the

genome plays an important role in the construction of proteins that distinguish a unique individual possessing varying characteristics and innate abilities.

In animal breeding, there are qualitative and quantitative traits. Qualitative traits such as colour are typically single genes. Conversely, most traits in dairy genetics are quantitative traits with varying extremes such as milk production or udder depth. These traits are not only influenced by management and environment, but also involve numerous genes acting in a much more complex interaction. For that reason, a full genome approach, as opposed to candidate gene approaches, is far more effective in unraveling the secrets behind these important economic traits.

Close to 99.9% of the nucleotides are exactly the same for all individuals. The remaining 0.1% still represents over 3 million nucleotides with unique coding among individuals in a population. A Single Nucleotide Polymorphism (or SNP pronounced snip) exists when a nucleotide at a specific location on the genome is different for at least 1% of the population. For example, a sequence of …AGATA**G**CT… might change to …AGATA**C**CT… for a portion of animals in a population (Figure 2). SNPs are spread over both gene coding regions and the more vast non-coding regions of the genome however SNPs in both regions are very important in explaining genetic variation. Though some SNPs have no effect on cell function, many SNPs either solely or in combination with other SNPs, can help explain significant genetic differences. By the completion of the bovine genome project, researchers had discovered approximately 2.44 million SNPs on the bovine genome (Eck et al, 2009).

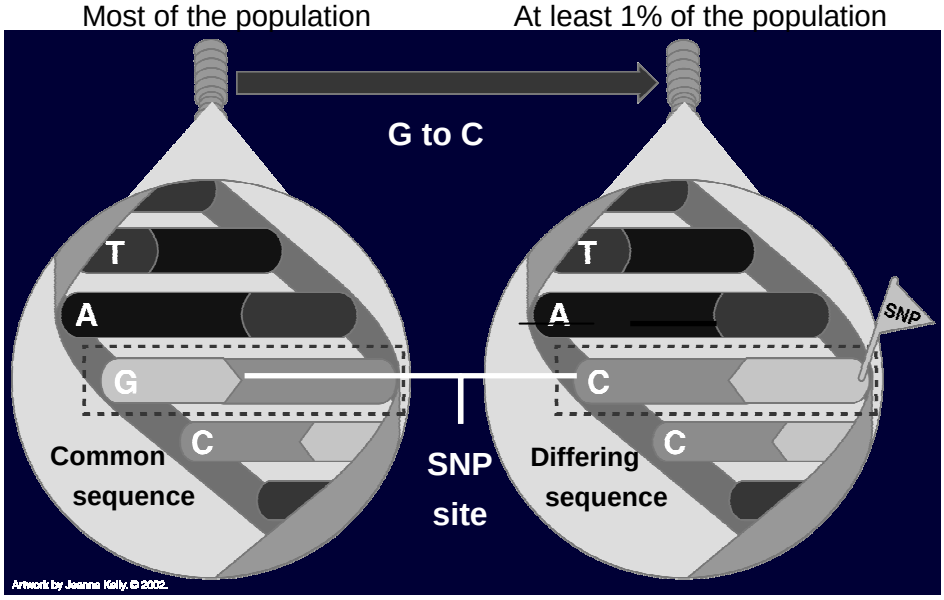

Figure 2: Identifying a SNP Site on the Genome (National Cancer Institute 2005)

As part of the bovine genome project, 29 bovine chromosomes were sequenced (including the X chromosome but not the Y chromosome). To better understand inheritance and transmission to the next generation, the 29 chromosomes sequenced were actually 29 paired chromosomes. An allele pair (also called a genotype) shows the difference between the nucleotides on the two paired chromosomes. Animals with two different nucleotides at the same location on the paired chromosomes (for example a G on one and an A on the other) are called heterozygous. This infers the animal received one nucleotide type from one parent and the other nucleotide type from the other parent. Furthermore, this animal with a GA-genotype would transmit a G-nucleotide to 50% of their offspring and an A-nucleotide to the other 50%. Animals with identical nucleotides at the same location on both chromosome pairs (i.e. GG-genotype or AA-genotype) are said to be homozygous. With a homozygous genotype, the animal received the same nucleotide type from both parents and would also transmit that same nucleotide type to all offspring.

