University of Alberta

An evaluation of linkage disequilibrium and population structure in domestic cattle

by

Stephanie Dawn McKay ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Animal Science

Agriculture Food and Nutritional Sciences

Edmonton, Alberta
Fall 2007

Your file Votre référence
ISBN: 978-0-494-33025-8
Our file Notre référence
ISBN: 978-0-494-33025-8

ABSTRACT

We have completed the construction of radiation hybrid (RH) maps, linkage maps and linkage disequilibrium maps of all the bovine autosomes. A new method for high-throughput RH mapping was developed and successfully utilized to map 4690 sequence tagged site markers on the Roslin-Cambridge 3000 rad RH panel. A subset of the markers that were RH mapped were also used to construct linkage maps and linkage disequilibrium maps. Linkage disequilibrium was assessed using r^2 among all pairs of syntenic markers within each surveyed breed from the *Bos taurus* and *Bos indicus* subspecies. Breeds included Angus, Brahman, Charolais, Dutch Black and White Dairy, Holstein, Japanese Black, Limousin and Nelore. Approximately 2670 markers spanning the entire bovine autosomal genome were used. It was found that r^2 declines rapidly at approximately 0.5 Mb in all breeds. These findings imply that the optimal number of markers needed for a whole genome association study is in the range of 30,000-50,000 loci. Finally, we have used this data set to study genetic structure between these eight breeds. Analysis of all eight breeds partitions only the divergence between taurus and indicus with statistical significance. The divergence between taurus and indicus breeds dominates that among all of the taurine breeds. Once indicus was removed from the analysis, Japanese Black cattle were determined to be divergent from all other taurine breeds.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

LD- linkage disequilibrium

Mbp – megabases

Kb - kilobases

SNP- single nucleotide polymorphism

WGRH$_{3000}$- 3000 rad bovine-hamster whole genome radiation hybrid panel

WGRH- whole genome radiation hybrid

RH – radiation hybrid

Btau 2.0 – second version of the bovine whole genome sequence assembly

SSC- *Sus scrofa* chromosome

BAC – bacterial artificial chromosome

STR- short tandem repeats

HSA- *Homo sapien* autosome

AI- artificially inseminated

OPA – oligo pooled assay

HSB- homologous synteny block

BTA- *Bos taurus* autosome

cM- centiMorgans

QTL – quantitative trait loci

MAF- minor allele frequency

mtDNA – mitochondria DNA

AMOVA- analysis of molecular variance

CHAPTER 1

General Introduction

1.1 Introduction

Whole genome association studies the use of polymorphic markers located across a genome to identify chromosomal segments associated with economically important traits. Measuring the extent of linkage disequilibrium (LD) across a genome is key to determining the optimal number of markers needed for execution of whole genome association studies. Linkage disequilibrium is the non random association of alleles at linked loci and has the potential to be used to find associations between haplotypes and phenotypes. Specific haplotypes, particular combination of alleles in a specific region of a chromosome (Kwok & Gu, 1999), can be responsible for the exhibition of phenotypes of interest. Linkage disequilibrium has immense potential when an uncommon haplotype arises in an affected individual. However, the length over which LD extends influences the number of markers needed for a whole genome association study. If LD is found to extend over very long distances then fewer markers would be needed for a whole genome association study. However, the region of the chromosome found to be associated with a particular trait of interest would be of considerable length. Since the extent of LD is somewhat variable across genomes, the number of markers and the spacing of markers needed to accurately measure LD will vary across a genome.

A majority of the characterization of LD in the bovine genome has involved microsatellite studies performed with dairy cattle. Farnir (2000) used 284 genome wide microsatellites to measure both syntenic and non-syntenic LD in a population of Dutch Black and White dairy cattle. They found LD to extend up to several tens of centimorgans (cM). While successive studies did utilize small numbers of animals and microsatellites (Tenesa *et al.*, 2003; Vallejo *et al.*, 2003), they did report findings similar to Farnir. Likewise, similar findings of long range LD has been reported for sheep and pigs (McRae *et al.*, 2002; Nsengimana *et al.*, 2004).

