STRUCTURAL VARIATION AT THE *KIT* LOCUS IS RESPONSIBLE FOR THE PIEBALD PHENOTYPE IN HEREFORD AND SIMMENTAL CATTLE

A Thesis Presented to the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment of the Requirements

for the Degree

Master of Science

---

by

LYNSEY WHITACRE

Dr. Jeremy F. Taylor, Thesis Advisor

JULY 2014

APPROVAL PAGE

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

STRUCTURAL VARIATION AT THE *KIT* LOCUS IS RESPONSIBLE FOR THE PIEBALD PHENOTYPE IN HEREFORD AND SIMMENTAL CATTLE

Presented by Lynsey Whitacre, a candidate for the degree of Master of Science, and hereby certify that in their opinion it is worthy of acceptance.

_____

Dr. Jeremy F. Taylor, Animal Science, UMC

_____

Dr. Robert D. Schnabel, Animal Science, UMC

_____

Dr. Jared E. Decker, Animal Science, UMC

_____

Dr. Gary S. Johnson, Veterinary Pathobiology, UMC

ACKNOWLEDGEMENTS

I would like to acknowledge those individuals responsible for distinguishing my master's degree as a highly experiential learning experience. Thanks to Dr. Jerry Taylor, Dr. Jared Decker, Dr. Bob Schnabel, Dr. JaeWoo Kim, Dr. Gary Johnson, Dr. Kevin Wells, Polyana Tizioto, Maria Haag, Holly Ramey, Jesse Hoff, and the other members of the Taylor Animal Genomics laboratory for providing a collaborative learning experience and supporting my efforts. Furthermore, special thanks to Dr. Juan Medrano (University of California-Davis), Dr. David Schwartz (University of Wisconsin-Madison), and Dr. Shiguo Zhou (University of Wisconsin-Madison) for collaborating with us and sharing bovine optical map data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

REVIEW OF LITERATURE

Introduction to coat color genetics

Variation in the coat color patterns of domesticated cattle has interested breeders for many centuries, as demonstrated by early Lascaux cave drawings that depict cattle with white spotting (Olson, 1999). These patterns were captured within breeds by selective breeding due to their uniqueness and their ability to provide an easy means of breed identification. Early breeders also relied on coat color patterns to evaluate animals due to a lack of scientific knowledge and measurements of economically important traits. Coat color phenotypes, and their associated genotypes, are described relative to 'wild-type' animals, which vary from reddish-brown to brownish-black with a tan muzzle ring. Wild-type animals also tend to be darker at the head and neck, feet, and hindquarters, and do not exhibit the white spotting, or piebald, phenotypes seen in domesticated cattle.

The primary loci responsible for coat color phenotypes in cattle are agouti (*A*), extension (*E*), albino (*C*), brown (*B*), dilution (*D*), spotted (*S*), and roan (*R*). The *A* and *E* loci are responsible for red and black coat colors, respectively. Black coat color is inherited as a dominant trait whereas red is inherited as a recessive trait. However, the *E* locus is epistatic to the *A* locus. Therefore, a dominant black allele at *E* produces a black coat even if the animal is homozygous for the recessive red allele at the *A* locus. The *B* locus determines brown coloring, *D* determines the dilution of the pigment, *R* results in

the intermixing of colored and white (roan), and the *Spotted* locus, the focus of this study, is responsible for several unique white spotting patterns including the Hereford pattern and color-sided pattern. Research thus far indicates that the *A* locus encodes the agouti signaling protein (*ASIP*), *E* encodes melanocortin 1 receptor (*MC1R*), *B* encodes tyrosinase related protein 1 (*TYRP1*), *D* encodes the silver gene (*PMEL*), and *R* encodes the *KIT* ligand gene (*KITLG*), as reviewed by Schmutz (2012). The gene encoded by the *Spotted* locus has yet to be determined.

Economic impacts of coat color

The aesthetic reasons for studying coat color patterns are obvious, but there are also several economic reasons, including breed traceability, adaptability to different environments, and susceptibility to damage from ultraviolet (UV) radiation. Coat color genes are strongly associated with and differentiate the many breeds of cattle throughout the world. This makes them good candidates for use in DNA-based breed traceability and authentication marker panels. For example, polymorphisms at the *E* locus have been used for the authentication of dairy products from French breeds and Parmigiano Reggiano cheese obtained from Reggiana milk (Fontanesi et al., 2010). Although DNA-based breed authentication tests are not currently abundant, there is the potential for breed-specific coat color polymorphisms to be utilized in the identification and authentication of both dairy and beef breed branded products.

Desirable coat color phenotypes typically vary geographically, as some coat colors are better adapted to certain environments. Dairy and beef cattle with lighter coats and darker skin pigmentation are better adapted to tropical climates because the lighter color decreases the absorption of UV rays and, thus, reduces heat stress. Studies have

shown that the percentage of coat color that is white in Holstein cattle has a beneficial impact on milk production and reproductive traits (Seo et al., 2007). Additionally, the lack of pigmentation on the eyelids of cattle has been associated with the increased occurrence and severity of infectious bovine keratoconjunctivitis, more commonly known as pinkeye (Ward & Nielson, 1979). Pinkeye is caused by a bacterial infection of *Moraxella bovis* in the eye and is an important economic factor in cattle production as pinkeye affected calves have reduced weaning weights of 20 to 30 pounds and post-weaning affected animals have lower performance in average daily gain, 365 day weight, and final weight at slaughter (Brown et al., 1998).

The most important economic impact associated with coat color patterns is due to ocular squamous cell carcinoma (OSCC), or cancer eye. OSCC is the most common malignant tumor affecting cattle in North America and often results in economic losses through the reduced longevity of breeding animals as well as carcass condemnation at slaughter as selling meat from cancer-stricken cattle is against U.S. Federal law. Ocular lesions caused by OSCC account for approximately 13% of all carcass condemnations in U.S. slaughterhouses (Russell et al., 1976).

Several studies have concluded that the absence of eyelid pigmentation is a significant predisposing factor to OSCC (Guilbert et al., 1948; Bonsma, 1949; Anderson et al., 1957). Herefords and other piebald cattle, which often retain no pigmentation around their eyes, are particularly predisposed to OSCC. Russell et al. (1976) observed a herd of Hereford cows over a 5-year period and detected that an eye lesion occurred at least once during the 5 years in 51% of the animals. Similarly, the incidence of OSCC in pertinently exposed piebald Simmental cattle is as high as 53% (Den Otter et al., 1995).

The presence of pigment surrounding the eyes has, thus, become an important phenotype to score, and is known as ambilateral circumocular pigmentation (ACOP). The heritability of this trait has been estimated in three herds of Hereford cattle to be between 0.41 and 0.50 (Anderson et al., 1957), which indicates that selection for ACOP should lead to an increase in lid pigment and therefore a decrease in the incidence of OSCC. A recent study in Simmental cattle found 12 quantitative trait locus (QTL) regions associated with ACOP. These 12 regions explained 45% of the phenotypic variation in eyelid pigmentation, and the number of progeny with ACOP increased from 22% to >50% when the sire had more favorable alleles for pigmentation than non-favorable alleles (Pausch et al., 2012). These data suggest that selecting bulls based on QTL alleles for ACOP could rapidly increase the number of progeny with ACOP, thereby contributing to a decrease in prevalence of OSCC. Furthermore, lesions do not usually appear until at least 4 years of age, but ACOP can be evaluated at birth, which allows the early selection of animals to retain for breeding.

Melanocyte development and melanogenesis

The molecular basis of coat color is due to the presence or absence of melanins in the hair follicle. Melanin is a pigment produced in the melanocyte by specialized organelles called melanosomes, which are transferred to the hair and skin via keratinocytes. During embryonic development, melanoblasts, precursor cells of melanocytes, migrate from their origin in the neural crest following a dorsolateral pathway until they reach the skin and colonize the interfollicular space, hair follicle, and dermis and differentiate into melanocytes (Seo et al., 2007; Jimbow et al., 1976). Mature melanocytes are then stimulated by UV radiation to produce melanin to protect the cell's

nucleus, and the DNA that it harbors, from damage. Two types of melanin are found in mammals, eumelanin and pheomelanin. Eumelanin is responsible for brown to black pigments, and pheomelanin is responsible for red to yellow pigments. Both types of melanin are produced via oxidation of the amino acid tyrosine, a reaction that is catalyzed by tyrosinase activity and followed by polymerization. Eumelanin is the more common of the two pigments, and pheomelanin appears to have arisen as a modification to the eumelanin pathway (Jimbow et al., 1976).

There are several genes known to be involved in regulating the development and migration of melanocytes including *MITF*, *KIT*, *KITLG*, and *ERBB3*. *MITF* is thought to be crucial to both the survival and differentiation of melanocytes and regulates *MC1R*, one of the primary proteins involved in melanogenesis (Levy et al., 2006). *KIT* is known to be associated with the early migration of at least three populations of stem cells including primordial germ cells, hematopoietic cells, and melanocytes (Fleischman, 1993). *KITLG* is the ligand for *KIT*. When bound to this specific ligand, *KIT* possesses tyrosine kinase activity. *ERRB3* is also a receptor tyrosine kinase. It is required for embryonic development and is necessary for melanocyte maturation (Buac et al., 2009). Additional genes implicated in coat color phenotypes, such as *PAX3*, a transcription factor known to be involved in melanogenesis, primarily affect the melanin production pathway, and not the development or migration of melanocytes (Kubic et al., 2008).

Piebaldism

Piebaldism is an autosomal dominant genetic trait characterized by patches of skin and hair that lack pigmentation. This condition is known to occur in humans, mice, horses, dogs, snakes, birds, cats, and cattle. It is most commonly associated with white

areas on the forehead, ventral chest/abdomen, and the extremities (Giebel & Spritz, 1991). Piebaldism is not related to albinism. Piebaldism appears to result from a complete lack of melanocytes in the affected skin or hair follicles, whereas albino animals maintain mature melanocytes in their skin and hair, but the pathway to melanin production is blocked.

Piebald patterns are widely recognized among cattle and have been recorded in *Bos taurus*, *Bos indicus*, and mixed breed types as well as in the yak and certain water buffalo (Rife, 1962; Searle, 1968). *Bos taurus* breeds that exhibit piebald phenotypes include Hereford, Simmental, Holstein, and Pinzgauer. The Hereford phenotype is characterized by white on the head, lower legs, ventral areas, and tail switch (Figure 1.1A). Holsteins are recognized by patterns of irregular spotting on the body that can vary from nearly all white to a very limited white spotting (Figure 1.1B). Fullblood Simmentals exhibit a coat color pattern that appears to be a combination of the Hereford and Holstein phenotypes, with white facial patterns and random white spotting on the body (Figure 1.1C). Pinzgauers exhibit the "color-sided" pattern in which white spots are seen on the tail, the ventral and dorsal parts of the body, and occasionally on the head (Figure 1.1D).

The *Spotted* locus and associated phenotypes

The region of the genome responsible for several of these unique piebald coat color phenotypes in cattle has historically been referred to as the *Spotted* locus. There are four hypothesized alleles: S^H, the Hereford or white face allele, S^P, the Pinzgauer or lineback allele, *s*, the recessive spotting allele, and S^+, the non-spotted allele (Grosz & MacNeil, 1999). The S^H allele is responsible for the Hereford pattern when it is present in

the homozygous state. This allele is incompletely dominant to the wild-type S^+ allele, and heterozygotes usually exhibit white only on the face. The *s* allele is recessive and only impacts coat color when the animal is homozygous for *s*. Homozygous *s/s* cattle exhibit irregular white spotting on the body as seen in the Holstein breed.

The *Spotted* locus has been mapped to a 20.8 Mb region on BTA6 by Grosz & MacNeil (1999). This region contains a gene that encodes the mast cell and stem cell growth factor receptor *KIT*. An intuitive candidate gene for the causal mutations underlying the four hypothesized alleles, *KIT* is a tyrosine kinase that, during embryonic development, mediates melanocyte migration from the neural crest along the dorsolateral pathway to a final destination in the skin. *KIT* is known to be associated with several white patterning phenotypes such as that caused by the *W* locus in mice (Geissler et al., 1988), human piebaldism (Giebel & Spritz, 1991), the white dominant and belted phenotype in pigs (Marklund et al., 1998; Rubin et al., 2012), and the Sabino and Tobiano patterns in horses (Brooks & Bailey, 2005; Brooks et al., 2007). Reinsch et al. (1999) also presented evidence for a QTL responsible for the extent of white spotting in Holsteins near *KIT*. Furthermore, a genome-wide association scan in a population of F_2 Holstein-Friesian and Jersey crossbred cows found significant white spotting QTLs on BTA6 in a region near *KIT*, BTA18 and BTA22 (Liu et al., 2009). Hayes et al. (2010) also demonstrated that the proportion of black in Holsteins is partially due to variation in *KIT*, *MITF*, and a locus on BTA8. These loci explained 24% of the total phenotypic variation in the proportion of coat color that was white in this population, with the largest proportion of variation being explained by *KIT*.

Despite strong associations between white spotting phenotypes and *KIT*, a casual mutation has only been identified for one white spotting phenotype in cattle. A serial translocation of 492 kb from a region on BTA6 containing *KIT* to BTA29 is responsible for color-sidedness in Belgian Blue and Brown Swiss cattle (Durkin et al., 2012). The mutations underlying the Hereford, Simmental, and Holstein white spotting phenotypes remain unknown, and the phenotypes are clearly regulated in a complex manner. Recently, in an attempt to pinpoint these mutations, the targeted sequencing of a 19,378 bp region containing the *KIT* gene in Hereford, Holstein, and Angus cattle was carried out. However, this study did not identify any putative functional coding mutations, suggesting that regulatory mutations or structural variation may be responsible for the patterning variation (Fontanesi et al., 2009).

Coat color phenotypes of other common breeds

The breeds putatively fixed for spotted phenotypes included Hereford, Simmental, and Holstein, all from the *Bos taurus* subspecies. It is important to note that there has recently been a movement away from traditionally spotted Simmental animals in North America due to the ability to register high percentage cross-bred animals as Simmental. As a result, the large majority of animals registered by the American Simmental Association are admixed and only a small proportion of fullblood Simmental bulls remain. These admixed individuals frequently exhibit a solid black coat color phenotype due to introgression of the black *MC1R* allele from Angus. Only fullblood Simmental animals were analyzed in this study. The other analyzed breeds included Angus, Jersey, Limousin, Maine-Anjou, N'Dama, Romagnola, and Shorthorn from the *Bos taurus* subspecies and Brahman, Gir, and Nelore from the *Bos indicus* subspecies. Additionally,

Beefmaster, a *Bos taurus/Bos indicus* hybrid and American bison (*Bison bison*) sequences were analyzed.

