Genomic tools for characterizing monogenic and polygenic traits in ruminants - using the bovine as an example

JF Taylor, RH Chapple, JE Decker, SJ Gregg, JW Kim, SD McKay, HR Ramey, MM Rolf, TM Taxis and RD Schnabel

S135B Animal Sciences Research Center, University of Missouri, Columbia, USA 65211

Next generation sequencing platforms have democratized genome sequencing. Large genome centers are no longer required to produce genome sequences costing millions. A few lanes of paired-end sequence on an Illumina Genome Analyzer, costing <$10,000, will produce more sequence than generated only a few years ago to produce the human and cow assemblies. The *de novo* assembly of large numbers of short reads into a high-quality whole-genome sequence is now technically feasible and will allow the whole genome sequencing and assembly of a broad spectrum of ruminant species. Next-generation sequencing instruments are also proving very useful for transcriptome or resequencing projects in which the entire RNA population produced by a tissue, or the entire genomes of individual animals are sequenced, and the produced reads are aligned to a reference assembly. We have used this strategy to examine gene expression differences in tissues from cattle differing in feed efficiency, to perform genome-wide single nucleotide polymorphism discovery for the construction of ultrahigh-density genotyping assays, and in combination with genome-wide association analysis, for the identification of mutations responsible for Mendelian diseases. The new 800K SNP bovine genotyping assays possess the resolution to map trait associations to the locations of individual genes and the 45 million polymorphisms identified in >180X genome sequence coverage on over 200 animals can be queried to identify the polymorphisms present within positional candidate genes. These new tools should rapidly allow the identification of genes and mutations underlying variation in cattle production and reproductive traits.

Introduction

The higher ruminants are believed to have rapidly speciated in the Mid-Eocene, resulting in five distinct extant families: Antilocapridae, Giraffidae, Moschidae, Cervidae, and Bovidae. While there are about 200 species of ruminants (Hackmann & Spain 2010), few have been domesticated (cattle, goats, sheep, bison, yaks, buffalo, deer, etc) and the phylogeny describing the evolutionary relationships among species has yet to be fully resolved (Decker *et al.* 2009). Similarly, the history of domestication and breed formation within a species such as cattle is

Corresponding author E-mail: taylorjerr@missouri.edu

poorly understood. In fact, the number and location of prehistoric domestication events for the extinct aurochs (*Bos primigenius*) has been controversial, and the ancestry of many of the modern breeds of cattle is unknown (Decker *et al.* 2009). Knowledge of these relationships is of more than academic and conservation interest since the evolutionary history of breeds and species will shed insight into the genes and pathways which control phenotypes which are either shared or diverged among species or breeds. For example, Angus and Shorthorn cattle both accumulate large amounts of intramuscular fat (marbling) and their close genetic relationship as breeds suggests that the genetic mechanisms underlying marbling (loci at which naturally occurring variants induce phenotypic effects and allele frequencies at these loci) are quite similar. On the other hand, Angus and Wagyu cattle are both black-hided and marble highly but are distantly related (Decker *et al.* 2009). Since the cattle phylogeny contains few black-hided breeds and the propensity to marble varies dramatically among the more recently formed breeds, there are two possible explanations for these phenomena. In the case of coat color, either independent mutations have occurred at a single gene (e.g., Melanocortin 1 receptor, *MC1R*) leading to black coat color (Klungland *et al.* 1995), or mutations in two independent genes within the same (melanocortin) pathway have occurred (Candille *et al.* 2007). On the other hand, marbling is inherited as a quantitative trait and for the breeds to have similar high means, either the same genes control variation in both breeds (with allele frequency distributions that generate similar means), or the breeds share a core of identical genes and also breed specific genes for which variation contributes to within breed variation, and the similarity of breed means. These hypotheses are testable by a number of different approaches. However, the point to be made here is that the superimposition of phenotypes on phylogenies (species or breed trees) allows the identification of experimental models which can contribute to the eludication of the genes and pathways responsible for variation within populations and the evolution of population specific phenotypes.

Genetic analysis within a species

Monogenic phenotypes

Hypotheses concerning the mode of inheritance and allelism (whether or not a mutation within the same gene is responsible for a phenotype) can be determined by pedigree analysis and by making experimental crosses. For example, we characterized a fatal movement disorder of Chinese Crested dogs that was clinically and pathologically indistinguishable from a condition known as canine multiple system degeneration that had previously been recognized in Kerry Blue Terriers (O'Brien *et al.* 2005). We showed that the disease segregated as an autosomal recessive and by making crosses between Chinese Crested and Kerry Blue Terrier carriers of the disease allele, we were able to show that mutations within the same gene were responsible for the disease in both breeds. Since both diseases are recessive and are likely due to the loss of gene function, this result also strongly suggests that the disease is identical in both breeds and not simply that the disorder was clinically and pathologically indistinguishable between the breeds. Because microsatellite loci (primarily dinucleotide repeat polymorphisms which could be amplified by PCR and scored on a fragment size analyzer) were plentifully available in dog, we were able to perform a linkage analysis within an extended family and localize the disease locus to a 15 Mb region of dog chromosome 1. This region was subsequently fine-mapped using additional markers and more distantly related affected dogs to reduce the size of the interval harboring the disease locus to less than 1 Mb. Because the sequence assembly for the dog genome (a Boxer) had just become available, we were able to identify and then

sequence exons from candidate genes (e.g., *PARK2*) within this region; however, we are yet to identify the causal mutation underlying this disease. The reasons for this are that: 1) The region contains 27 genes and a number of these could be viable candidates, 2) Although we targeted exons for sequence analysis, we really have no idea what kind of mutation we are looking for – it could certainly be noncoding or a duplication of a coding region – neither of which would be detected by our exon sequencing strategy, and 3) the region is large and until very recently there have been no straightforward or cost-effective methods for sequencing large targeted regions of DNA from individuals.