The extent of LD work in beef cattle has been limited to a single publication (Odani *et al.*, 2006). Linkage disequilibrium was characterized and compared between Japanese Black and Japanese Brown beef cattle. Genome wide LD analysis was performed with 246 autosomal microsatellite markers in Japanese Black cattle and 156 autosomal microsatellites in Japanese Brown cattle. In both breeds, significant LD was observed for marker pairs separated by less than 40 cM. While both of these breeds are beef cattle, they are selectively bred for different economically important traits. The variations that Odani (2006) found in LD between breeds was attributed to variations in selective breeding programs.

The objective of this thesis was to generate tools that could be used to produce more efficient whole genome association studies. This thesis is comprised of three genome wide studies of the autosomal bovine genome. The first study details implementation of a new methodology for high throughout mapping and successive generation of radiation hybrid and linkage maps for the autosomal bovine genome. This methodology was used to generate maps that were used to determine accurate ordering, necessary for the successive chapters. The second study characterized the extent of linkage disequilibrium in eight breeds of cattle. The final study explores population structure and diversity within and between the same eight breeds of cattle. Implementation of the conclusions from each chapter could result in whole genome association studies that are more accurate and efficient.

1.2 Genome wide radiation hybrid and linkage maps

High density linkage and physical maps are extremely labor intensive to construct. Generation of genotypes, for linkage mapping or typing a radiation hybrid panel for whole genome mapping requires many thousands of analyses. In order to improve the throughput and accuracy of gene mapping approaches we have developed a new methodology using instrumentation designed for large scale single nucleotide polymorphism (SNP) analysis. An Illumina® BeadStation was used to simultaneously type large numbers of markers on the Roslin-Cambridge 3000 rad bovine-hamster whole genome radiation hybrid panel ($WGRH_{3000}$). In five multiplex reactions, 4690 sequence tagged site markers were successfully typed on the $WGRH_{3000}$ panel DNA. These sequence tagged sites harbored SNPs that were developed as a result of the bovine genome sequencing initiative. Typically, the most time consuming and expensive part of creating high density RH-maps is genotyping the markers on the RH panel when using conventional approaches. Using the high-throughput approach described in Chapter 3, a high-density map of an entire mammalian genome can be generated in a matter of days. With the method described in this thesis we have developed bovine WGRH and linkage maps with direct comparison to the bovine whole genome sequence assembly (Btau 2.0).

1.3 A bovine whole genome linkage disequilibrium map

Generation of low resolution LD maps with evenly spaced single nucleotide polymorphisms across the genome has the potential to provide an enhanced understanding of the structure of LD across a genome. Chapter 4 characterizes whole genome linkage disequilibrium in eight breeds of cattle. This resource can be used in studies to associate genetic variation with economically important traits while simultaneously allowing further analysis of long range linkage disequilibrium in cattle. Linkage disequilibrium was measured using the squared correlation coefficient between locus allele frequencies, r^2, among all pairs of markers on each chromosome within each breed of cattle from *Bos taurus* and *Bos indicus* subspecies, including Angus, Brahman, Charolais, Dutch Black and White Dairy, Holstein, Japanese Black, Limousin and Nelore. Approximately 2670 markers spanning the entire bovine autosomal genome were utilized. We found that r^2 declines rapidly at approximately 0.5 megabases (Mb) in all breeds. Our results, based on much larger marker sets and across eight breeds of cattle are contradictory to other previously published reports. These findings imply that the optimal number of markers needed for an optimal whole genome association study is 30,000-50,000.

1.4 An assessment of population structure with eight breeds of cattle

Population structure and breed diversity analyses have provided insight into the origin and evolution of cattle. We used a high-density single nucleotide polymorphism panel to examine population structure and diversity among eight breeds representing both subspecies of cattle. To examine population structure, we utilized the linkage model in the program STRUCTURE and we also used Fst estimates to construct a neighbor-joining tree to represent the phylogenetic relationship among these breeds. In an analysis of the entire data set, STRUCTURE initially differentiated the *Bos taurus* and *Bos indicus* breeds but did not partition any of the *Bos taurus* breeds. By removing the *Bos indicus* breeds from the analysis, beef versus dairy and European versus Asian substructures were detected among the *Bos taurus* breeds.