The spotted coat color patterns of Hereford, Simmental, and Holstein cattle were previously described. The Angus, Jersey, Limousin, Maine-Anjou, N'Dama, Romagnola, and Shorthorn breeds are all non-spotted. Angus cattle exhibit a solid black coat color (Figure 1.2A). Jersey cattle are typically light brown in color with darker regions at the extremities and a light grey muzzle (Figure 1.2B). Fullblood Limousin cattle are generally a solid golden-red color, but registered Limousin may also be black due to admixture with Angus (Figure 1.2C). Historically, Maine-Anjou cattle were red with white on the underside of the belly and on the back legs (Figure 1.2D), but, much like Simmental and Limousin, the majority of modern Maine-Anjou cattle have been admixed, resulting in a black coat (Figure 1.2E). N'Dama cattle exhibit a fawn coat color with darker extremities (Figure 1.2F). Romagnola cattle have black skin with light to dark grey coats (Figure 1.2G), and, finally, Shorthorn cattle have a variable coat color ranging from red to white with the intermediate being roan which is a red coat with white spots or patches (Figure 1.2H). The white spotting occasionally seen in the Shorthorn, Belgian Blue and fullblood Maine-Anjou cattle is genetically independent of the Hereford and Simmental piebald phenotypes and is caused by a mutation within *KITL* (Seitz et al., 1999). The Maine-Anjou white spotting phenotype is only present in homozygous recessive animals, unlike the Hereford white spotting phenotype (Olson, 1999).

While piebald phenotypes have been documented in *Bos indicus* cattle, none of the Brahman, Gir or Nelore breeds typically exhibit spotted coat color patterns. Brahman

cattle range in color from grey to red with varying shades of darkness (Figure 1.3A). Gir cattle range from solid red to shades of red and white to dark grey and white (Figure 1.3B). Nelore cattle exhibit substantially less variation and are primarily light grey although some also exhibit darker grey regions around the neck and legs (Figure 1.3C).

The Beefmaster breed is a composite *Bos taurus/Bos indicus* hybrid that is, on average, 25% Shorthorn, 25% Hereford, and 50% Brahman. This genetic combination results in a highly variable coat color phenotype and ranges from solid black or solid red to black or red with white on the face or underbelly (Figure 1.4). White spotting in the coat of a Beefmaster can have its genetic origin from either the Shorthorn or Hereford breeds. Piebald patterns, recognized by the obligatory white face, presumably result from a Hereford allele at the *Spotted* locus whereas spotting only behind the navel or on the hind legs is presumably of Shorthorn origin. Finally, American bison exhibit a dark brown coat color phenotype (Figure 1.5A). However, when crossed with spotted cattle breeds, bison x cow crosses have been recorded to exhibit a white face or partial piebald phenotype (Figure 1.5B). The American bison and taurine and indicine cattle are distantly genetically related and are from the Bovini tribe, a division of the Bovinae sub-family and Bovidae family.

Figure 1.1. Representative coat color phenotypes of four spotted cattle breeds: A) Hereford bull. B) Holstein cow. C) Simmental bull. D) Pinzgauer bull.

Figure 1.2. Representative coat color phenotypes of *Bos taurus* non-spotted cattle breeds: A) Angus bull. B) Jersey cow. C) Limousin bull. D) Fullblood Maine-Anjou bull. E) Modern, admixed Maine-Anjou bull. F) N'Dama bull. G) Romagnola bull. H) Two Shorthorn animals.

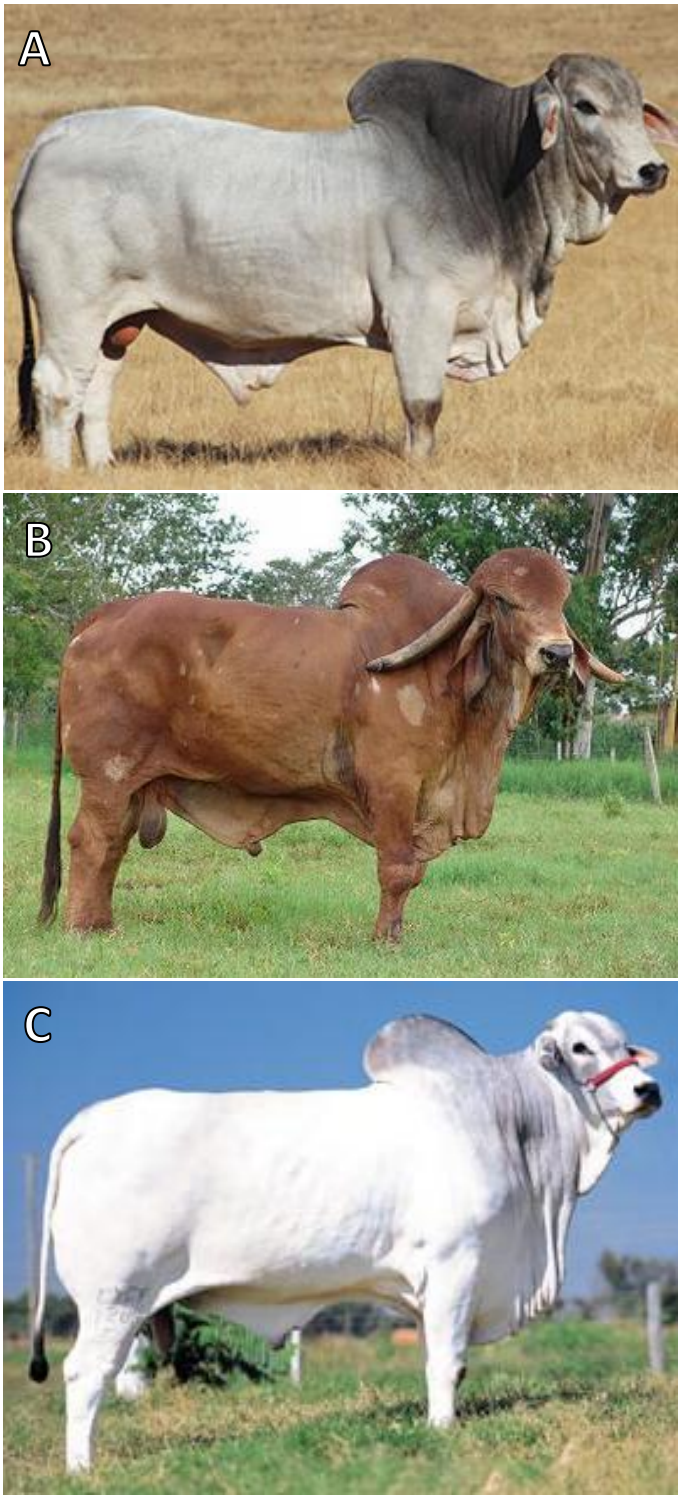

Figure 1.3. Representative coat color phenotypes of *Bos indicus* non-spotted cattle breeds: A) Brahman bull. B) Gir bull. C) Nelore bull.

Figure 1.4. Representative coat color phenotypes of Beefmaster cattle: A) Spotted bull. B) Solid red bull with a white underbelly. C) Solid reddish-yellow bull. D) Solid black bull.

Figure 1.5. Representative coat color phenotypes of: A) American bison. B) Piebald F_1 American bison x Hereford cow.

CHAPTER TWO

STRUCTURAL VARIATION AT THE *KIT* LOCUS IS RESPONSIBLE FOR THE PIEBALD PHENOTYPE IN HEREFORD AND SIMMENTAL CATTLE

Materials and Methods

The research presented here was conducted based upon the postulate that since the Hereford breed is fixed for the white spotting pattern, the phenotype is the result of a historic selective sweep that took place during breed development. The sweep fixed the spotted allele and simultaneously fixed alleles at nearby hitchhiking loci to produce a small region on BTA6 that is devoid of variation in Hereford. In order to identify the causal mutation, we compared DNA sequences between spotted and non-spotted breeds of cattle. The process of identifying and verifying candidate polymorphisms for the S*potted* locus was achieved in four steps: 1) identification of selective sweep regions exclusive to the Hereford breed, 2) local *de novo* genome assembly of the selective sweep regions in Hereford with, and Angus without, the sweeps, 3) detection of variants predicted to be fixed for alternative alleles between Hereford and Angus individuals, and 4) determination and assessment of structural variation in 29 spotted animals and 55 non-spotted animals from 14 cattle breeds and two species.

Genotyping, imputation, and SNP filtering

Genotyping is the molecular process of scoring the specific alleles that are inherited by an individual from its parents at select single nucleotide polymorphisms (SNPs) or other loci. Since 2009, high-throughput genotyping in cattle has been performed by analyzing the individual's DNA using biochemical assays, such as the Illumina BovineSNP50 or BovineHD Genotyping BeadChips. The BovineHD BeadChip contains approximately 777,000 SNPs evenly distributed throughout the genome with a mean gap size of 3.4 kb between SNPs. It is an accurate technology with an average call rate greater than 99% for all SNPs across common bovine breeds. Similarly, the BovineSNP50 assay contains approximately 54,000 SNPs with a mean gap size of 49.4 kb and also obtains average call rates greater than 99% (Matukumalli et al., 2009).

Twenty-four fullblood Hereford animals were genotyped with the Illumina BovineHD assay and an additional 787 fullblood Hereford animals were genotyped with the Illumina BovineSNP50 assay. Analysis of fullblood individuals is vital for the detection of selective sweeps in the region of the *Spotted* locus because of the dominant nature of the spotted phenotype. Introgression of alleles from other breeds may lead to variation at *Spotted* and the inability to detect a selective sweep region despite the fact that all genotyped animals possessed the white face phenotype. All animals originally genotyped with the BovineSNP50 assay had their genotypes imputed to the set of BovineHD genotypes scored in the 24 fullblood animals and an additional 491 high-percentage Hereford animals (on average 89.2% Hereford by pedigree) previously genotyped with the BovineHD assay.

Genotype imputation uses short stretches of common haplotypes, putatively identical by descent, to estimate the alleles at several SNPs that have not been directly genotyped. This allows the analysis of SNPs that have not directly been genotyped and is particularly useful for comparing the genomic architecture of animals genotyped with different assays that contain overlapping loci. BEAGLE v3.3.2 was utilized for imputation in this research due to its computational efficiency and accuracy. BEAGLE focuses only on genotypes for a relatively small number of nearby flanking SNPs when estimating each missing genotype whereas some more sophisticated imputation tools focus on all observed genotypes when estimating each missing genotype, but are far more computationally demanding (Browning & Browning, 2009). SNPs were filtered on call rate ≥85%, which left a total of 5,086 HD SNPs within the critical region for *Spotted* identified by Grosz & MacNeil (1999) between *BM4528* at 64,727,183 bp and *ELO3* at 85,504,512 bp on BTA6 (UMD3.1 assembly).

Identification of selective sweep regions from SNP data

Based on the hypothesis that the *Spotted* locus must have been under intense selection for the Hereford pattern to become fixed within the breed and evidence of a 209 kb selective sweep at 70.65 Mb on BTA6 found in 50K genotyping data by Ramey et al. (2013), 811 fullblood Herefords with imputed BovineHD genotypes were analyzed for sweeps within the 20.8 Mb region on BTA6 where the *Spotted* locus was mapped. Selective sweep regions were defined as regions greater than 75 kb, containing at least 25 consecutive markers all with a minor allele frequency (MAF) less than 1% in the sample. While selective sweeps are traditionally classified as regions of the genome that have lost all genetic variation (i.e., MAF = 0), MAF was allowed to vary between 0 and 0.01 in a

sweep region to account for genotyping errors, the possibility of new mutations, and assembly errors that may erroneously assign a variable marker to a sweep region.

Sequencing, *de novo* assembly, and variant calling

The genomes of five Angus bulls and one Hereford cow were sequenced using the Illumina Genome Analyzer IIx and HiSeq2000 technologies, which generated paired-end or mate-pair reads of approximately 100 bp. Genome assembly is generally classified as either reference-guided assembly or *de novo* assembly, the former of which is dependent on a pre-existing reference assembly for the species of interest or a closely related species. Due to multiple gaps in the UMD3.1 *Bos taurus* reference assembly and evidence that non-Hereford bacterial artificial chromosome (BAC) end sequences may have been incorporated into this particular region of the reference assembly, local *de novo* assemblies were generated for the Hereford selective sweep region. The Hereford *de novo* assembly was constructed using Illumina paired-end reads and mate-pair reads from the *Bos taurus* reference individual while three Angus *de novo* assemblies were generated by pooling Illumina paired-end reads and mate-pair reads from three sire-son pairs. We combined reads from sire-son pairs in the Angus assemblies in order to increase the extent of sequence coverage of the region and improve our ability to detect structural differences between Hereford and Angus.

All four of the *de novo* assemblies were generated by first aligning all available reads to the reference genome and extracting the read pairs for which both the forward and reverse read mapped to the region as well as the read pairs for which only one read mapped to the region and its pair was either unmapped or mapped to a different location in the genome. This resulted in a total number of unique reads per assembly ranging from

672,044 to 925,492 (Table 2.1). The reads were then assembled using the Maryland Super-Read Cabog Assembler (MaSuRCA) v2.1.0. MaSuRCA is a unique *de novo* genome assembly program that incorporates the strengths of both the overlap-layout-consensus and de Bruijn graph approaches into the building of contigs and scaffolds (Zimin et al., 2013).

Once the assemblies had been generated and quality checked using amosvalidate (Phillippy et al., 2008), each of the Angus *de novo* assemblies was individually aligned to the Hereford *de novo* assembly using MUMmer (Kurtz et al., 2004). MUMmer was also used to call variants (SNPs and indels) between the three alignments. These three sets of variants were then compared and variants that were not present in all three alignments were discarded because, putatively, all Angus must differ from Hereford at the causal locus. The consolidated list of candidate variants was then compared to that generated by the alignment of a second Hereford genome sequence to the *de novo* Hereford assembly. At this point, only variants for which all 5 Angus and both Herefords were fixed for different alleles were retained.

Identification of duplications

The detection of novel structural variation from genomic sequences is arduous and often convoluted by the presence of errors in the reference genome assembly. Some detection methods rely on paired-end read mapping, which detects structural variants by comparing the actual distance between mapped read pairs to the estimated distance based on the average library insert size. Translocations and inversions can be detected in a similar manner using mapped mate-pair reads, which have a larger average insert size. Duplications, or copy number variants (CNVs), can be detected using read depth of

coverage (DOC) data that are generated after aligning next generation sequence (NGS) reads. An increase in the detected DOC is indicative of duplications or repetitive elements that are collapsed into one copy by the computational process of assembling the reads, which cannot distinguish the copies from one another due to their nearly identical sequences. This method relies first on estimating the average read depth, or coverage, across the whole genome, which for our data generally ranged from 5X to 40X. Next, regions of the genome that appear at least two times greater than the average genome coverage are identified as duplications and are compared to homologous regions of other genomes to distinguish polymorphic duplications, or CNVs (Yoon et al., 2009; Bickhart et al., 2012).

A total of 84 genomes, 29 from spotted animals and 55 from non-spotted animals, from 14 bovine breeds and two species were available to us in a separate project dataset (Table 2.2). These genome sequences were aligned to the UMD3.1 assembly using the NextGENe software (SoftGenetics, LLC), and a custom perl script was written to extract base-by-base sequence coverage information from the alignment of the reads to a 5 Mb region from 70 to 75 Mb, which contained all of the selective sweep regions and also extended through the entirety of *KIT*. The coat color phenotype of each individual was scored from photographs of each of the sequenced animals, and the average sequence coverage of each individual's genome was calculated (Table 2.3). Duplications across the region were identified when large stretches of consecutive bases were covered between 2 and 10 times the average genome DOC. Duplications greater than 10 times the average genome DOC were disregarded because they presumably represent common repetitive elements within the genome. The copy number across 100 bp bins was calculated by

dividing the average coverage across the bin by the average genome coverage such that single copy regions displayed a copy number of one.