■ The SNP Chip

To read the nucleotide type at a specific SNP location, researchers first built chemical probes capable of searching the entire genome for a particular SNP

of interest. These probes cannot determine a long string of nucleotides, but only the next nucleotide in a sequence. Fortunately, a tremendous advancement in genomic testing called the “SNP Chip” allows researchers to conduct thousands of simultaneous probes on multiple animals in a chemical reaction called an assay. Current assays can assess up to 200,000 SNPs on 12 different animals. The next wave of high density panels coming in 2010 will allow researchers to read up to 700,000 SNPs per animal. (Harrison, 2009)

SNPs are very important due to their quantity, low cost to read thousands of genotypes in a single assay, and their relationship to genes or in reshaping the proteins produced by genes that determine cellular function. With a close proximity to genes, SNPs are known to express genes, either one-to-one or possibly as multiple SNPs in combination to reflect a gene. By extension, it makes sense that a large quantity of SNPs would identify a large number of genes.

The objective of high density panels is to densely cover the whole genome with SNPs such that there’s a high probability for many SNPs to be in close proximity to most genes of interest, and thereby explain the majority of genetic variation in a population. Early large density SNP panels in dairy research ranged from 1K and 10K SNPs, and though the results showed some promise, researchers concluded a denser panel would be required to capture the linkage disequilibrium necessary for whole genome mapping (Sargolzaei, et al, 2008).

■ Illumina BovineSNP50 Beadchip

In 2007, a collaboration of research institutions including USDA, University of Missouri and the University of Alberta, developed a high density 60,000 Bead Illumina iSelect® assay called the “Illumina Bovine50SNP Beadchip”. The resulting 58,000 SNPs were selected to be equally spaced along the genome. With the large quantity of SNPs, it was anticipated there would be significant genetic variation explained.

To estimate and validate the importance of the selected SNPs in explaining the genetic variance for all traits, researchers needed to test a large number of animals in the population. A consortium of 7 A.I. centers in North America was established including Canadian-based A.I. units Semex and Alta Genetics. In the first phase of the project, the A.I. units provided semen for over 5,000 Holstein bulls for testing using the Illumina Bovine50SNP Beadchip. The chip uncovers the bulls’ genotype for each of the 58,000 SNPs. The genotype would be a combination of two possible nucleotide types for the specific SNP on the paired chromosome. The SNP is identified with a 0, 1 or 2 representing the number of copies of the one of the two types.

For example, if a specific SNP was either a C or T in the population then the allele combination would be 0 - CC (0 copies of T), 1 - CT (1 copy of T) and 2 – TT (2 copies of T). Thus, each tested animal receives a result containing 58,000 values (0, 1 or 2 for each SNP).

The SNP results are just a bunch of codes until they are matched up with the intelligence of the elaborate genetic evaluation systems in Canada and the United States. Most dairy bulls have been proven through a random-sampling young sire proving program where the performance of all daughters is measured for many traits. The performance of each daughter is adjusted for non-genetic factors such as age at calving and level of herd management, and then corrected for the genetic merit of the dam. The daughter information for a sire is combined to derive a relatively pure estimate of the genetic worth of each sire for all traits. In Canada, there are 63 traits or indexes that are genetically analyzed by the Canadian Dairy Network.

By taking a single trait such as EBV Fat Kg, researchers estimated the relationship between the 58,000 SNPs and the EBVs for all tested bulls. An intense mathematical algorithm derives 3 possible SNP solutions (for 0, 1 or 2) for the 58,000 SNPs, so it results in 174,000 SNP estimates. A bull's genomic estimation for a trait would be calculated by summing the SNP estimates that correspond with the bulls' genotypes. The process is repeated for all other traits in dairy cattle selection. In the genomic predictions, the actual number of SNPs evaluated was reduced from 58,000 to 43,385 due to duplicative or non-informative SNPs (Wiggans, 2009).