1.5 Literature Cited

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., and Georges M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10:** 220-7.

McKay S. D., Schnabel R. D., Murdoch B. M., Aerts J., Gill C. A., Gao C., Li C., Matukumalli L. K., Stothard P., Wang Z., Van Tassell C. P., Williams J. L., Taylor J. F., and Moore S. S. (2007). Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Anim Genet* **38:** 120-5.

McRae A. F., McEwan J. C., Dodds K. G., Wilson T., Crawford A. M., and Slate J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160:** 1113-22.

Nsengimana J., Baret P., Haley C. S., and Visscher P. M. (2004). Linkage disequilibrium in the domesticated pig. *Genetics* **166:** 1395-404.

Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., and Sasaki Y. (2006). Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Anim Genet* **37:** 139-44.

Tenesa A., Knott S. A., Ward D., Smith D., Williams J. L., and Visscher P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *J Anim Sci* **81:** 617-23.

Vallejo R. L., Li Y. L., Rogers G. W., and Ashwell M. S. (2003). Genetic diversity and background linkage disequilibrium in the North American Holstein cattle population. *J Dairy Sci* **86:** 4137-47.

CHAPTER 2

Literature Review

2.1 Introduction

This thesis is comprised of a structural genomics analysis of the bovine autosomal genome. New high throughput methodology was utilized in construction of radiation hybrid (RH) and linkage maps for all 29 bovine autosomes. Information obtained from construction of RH and linkage maps was used for construction of a whole genome linkage disequilibrium (LD) map. A comparative whole genome linkage disequilibrium map was constructed comprised of 2701 SNP markers in eight breeds of cattle. The implications of the level of LD present across the cattle genome is discussed.

When specific alleles at two different loci on the same chromosome are found together more often than would be expected, these are said to be in linkage disequilibrium. We can use LD to aid in narrowing down regions of chromosomes associated with economically important traits. If a population that displayed a trait of interest were genotyped with markers evenly spaced throughout the genome, we could measure the LD between every pair of marker on each chromosome in each member of the population. If the causative mutation for the economically important trait is in disequilibrium with a nearby locus, then the region of the chromosome associated with the economically important trait would span this locus. Of major concern is the situation where the loci in disequilibrium are incorrectly mapped, then the region of association would be inaccurately identified.

In order to correctly characterize LD accurate map locations are needed. There are two primary types of maps, physical and genetic, and examples of each type of mapping are radiation hybrid and linkage maps, respectively. Generally, linkage maps are advantageous for overall ordering of loci along a chromosome while radiation hybrid maps are often better for fine mapping closely linked loci (Mellersh *et al.*, 2000). Radiation hybrid maps are generated by irradiating a donor cell of interest with a lethal amount of radiation, thus breaking chromosomes into smaller pieces. Cell lines containing broken donor chromosomes are then fused with a recipient rodent cell line that is deficient for a selectable marker. Selection events then target and remove both rodent and donor specific cell lines, effectively leaving radiation hybrids which contain cattle chromosomal fragments in addition to the fragment from the cattle genome containing the selectable marker (Figure 2.1). The hybrid cell lines are grown individually, cloned and their DNA is extracted. Low resolution RH panels have 90-94 hybrids in addition to a positive and negative control. Traditional means of RH mapping consist of duplicated PCRs with the RH panel and primers from your marker of interest. The presence or absence of the marker in each hybrid is noted and maps are built accordingly. Conversely, linkage maps are generated by genotyping families with polymorphic markers and map distances are generated by measuring recombination events.