When individuals belong to extended nuclear families, linkage analysis tests the concordance between phenotype and genotype to establish linkage of the disease locus to a specific chromosomal region. For example, if the disease is known to be inherited as a fully penetrant autosomal recessive we would expect all affected progeny within a family to have inherited the same combination of chromosomal fragments from each of their parents in the region of the disease locus and all unaffected progeny to have inherited different chromosomal combinations. Because only a single meiosis separates parents and progeny there is only a very limited opportunity for recombination to rearrange the parental chromosomal haplotypes (combination of alleles present on a single chromosome or region of a chromosome) present in the progeny and the size of the chromosomal region harboring the disease locus is determined by the number of affected progeny included in the mapping population. Additionally, the resolution of detection of recombination breakpoints is determined by the resolution of the marker map used to perform the linkage analysis. Until very recently, microsatellites were used for almost all linkage analyses and these markers could only be multiplexed in groups of 6-8, meaning that 40-50 separate PCR reactions had to be performed on each individual's DNA to achieve a genome-wide marker map density of one marker per 10 Mb. Not only is microsatellite genotyping slow and tedious, but it is expensive at a cost of $0.25-$1 per produced genotype.

The evolutionary history of the majority of domesticated livestock species differs dramatically from that of human. Modern humans arose from a small effective population size some 100,000 years ago and the population has recently been rapidly expanding, leading to a relatively large current effective population size ($N_e \approx 7{,}000$, Tenesa *et al.* 2007). On the other hand, the domestication of all livestock species occurred within the last 10,000 years (Loftus *et al.* 1994, Giuffra *et al.* 2000), breed societies were formed only within the last 200 years, and the population bottlenecks created by these events have led to relatively small current effective population sizes in cattle ($N_e \approx 100$-200, Bovine HapMap Consortium 2009). As a consequence, many monogenic traits, such as disease are caused by novel rare mutations within human populations, but within a domesticated ruminant species, the majority are caused by a single founder event and all copies of the disease allele within a population are identical by descent to the allele present in the founder individual. Since the original mutation occurred on a single chromosome (with a specific haplotype surrounding the disease allele) which has subsequently been rearranged by recombination within descendants, we expect that all individuals with a fully penetrant autosomal recessive disease are actually homozygous for a small core haplotype that is identical to that present in the founder individual. The size of this haplotype can be quite small if the mutation is old and many generations of recombination have occurred since the occurrence of the original mutation and microsatellite mapping usually does not possess the resolution to detect these core haplotypes.

Recently, high density single nucleotide polymorphism (SNP) genotyping assays have been developed for several ruminant species including cattle (Matukumalli *et al.* 2009) and sheep (Magee *et al.* 2010). While these first generation assays allow the detection of genotypes at ~50,000 loci, second generation assays allowing the detection of ~777,000 SNPs now

available for cattle (http://www.illumina.com/documents/%5Cproducts%5Cdatasheets%5Cdatasheet_ bovineHD.pdf). These assays allow a completely different approach to the mapping of genes influencing monogenic and polygenic loci within species based upon the linkage disequilibrium (LD) between loci that are physically close together on a chromosome. LD is usually measured as the squared correlation coefficient between alleles that are present at two different loci, and the extent of this correlation is influenced by the age of the two mutations and the evolutionary history of the population. In populations such as human, which have large effective population sizes, the correlation between two loci that are old and have moderately high allele frequencies is quite low, whereas in domesticated ruminant populations, these correlations are much larger (Bovine HapMap Consortium 2009). Thus, we would expect that the core haplotypes harboring a disease mutation in, e.g., cattle populations would be quite large and that the resolution of the BovineSNP50 assay (Matukumalli *et al.* 2009) would be sufficient to detect regions of homozygosity in affected animals that might be only quite distantly related. This is, in fact, exactly the case, and the assay allows the rapid localization of recessive disease mutations by case control genome-wide association (GWA) analysis which seeks to identify genomic regions in which all affected individuals are homozygous for a core haplotype, and unaffected individuals are not homozygous for this haplotype (Charlier *et al.* 2008, Meyers *et al.* 2010).

The approach can be modified in a rather interesting way to detect regions of the genome that have undergone recent selective sweeps within a population. Strong selection for a phenotype determined by genotype at a single locus will result in the rapid fixation of a single allele at the causal locus. However, because only relatively few generations of selection (from an evolutionary perspective) are required to drive the selected allele to fixation, we expect LD to drag relatively large flanking chromosomal regions to fixation in the process. Thus, the signature of a selective sweep is the loss of variability within the genome in a region flanking a strongly selected allele in all individuals within a population (Nielsen *et al.* 2005). As time passes following the selective sweep new mutations or migration into the population will allow the accumulation of variation in the region, however, variation will remain reduced within the selected region for a very large number of generations. For example, the locus responsible for the presence or absence of horns in cattle (the polled locus) was first mapped to the centromeric end of chromosome 1 (BTA1) by Georges *et al.* (1993) and was subsequently fine mapped to a 1 Mb region of this chromosome (Drögemüller *et al.* 2005, Wunderlich *et al.* 2006). Horns are inherited as a recessive with the horned phenotype being ancestral and the polled allele being derived. Strong selection within a breed to produce only polled cattle should have produced a selective sweep for the polled allele leaving a strong signature of selection detectable as loss of polymorphism within the region of the genome harboring this gene. To test this, we examined the minor allele frequencies (MAF, frequency of the rarer of the two alleles present at each SNP) of 54,442 SNPs scored in 3,668 registered Angus bulls to identify genomic regions in which losses of diversity suggested the presence of strong selective sweeps. Strong evidence of a selective sweep was found between 1.71-2.01 Mb (UMD3.0 assembly) on BTA1 where 11 consecutive SNP spanning 301 kb had MAF <0.005 (Figure 1). While 7,964 of the 54,442 tested SNP (14.6%) had MAF < 0.005, the probability that 11 consecutive SNP would have small MAF is vanishingly small (despite the fact that they are not inherited independently) and this suggests that a strong recent selective sweep was focused on this region of the Angus genome. Not coincidentally, this region is located within the 1 Mb region identified as harboring the polled locus by Drögemüller *et al.* (2005).