To validate the gain in accuracy from genomics, researchers estimated the SNP effects on an independent group of predictor bulls born prior to year 2000, and applied these effects to a separate group of bulls born from 2000 to 2004 (Van Tassel et al, 2009). Genomic values were estimated for all 2000-2004 born bulls based on a combination of parentage and genomic information. These values were then compared with actual sire proofs from 2009 for the same bulls based on daughter information. The objective was to evaluate how much gain in accuracy was achieved by the inclusion of genomic information compared to parent average alone. Table 1 shows the gain in reliability (accuracy) in Canada and the United States by the inclusion of genomic information. The highest gains were achieved in Fat % and Protein %. This validation study confirmed the power of genomic prediction and initiated the genomic revolution in the dairy genetics industry that continues to unfold.

Table 1: Increase in Accuracy from Genomics

Trait	Gain in Reliability in Canada [1]	Gain in Reliability in U.S. [2]
Milk	24	26
Fat	29	32
Fat %	40	50
Protein	21	24
Protein %	33	38
SCS	17	23
Conformation	21	20

1. Schenkel et al, 2009
2. Van Tassell et al, 2009

The development of dairy genomic values has involved an extensive cooperative network between the United States and Canada involving USDA, University of Guelph, Canadian Dairy Network, Holstein USA and Holstein Canada. The agreements include the sharing of genotypes for all tested animals in North America as well as technical exchanges to perfect the science and technology on both sides of the border.

All females and males benefit from gained accuracy due to the inclusion of genomic information. There are varying degrees in accuracy gain dependent on the animals' age and amount of information included in the traditional genetic evaluation (Table 2). Logically, the role of genomics is lesser for an animal performance and progeny compared to a young heifer or bull with no progeny or performance. It is easy to see how genomics will allow the youngest animals in a population to be used with greater certainty and reliability. For young bulls, the traits with the greatest gain in accuracy from genomics are SCS (+30), Production traits (+29), Calving Ability (+29), Conformation (+27) and Mammary System (+27), whereas for proven bulls, it is the lower heritable traits such as Herd Life (+13), Daughter Calving Ability (+9), Daughter Fertility (+8) and Milking Temperament (+7) (Van Doormaal, 2010).

Table 2: Gains in Reliability by Animal Sub-group in Canada (Van Doormaal, 2010)

Sub-Group	Average Reliability (%)		
	Traditional	Genomics	Gain
Young Bulls & Heifers (2007-2009)	34	61	27
Younger Cows in 1st or 2nd Lactation	54	68	14
Foreign Cows with MACE in Canada	43	65	22
1st Crop Proven Sires in Canada	86	89	3
Foreign Sires with MACE in Canada	70	80	10

■ The Genomic Revolution

The sequencing of the bovine genome will revolutionize the dairy industry by making greater genetic progress in milk production, feed efficiency, reproduction, health and disease resistance leading to a more efficient and sustainable production system with a smaller environmental footprint.

There are several factors which have enabled the dairy industry to quickly take advantage of these genomic advancements compared to other animal-based sectors. A high proportion of dairy herds participate in the genetic improvement programs via animal registration, DHI (production and fitness traits) and conformation appraisal. Since dairy cows tend to have longer active lives as production units, data management and analysis are fundamental components on modern dairy farms. And with the maturity of the dairy genetic industry, a large quantity of traits are measured and collected on a national basis. These phenotypic observations are adjusted for management and environmental factors to derive unbiased genetic estimates, which are vital components for the estimation of genomic values for all traits.

Genomics will allow the top genetics to be more accurately identified earlier in life. The expanded use of young genomic bulls and virgin heifers will set the pace for the next generation, leading to a reduced generation interval and more rapid genetic progress for selected traits. This is especially true for lower heritable traits of economic importance such as fertility and health traits.

The gain in accuracy continues to increase as more predictor bulls are genotyped. It is anticipated that the average reliability gain might continue to increase slightly until the number of predictor bulls is close to the number of SNP effects. Table 3 was taken from a study by the US Department of Agriculture showing how the accuracy gains from genomics increases with the number of predictor bulls that have been genotyped.

Table 3: Accuracy Gains with More Animals Genotyped (Van Tassell et al, 2009)

Number of Predictor Bulls	Number of Predictee Bulls	Average Reliability Gain (27 traits)
2130	261	17
3576	1759	23
4422	2035	29
6184	7330	30

Preliminary data shows the linkage disequilibrium in bovine is only slightly larger than humans. It is estimated that a high density panel consisting of up to 300,000 SNPs for cattle will be necessary for optimal marker coverage through the entire genome (Eck et al, 2009).