2.2 Measures of linkage disequilibrium

Allele frequencies are utilized in the three common calculations used to measure linkage disequilibrium, D, D' and r^2. D is a measure of the magnitude of linkage disequilibrium. More specifically, D is the difference between the observed frequency of a two-locus haplotype and the frequency it would be expected to show if the alleles were segregating at random in a population (Ardlie *et al.*, 2002).

$$D = P_{AB} - P_A \times P_B$$

Two loci, A and B, are represented. Each locus is assumed to have two alleles. These two alleles can result in four possible haplotypes. Therefore, D can be calculated for each of the four possible haplotypes. P_{AB} represents the observed frequency of the haplotype. P_A is the allele frequency at the first locus. P_B is the allele frequency at the second locus. For these two loci, you would have four possible haplotypes. If D differs significantly from zero then linkage disequilibrium is present. However, the degree of LD (D_t) between loci is dependent on recombination (θ) and the amount of time (t) that has elapsed since the mutation has arisen: $D_t = D_0 (1-\theta)^t$. Therefore, D will generally tend to be smaller when two loci are further apart and will decrease through time as a result of recombination.

Another common measure of LD is *D'*. *D'* is obtained by dividing D by its maximum possible value (D_{max}) given the allele frequencies at the two loci (Ardlie *et al.*, 2002):

$$D' = D/D_{max}.$$

If D' equals 1 then there has been no recombination between the two loci and LD is complete. If D' is less than 1 then the two loci have been interrupted by recombination, at some point in time. As D' approaches zero, LD becomes nonexistent. D' is strongly dependent on sample size which makes comparisons difficult. When loci have common alleles, D' is strongly inflated in small samples but when loci have rare alleles, D' is even more inflated. Furthermore, D' does attempt to avoid dependence on allele frequency but it is not independent of allelic frequencies (Lewontin, 1988). Therefore D' should not be used to compare LD between studies or to measure the extent of LD. D' should primarily be used as a means to indicate if recombination has occurred.

A third common measure of linkage disequilibrium is the correlation between two loci, r^2. The correlation is determined by dividing D by the product of the four allele frequencies at the two loci (Ardlie *et al.*, 2002)

$$r^2 = D^2/p_1p_2q_1q_2.$$

The squared correlation coefficient between locus allele frequencies, r2, is strongly influenced by the order in which the mutations arose and not necessarily the physical distance between loci (Daly *et al.*, 2001). This measure quantifies the amount of information that can be inferred about one locus from another (Pritchard & Przeworski, 2001; Zhao *et al.*, 2005) and can be used to estimate the number of loci needed for association studies (Kruglyak, 1999; Pritchard & Przeworski, 2001).

Early LD publications utilized D' to characterize their LD findings. Only recently with the completion of multiple mammalian genome sequences has r^2 been primarily used to characterize LD. When both D' and r^2 have been used in LD studies, substantial variations between the measures have been observed. It is believed that allele frequencies play a role in this (Boyles *et al.*, 2005). When two markers in linkage disequilibrium have equal allele frequencies, D' will equal r^2. However, when allele frequencies between two markers are not equal, r^2 will not reach one and D' values will be inflated. An example of this might be when one of the makers is a rare allele. The use of r^2 is therefore now a more common measure of LD in part because it is not prone to errors resulting from difference in population size and allele frequencies.

2.3 Factors affecting linkage disequilibrium

Linkage analysis is used to identify genes that are believed to be involved with quantitative traits by investigating the transmission of alleles in families known to possess this trait. Using single nucleotide polymorphisms (SNPs), naturally occurring variations in DNA sequences, to follow the allelic transmission between generations, we can search for association between polymorphisms and quantitative traits. However, linkage disequilibrium mapping has the potential to increase the mapping resolution of quantitative traits (Jorde, 2000). Linkage disequilibrium occurs when two or more alleles occur together more frequently than would be expected given the distance between these alleles. Quantifying the extent of LD in a genome is a necessary first step in order to determine the optimal number of markers required for mapping quantitative traits by linkage disequilibrium.