Duplications in the identified Hereford selective sweep regions were recorded as well as duplications that were specific to spotted breeds. Duplications were also analyzed to determine if the copy number was variable or fixed within each breed. Finally, bovine optical map data generated by Zhou et al. (unpublished data) from the Hereford reference individual was also used to determine if the detected duplicated regions represented tandem repeats, or if the repeats appeared to be dispersed throughout the genome. Optical mapping is a biological method for validating structural variation detected by the previously described computational algorithms. Optical mapping constructs accurate, high-resolution physical maps of chromosomes via the restriction enzyme digestion of large fragments of genomic DNA (Schwartz et al., 1993). After the assembly of restriction fragments into a genome-wide physical map, complex genomic structures and discrepancies with the reference genome sequence may be revealed. Contigs generated from the assembly of restriction fragments can span tens of megabases, alleviating the computational challenges associated with the assembly of repetitive regions of the genome.

Results

Identification of selective sweep regions in Hereford cattle

Analysis of MAF on BTA6 from 64.7 to 85.5 Mb in the 811 fullblood Hereford animals identified four stretches of homozygosity, each individually greater than 75 kb with all contiguous SNPs possessing MAF < 1% (Table 2.4; Figure 2.1). All four of the

predicted sweep regions were localized to a 4.1 Mb portion of the 20.8 Mb region characterized by Grosz & MacNeil (1999). The MAF pattern for the loci located between the third and fourth sweep regions is distinctive with MAF either between 0 and 0.1 or between 0.4 and 0.5, with no loci having a MAF between 0.1 and 0.4 (Figure 2.1). Two of the four Hereford selective sweep regions overlap annotated bovine genes (Figure 2.2). The largest of the selective sweep regions is approximately 415 kb and extends into *KIT*. The region begins approximately 354 kb upstream of the locus and extends for 77 kb through the first exon and into a portion of the very large first intron. However, immediately following the end of this putative selective sweep region begins the region with a peculiar MAF pattern. The next two largest sweep regions are approximately 118 kb and 92 kb. The 118 kb sweep region overlaps with part of *FIP1L1*, a factor interacting with *PAPOLA* and *CPSF1*, and part of *LNX1*, a ligand of numb-protein X1.

De novo assembly and characterization of breed specific variants within selective sweep regions

Local *de novo* assembly of three Angus sire-son pairs was conducted across 2 Mb on BTA6. As previously stated, amosvalidate was used to detect misassemblies based on alignment breakpoint, read coverage, mate-pair inconsistencies, and micro-heterogeneities across multiple overlapping reads (Phillipy et al., 2008). No significant missassemblies were detected. The first Angus assembly resulted in 196 contigs and 134 scaffolds. The N50 statistics for this assembly were 20,250 bases and 28,657 bases for contigs and scaffolds, respectively. N50 is defined as the length, N, for which 50% of all the bases in the assembly are contained in contigs or scaffolds that are equal to or larger than N (Miller et al., 2010). The second Angus assembly resulted in 329 contigs with a

N50 of 12,416 bases and 213 scaffolds with a N50 of 20,402 bases. The final Angus assembly resulted in 297 contigs with a N50 of 13,448 bases and 144 scaffolds with a N50 of 29,811 bases. The Hereford assembly of a single individual resulted in 371 contigs with a N50 of 8,941 bases and 200 scaffolds with a N50 of 24,508 bases (Table 2.1).

Pairwise alignment of the Angus *de novo* assemblies to the Hereford *de novo* assembly revealed 323 predicted fixed differences within the Hereford selective sweep regions. Thirty-four of the fixed differences were due to indels, and the remaining 289 putative fixed differences were SNPs. None of the indels occurred within an annotated gene or known promoter region and only two of the SNPs were in an annotated gene (Table 2.5); both in intron 1 of *KIT*.

Identification of duplications

Genomic data were also analyzed for structural variants within the Hereford selective sweep regions. There was no evidence of large indels, inversions, or translocations in Hereford *versus* the Angus genome alignments. However, when all 84 genomes were analyzed, several breed specific duplications were revealed. When the average copy number by breed was analyzed two major conclusions were drawn. First, all breeds and both species appear to have a repetitive element within intron 1 of *KIT* resulting in a high copy number between approximately 71.80 Mb and 71.83 Mb (Figure 2.3). Second, Hereford and Simmental cattle possess two duplications at the same genomic locations (Figure 2.3A,B), which were not detected in any of the other breeds or bison (Figure 2.3C-O). These duplications were universally detected in all 14 Hereford cattle (Figure 2.4). The first apparent duplication is approximately 4.5 kb in size and is

duplicated approximately 6 times. The repeated motif within the reference assembly begins at 71.7470 Mb and extends to 71.7515 Mb, approximately 50 kb upstream of *KIT*. The second duplication is approximately 15 kb in size and is also duplicated approximately 6 times. This repeated motif begins at 71.810 Mb and extends to 71.825 Mb within the reference assembly and is located within the first intron of *KIT* between the two repetitive elements present in all other breeds and bison. All Simmental cattle also had DOC plots remarkably similar to those of the 14 Herefords (Figure 2.5). The first repeat region, from 71.7470 to 71.7515 Mb was duplicated three times, half the number of duplications seen in Hereford cattle. The second repeat region, from 71.810 to 71.825 Mb was duplicated six times, and was identical in copy number and location to the Hereford duplication.

The breed average sequence DOC plots reveal that these duplications were not detected in any of the remaining 64 animals (Figures 2.6-2.18), except for one Beefmaster. Animal 86423, a white-faced Beefmaster bull, had a DOC plot that was nearly identical to that of the Simmental cattle with a copy number of three at each of the Hereford/Simmental duplication sites (Figure 2.7) indicating that this animal is heterozygous for a Hereford allele at *KIT*. Brahman, Gir, and Nelore cattle all have an increased copy number in intron one, but the duplication boundaries and DOC signature appear to be different from that present within Simmental and Hereford (Figure 2.3F, G, L). Bison also have duplications in this region of intron one and, despite being diverged from cattle by at least 1 million years (Ritz et al., 2000; Decker et al., 2009), also have the same repetitive regions at 71.80 and 71.83 Mb (Figure 2.3O).

Genomic organization of *KIT* CNVs

In the process of aligning the sequence data to the reference assembly, reads which align with equal stringency to several genomic locations are assigned to each of those locations. Consequently, repetitive elements that have been assembled into multiple genomic regions will be revealed as a DOC signature that is greater than the flanking single copy regions. Conversely, regions of the genome that have been collapsed in the assembly process because they are repetitive will also produce a DOC signature that is indicative of a copy number variant. Therefore, DOC signatures indicating the presence of a repeated motif do not reveal whether these motifs are dispersed throughout the genome or are present as a tandemly repeated motif. Consequently, we sought to address this issue by two approaches. First, was to examine the bovine optical map for concordance to the assembly in the region harboring the *KIT* locus. The bovine optical map was produced from DNA from the reference assembly Hereford and was found to be discordant with the UMD3.1 reference assembly in the region of *KIT* (Figure 2.19; Zhou et al., unpublished data). The optical map reveals that the intron 1 CNV is duplicated in tandem and that ~110 kb of consecutive sequence is missing from the reference assembly. The genomic organization of the smaller upstream duplication was not resolved using the optical map because the repeated motif does not contain a restriction enzyme cut site.

We resolved the genomic organization of the first duplication by identifying mate-pair reads for which one read of each pair mapped to one of the two repeats and the second read mapped to different regions of the genome. If a significant number of such pairs could be identified it would provide evidence that the duplications were dispersed

throughout the genome rather than being tandem repeats on BTA6. All reads that aligned to the duplicated region were extracted and then filtered to retain only reads from mate-pair libraries. The genomic location where the read and its mate pair mapped were then extracted from the alignment data. These data were parsed using a custom Perl script to determine how often a mate pair read mapped more than 50 Mb away – essentially to a separate chromosome. Of the 1,465 reads from mate pair libraries that aligned to the duplicated region only 163 (~11.1%) had mates that mapped greater than 50 Mb away. Additionally, the 163 reads did not map to a common location, strongly suggesting this repeat is also a tandem duplication.

Discussion

Four putative selective sweep regions were detected within a small 4.1 Mb region of the larger 20.8 Mb candidate region for the *Spotted* locus identified by Grosz & MacNeil (1999). The first two detected selective sweep regions overlap the beginning and end of a selective sweep previously identified in Hereford using lower density SNP data (Ramey et al., 2013). The largest selective sweep region detected in Hereford points to an intuitive candidate for the *Spotted* locus. *KIT*, a tyrosine kinase, is responsible for the migration of melanocytes from their origin in the neural crest to their final destination in the dermis (Fleischman, 1993). The region directly following this sweep region also demonstrates MAF that are inconsistent with the expectation that SNP MAFs are independently and identically distributed as uniform on the interval [0 – 0.5]. The complete absence of SNP with MAF between 0.1 and 0.4 in the region between sweeps 3 and 4 immediately suggests that this region of the genome may be misassembled due to

the presence of a collapsed repeat. This observation motivated the analysis of both point mutations with alternate alleles fixed in spotted *versus* non-spotted breeds and analyses of structural variation, primarily by copy number.

Three hundred twenty-three SNPs and indels were predicted to be fixed for alternate alleles between Hereford and Angus based on the local *de novo* assemblies, but none were especially strong candidates for the *Spotted* mutation because they were all in intergenic or non-coding regions. However, the duplications involving *KIT* are strong candidates considering that they are present only in spotted animals. The white-faced Beefmaster (animal 86423) presents strong evidence for the causality of this allele. This bull is the only white-faced Beefmaster and the only Beefmaster exhibiting increased copy number at both of the duplicated regions. He exhibits a copy number of 3 at each of these locations which is consistent with having one Hereford chromosome and one Shorthorn chromosome, as the Hereford copy number is 6 (3 copies on each chromosome). The remaining Beefmaster animals were solid, except one that exhibited a white patch on the underline. This animal did not have either duplication, and since this white spotting exists without a white facial pattern, it is of Shorthorn origin.

Both duplications are located within a selective sweep in the Hereford breed, and, thus, have a fixed copy number. However, they do not appear to be within a selective sweep in Simmental and copy number appears to be variable (Figure 2.5). This suggests that the mutations both originated in Simmental, the older of the two breeds, and that a high copy (functionally strong) allele was selected to fixation in the Hereford breed. We hypothesize that the duplications present in Hereford and Simmental cattle reduce the transcriptional efficiency of *KIT*, causing retardation of melanocyte migration. While

functionally the mutation may behave additively, reduced *KIT* expression results in failure of melanocytes to reach the face – the final deposition site – and a dominant phenotype.

Whether the upstream duplication or intron 1 duplication is the causal mutation is not obvious. The upstream duplication is approximately 50 kb from the transcriptional start site of *KIT*. The regulation of *KIT* has been reported to be complex and involves *cis* regulatory elements immediately proximal to *KIT* as well as sequences located some distance upstream of the *KIT* coding region. A 9 kb hypersensitive cluster located 154 kb upstream of KIT has been identified in mouse melanocytes, but not in other cells, suggesting tissue-specific long-range regulation of gene expression (Berrozpe et al., 1999). Structural variation surrounding this hypersensitive region is responsible for two white-spotting patterns in mice, W^{57}, which results from an 80 kb deletion, and W^{bd}, which results from a 2.8 Mb inversion (Kluppel et al., 1997). Both of these mutations appear to interfere with distinct steps in melanocyte development, making it plausible for the duplicated sequence upstream of *KIT* in Hereford and Simmental cattle to interrupt a long-range regulatory element, resulting in the piebald phenotype. However, the locations of long-range regulatory elements have not been reported in cattle, and their identification by sequence conservation across species is only moderately accurate (Noonan & McCallion, 2010).

The second duplication adds a large amount of sequence to intron 1, possibly resulting in reduced transcriptional efficiency due to the increase in the number of nucleotides that the transcriptional machinery must read through to generate a transcript. Expression of *KIT* is required for melanocyte migration from the neural crest along a

dorsolateral pathway in a rostral to caudal sequence to their final destination in the dermis. In mice, this migration begins around embryonic day 11 and commences approximately by embryonic day 14-15 (Rawles, 1947; Mayer, 1973). The limited time allotted for melanocyte migration coupled with an increase in the amount of time necessary to generate a *KIT* transcript could presumably result in inadequate *KIT* expression and hinder melanocyte migration.

However, other non-spotted breeds possess different duplications that also add sequence to the first intron. Brahman, Gir, Nelore, and Bison all show increased copy numbers throughout regions of intron one and, based on their coat color phenotypes, do not appear to have disturbed melanocyte migration. This presents evidence for the upstream duplication as being the causal variant, however, if the upstream duplication interrupts a tissue-specific long-range enhancer, a reduction in *KIT* expression would only be observed in melanocytes and this would require extensive experimentation using developing embryos to validate. On the other hand, if the intronic duplication were the causal mutation, we would expect to see reduced transcription of *KIT* in all tissues. Allele-specific *KIT* expression analyses in animals homozygous and heterozygous for Hereford and Angus alleles might then determine if the transcriptional efficiency of KIT is reduced from alleles that contain the intron 1 duplications (Hereford alleles).

Regardless, we conclude that structural variation involving the *KIT* locus is responsible for the piebald pattern in Hereford and Simmental cattle. Dorsal spotting on Simmental and Holstein cattle and white patterning on Maine-Anjou and Shorthorn cattle are not caused by either of the copy number variants detected in Hereford and Simmental. The irregular spotting on Simmental and Holstein is an independently inherited recessive

trait and there is no evidence of either of the duplications being present in Holstein cattle. Simmental x Holstein F_1 animals exhibit a white face and white spots on the body whereas Hereford x Holstein F_1 animals only show a white face and never show more white spotting than would be present on a fullblood Hereford (Olson, 1981). This suggests that Simmental and Holstein possess the recessive body spotting allele, which is distinct from the allele causing white facial patterns in Hereford and Simmental. A third mutation is likely responsible for the white spotting on fullblood Maine-Anjou cattle and in some Shorthorn cattle, neither of which show evidence for the presence of either duplication.

Table 2.1. Statistics for local *de novo* assemblies of BTA6 from 70.4 to 72.4 Mb.

Assembly	Total Unique Reads	Number of Contigs	N50 – Contigs (bp)	Number of Scaffolds	N50 – Scaffolds (bp)
Angus 1	925,492	196	20,250	134	28,657
Angus 2	672,044	329	12,416	213	20,402
Angus 3	790,208	297	13,448	144	29,811
Hereford	589,850	371	8,941	200	24,508

Table 2.2. Coat color phenotype and number of individuals analyzed for structural variation involving *KIT* by breed.

Breed	Breed Code	Coat Color Phenotype	Number of Animals
Angus	AN	non-spotted	10
Bison	BB	non-spotted	3
Beefmaster	BEFM	variable	8
Brahman	BR	non-spotted	5
Gir	GIR	variable	4
Hereford	HFD	spotted	14
Holstein	HOL	spotted	8
Jersey	JER	non-spotted	3
Limousin	LM	non-spotted	6
Maine-Anjou	MAAN	variable	5
N'Dama	NDAM	non-spotted	1
Nelore	NEL	non-spotted	6
Romagnola	RMG	non-spotted	4
Shorthorn	SH	variable	1
Simmental	SIM	spotted	6
Total			84

Table 2.3. Spotted phenotype and average whole genome sequence coverage for 84 animals analyzed for structural variation involving *KIT*.