The technology is advancing at an incredible pace. The BovineHD high density SNP panel, which includes more than 500,000 SNPs, will be available on the market in the first quarter of 2010 (AllBusiness 2010). Higher density panels provide more accurate genomic predictions however the limiting factor will be the low number of genotyped predictor bulls in contrast to the large number of SNP effects to be estimated. Novel approaches are being considered to fill in (or impute) missing genotypes by analyzing the relationship of SNPs on the low density, 50K and BovineHD panels. The objective will be to maximize the amount of information and accuracy wherever possible. In addition, improvements in technology are expected to reduce the cost of testing over time. Getting more for less will be the catch phrase in genomic testing as it moves forward.

Illumina plans to introduce a low density chip in the first quarter of 2010 that contains 3,000 SNPs (genomeweb 2010). By selecting the most informative SNPs from the larger panel to predict the most important economic traits, the low density panel will try to find a balance between low cost and optimal return. It will bring genomic testing to a much broader base of dairy herds. Many dairy farms are expected to choose regular testing of young herd entries as a pre-screen culling and mating tool to optimize their herd genetics and performance.

■ Summary

There are obvious benefits but also some potential pitfalls that could come from genomics. Genomics has the potential to identify new genetic lines into mainstream breeding that will act to control inbreeding levels. Conversely, any technology which speeds up genetic progress typically places more focus

on the best genetics with similar bloodlines and therefore acts as a catalyst for higher inbreeding. Inbreeding levels must be controlled for genomics to have a positive effect on the future of dairy cattle breeds.

Another concern will be any shortfall in the genomic predictions and how far along the breeding companies are in subsequent generations before a corrective factor can be applied. The past has shown us that some animals have not maximized their genetic potential due to limitations in other traits or management practices. The full genome approach should help to minimize this concern however it will continue to be important to analyze differences in genomic predictions before and after milking daughters are included. The potential gain from genomics far outweighs any possible concerns. The success of genomics will be in how it is managed. It will most certainly have a lasting and positive effect on the dairy industry in Canada and around the world.

■ References

AllBusiness. 2010. Illumina launches high-density bovine beadchip. Article published at www.allbusiness.com.

Burt, D.W. et al. 2009. The cattle genome reveals its secrets. J. Biology 8:36.

Eck, S.H. et al. 2009. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biology 10:R82.

Genomeweb. 2010. Illumina to debut MaizeSNP50 chip, low-density bovine panel this quarter. Article published at www.genomeweb.com.

Harrison, A. 2009. Cow or Ùber Cow. Illumination Fall 2009 University of Missouri. Article published at http://illumination.missouri.edu/f09/uber_cow.

National Cancer Institute. 2005. Understanding cancer series: genetic variation (SNPs). Images available at www.cancer.gov.

National Human Genome Research Institute. 2009. USDA and NIH funded international science consortium publishes analysis of domestic cattle genome sequence. Article published at www.genome.gov/27531571.

Sargolzaei, M., et al. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Science 91:2106.

Schenkel, F.S. et al. 2009. Reliability of Genomic Evaluation of Holstein Cattle in Canada. Interbull International Workshop, Genomic Information in Genetic Evaluations, Uppsala, Sweden, January 26-29, 2009.

The Bovine Genome Sequencing & Analysis Consortium, et al. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 324 522-528.

The Bovine HapMap Consortium. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324 528-532.

Tellam, R.L. et al. 2009. Unlocking the bovine genome. BMC genomics 10:193.

Van Doormaal, B. 2010. Genomics REL Gains Tables – January 2010.pdf. Personal communication.

Van Tassell, C. et al. 2009. Walking the cattle continuum: Moving from the BovineSNP50 to higher and lower density SNP panels. Presentation published at: www.illumina.com/documents/.../presentations/2009_08_tassell_curt.pdf.

Wiggans, G.R. 2009. Why genomics works and what is coming. Presentation published at: http://aipl.arsusda.gov/publish/presentations/MISC09/DiscoverConf09_grw.ppt.

Woollacott, E. 2009. Bovine genome project could trigger farming revolution. Article published on TG Daily at www.tgdaily.com/general-sciences-features/42153-bovine-genome-project-could-trigger-farming-revolution