Meiotic recombination and mutation are the primary recurrent processes affecting the level of linkage disequilibrium. When loci are far apart, recombination is more likely to occur between the loci and thus breakdown haplotypes resulting in a decrease in linkage disequilibrium. In particular, if we could predict high or low rates of recombination, hot and cold spots, then we could more accurately predict and compare gene location and order as well as obtain a greater understanding for the molecular mechanisms underlying recombination (McVean *et al.*, 2004; Petes, 2001). Theoretically, recombination hot spots are formed when one of three different histone modifications yields a double-stranded break of the DNA that in turn leads to the generation of recombination through employment of Holliday junctions. The Holliday model of crossing over explains crossovers and accounts for gene conversion. Holliday junctions occur when single strands of DNA in each of two paired homologous chromosomes break at the same site. The DNA unwinds at both sites and pairs with their respective complementary sequences in the other chromosome. The segment of double stranded DNA containing strands that originate from different DNA molecules is called heteroduplex DNA. Heteroduplex DNA is formed as the Holliday junction migrates along the chromosome. An estimated 30-50,000 hot spots are thought to exist throughout the human genome (Myers *et al.*, 2006), with some 25,000 currently characterized. This corresponds to one hot spot located, on average, every 50-100 kb. Furthermore, the average amount of recombination in the human genome is estimated to be 0.075 cM or 1 crossover per every 1300 meioses.

Two levels of regulation determine the frequency of recombination rates, chromosomal structure and DNA sequence. With respect to chromosome structure we know that females display a higher recombination rate than males. Recombination near the centromere is greatly reduced in males, however, recombination near the telomere is greatly increased in both sexes, especially males. DNA sequence, regions containing high amounts of G and C content have shown a weak association to increased recombination rates. Only recently has evidence been presented to suggest that a particular sequence is associated with recombination hotspots (Myers *et al.*, 2006). The sequence CCTCCCT may be associated to hotspot occurrence in the human genome. If SNPs occurred within this particular sequence then recombination activity would be strongly reduced. Characterization of human recombination hotspots involved use of some 1.6 million SNPs until high density analysis is possible we will be unable to detect hotspots in other mammalian species. Only now is a similar quantity of SNPs becoming publicly available in cattle. With the recent advances in high-throughput genotyping methodologies, a project of this magnitude is now feasible in other species such as dogs and cattle.

Additional factors are known to drive the extent and distribution of linkage disequilibrium (Ardlie *et al.*, 2002). Genetic drift is the random change in allele frequencies within a population over time. It can result in a reduction in the frequency of particular haplotypes from the population thus increasing linkage disequilibrium. The results of this are more often seen in small populations where random mating still results in inbreeding and there is a chance that individuals share alleles that are identical by descent. When two alleles in an individual are identical by descent, both alleles can be traced back to a single copy in a distant ancestor. In inbred populations, we expect common chromosomal segments to be shared by individuals. The result is an increase in haplotype sharing and a subsequent increase in linkage disequilibrium. Aspects of population structure, namely inbreeding, lowers diversity levels and increases LD.

While several reasons exist that can explain long blocks of LD, a possible reason is an extreme founder effect or a population bottleneck (Reich *et al.*, 2001). The population may have been so small that a few ancient haplotypes have produced the majority of the current haplotypes. There are two means by which natural selection affects the extent of linkage disequilibrium in a genome, hitchhiking and epistasis. Hitchhiking is when a neutral allele is rapidly swept to a high frequency because it is linked to a selectively favored allele. This builds up LD between alleles through selection of rare genetic variants. When the rare genetic variant is selected for, alleles at linked loci tag along and will increase in frequency. Such direct positive selection is otherwise known as a selective sweep. Over time, as the alleles become fixed, LD between the rare variant and the hitchhiking allele will decrease. Epistasis is when one allele masks the phenotypic effects of another allele. Through epistatic selection, LD is maintained through association of particular alleles at different loci.

Recent admixture events can result in substantial linkage disequilibrium over long distances. Admixture is the interbreeding between genetically differentiated populations. Initially, LD is relative to the variation in allele frequency between the admixed populations. Furthermore, in recently admixed populations LD is unrelated to the distance between the markers. In successive generations, LD between unlinked markers will disperse while LD between nearby markers will be slowly broken down by recombination over time.