The largest region of the Angus genome (555 kb) in which 10 consecutive SNPs possessed MAF <0.008 was on BTA12 from 25.88-26.43 Mb and contains 8 annotated genes which clearly warrant further investigation as to their involvement in the determination of various

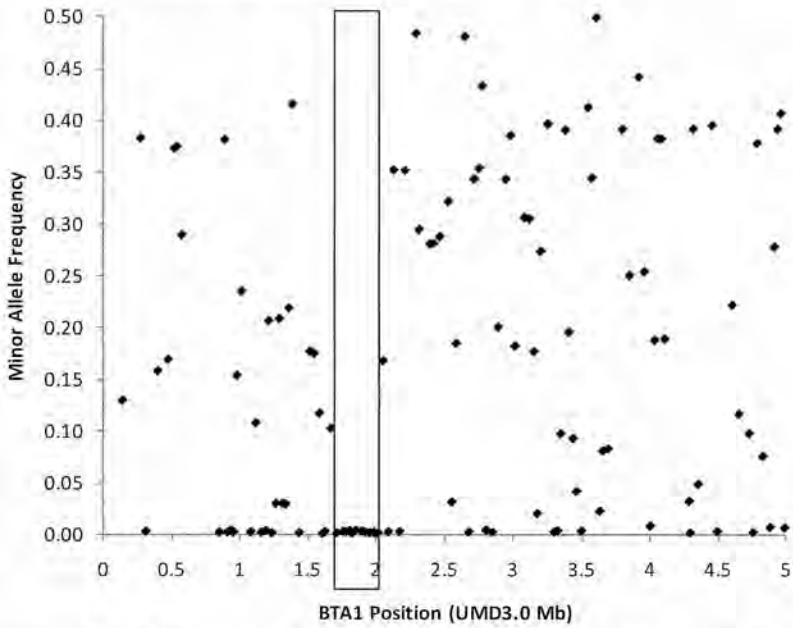


Fig. 1. Minor allele frequency for 116 BovineSNP50 SNP loci mapped to the first 5 Mb of bovine chromosome 1 in the UMD3.0 sequence assembly scored in 3,668 registered Angus bulls. The plot reveals a region of 11 consecutive SNPs spanning 301 kb with minor allele frequency less than 0.005 within the 1 Mb region previously shown to harbor the horn/poll locus (Drögemüller *et al.* 2005).

phenotypes. At the center of this interval is *CNGB1* in which mutations have been found to be responsible for an autosomal recessive form of retinitis pigmentosis in human (Bareil *et al.* 2001). Interestingly, the region harboring the *MC1R* locus on BTA18 which is responsible for black/non-black coat color possessed 5 consecutive SNP with MAF < 0.008 and spanned 320 kb suggesting that selective sweeps of similar intensity occurred for black coat color and polled in Angus cattle. The fact that MAF $\neq 0$ in these intervals is most likely due to genotyping error which is similar in magnitude to these allele frequencies; however, it could also be due to the incomplete elimination of the recessive alleles at these loci, the accumulation of new mutations or the introgression of new haplotypes – all of which are testable hypotheses.

Polygenic phenotypes

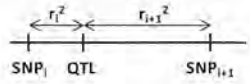
There have been many linkage and LD mapping experiments performed in cattle to identify loci that contribute to variation in quantitative (polygenic) traits such as growth, carcass quality, feed efficiency and fertility using families produced either within breeds (McClure *et al.* 2010), by crosses between breeds (Kim *et al.* 2003) or by the analysis of populations of individuals in the absence of pedigree information (Barendse *et al.* 2007). Linkage analysis to detect quantitative trait loci (QTLs) requires the assembly of families of individuals in which the phenotypes of progeny which inherited differing combinations of parental alleles are statistically contrasted to detect the presence of nearby genes of large effect on the phenotype. As with linkage analysis to detect monogenic trait loci, the resolution of the chromosomal intervals detected to harbor QTL is usually quite poor with confidence intervals often spanning 50% or

more of a chromosome due to limits to the numbers of progeny produced within families (Kim *et al.* 2002). However, linkage analysis of QTLs suffers from two additional shortfalls (Sellner *et al.* 2007). First, the only loci that can be detected are those for which the parents happen to be heterozygous, which means that many QTL will go undetected within any one experiment simply due to lack of parental heterozygosity. Second, the magnitude of QTLs that can be detected (statistical power of the experiment) is also limited by family size. The majority of early linkage analyses which employed microsatellite genotyping of half-sib, back cross or F_2 families used only a few hundred progeny and typically detected no more than 3-5 QTLs per analyzed trait. The largest such analysis was performed by McClure *et al.* (2010) who scored 402 marker loci (predominantly microsatellites) in 38 Angus half-sib families comprising 1,622 steers and an extended pedigree of 1,769 Angus sires and detected an average of 48.1 QTL per analyzed trait. This result is in remarkable agreement with the estimate of 50-100 genes predicted to underlie variation in quantitative traits in dairy populations assuming that the polymorphisms in these genes were neutral with respect to fitness (Hayes & Goddard 2001).

GWA analysis utilizes SNPs evenly spaced throughout the genome to detect the presence of nearby QTLs. In its simplest form, the analysis is performed one SNP at a time by performing an F-test to establish if the mean phenotype differs among individuals with different SNP genotypes. Because the presence of LD requires that alleles at the QTL and SNP locus be correlated, the distribution of QTL genotypes present within each of the SNP genotype classes differs (Figure 2). Consequently, even though the flanking SNP locus itself generally has no effect on phenotype, a test to determine whether the phenotypic mean differs among individuals with different SNP genotypes will be significant if there is a large effect QTL nearby that is in strong LD with the tested SNP. This turns out to be a rather important assumption because manipulation of the formulae in Figure 2C shows that for two loci to be in very strong LD it is necessary (but not sufficient) that they have very similar allele frequencies. However, the vast majority of genotyping assays are designed to include only SNPs that have high MAF in the populations in which the assay is intended to be used. Therefore, by definition, these assays are designed to detect only those common variants that underlie phenotype within any genotyped population. This, at least in part, explains the missing heritability in human GWA studies (Maher 2008) where common SNP variants are used to detect rare causal variants (and they don't!). Similarly, the extent of LD estimated within the genomes of species using these assays (e.g., BovineHap Project 2009) is misleading because what is actually being estimated is the linkage disequilibrium between common variants separated by specific physical distances. While the true distributions of QTL effects that underlie quantitative traits in livestock are unknown, in all likelihood they are biased towards common variants. Hayes & Goddard (2001) empirically estimated that 17 and 35% of the largest effect QTL explained 90% of the genetic variance in dairy and swine, respectively. Thus, 50K common SNPs appear to be sufficient to perform GWA studies within breeds of ruminants for the purpose of identifying the genes of large effect which explain the majority of genetic variation within quantitative traits.