ID Number	Breed Code	Spotted Phenotype	Average Coverage
186	AN	--	43.12
219	AN	--	33.81
261	AN	--	40.16
294	AN	--	38.61
342	AN	--	21.03
407	AN	--	22.84
728	AN	--	21.33
32065	AN	--	29.68
32103	AN	--	26.75
34122	AN	--	27.23
2406	BB	--	30.02
20079	BB	--	35.90
20087	BB	--	28.55
86417	BEFM	white patch on underline	28.45
86418	BEFM	--	24.39
86419	BEFM	--	25.91
86420	BEFM	--	27.99
86421	BEFM	--	21.56
86423	BEFM	white face only	26.80
86424	BEFM	--	29.00
86425	BEFM	--	28.58
1918	BR	--	7.48
1923	BR	--	5.59
3009	BR	--	7.06
3014	BR	--	9.79
3037	BR	--	7.81
33534	GIR	--	8.77
33537	GIR	--	13.36
33539	GIR	--	14.69
33540	GIR	--	6.56
20830	HFD	Hereford pattern	29.02
34213	HFD	Hereford pattern	27.60
51368	HFD	Hereford pattern	24.99
51516	HFD	Hereford pattern	26.80
69621	HFD	Hereford pattern	25.98
69767	HFD	Hereford pattern	27.23
69898	HFD	Hereford pattern	26.86
69936	HFD	Hereford pattern	26.98

ID Number	Breed Code	Spotted Phenotype	Average Coverage
69958	HFD	Hereford pattern	27.63
87927	HFD	Hereford pattern	25.11
87930	HFD	Hereford pattern	24.70
87931	HFD	Hereford pattern	23.61
87932	HFD	Hereford pattern	24.65
87955	HFD	Hereford pattern	18.09
986	HOL	white spotting on body	4.92
1051	HOL	white spotting on body	5.43
1510	HOL	white spotting on body	5.04
2081	HOL	white spotting on body	6.21
18490	HOL	white spotting on body	6.36
23200	HOL	white spotting on body	9.14
47325	HOL	white spotting on body	6.32
47371	HOL	white spotting on body	4.73
33675	JER	--	7.92
33677	JER	--	6.04
33681	JER	--	13.36
7380	LM	--	8.73
7397	LM	--	5.96
7414	LM	--	6.58
7436	LM	--	8.89
7440	LM	--	11.66
7713	LM	--	11.64
87957	MAAN	white on hind legs	30.44
87958	MAAN	--	18.91
87959	MAAN	--	31.05
87960	MAAN	--	28.33
87961	MAAN	--	21.05
33749	NDAM	--	23.13
33772	NEL	--	6.96
33775	NEL	--	5.64
33780	NEL	--	7.68
33781	NEL	--	6.34
33784	NEL	--	13.01
33794	NEL	--	5.06
33841	RMG	--	5.52
33844	RMG	--	10.18
33845	RMG	--	7.41
33846	RMG	--	6.64
27422	SH	--	27.29
71407	SIM	white spotting on body and face	27.30

ID Number	Breed Code	Spotted Phenotype	Average Coverage
71411	SIM	white spotting on body and face	27.91
71415	SIM	white spotting on body and face	26.16
71420	SIM	white spotting on body and face	25.87
71657	SIM	white spotting on body and face	32.39
71658	SIM	white spotting on body and face	26.55

Table 2.4. Putative Hereford selective sweep regions detected in 811 fullblood Hereford cattle genotyped with the Illumina BovineSNP50 and BovineHD assays and imputed to the BovineHD content.

Sweep Region	Start coordinate (UMD3.1)	End coordinate (UMD3.1)	Block Size (bp)	Fixed Different SNPs (HFD vs. AN)	Fixed Different Indels (HFD vs. AN)
1	70,553,810	70,672,607	118,797	0	0
2	70,840,395	70,932,645	92,250	211	25
3	71,442,125	71,857,162	415,037	78	9
4	74,550,731	74,626,903	76,172	0	0

Table 2.5 Predicted fixed sequence differences detected between the Hereford and Angus local *de novo* genome alignments.

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001639	10,692	70,840,395	A	G
7180000001639	14,585	70,844,288	T	.
7180000001639	14,586	70,844,289	T	.
7180000001732	1,329	70,846,376	A	C
7180000001732	6,802	70,851,849	T	.
7180000001732	10,565	70,855,612	C	T
7180000001733	3,578	70,862,917	A	G
7180000001733	3,733	70,863,072	C	.
7180000001733	3,734	70,863,073	A	.
7180000001733	5,528	70,864,867	C	T
7180000001733	6,356	70,865,695	A	C
7180000001733	6,536	70,865,875	.	C
7180000001733	6,536	70,865,875	.	T
7180000001733	6,974	70,866,313	C	T
7180000001322	4,326	70,867,758	C	G
7180000001322	3,548	70,868,536	.	A
7180000001322	3,037	70,869,047	G	A
7180000001322	2,100	70,869,984	G	A
7180000001335	4,345	70,872,907	C	G
7180000001335	4,133	70,873,119	G	A
7180000001335	2,878	70,874,374	T	G
7180000001335	1,864	70,875,388	C	.
7180000001335	1,355	70,875,897	C	G
7180000001699	12,744	70,877,642	A	G
7180000001699	12,061	70,878,325	T	C
7180000001699	11,587	70,878,799	C	G
7180000001699	11,510	70,878,876	C	T
7180000001699	6,563	70,883,823	G	A
7180000001699	6,470	70,883,916	G	A
7180000001699	5,993	70,884,393	G	T
7180000001699	5,978	70,884,408	C	T
7180000001699	4,098	70,886,288	C	T
7180000001699	2,898	70,887,488	G	A
7180000001699	2,329	70,888,057	T	G
7180000001699	1,928	70,888,458	G	T
7180000001699	1,867	70,888,519	T	C
7180000001699	1,647	70,888,739	.	A
7180000001699	1,647	70,888,739	.	G
7180000001699	1,059	70,889,327	G	T
7180000001699	1,057	70,889,329	G	T
7180000001699	741	70,889,645	C	T

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001363	2,476	70,892,929	A	C
7180000001363	3,879	70,894,332	C	T
7180000001363	3,958	70,894,411	A	G
7180000001363	4,000	70,894,453	T	C
7180000001363	4,029	70,894,482	A	G
7180000001363	4,038	70,894,491	T	C
7180000001363	4,052	70,894,505	G	A
7180000001363	4,137	70,894,590	C	A
7180000001363	4,173	70,894,626	T	C
7180000001363	4,184	70,894,637	C	G
7180000001363	4,250	70,894,703	G	A
7180000001363	4,565	70,895,018	G	A
7180000001363	4,593	70,895,046	C	T
7180000001363	4,610	70,895,063	T	C
7180000001363	4,759	70,895,212	T	C
7180000001363	4,781	70,895,234	G	A
7180000001317	535	70,895,954	G	C
7180000001317	591	70,896,010	A	C
7180000001317	787	70,896,206	C	T
7180000001317	2,886	70,898,305	A	G
7180000001317	2,912	70,898,331	G	A
7180000001317	2,914	70,898,333	A	.
7180000001317	2,915	70,898,334	G	T
7180000001317	3,001	70,898,420	C	T
7180000001317	3,123	70,898,542	T	C
7180000001317	3,126	70,898,545	C	T
7180000001317	3,127	70,898,546	G	A
7180000001317	3,160	70,898,579	G	A
7180000001317	3,209	70,898,628	A	G
7180000001317	3,240	70,898,659	G	A
7180000001317	3,303	70,898,722	G	T
7180000001317	3,713	70,899,132	T	C
7180000001317	3,714	70,899,133	G	T
7180000001317	3,778	70,899,197	T	C
7180000001317	4,061	70,899,480	T	C
7180000001317	4,237	70,899,656	C	A
7180000001317	4,243	70,899,662	A	G
7180000001317	4,410	70,899,829	C	T
7180000001317	4,526	70,899,945	.	A
7180000001317	4,533	70,899,952	T	G
7180000001317	4,665	70,900,084	A	G
7180000001317	4,791	70,900,210	G	T
7180000001317	4,855	70,900,274	A	G
7180000001317	4,878	70,900,297	C	T

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001317	5,275	70,900,694	T	C
7180000001317	5,523	70,900,942	A	.
7180000001317	5,692	70,901,111	C	A
7180000001317	5,755	70,901,174	T	A
7180000001317	5,762	70,901,181	A	T
7180000001317	5,768	70,901,187	A	T
7180000001317	5,819	70,901,238	G	A
7180000001317	5,897	70,901,316	A	G
7180000001317	6,013	70,901,432	C	T
7180000001317	6,014	70,901,433	C	G
7180000001317	6,373	70,901,792	C	T
7180000001317	7,087	70,902,506	G	C
7180000001686	2,798	70,902,841	G	A
7180000001686	2,790	70,902,849	A	G
7180000001686	2,779	70,902,860	T	C
7180000001686	2,195	70,903,444	.	A
7180000001686	2,195	70,903,444	.	G
7180000001686	2,074	70,903,565	C	T
7180000001686	1,835	70,903,804	C	T
7180000001686	1,056	70,904,583	A	G
7180000001686	280	70,905,359	T	C
7180000001686	263	70,905,376	G	C
7180000001701	303	70,906,561	T	A
7180000001701	350	70,906,608	A	G
7180000001701	478	70,906,736	T	A
7180000001701	856	70,907,114	C	A
7180000001701	894	70,907,152	T	C
7180000001701	1,000	70,907,258	C	T
7180000001701	1,035	70,907,293	T	C
7180000001701	1,065	70,907,323	T	G
7180000001701	1,267	70,907,525	G	A
7180000001701	1,293	70,907,551	T	C
7180000001701	1,320	70,907,578	A	T
7180000001701	1,415	70,907,673	A	C
7180000001701	1,418	70,907,676	T	C
7180000001701	1,551	70,907,809	T	C
7180000001701	1,596	70,907,854	T	A
7180000001762	23	70,908,092	A	G
7180000001762	35	70,908,104	T	.
7180000001762	36	70,908,105	C	G
7180000001762	38	70,908,107	C	.
7180000001762	44	70,908,113	T	C
7180000001762	50	70,908,119	T	C
7180000001762	397	70,908,466	T	.

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001762	398	70,908,467	C	.
7180000001762	399	70,908,468	T	.
7180000001762	400	70,908,469	A	.
7180000001762	401	70,908,470	C	.
7180000001762	402	70,908,471	A	.
7180000001762	403	70,908,472	C	.
7180000001762	404	70,908,473	A	.
7180000001762	405	70,908,474	A	.
7180000001762	406	70,908,475	G	.
7180000001762	407	70,908,476	A	.
7180000001762	408	70,908,477	A	.
7180000001762	409	70,908,478	A	.
7180000001762	410	70,908,479	C	.
7180000001762	411	70,908,480	G	.
7180000001762	412	70,908,481	T	.
7180000001762	413	70,908,482	G	.
7180000001762	414	70,908,483	C	.
7180000001762	415	70,908,484	A	.
7180000001762	416	70,908,485	T	.
7180000001762	417	70,908,486	T	.
7180000001762	418	70,908,487	G	.
7180000001762	419	70,908,488	C	.
7180000001762	420	70,908,489	A	.
7180000001762	421	70,908,490	A	.
7180000001762	422	70,908,491	G	.
7180000001762	423	70,908,492	A	.
7180000001762	424	70,908,493	G	.
7180000001762	427	70,908,496	C	.
7180000001762	428	70,908,497	T	.
7180000001762	429	70,908,498	G	.
7180000001762	668	70,908,737	T	C
7180000001762	716	70,908,785	C	A
7180000001762	991	70,909,060	A	G
7180000001762	1,127	70,909,196	T	C
7180000001762	1,279	70,909,348	A	G
7180000001762	1,344	70,909,413	C	G
7180000001762	1,354	70,909,423	G	A
7180000001762	1,456	70,909,525	C	.
7180000001762	1,457	70,909,526	A	.
7180000001762	1,458	70,909,527	C	.
7180000001762	1,459	70,909,528	T	.
7180000001762	1,460	70,909,529	G	.
7180000001762	1,461	70,909,530	C	.
7180000001762	1,495	70,909,564	T	C

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001762	1,542	70,909,611	C	T
7180000001762	1,555	70,909,624	G	T
7180000001762	1,567	70,909,636	T	A
7180000001762	1,584	70,909,653	T	C
7180000001762	1,598	70,909,667	A	T
7180000001762	1,943	70,910,012	C	A
7180000001762	2,180	70,910,249	C	T
7180000001762	2,219	70,910,288	A	C
7180000001762	2,251	70,910,320	A	G
7180000001762	2,289	70,910,358	T	C
7180000001762	2,551	70,910,620	C	G
7180000001762	2,735	70,910,804	G	A
7180000001762	2,974	70,911,043	T	C
7180000001762	3,330	70,911,399	A	C
7180000001762	3,336	70,911,405	G	A
7180000001762	3,382	70,911,451	T	C
7180000001762	3,434	70,911,503	A	G
7180000001762	3,567	70,911,636	G	A
7180000001762	3,713	70,911,782	C	G
7180000001762	3,804	70,911,873	G	A
7180000001762	4,384	70,912,453	T	G
7180000001762	4,601	70,912,670	G	A
7180000001762	4,661	70,912,730	A	C
7180000001762	4,908	70,912,977	T	C
7180000001762	5,078	70,913,147	A	G
7180000001762	5,191	70,913,260	G	T
7180000001762	5,434	70,913,503	C	T
7180000001762	5,494	70,913,563	G	A
7180000001762	5,816	70,913,885	G	A
7180000001762	5,920	70,913,989	G	C
7180000001762	6,014	70,914,083	G	A
7180000001762	6,238	70,914,307	T	C
7180000001762	7,215	70,915,284	G	T
7180000001762	7,365	70,915,434	C	T
7180000001762	7,409	70,915,478	A	G
7180000001762	7,430	70,915,499	A	.
7180000001762	7,480	70,915,549	A	C
7180000001762	7,655	70,915,724	C	T
7180000001762	7,761	70,915,830	T	C
7180000001762	8,318	70,916,387	C	T
7180000001762	8,518	70,916,587	T	.
7180000001762	8,519	70,916,588	A	.
7180000001762	8,520	70,916,589	C	.
7180000001762	8,521	70,916,590	C	.