2.4 Linkage disequilibrium in cattle

Genome wide linkage disequilibrium in the bovine has been performed primarily with dairy cattle. The most noteworthy publication used a panel of 284 genome wide microsatellites with a Dutch Black and White Dairy cattle population that resulted in 276,048 genotypes (Farnir *et al.*, 2000). Genotypes for 581 maternal gametes were used to estimate LD using Lewontin's normalized D' measure. The extent of LD was initially measured between syntenic markers, that is markers located on the same chromosome but that are not necessarily linked. The results showed high levels of LD not only between closely linked markers but for markers located as much as 40 cM apart. To confirm these findings a second set of 627 cattle, believed to be more representative of the Dutch Black and White general population, were genotyped for eight microsatellites located on different autosomes. In addition, 175 of these 627 cattle were further genotyped for an added nineteen microsatellites, sixteen of which were located on chromosome 14 while the other three microsatellites were located on chromosome 6. For markers with an inter-marker distance less than 5 cM, D' averaged 0.46 while markers greater that 30 cM apart produced an average D' value of 24%. Moreover for nonsyntenic markers, the average D' measurement was 20%. These values confirmed the initial study.

This study is important because it initially characterized long range linkage disequilibrium in cattle as extending over several tens of cM. A finding of linkage disequilibrium extending up to 40 cM is significant, especially when compared to what has been measured in human where LD has been shown to extend up to a maximum of 1 cM. The findings in cattle are however not surprising, given the degree of relatedness between dairy cattle worldwide. The effective population size of dairy cattle breeds is estimated to be less than 50 animals (Boichard *et al.*, 1996). This small effective population size can result in an increase in inbreeding. Inbreeding in turn results in a loss of haplotypes and an increase in linkage disequilibrium as more of any individual's genome within the population will be identical by descent which will cause an increase in LD. Linkage disequilibrium extending over great distances implies that fewer markers will be needed to produce linkage disequilibrium maps in cattle and strong associations between DNA markers and QTLs could be seen over greater distances.

To date Farnir *et al*, (2000) is still one of the most comprehensive LD studies in cattle because the microsatellites used covered all of the bovine autosomes. In addition a large number of animals were used. As is the case with almost every bovine LD paper published to date, microsatellites were used and D' was as a measure of LD. At the time the Fanir et al. (2000) paper was published, few bovine SNPs had been characterized so microsatellites were used instead.

SNPs have now become the marker of choice for many genetics studies including LD. SNPs have lower error rates, higher call rates and are more cost effective than microsatellites. This is coupled with high throughput capabilities (Goode & Jarvik, 2005) enabling many thousands of SNPs to be analyzed simultaneously. The study described in chapter 4 of this thesis was performed using SNPs. In order to directly compare results between LD studies, one has to genotype the same set of animals and some of the same markers. This is not always possible because it may not be cost effective or the DNA resources may have been depleted. Therefore, when entering into a genome wide LD study using SNPs and multiple breeds of cattle we knew that we would not be able to directly compare our results to Farnir *et al* (2000) due to the differences in the markers analyzed. Nevertheless it was important to include a subset of the animals used by Farnir *et al.* (2000)for comparative purposes with the other breeds of cattle utilized in our study.

In a second genome wide LD study with dairy cattle, highly heterozygous Holstein bulls as unrelated as possible, were used to quantify the level of genetic diversity in United States Holstein cattle (Vallejo *et al.*, 2003). In doing so, 23 Holstein bulls were genotyped with 54 microsatellite loci that spanned most of the bovine autosomal genome. The investigators found extensive LD in the United States Holstein population and agree with Farnir's findings of LD between syntenic loci extending over several tens of centimorgans.

Generally, using unrelated animals or trios are the two approaches used when setting up pedigrees for use in LD studies. Trios usually consist of DNA from the sire, dam and an offspring. Using animals as unrelated as possible yields a higher number of unique chromosomes, and has the potential to yield a more global representation of the breed. This is seen in the Vallejo (2003) study. On the other hand, trios, which consist of father, mother and an offspring, are used so that phase can be determined. Phase is the determination of which allele came from which parent. However when working with livestock species, large half sibling families are often readily available. We can take advantage of this by genotyping small families consisting of a grandsire, sire and multiple half sib offspring. This will enable us to obtain phase. If the small families are as unrelated as possible and if we extract and use only the maternal haplotypes for analysis, our results will still represent a substantial number of unique chromosomes and have the potential to be representative of the breed as a whole.