We used the BovineSNP50 assay to genotype 3,240 animals from 5 beef breeds (Table 1) with Warner-Bratzler shear force (WBSF) measures of beef tenderness. We also genotyped 7 SNPs in a 150 kb region spanning calpastatin (*CAST*) on BTA7 and 43 SNPs in a 208 kb region spanning calpain (*CAPN1*) on BTA29. Both genes have previously been shown to be associated with WBSF in several beef breeds. Because pedigree relationships within populations can result in the stratification of samples into families that result in spurious associations in GWA (MacLeod *et al.* 2010), we performed a more sophisticated analysis in which a genomic relationship matrix was estimated for each breed group using the animal's genotypes and all SNP effects were simultaneously estimated by best linear unbiased prediction (VanRaden 2008). Figure 3 shows the standardized estimated SNP allele substitution effects in the 4 breeds with

A)



B)

QTL genotype	QQ	Qq	qq
Genotypic Mean	a	d	-a

C)

Allele frequency		P_i	$1-P_i$	
	Alleles	A_i	B_i	
P_Q	Q	$P_Q P_i + D$	$P_Q(1-P_i) - D$	$D = Cov(SNP_i, QTL)$
$1-P_Q$	q	$(1-P_Q)P_i - D$	$(1-P_Q)(1-P_i) + D$	$r_i^2 = D^2/[P_Q(1-P_Q)P_i(1-P_i)]$

D)

SNP Genotype		A_iA_i		...		B_iB_i	
QTL Genotype	QQ	Qq	qq	...	QQ	Qq	qq
Frequency	$(P_Q P_i + D)^2$	$2(P_Q P_i + D)[(1-P_Q)P_i - D]$	$[(1-P_Q)P_i - D]^2$	...	$[P_Q(1-P_i) - D]^2$	$2[P_Q(1-P_i) - D] \times [(1-P_Q)(1-P_i) + D]$	$[(1-P_Q)(1-P_i) + D]^2$

Fig. 2. Affect of LD between loci on multilocus genotype frequencies. A) Representation of a chromosomal architecture with two SNPs flanking a QTL with LD r_i^2 and r_{i+1}^2 between each SNP and the QTL, respectively, B) QTL is defined such that genotypes have different mean phenotypes (genotypic values), C) Manifestation of LD is the overrepresentation of two haplotype classes and underrepresentation of the remaining two haplotype classes by an amount D (the covariance between alleles at the two loci) relative to the expectation under independence of alleles at the two loci, and D) Effect of LD on QTL genotype frequencies within each of the SNP genotype classes.

Table 1. Estimates of genetic parameters for Warner Bratzler Shear Force in 5 beef breeds. Each breed was separately analyzed using an animal model incorporating a gender × herd-of-origin × slaughter contemporary group and using a genomic relationship matrix estimated from 40,645 SNPs.

		Warner-Bratzler Shear Force (kg)		
Breed	N	σ_A^2	σ_E^2	h^2
Angus	651	0.2184	0.2036	0.52
Charolais	695	0.2275	0.2664	0.46
Hereford	1,095	0.1500	0.7325	0.17
Limousin	283	0.0723	0.7227	0.09
Simmental	516	0.0580	0.6917	0.08
Total	3,240			

largest sample sizes. Remarkably, these plots show little concordance between the regions harboring WBSF quantitative trait loci (QTL) across breeds with the exception of the effect of *CAPN1* which is large in all breeds. Similar results have been seen when performing GWA for milk traits in Jersey and Holstein populations (Dorian Garrick, *pers. comm.*) and suggest either that different QTL are responsible for trait variation in different breeds, or that the resolution of

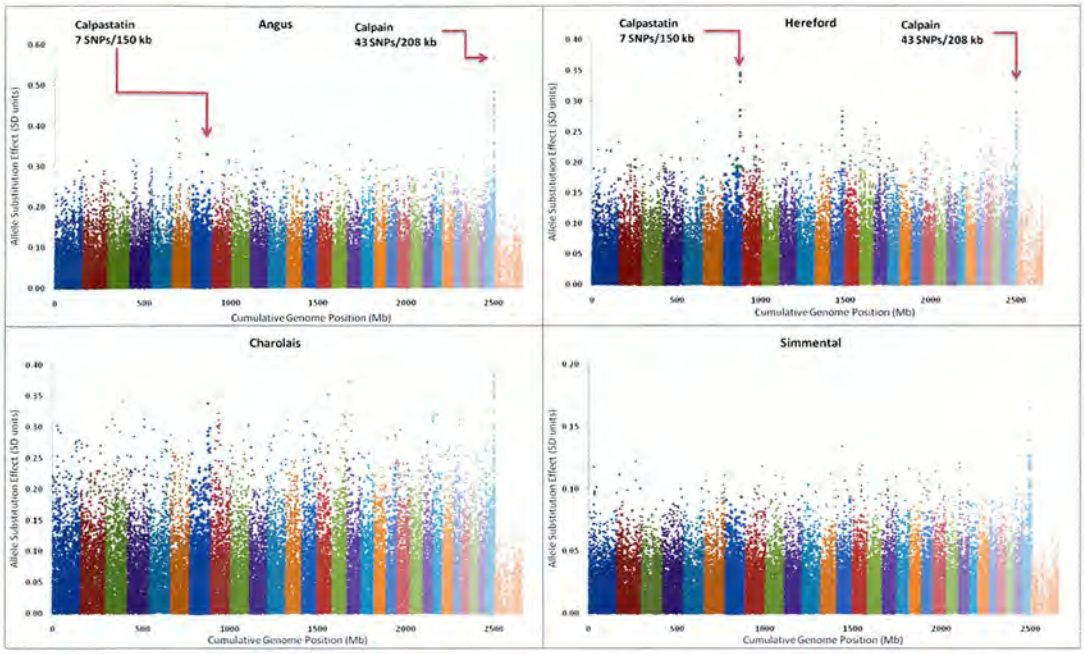


Fig. 3. Manhattan plots of SNP allele substitution effects for WBSF for 40,645 SNPs in 4 breeds. Genomic locations of CAST (BTA7) and CAPN1 (BTA29) are indicated by red arrows.

the BovineSNP50 assay is not sufficient to detect the QTLs which are for the most part identical across breeds, but differ in LD with nearby SNPs due to breed specific differences in MAF. We believe the latter explanation to be the most likely, since the average genome-wide SNP spacing in this study was 65 kb except for the region on BTA29 harboring *CAPN1* which had an average SNP spacing of 4.8 kb and produced one strongly associated SNP in all 5 breeds with the same allele being desirable across all breeds. Nevertheless, this is a very important issue which must be resolved before significant further efforts are made to identify the genes that underlie quantitative traits in domesticated ruminant species.