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001762	8,588	70,916,657	T	C
7180000001762	8,949	70,917,018	T	C
7180000001762	9,052	70,917,121	C	T
7180000001762	9,461	70,917,530	T	C
7180000001762	9,691	70,917,760	A	G
7180000001762	9,720	70,917,789	C	T
7180000001762	9,850	70,917,919	A	G
7180000001762	10,055	70,918,124	A	C
7180000001762	10,146	70,918,215	.	A
7180000001762	10,147	70,918,216	T	A
7180000001762	10,443	70,918,512	C	A
7180000001762	10,523	70,918,592	C	T
7180000001762	10,590	70,918,659	.	T
7180000001762	10,644	70,918,713	G	T
7180000001762	10,688	70,918,757	A	G
7180000001762	10,706	70,918,775	G	T
7180000001762	10,799	70,918,868	G	.
7180000001762	10,800	70,918,869	T	.
7180000001762	10,801	70,918,870	T	.
7180000001762	10,802	70,918,871	T	.
7180000001762	10,803	70,918,872	T	.
7180000001762	10,826	70,918,895	A	C
7180000001762	11,097	70,919,166	T	G
7180000001762	11,417	70,919,486	T	C
7180000001762	11,436	70,919,505	T	C
7180000001762	11,557	70,919,626	T	C
7180000001762	11,563	70,919,632	T	G
7180000001762	12,008	70,920,077	C	G
7180000001762	12,114	70,920,183	C	T
7180000001762	12,144	70,920,213	C	A
7180000001762	12,194	70,920,263	T	C
7180000001762	12,360	70,920,429	.	T
7180000001762	12,360	70,920,429	.	A
7180000001762	12,374	70,920,443	A	G
7180000001762	12,410	70,920,479	A	G
7180000001762	12,429	70,920,498	.	A
7180000001762	12,429	70,920,498	.	A
7180000001762	12,429	70,920,498	.	T
7180000001762	12,429	70,920,498	.	A
7180000001762	12,429	70,920,498	.	A
7180000001762	12,429	70,920,498	.	A
7180000001762	12,566	70,920,635	G	A
7180000001354	6,044	70,922,473	C	T
7180000001354	6,001	70,922,516	T	C

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001354	5,696	70,922,821	T	G
7180000001354	5,267	70,923,250	A	T
7180000001354	5,220	70,923,297	T	G
7180000001354	4,896	70,923,621	A	G
7180000001354	4,801	70,923,716	C	G
7180000001354	4,742	70,923,775	A	G
7180000001354	4,679	70,923,838	C	T
7180000001354	4,603	70,923,914	C	T
7180000001354	4,484	70,924,033	G	A
7180000001354	4,402	70,924,115	.	G
7180000001354	4,402	70,924,115	.	C
7180000001354	3,830	70,924,687	.	T
7180000001354	3,824	70,924,693	G	A
7180000001354	3,787	70,924,730	C	T
7180000001354	3,733	70,924,784	T	C
7180000001354	3,634	70,924,883	A	G
7180000001354	3,302	70,925,215	G	A
7180000001354	3,194	70,925,323	A	G
7180000001354	3,160	70,925,357	T	C
7180000001354	3,142	70,925,375	T	C
7180000001354	3,132	70,925,385	G	A
7180000001354	2,562	70,925,955	T	C
7180000001354	2,405	70,926,112	C	T
7180000001354	2,082	70,926,435	G	A
7180000001354	2,043	70,926,474	A	G
7180000001354	1,938	70,926,579	.	A
7180000001354	1,938	70,926,579	.	A
7180000001354	1,938	70,926,579	.	G
7180000001354	1,400	70,927,117	C	T
7180000001798	1,961	70,932,235	A	G
7180000001798	1,867	70,932,329	A	G
7180000001671	143	71,452,534	G	T
7180000001671	2,470	71,454,861	A	G
7180000001671	2,720	71,455,111	C	T
7180000001814	5,523	71,460,025	A	G
7180000001814	1,712	71,463,836	G	T
7180000001814	1,275	71,464,273	T	C
7180000001813	2,595	71,469,540	C	T
7180000001813	843	71,471,292	A	G
7180000001812	3,680	71,475,922	T	C
7180000001812	3,594	71,476,008	T	C
7180000001812	3,562	71,476,040	A	G
7180000001812	2,338	71,477,264	A	G
7180000001812	2,296	71,477,306	T	G

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001812	1,676	71,477,926	A	G
7180000001812	1,583	71,478,019	A	G
7180000001812	1,545	71,478,057	A	T
7180000001812	888	71,478,714	G	A
7180000001812	879	71,478,723	T	C
7180000001811	29,635	71,483,178	.	A
7180000001811	4,662	71,508,045	A	C
7180000001811	4,634	71,508,073	T	C
7180000001637	5,540	71,529,755	T	.
7180000001637	5,539	71,529,756	A	.
7180000001637	2,910	71,532,385	C	T
7180000001810	35,942	71,537,263	A	T
7180000001810	10,474	71,562,731	.	T
7180000001810	10,075	71,563,130	A	G
7180000001810	9,622	71,563,583	A	C
7180000001810	9,609	71,563,596	G	A
7180000001810	9,468	71,563,737	A	G
7180000001810	7,635	71,565,570	A	G
7180000001810	7,201	71,566,004	A	C
7180000001810	6,137	71,567,068	.	A
7180000001810	6,083	71,567,122	T	C
7180000001810	6,072	71,567,133	C	T
7180000001810	6,009	71,567,196	.	A
7180000001810	6,009	71,567,196	.	T
7180000001810	4,913	71,568,292	G	A
7180000001810	4,824	71,568,381	G	A
7180000001810	4,105	71,569,100	A	G
7180000001735	2,708	71,578,891	C	A
7180000001735	2,936	71,579,119	A	C
7180000001735	2,995	71,579,178	A	G
7180000001735	4,391	71,580,574	G	A
7180000001735	4,395	71,580,578	.	A
7180000001735	5,399	71,581,582	G	A
7180000001735	5,672	71,581,855	C	G
7180000001735	6,161	71,582,344	A	C
7180000001735	6,172	71,582,355	T	G
7180000001735	7,184	71,583,367	.	A
7180000001735	7,184	71,583,367	.	C
7180000001735	7,184	71,583,367	.	A
7180000001735	7,687	71,583,870	G	A
7180000001735	8,251	71,584,434	T	C
7180000001735	8,383	71,584,566	.	T
7180000001735	8,383	71,584,566	.	G
7180000001657	3,421	71,588,983	A	G

Contig	Contig Position	BTA6 Position (UMD3.1)	HFD Allele	AN Allele
7180000001657	2,569	71,589,835	C	T
7180000001657	1,797	71,590,607	C	T
7180000001736	845	71,593,307	G	A
7180000001736	1,053	71,593,515	T	C
7180000001736	5,915	71,598,377	G	A
7180000001736	6,063	71,598,525	G	T
7180000001736	6,178	71,598,640	G	A
7180000001736	6,447	71,599,517	C	T
7180000001736	6,647	71,599,717	A	C
7180000001736	7,536	71,600,606	G	A
7180000001736	7,701	71,600,771	C	T
7180000001736	7,717	71,600,787	C	G
7180000001736	8,638	71,601,708	A	G
7180000001736	8,640	71,601,710	G	A
7180000001736	14,241	71,607,311	G	C
7180000001777	609	71,616,142	T	C
7180000001734	12,969	71,649,309	G	C
7180000001728	725	71,671,852	G	A
7180000001728	9,668	71,680,795	T	C
7180000001727	1,466	71,684,157	A	C
7180000001367	3,008	71,692,345	G	T
7180000001799	1,264	71,698,815	A	G
7180000001799	1,300	71,698,851	A	C
7180000001794	6,038	71,709,067	A	G
7180000001660	890	71,722,666	C	T
7180000001660	1,837	71,723,613	G	T
7180000001660	7,907	71,729,683	T	.
7180000001809	817	71,747,438	A	G
7180000001679	633	71,759,234	A	C
7180000001679	3,135	71,761,736	T	C
7180000001765	529	71,767,518	A	C
7180000001765	5,620	71,772,609	C	A
7180000001765	6,044	71,773,033	A	G
7180000001773	3,543	71,831,977	G	A
7180000001400	1,458	71,841,452	A	T

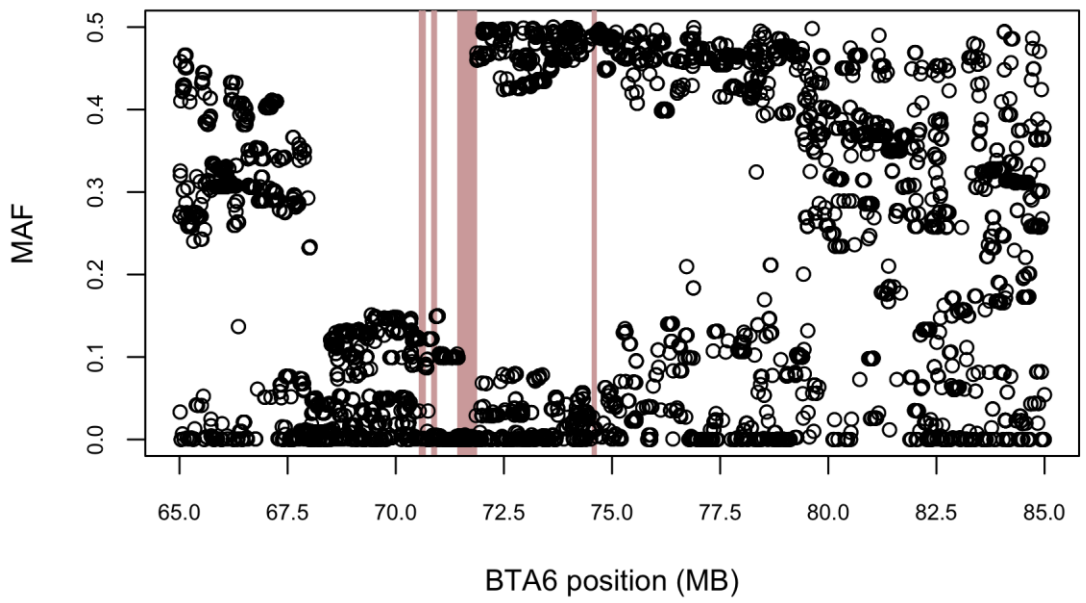

Figure 2.1. Minor allele frequencies (MAF) for 5,086 Illumina BovineHD SNPs spanning putative selective sweep regions in Hereford cattle. Highlighted regions are 75 kb or greater in which all contiguous loci have MAF < 0.01.

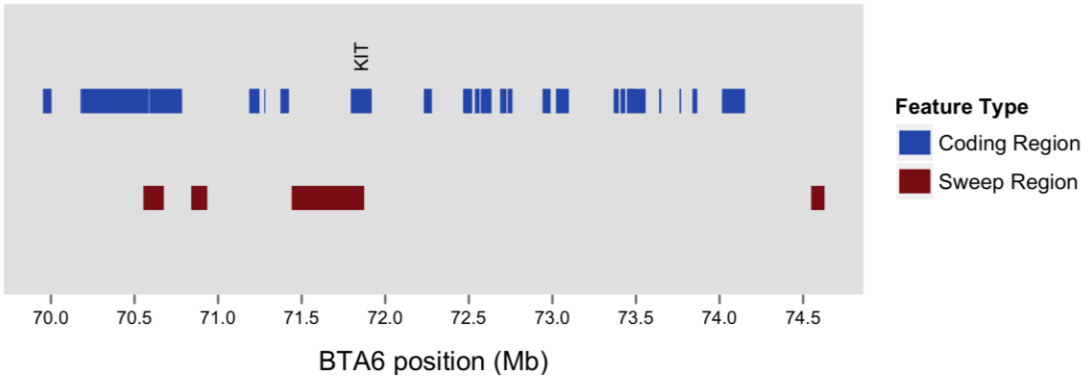

Figure 2.2. Gene annotation for putative Hereford selective sweep regions.

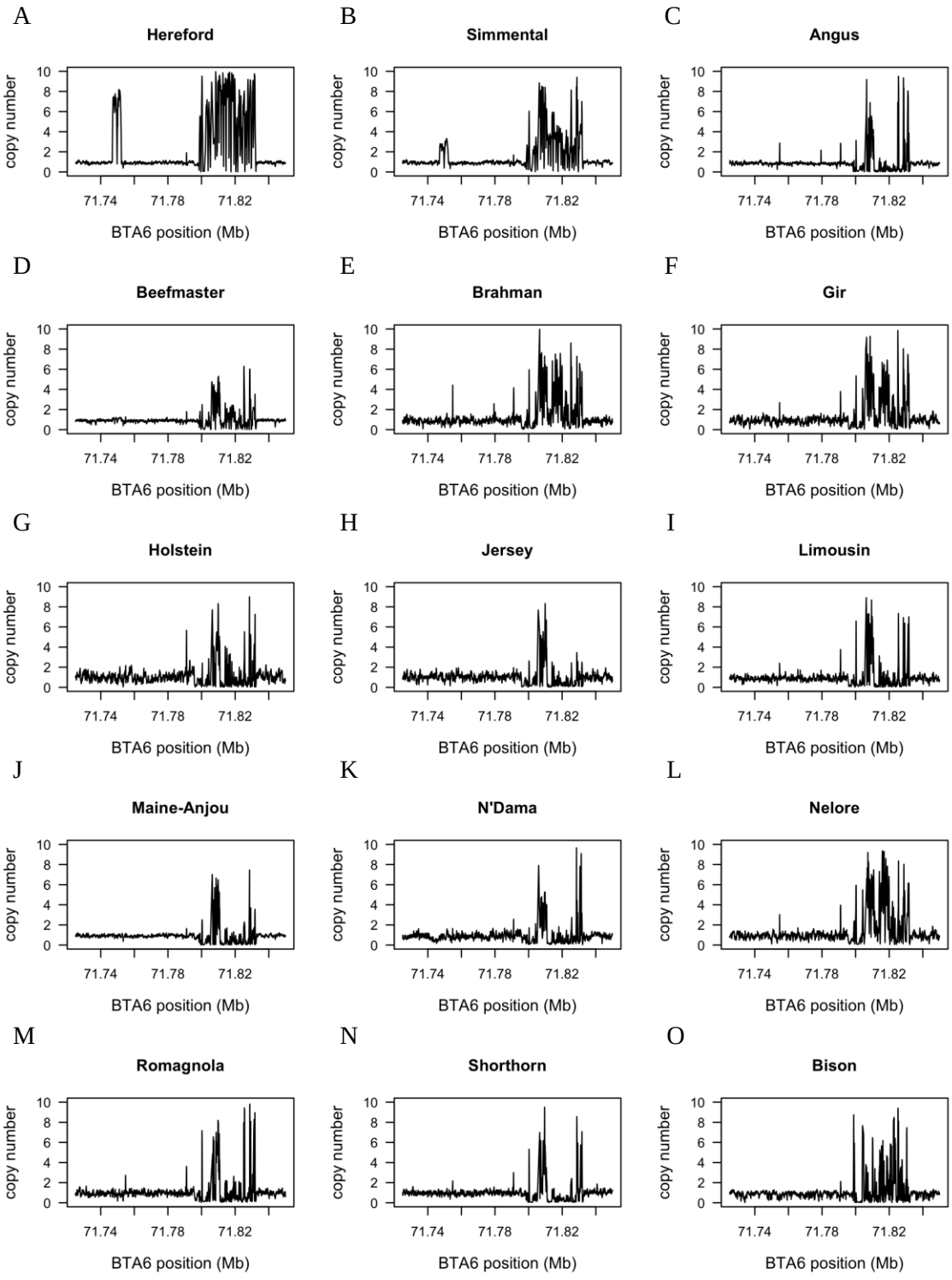

Figure 2.3. Breed average BTA6 copy number from 71.72 to 72.84 Mb across 100 bp bins. A) Hereford (n = 14). B) Simmental (n = 6). C) Angus (n = 10). D) Beefmaster (n = 8). E) Brahman (n = 5). F) Gir (n = 4). G) Holstein (n = 8). H) Jersey (n = 3). I) Limousin (n = 6). J) Maine-Anjou (n = 5). K) N'Dama (n = 1). L) Nelore (n = 6). M) Romagnola (n = 4). N) Shorthorn (n = 1). O) Bison (n = 3).

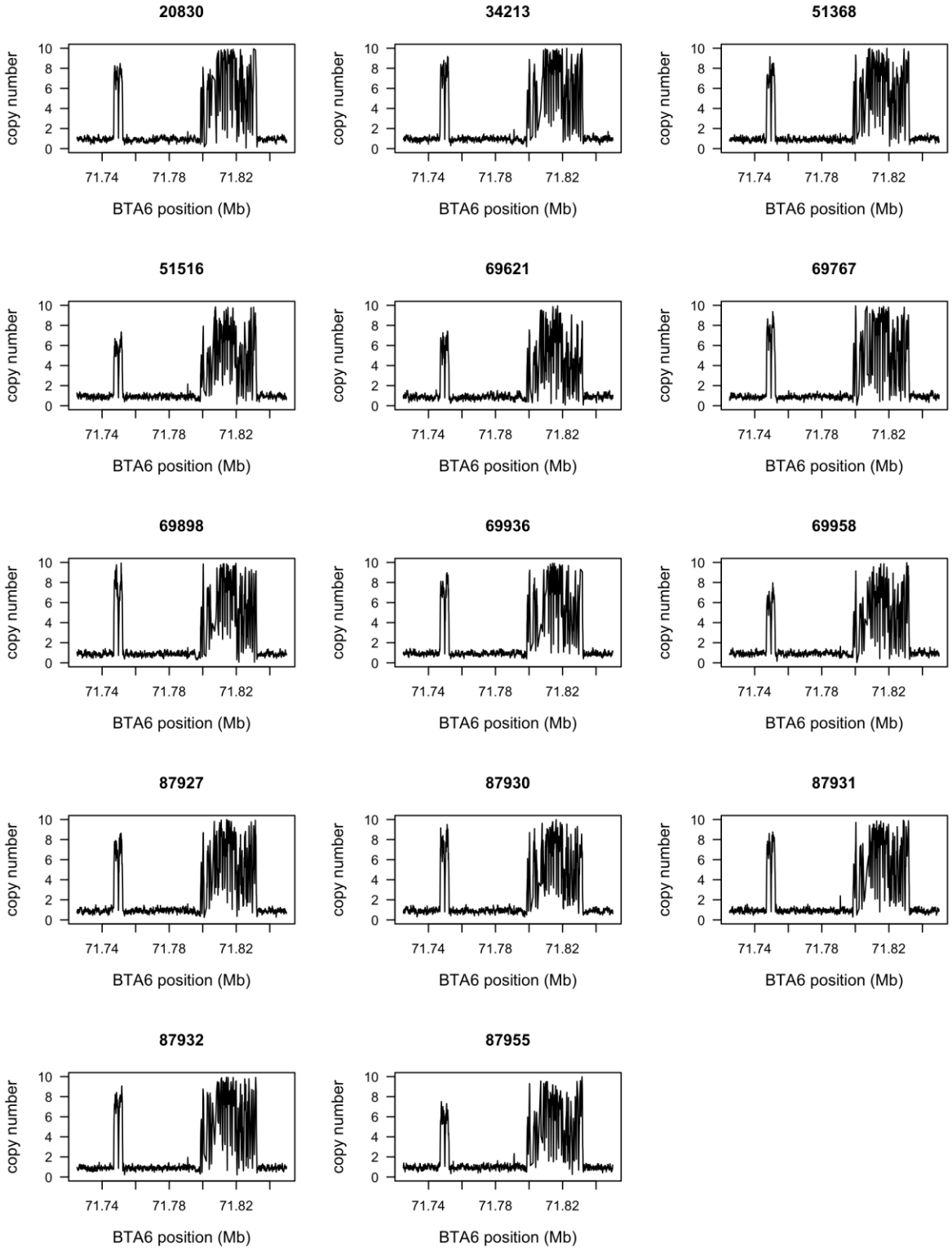

Figure 2.4. Hereford copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

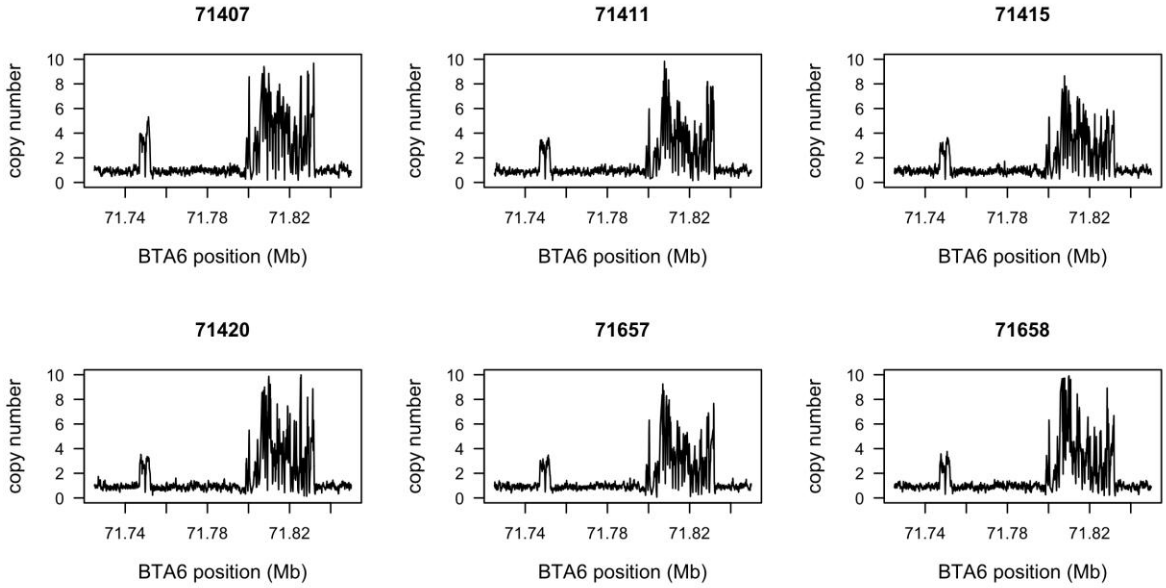

Figure 2.5. Simmental copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by100 bp bins.