Gene hunting

The process of identifying the genes and mutations that underlie QTL has been hampered by the lack of genome sequences from which to identify suitable candidate genes, the large size of the QTL regions identified by linkage analysis and the inability to sequence large regions of genomic DNA to hunt for polymorphisms which may be responsible for trait variation (Sellner *et al.* 2007). Following 20 years of QTL mapping in cattle, very few genes and mutations underlying QTL have been identified – *DGAT1* and *ABCG2* with effects on milk traits in dairy cattle and perhaps *CAST* and *CAPN1* with effects on beef tenderness, although the causal mutations within these genes do not appear to have been identified. This situation appears to be about to quickly change. Genome sequences assembled from long-read Sanger sequencing have been produced for the cow (Bovine Genome Sequencing and Analysis Consortium 2009; Zimin *et al.* 2009) and genome projects are underway in buffalo, sheep and deer. High-density genotyping assays have been developed for sheep and cattle which allow

large numbers of samples to be rapidly genotyped for large numbers of SNPs. These assays allow the localization of QTL by LD analysis which generally results in far smaller QTL regions than are produced in linkage analyses. Finally, next-generation sequencing technologies are revolutionizing many aspects of biological and genomic research, and in particular, make it possible to very simply sequence large chromosomal intervals to seek polymorphisms which may underlie monogenic or polygenetic traits. By simply resequencing the entire genome of a disease-affected individual and focusing only on the sequences that align to the region harboring the disease locus, all mutations present within the region (relative to the reference sequence) can be identified. While this approach may sound wasteful, it is a very effective way to rapidly identify candidate mutations and allows the examination of the candidate region for duplications and deletions. To identify candidate mutations underlying an autosomal recessive neurological disease in dogs, we first mapped the disease locus by GWA to a small region of canine chromosome 4 and then produced a 9X average depth sequence coverage of the entire genome of an affected dog using just four lanes of a single Illumina Genome Analyzer IIx (GA IIx) flow cell at a total cost of under $10,000. The dog was homozygous by descent for the disease causing region of chromosome 4 and Figure 4 shows a view of the sequence pile up for the *LOC489223* gene from this dog when aligned by NextGENe (http://www.softgenetics.com/NextGENe.html) to the CanFam2 Boxer dog assembly. This figure shows two mutations leading to amino acid substitutions within this gene which become candidates for the disease-causing mutation. The final step to this analysis is to genotype mutations detected within the candidate region to establish (by concordance with disease phenotype) which of the detected polymorphisms is causal. If the candidate region is large, there may be many hundreds or even thousands of detected polymorphisms and currently, there is no inexpensive genotyping platform which allows simultaneously assaying this number of polymorphisms in a few hundred individuals to establish the identity of the causal polymorphism. This appears to be the single remaining limitation to the detection of the genes and polymorphisms which underlie disease and quantitative trait variation.

Genomic selection

Fortunately, for the purpose of implementing marker-assisted selection of livestock for almost any trait (including fertility) it is not necessary to identify the genes which underlie genetic variation in the trait. Probably the most important breakthrough in genetic improvement in the last 25 years has been the recent demonstration that Genomic Selection (GS) first proposed by Meuwissen *et al.* (2001) can be effectively implemented within breeds of cattle using the BovineSNP50 assay. GS is a methodology to predict animals' breeding values from high-density SNP panels which utilizes a two-stage approach in which animals with phenotypes and genotypes are first used in a training analysis to establish relationships between individual SNPs and trait variation (the normalized values of these SNP effects are shown in Figure 3) and then the inferred breeding value prediction equations are validated in independent populations. In subsequent generations, the breeding values of animals may be estimated at birth from their BovineSNP50 genotypes and the prediction equations. This technology has revolutionized dairy cattle breeding worldwide and genetic progress in milk production is expected to double due to the decrease in generation interval that has been achieved by a reduced need to progeny test young bulls and the high accuracies of the molecular estimates of breeding value (Hayes *et al.* 2009, VanRaden *et al.* 2009).

The technology is also being deployed within the US beef industry, however, the much lower use of artificial insemination (AI) within the industry and the broader composition of