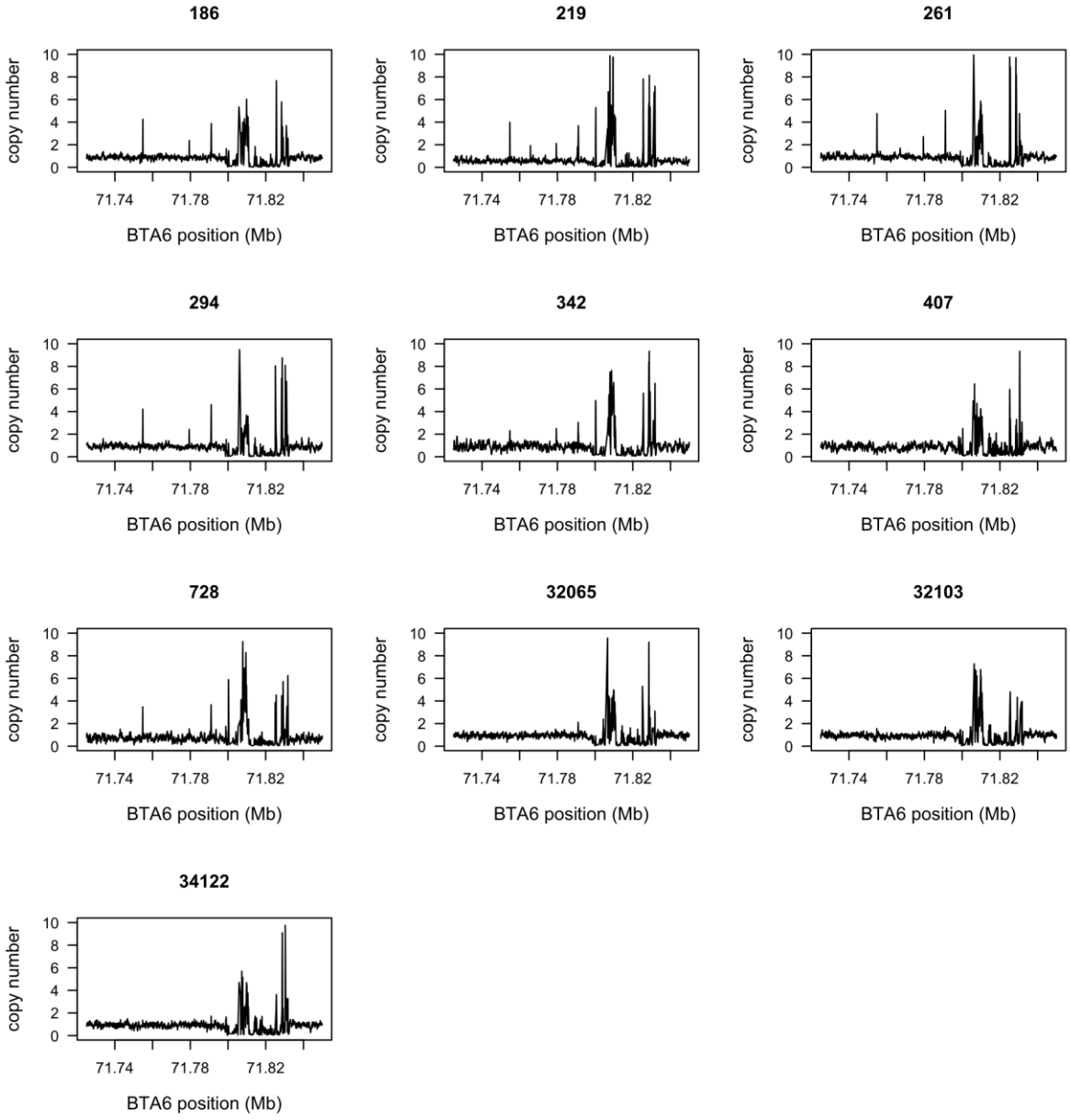

Figure 2.6. Angus copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

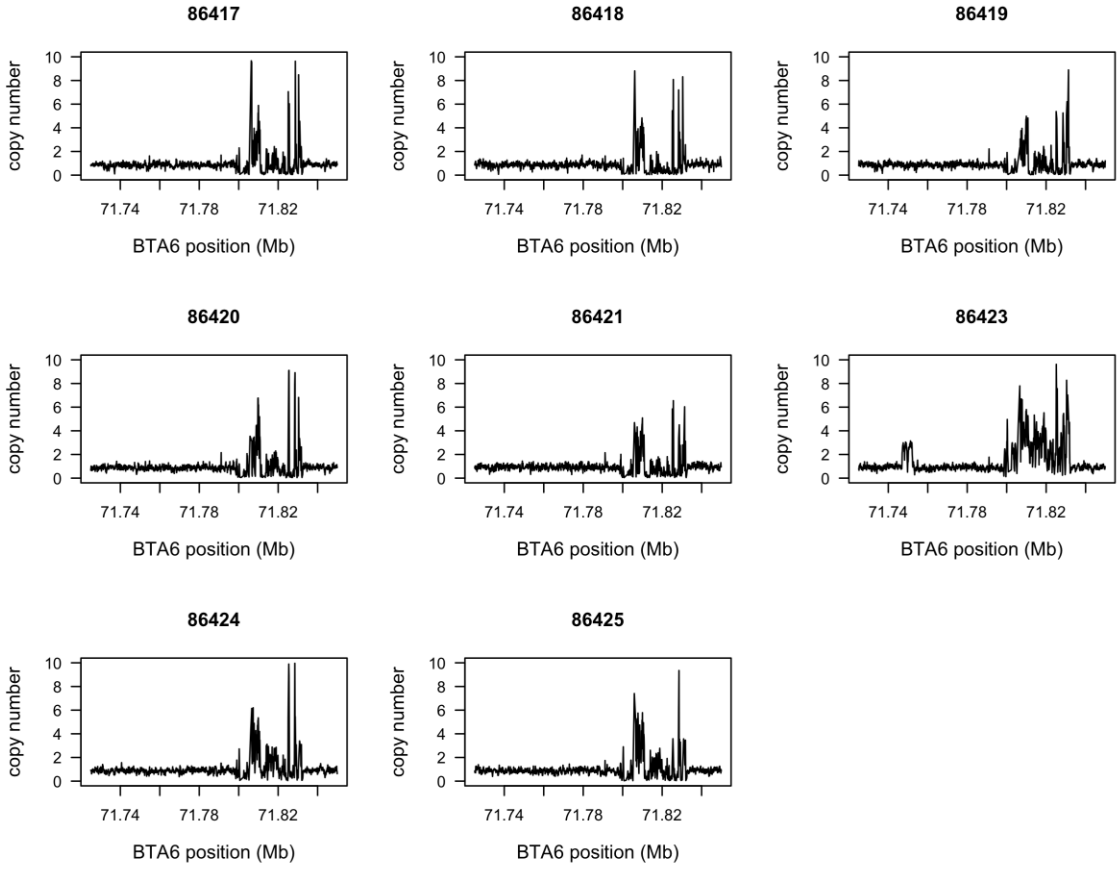

Figure 2.7. Beefmaster copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

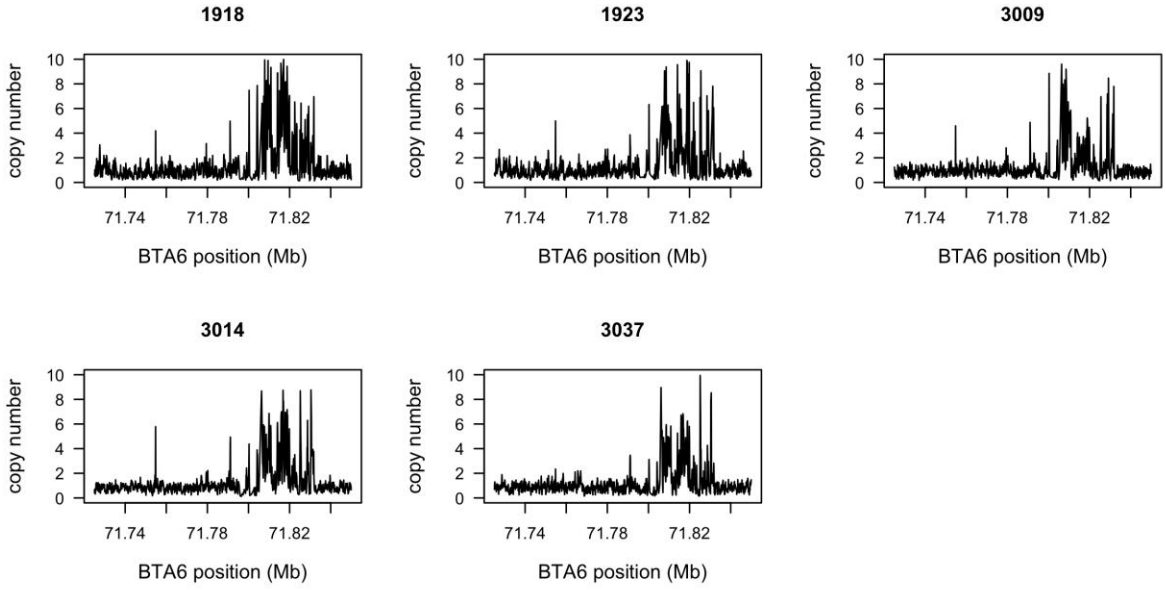

Figure 2.8. Brahman copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

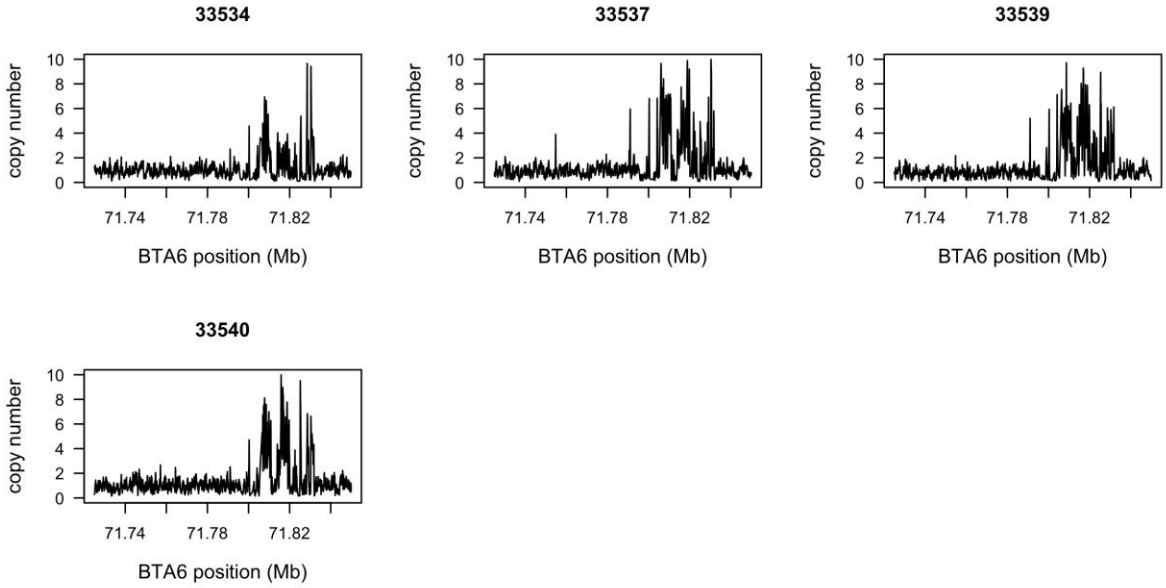

Figure 2.9. Gir copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

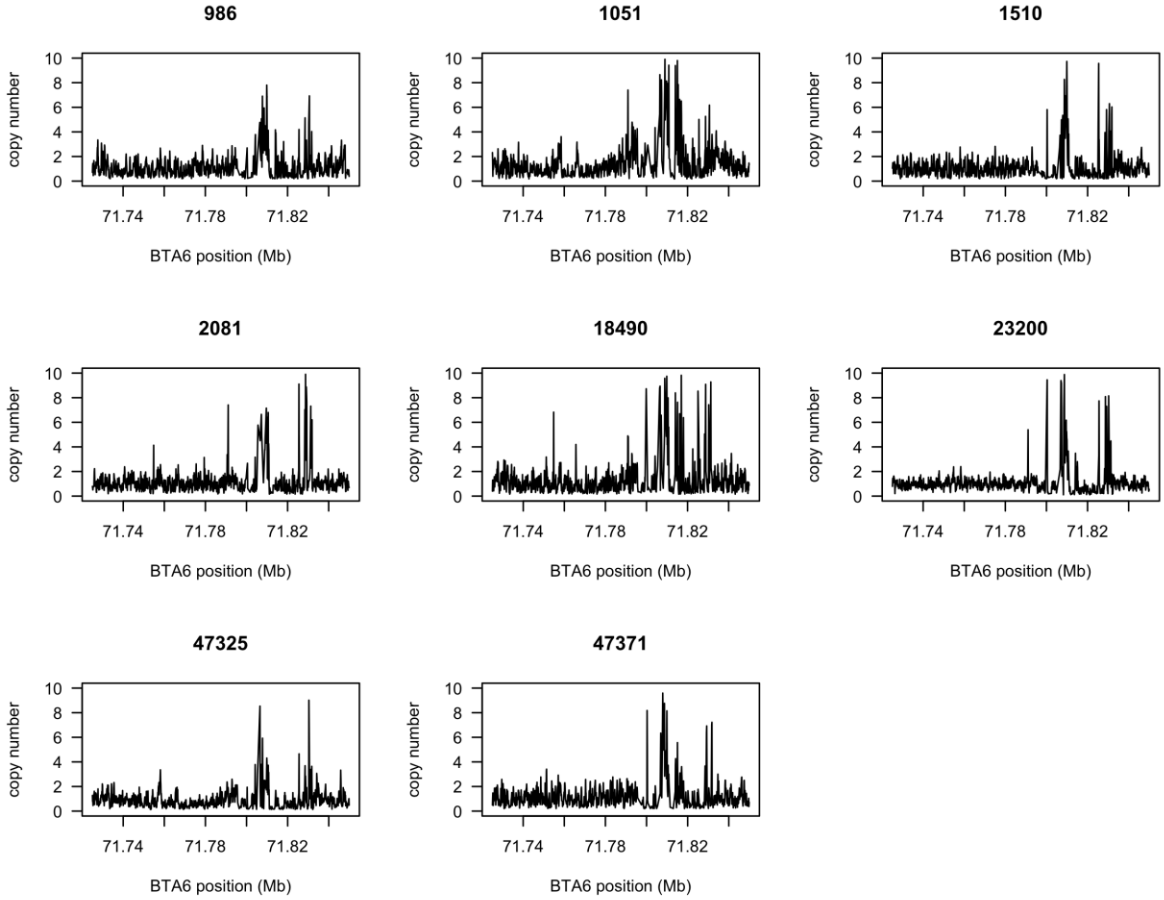

Figure 2.10. Holstein copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

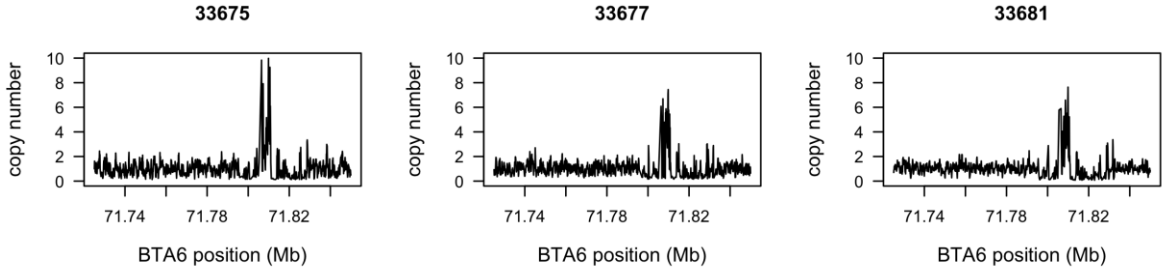

Figure 2.11. Jersey copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

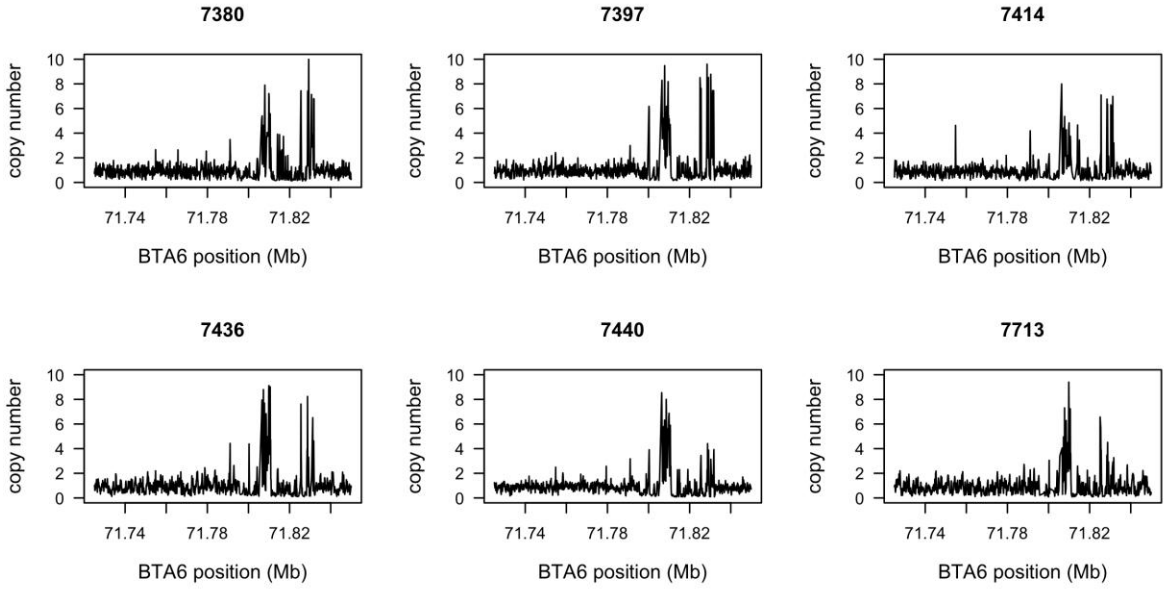

Figure 2.12. Limousin copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

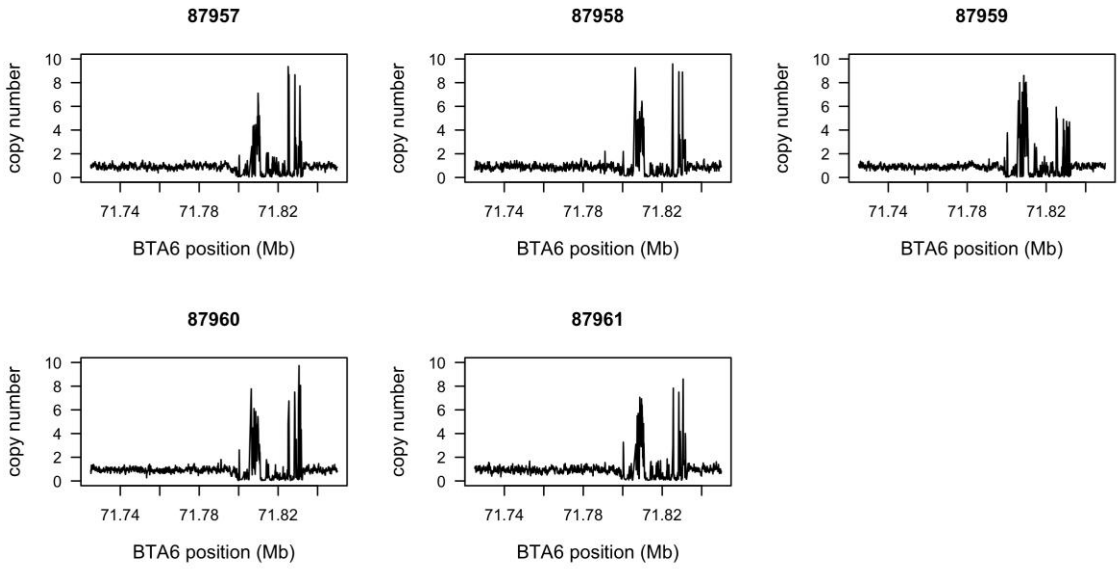

Figure 2.13. Maine-Anjou copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