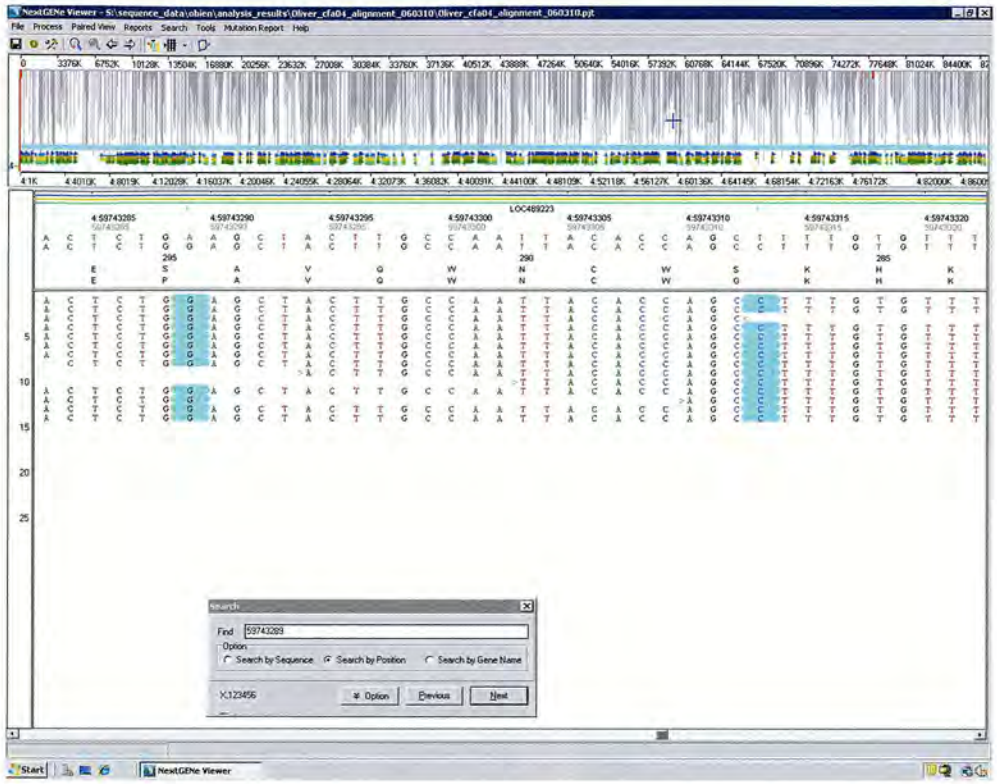


Fig. 4. Screenshot of a NextGENe browser view showing a 14X depth of coverage of sequence produced from a dog homozygous for genomic region on canine chromosome 4 harboring an allele causal for an autosomal recessive neurological disease reveals two charged amino acid substitutions in *LOC489223*. The dog was whole genome sequenced to an average depth of 9X and the only sequence used for mutation discovery was that which aligned to the region of the Boxer reference sequence which was established by LD analysis as harboring the disease causing allele.

breeds employed for U.S. beef production have slowed adoption relative to the dairy industry. First, it has been very difficult to assemble the large training populations needed to develop models with high predictive power. In Angus, the numerically most important beef breed in the US, we have now genotyped only 3,668 registered bulls – far less than the 8,000 Holsteins genotyped when GS was deployed within the dairy industry. Second, the majority of commercial beef animals are bred by natural service and the cost of genotyping young bulls must be amortized over a much smaller number of progeny that are produced by AI sires. It has been vital in the beef industry to reduce the cost of DNA testing to enable GS to be deployed broadly within the industry. When we first developed molecular breeding value prediction equations using genetic evaluations provided by the American Angus Association (AAA) and 41,028 BovineSNP50 SNPs genotyped in 1,710 registered Angus bulls, we developed a reduced set of 384 SNPs which were predictive of breeding values for marbling, ribeye muscle area, backfat thickness and yearling weight to reduce the cost of DNA testing. Validation was performed by genotyping an independent set of 285 bulls and estimating genetic correlations between molecular estimates of breeding value and estimates produced in a multivariate mixed linear model analysis using progeny data (MacNeil *et al.* 2010). Despite the small numbers of

tested SNPs, these correlations were 0.65, 0.58, 0.50 and 0.54, respectively. Based upon these results, the AAA now delivers "combined" estimates of genetic merit based upon all sources of available data (molecular, pedigree and phenotype) to the US beef industry. GS has swept through the US beef and dairy industries within less than three years of the development of the BovineSNP50 assay and the technology has been seamlessly adopted by both industries. However, the Angus national genetic evaluation system is now run weekly rather than bi-annually to allow rapid delivery of DNA test results to producers. This rate of technology adoption is almost unprecedented within animal agriculture and is driven by two factors: 1) the technology has been demonstrated to work in dairy cattle, and 2) it provides information that producers desire and for which they are willing to pay.

Genetic analysis across species

Because we cannot make viable crosses between the majority of ruminant species, we cannot use traditional mapping approaches to identify the genes responsible for the phenotypic differences that have evolved following speciation events. In fact, until recently it has been difficult to even reconstruct the evolutionary history of species such as the ruminants which rapidly radiated. Decker *et al.* (2009) have shown that tools developed to detect variation within a species such as the BovineSNP50 assay can be accurately applied as tools to explore orthologous single nucleotide sequences among closely related species. While SNPs remain variable within a species for no more than 1-2 million years with one allele becoming fixed either due to drift or selection, it appears that recurrent mutations occur at the same loci within different species and that different alleles become fixed within different lineages. Thus, genotyping tools such as the BovineSNP50 assay are capable of detecting the nucleotide that is present at the position in an outgroup species' genome that is orthologous to each SNP within the bovine genome. Remarkably, as genetic distance from cow increases, these nucleotides are not all identical (representing the nucleotide present in the genome of the common ancestor of all advanced ruminants). Of the 40,843 bovine SNP used to study the evolution of 61 higher ruminant species, Decker *et al.* (2009) found that 21,019 were phylogenetically informative among the non-cattle species. This result suggests that there are likely to be a very large number of differences between the genome sequences of ruminants and even if whole genome sequences existed for every ruminant species, sequence-based GWA which attempted to identify mutations concordant with species' phenotypes are likely to reveal large numbers of loci consistent with the phenotype differences among species. Despite this, many causal mutations responsible for the differences among species (along with many false positives) will be among the set of congruent genotypes and may point to gene targets for study within hybridized species (e.g., Bison × Cattle) or for mutation studies within transgenic models.

The more usual form of analysis will be to compare gene content between species to identify othologs shared between all ruminants *versus* those that are lineage specific (including duplicated genes) and to identify genes putatively under selection as manifested by differing rates of synonymous and non synonymous substitutions. This approach was employed to identify innate immunity genes specific to the cattle lineage, some of which appear to be under strong selection within the species (Bovine Genome Sequencing and Analysis Consortium 2009). However, there are a number of problems inherent to this form of analysis. The divergence between species can make it difficult to establish gene orthology (same gene descended from a common ancestor) – although this is not likely to be an issue for the higher ruminants which are diverged by about 29 million years. A more important problem is that the majority of ruminant genes can only be identified by prediction programs or through their similarity to better

studied human genes. Naturally, this means that lineage specific genes are those that are most likely to be missed in this form of analysis. Finally, with only a single representative sequenced within each species, we have no idea about which sites are variable within a species and this makes it difficult to estimate the synonymous *versus* nonsynonymous substitution rates within genes. However, the next few years are likely to result in the generation of enormous amounts of genome and transcriptome sequence within entire clades of species which will change the way that we study biology.

Tools and reagents

SNPs and SNP chips

The identification of SNPs within species for which very little genomic information is available is now relatively straight forward. By the next-generation sequencing of reduced representation libraries produced either by restriction digestion of genomic DNA and fragment size selection (Van Tassell *et al.* 2008) or tissue transcript libraries from individual animals or pools of animals, it is now possible to rapidly identify hundreds of thousands of SNPs and simultaneously estimate MAF. More recently, to generate SNPs for the design of 800K SNP Illumina and Affymetrix assays, we used an Illumina GA IIx to sequence 5 mate-pair and 5 paired-end genomic DNA libraries for each of: 1) a pool of 10 Brahman (*Bos taurus indicus*), 2) a pool of 15 Hanwoo (Korean *Bos taurus taurus*), and 3) 3 individual high-impact Angus bulls (Table 2). Sequence data were trimmed, filtered and aligned to the UMD3.0 sequence assembly for polymorphism discovery using NextGENe and resulted in the discovery of more than 20 million putative SNPs (Table 2). Other public efforts led to the identification of over 45 million polymorphisms in 200 individuals or pools of individuals sequenced to varying depths. The identification of the same SNPs within different individuals or breeds testifies to the validity of these loci and avoids the problem of sequencing errors being identified as SNPs. Such SNPs provide the foundation for the design of high density genotyping assays which are straightforward but very expensive to develop using Illumina Infinium or Affymetrix Axiom chemistry due to the high cost of oligonucleotide synthesis. Thus, it remains to be seen whether these assays will have broad species utilization in the future, or if the cost of genotyping will decrease to the point that genome-wide genotypes will be produced by sequencing.

Table 2. Generation of whole genome sequence data from 3 cattle breeds within the authors' laboratory at the University of Missouri.

Library	GAIIx Lanes	Post-Filter Reads (million)	Total Bases (billion)	Genome Coverage (2.685Gb = 1X)	Average Read Length (bp)	Unfiltered SNPs and Indels (million)
Brahman (N = 10)	13	454.4	32.942	12.27	72.12	19.9[a]
Hanwoo (N = 15)	16	497.5	35.221	13.11	70.97	18.7[a]
Angus (N = 3)	34	1,179.3	97.304	36.23	78.92	14.8[a]
B/R New Design 036	9	310.0	23.565	8.77	75.18	
GDAR SVF Traveler 234D	17	591.0	52.813	19.66	87.50	7.2[b]
N Bar Emulation EXT	9	278.3	20.926	7.79	74.08	
All Libraries	98	2,131.2	165.467	61.61	75.97	

[a]Delivered to Affymetrix and/or Illumina based upon 47X total genome coverage
[b]Based upon the current 19.66X coverage for this animal

Also of importance is the fact that these SNPs and sequences will nearly all be placed within the public domain over the coming 12 months which will produce a public resource of very considerable value. Researchers interested in the diversity within specific genes can query these data to extract the information they need without the need for expensive and costly resequencing projects.

De novo genome sequences and sequence annotation

The current sequence assemblies for agricultural species are all in early iterations and contain significant errors including contigs assembled to the wrong chromosomes, inverted scaffolds and rudimentary annotations. Currently, the animal used to produce the bovine genome sequence assembly (Hereford, L1 Dominette 01449) is being sequenced to a much greater depth on an Illumina GA IIx using mate-pair and paired-end libraries to provide a much greater depth of sequence coverage which will be reassembled along with the existing Sanger reads by the Salzberg group at the University of Maryland. However, the annotation of the assembly requires a great deal of work and transcript libraries produced from a large number of tissues at different stages of development need to be sequenced in RNA-seq experiments to identify the genes and splice variants present within the genome (Mortazavi *et al.* 2008).

It now appears to be feasible to generate *de novo* genome sequence assemblies from short-read sequencing technologies (Ruiqiang *et al.* 2010) and the increasing read-lengths and decreasing costs per Gb of sequence will undoubtedly lead to a rapid increase in the number of *de novo* sequence assemblies produced for ruminants. Furthermore, we will also begin to see many more 100 genome or 1,000 genome projects for individual species which will assist us to identify important functional variants and genomic regions that are under strong selection. Knowledge of the regions within a genome that are variable will also greatly assist in the comparison of genomes between species.

Other applications of next-generation sequencers

Next-generation sequencing instruments are powerful tools for examining genome-wide phenomena. In addition to sequencing DNA and RNA populations, methodologies have been developed which allow the capture (and identification by sequencing) of genomic regions to which proteins such as transcription factors bind (CHIP-seq, Johnson *et al.* 2007), which are methylated (e.g., reduced representation bisulphate sequencing, Meissner *et al.* 2007), or which are physically close together within cells derived from specific tissue types (Hi-C, Lieberman-Aiden *et al.* 2009). These tools will revolutionize our understanding of genome organization and function.

Conclusions

Next-generation sequencing technologies are having enormous impacts in fundamental biology and applied animal agriculture. As competing technologies emerge and the cost of sequencing decreases, we will begin to see the genomes and transcriptomes of closely related- ruminants sequenced, and comparative analyses will point to the genes, structural differences and polymorphisms responsible for the evolution of phenotypic differences between species. Furthermore, many representatives from different breeds and species will be sequenced, which will lead to a much better understanding of the fundamental causes of genetic variation within a species.

While this information will enable genetic improvement within species utilizing existing genetic variation, it will also guide the engineering of transgenic animals with increased adaptation to changing production environments, disease resistance, reproductive and productive capabilities (Fahrenkrug *et al.* 2010). Both approaches to animal improvement will be needed to meet the world's human dietary needs in the very near future.

Acknowledgements

We thank the Circle A Ranch and the MFA, Inc. for providing samples and data, and the American Angus Association for providing pedigree data and expected progeny differences for Angus sires. JFT is supported by National Research Initiative Grant no. 2008-35205-04687 from the USDA Cooperative State Research, Education, and Extension Service, Agriculture and Food Research Initiative grant number 2009-65205-05635 from the USDA National Institute of Food and Agriculture and grant 13321 from the Missouri Life Science Research Board.

References

Bareil C, Hamel CP, Delague V, Arnaud B, Demaille J & Claustres M 2001 Segregation of a mutation in CNGB1 encoding the beta-subunit of the rod cGMP-gated channel in a family with autosomal recessive retinitis pigmentosa. *Human Genetics* **108** 328-334.

Barendse W, Reverter A, Bunch RJ, Harrison BE, Barris W & Thomas MB 2007 A validated whole-genome association study of efficient food conversion in cattle. *Genetics* **176** 1893-1905.

Bovine Genome Sequencing and Analysis Consortium 2009 The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324** 522-528.

Bovine HapMap Consortium 2009 Genome Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* **324** 528-532.