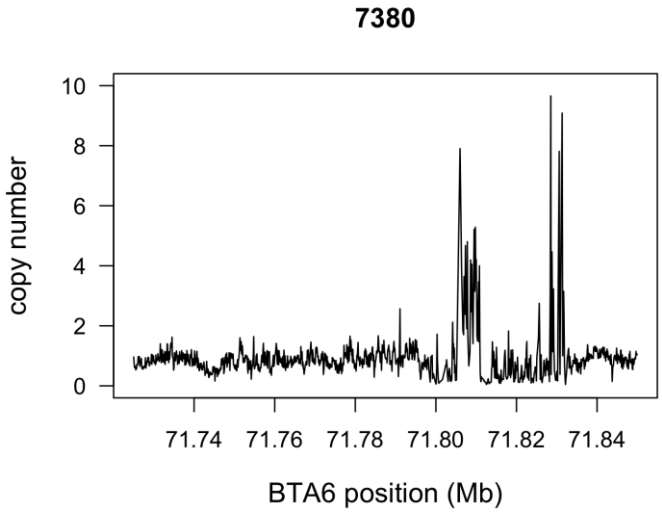

Figure 2.14. N'Dama copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

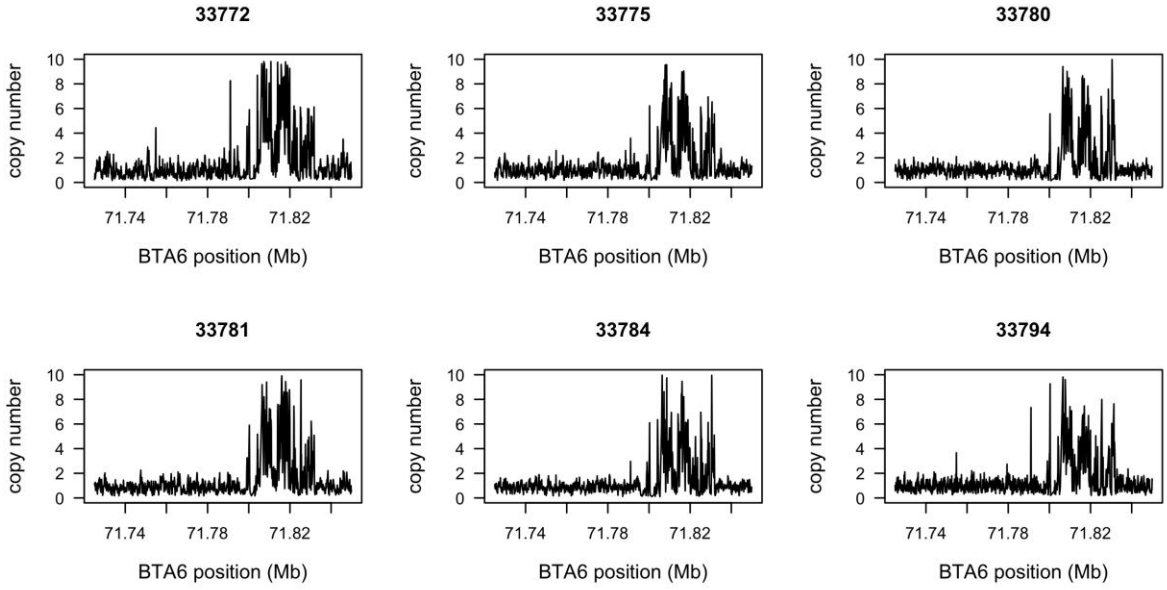

Figure 2.15. Nelore copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

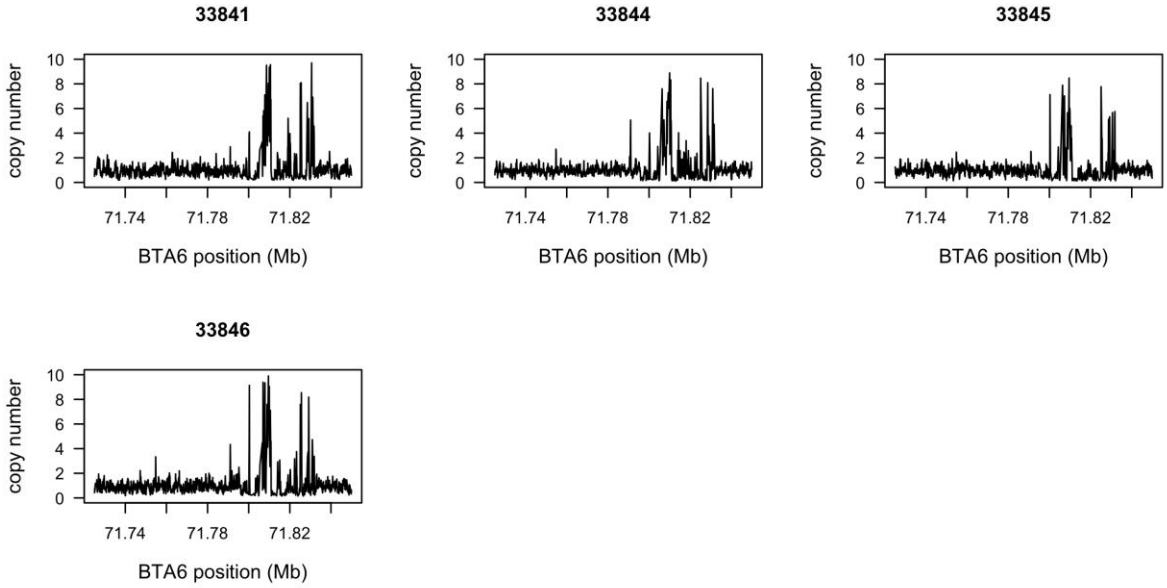

Figure 2.16. Romagnola copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

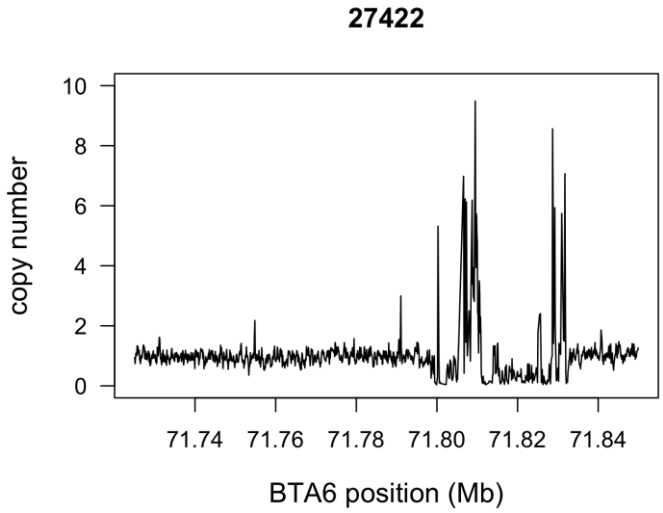

Figure 2.17. Shorthorn copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

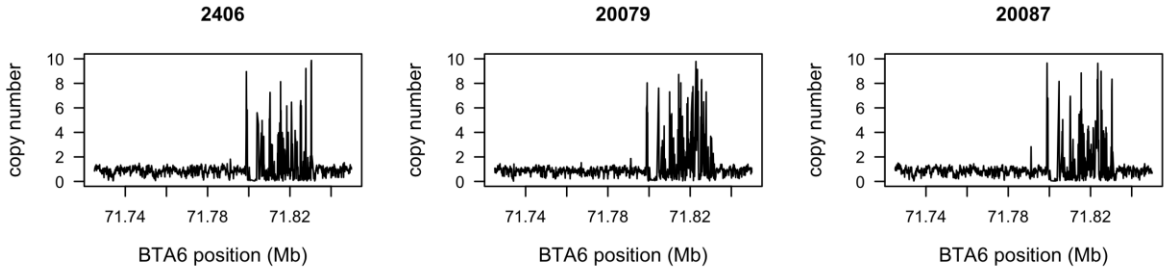

Figure 2.18. Bison copy number variation from 71.72 to 72.84 Mb on BTA6 averaged by 100 bp bins.

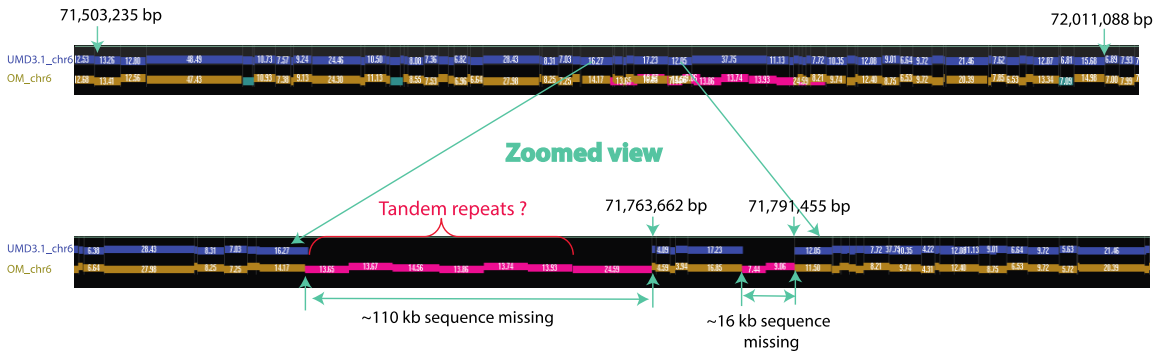

Figure 2.19. Discordances between the bovine optical map and UMD3.1 for BTA6 from 71.5 to 72.0 Mb (Zhou et al., unpublished data) indicates the *KIT* intron 1 copy number variant is a tandem repeat.

REFERENCES

Anderson, D. E., Chambers, D. & Lush, J. L. Studies on Bovine Ocular Squamous Carcinoma (''Cancer Eye") III . Inheritance of Eyelid Pigmentation. *Journal of Animal Science* **16**, 1007–1016 (1957).

Berrozpe, G., Timokhina, I., Yukl, S., Tajima, Y., Ono, M., Zelenetz, A. D. & Besmer, P. The W(sh), W(57), and Ph Kit expression mutations define tissue-specific control elements located between -23 and -154 kb upstream of Kit. *Blood* **94**, 2658–2666 (1999).

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., Song, J., Schnabel, R. D., Ventura, M., Taylor, J. F., Garcia, J. F., Van Tassell, C. P., Sonstegard, T. S., Eichler, E. E. & Liu G. E. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research* **22**, 778-790 (2012).

Bonsma, J. C. Breeding cattle for increased adaptability to tropical and sub-tropical environments. *The Journal of Agricultural Science* **39,** 204-221 (1949).

Brooks, S. A., Lear, T. L., Adelson, D. L. & Bailey, E. A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. *Cytogenetic and Genome Research* **119**, 225–230 (2007).

Brooks, S. A. & Bailey, E. Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. *Mammalian Genome* **16**, 893–902 (2005).

Brown, M. H., Brightman, A. H., Fenwick, B. W. & Rider, M. A. Infectious bovine keratoconjunctivitis: a review. *Journal of Veterinary Internal Medicine* **12**, 259–266 (1998).

Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223 (2009).

Buac, K., Xu, M., Cronin, J., Weeraratna, A. T., Hewitt, S. M & Pavan W. J. NRG1/ERBB3 signaling in melanocyte development and melanoma: inhibition of differentiation and promotion of proliferation. *Pigment Cell and Melanoma Research* **22**, 773–784 (2009).

Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., Cooper, A., Vilkki, J., Seabury, C. M., Caetano, A. R., Johnson, G. S., Brenneman, R. A., Hanotte, O., Eggert, L. S., Wiener, P., Kim, J., Kim, K., Sonstegard, T. S., Van Tassell, C. P., Neibergs, H. L., McEwan, J. C., Brauning, R., Coutinho, L. L., Babar, M. E., Wilson, G. A., McClure, M. C., Rolf, M. M., Kim, J., Schnabel, R. D., & Taylor, J. F. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Science* **106**, 18644-18649 (2009).

Den Otter W, Hill, F. W., Klein, W. R., Everse, L. A., Ruitenberg, E.J., Van der Ven, L. T., Koten, J. W., Steerenberg, P. A., Faber, J. A. & Rutten V. P. Ocular squamous cell carcinoma in Simmental cattle in Zimbabwe. *American Journal of Veterinary Research* **56**, 1440–1444 (1995).

Durkin, K. Coppieters W, Drögemüller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L, Kamatani Y, Karim L, Lathrop M, Moser S, Oldenbroek K, Rieder S, Sartelet A, Sölkner J, Stålhammar H, Zelenika D, Zhang Z, Leeb T, Georges M, Charlier C. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81-86 (2012).

Fleischman, R. A. From white spots to stem cells: the role of the Kit receptor in mammalian development. *Trends in Genetics* **9**, 285-290 (1993).

Fontanesi, L., Tazzoli, M., Russo, V. & Beever, J. Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. *Animal Genetics* **41**, 295–303 (2009).

Fontanesi, L., Scotti, E. & Russo, V. Analysis of SNPs in the KIT gene of cattle with different coat colour patterns and perspectives to use these markers for breed traceability and authentication of beef and dairy products. *Italian Journal of Animal Science* **9**, 217–221 (2010).

Geissler, E. N., Ryan, M. A. & Housman, D. E. The Dominant-White Spotting (W) Locus of the Mouse Encodes the c-kit Proto-Oncogene. *Cell* **55**, 185–192 (1988).

Giebel, L. B. & Spritz, R. A. Mutation of the KIT (mast/stem cell growth factor receptor) protooncogene in human piebaldism. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 8696–8699 (1991).

Grosz, M. D. & MacNeil, M. D. The "spotted" locus maps to bovine chromosome 6 in a Hereford-Cross population. *The Journal of Heredity* **90**, 233–236 (1999).

Guilbert, H. R., Wahid, A., Wagnon, K. A. & Gregory, P.W. Observations on pigmentation of eyelids of Hereford cattle in relation to occurrence of ocular epitheliomas. *Journal of Animal Science* **7**, 426-429 (1948).

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**, e1001139 (2010).

Jimbow, K., Quevedo, W. C., Fitzpatrick, T. B., & Szabo, G. Some aspects of melanin biology: 1950 – 1975. *Journal of Investigative Dermatology*. **67**, 72-89 (1976).

Kluppel, M., Nagle, D. L., Bucan, M., & Bernstein, A. Long-range genomic rearrangements upstream of Kit dysregulate the developmental pattern of Kit expression

in W57 and Wbanded mice and interfere with distinct steps in melanocyte development. *Development* **124,** 65-77 (1997).

Kubic, J. D., Young, K. P., Plummer, R. S., Ludvik, A. E., & Lang, D. Pigmentation PAX-ways: the role of Pax3 in melanogenesis, melanocyte stem cell maintenance, and disease. *Pigment Cell & Melanoma Research* **21**, 627-645 (2008).

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. L. Versatile and open software for comparing large genomes. *Genome Biology* **5**, R12 (2004).

Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine* **12**, 406–414 (2006).

Liu, L., Harris, B., Keehan, M. & Zhang, Y. Genome scan for the degree of white spotting in dairy cattle. *Animal Genetics* **40**, 975–977 (2009).

Marklund, S., Kijas, J., Rodriguez-Martinez, H., Rönnstrand, L., Funa, K., Moller, M., Lange, D., Edfors-Lilja, I. & Andersson, L. Molecular basis for the dominant white phenotype in the domestic pig. *Genome Research* **8**, 826–833 (1998).

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S. & Van Tassell, C. P. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One*. 4:e5350 (2009).Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology* **12**, R112 (2011).

Mayer, T. C. The migratory pathway of neural crest cells into the skin of mouse embryos. *Developmental Biology* **34,** 39-46 (1973).

Miller, J. R., Koren, S., & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327 (2010).

Noonan, J. P. & McCallion, A. S. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics* **11,** 1-23 (2010).

Olson, T. A. The genetic basis for piebald patterns in cattle. *Journal of Heredity* **72**, 113-116 (1981).

Olson, T. A. Genetics of Colour Variation. In *The Genetics of Cattle*. pp 33–54. CAB International, Wallingford, England. (1999).

Pausch, H., Wang, X., Jung, S., Krogmeier, D., Edel, C., Emmerling, R., Götz, K. U. & Fries, R. Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. *PLoS One* **7**, e36346 (2012).

Phillippy, A. M., Schatz, M. C., Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology* **9**, R55 (2008).

Ramey, H. R., Decker, J. E., McKay, S. D., Rolf, M. M., Schnabel, R. D., & Taylor, J. F. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics* **14**, 382-400 (2013).

Rawles, M. E. Origin of pigment cells from the neural crest in the mouse embryo. *Physiological Zoology* **20,** 248-265 (1947).

Reinsch, N., Thomsen, H., Xu, N., Brink, M., Looft, C., Kalm, E., Brockmann, G. A., Grupe, S., Kühn, C., Schwerin, M., Leyhe, B., Hiendleder, S., Erhardt, G., Medjugorac, I., Russ, I., Förster, M., Reents, R. & Averdunk, G. A QTL for the degree of spotting in cattle shows synteny with the KIT locus on chromosome 6. *The Journal of Heredity* **90**, 629–634 (1999).

Rife, D. C. Color and horn variations in water buffalo: the inheritance of coat color, eye color, and shape of horns. *Journal of Heredity* **53**, 239-246 (1962).

Ritz, L. R., Glowatzki-Mullis, M. L., MacHugh, D. E., & Gaillard, C. Phylogenetic analysis of the tribe Bovini using microsatellites. *Animal Genetics* **31**, 178-185 (2000).

Rubin, C. J., Megens, H. J., Martinez Barrio, A., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C. B., Archibald, A. L., Fredholm, M., Groenen, M. A. & Andersson, L. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19529–19536 (2012).

Russell, W. C., Brinks, J. S., & Kainer, R.A. Incidence and heritability of ocular squamous cell tumors in Hereford cattle. *Journal of Animal Science* **43**, 1156–1162 (1976).

Schmutz, S. M. Genetics of coat color in cattle. In *Bovine Genomics*. pp 20-33. Wiley-Blackwell, Oxford, United Kingdom. (2012).

Schwartz, D. C., Li, X., Hernandex, L. I. Ramnarain, S. P., Huff, E. J., & Wang, Y. K. Ordered restriction mapping of Saccharomyces cerevisiae chromsomes constructed by optical mapping. *Science* **262,** 110-114 (1993).

Searle, A. G. Comparative genetics of coat colour in mammals. Logos Press Ltd. London, England (1968).

Seitz, J. J., Schmutz, S. M., Thue, T. D. & Buchanan, F. C. A missense mutation in the bovine MGF gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle. *Mammalian Genome* 10:710-712 (1999).

Seo, K., Mohanty, T. R., Choi, T. & Hwang, I. Biology of epidermal and hair pigmentation in cattle: a mini-review. *Veterinary Dermatology* **18**, 392–400 (2007).

Ward, J. K. & Nielson, M. K. Pinkeye (bovine infectious keratoconjunctivitis) in beef cattle. *Journal of Animal Science* **49** 361–366 (1979).

Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**, 1586-1592 (2009).

Zhou, S., Schwartz, D. C., & Medrano, J. F. Bovine optical map. Unpublished data.

Zimin, A., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., Yorke, J. A. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-2677 (2013).