Candille SI, Kaelin CB, Cattanach BM, Yu B, Thompson DA, Nix MA, Kerns JA, Schmutz SM, Millhauser GL & Barsh GS 2007 A -defensin mutation causes black coat color in domestic dogs. *Science* **318** 1418-1423.

Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, Frennet JC, Hanset R, Hubin X, Jorgensen C, Karim L, Kent M, Harvey K, Pearce BR, Simon P, Tama N, Nie H, Vandeputte S, Lien S, Longeri M, Fredholm M, Harvey RJ & Georges M 2008 Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40** 449-454.

Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, Chen K, Cooper A, Vilkki J, Seabury CM, Caetano AR, Johnson GS, Brenneman RA, Hanotte O, Eggert LS, Wiener P, Kim JJ, Kim KS, Sonstegard TS, Van Tassell CP, Neibergs HL, McEwan JC, Brauning R, Coutinho LL, Babar ME, Wilson GA, McClure MC, Rolf MM, Kim J, Schnabel RD & Taylor JF 2009 Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Science U S A.* **106** 18644-18649.

Drögemüller C, Wöhlke A, Mömke S & Distl O 2005 Fine mapping of the polled locus to a 1-Mb region on bovine chromosome 1q12. *Mammalian Genome* 16 613-620.

Fahrenkrug SC, Carlson DF, Doran T, Van Eenennaam A, Galli C, Hackett PB, Li N, Maga EA, Murray JD, Taylor JF, Wheeler M, Whitelaw B & Glenn B 2010 Precision genetics for complex objectives in animal agriculture. *Journal of Animal Science* doi:10.2527/jas.2010-2847.

Georges M, Drinkwater R, King T, Mishra A, Moore SS, Nielsen D, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE & Hetzel J 1993 Microsatellite mapping of a gene affecting horn development in Bos taurus. *Nature Genetics* **4** 206-210.

Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT & Andersson L 2000 The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154** 1785-1791.

Hackmann TJ & Spain JN 2010 Invited review: Ruminant ecology and evolution: Perspectives useful to ruminant livestock research and production. *Journal of Dairy Science* **93** 1320-1334.

Hayes B & Goddard ME 2001 The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33** 209-229.

Hayes BJ, Bowman PJ, Chamberlain AJ & Goddard ME 2009 Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92** 433-443.

Johnson DS, Mortazavi A, Myers RM & Wold B 2007 Genome-Wide Mapping of *in vivo* protein-DNA interactions. *Science* **316** 1497-1502.

Kim J-J, Davis SK & Taylor JF 2002 Application of non-parametric bootstrap methods to estimate confidence intervals for QTL location in a beef cattle QTL experimental population. *Genetics Research* **79** 259-263.

Kim J-J, Farnir F, Savell J & Taylor JF 2003. Detection of QTL for growth and beef carcass fatness traits in a cross

between *Bos taurus* (Angus) and *Bos indicus* (Brahman) cattle. *Journal of Animal Science* **81** 1933-1942.

Klungland H, Våge DI, Gomez-Raya L, Adalsteinsson S & Lien S 1995 The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. *Mammalian Genome* **6** 636-639.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES & Dekker J 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326** 289-293.

Loftus RT, MacHugh DE, Bradley DG, Sharp PM & Cunningham P 1994 Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Science U S A.* **91** 2757–2761.

MacLeod IM, Hayes BJ, Savin KW, Chamberlain AJ, McPartlan HC & Goddard ME 2010 Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. *Journal of Animal Breeding and Genetics* **127** 133-142.

MacNeil MD, Northcutt SL, Schnabel RD, Garrick DJ & Taylor JF 2010 Genetic correlations between carcass traits and molecular breeding values in Angus cattle. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*. August 1-6, Leipzig, Germany.

Magee DA, Park SDE, Scraggs E, Murphy AM, Doherty ML, Kijas JW, International Sheep Genomics Consortium & MacHugh DE 2010 Technical note: High fidelity of whole-genome amplified sheep (Ovis aries) DNA using a high-density single nucleotide polymorphism array-based genotyping platform. *Journal of Animal Science* published online Jun 18, 2010.

Maher B 2008 Personal genomes: The case of the missing heritability. *Nature* **456** 18-21.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS & Van Tassell CP 2009 Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4** e5350.

McClure MC, Morsci NS, Schnabel RD, Kim JW, Yao P, Rolf MM, McKay SD, Gregg SJ, Chapple RH, Northcutt SL & Taylor JF 2010 A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Animal Genetics* May 10. [Epub ahead of print].

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R & Lander ES 2007 Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454** 766-771.

Meuwissen TH, Hayes BJ & Goddard ME 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157** 1819-1829.

Meyers SN, McDaneld TG, Swist SL, Marron BM, Steffen DJ, O'Toole D, O'Connell JR, Beever JE, Sonstegard TS & Smith TP 2010 A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC Genomics* **11** 337.

Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5** 621-628.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG & Bustamante C 2005 Genomic scans for selective sweeps using SNP data. *Genome Research* **15** 1566-1575.

O'Brien DP, Johnson GS, Schnabel RD, Khan S, Coates JR, Johnson GC & Taylor JF 2005 Genetic mapping of canine multiple system degeneration and ectodermal dysplasia loci. *Journal of Heredity* **96** 727-734.

Ruiqiang L, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T, Yiu S, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J & Wang J 2010 The sequence and de novo assembly of the giant panda genome. *Nature* **463** 311-317.

Sellner EM, Kim JW, McClure MC, Taylor KH, Schnabel RD & Taylor JF 2007 Applications of Genomic Information in Livestock. *Journal of Animal Science* **85** 3148-3158.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME & Visscher PM 2007 Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17** 520–526.

VanRaden PM 2008 Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91** 4414–4423.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF & Schenkel FS 2009 Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92** 16-24.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC & Sonstegard TS 2008 Simultaneous SNP discovery and allele frequency estimation by high throughput sequencing of reduced representation genomic libraries. *Nature Methods* **5** 247-252.

Wunderlich KR, Abbey CA, Clayton DR, Song Y, Schein JE, Georges M, Coppieters W, Adelson DL, Taylor JF, Davis SL & Gill CA 2006 A 2.5-Mb contig constructed from Angus, Longhorn and horned Hereford DNA

spanning the polled interval on bovine chromosome 1. *Animal Genetics* 37 592-594.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA & Salzberg SL 2009 A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology **10** R42.