Inbreeding in the US Holstein population has continued to increase over the last several decades (Funk, 2006). Since artificial insemination became widespread in the early 1950's, the majority of the historical recombination events in the United States Holstein population can be traced back to four original sires (Ashwell & Van Tassell, 1999). Since LD is based on historical recombination within families, genotyping only 23 animals as carried out by Vallejo (2003) can still result in a global representation of the breed, as long as the appropriate animals were genotyped. However, with the US Holstein population, a common misunderstanding is that genotyping the top producing sires will trace back to the four original sires in which a majority of all historical recombination events are seen. This is not the case. If the top producing sires were chosen for genotyping, the four sires that represent a majority of the historical recombination events would not all be represented. This is because the four sires that represent a majority of the historical recombination events are not all top producing sires. Therefore, genotypes would represent the top producing sires but they would not comprise a global representation of the breed. In addition, Vallejo utilized only 54 microsatellites. There are multiple chromosomes where only one marker was used. Utilizing additional markers would result in the development of a more comprehensive examination of LD and provide greater insight into genetic diversity in the US Holstein population.

The extent of linkage disequilibrium in the United Kingdom dairy cattle population was measured in part to determine reasonable LD mapping methods and required marker density for effective LD mapping (Tenesa *et al.*, 2003). Genotyping with 50 Holstein bulls was performed with six markers located on BTA2 and an additional seven markers located on BTA6. The results differ from Farnir in that the average Given that D' value was 44%, the D' measurement for nonsyntenic loci was estimated to be 39%. The investigators reported significant LD only for distances smaller than 10.3 cM and significant associations were never found between nonsyntenic loci. D' is dependent upon sample size, differences between findings might be attributed to this parameter.

Tenesa et al. (2003) also claimed the ability to infer haplotypes without pedigree information. Essentially the authors suggest that statistical methods are sufficient to enable a smaller number of animals to be genotyped. The alternative approach is to genotype large families, in part for accurate estimation of haplotypes. Additionally they attribute the variation between their findings and Farnir's to sample size and relatedness of the animals genotyped. If they had used available pedigree information they could have selected appropriate animals to genotype that would have resulted in a comprehensive representation of the population and more characteristic measures of LD.

More recently an LD map of the bovine X chromosome was published (Sandor *et al.*, 2006). Observations in humans have typically found less polymorphism and higher LD on the X chromosome (Schaffner, 2004). However, in cattle, there is evidence to suggest that the measure of LD on the X chromosome is considerably higher than previously expected. They genotyped 929 bulls from 22 paternal half sib Holstein-Friesian families with 22 X chromosome specific microsatellites and 3 microsatellites located in the pseudoautosomal region. Measures of LD were reported as r^2 being 1.75 times larger compared to autosomal measures. Generally, less polymorphism on the X chromosome is thought to be due to higher genetic drift, reduced mutation rate in females compared to males and purifying selection due to male hemizygosity. Here, the higher level of LD on the X chromosome is thought to result from higher genetic drift and contributions from other undetermined factors.

Only recently has LD been measured in beef cattle (Odani *et al.*, 2006). Two breeds of Japanese beef cattle were utilized in an attempt to characterize LD. One Japanese black sire and his 162 half-sib progeny were genotyped with 246 autosomal microsatellite markers. Likewise an additional 406 half-sib Japanese brown cattle were genotyped with 156 autosomal microsatellites. The average number of alleles observed was 6.3 for

Japanese black cattle and 6.7 for Japanese Brown cattle. Map locations from the Shirakawa-USDA map (Ihara *et al.*, 2004) was used with a total map length of 2820 cM (Haldane) and an average marker interval of 12.4 cM for Japanese Black cattle. The total map length for Japanese brown cattle was 2795 cM with an average inter-marker interval of 20.3 cM. For syntenic markers, mean values of D′ were 0.163 for Japanese Brown and 0.251 for Japanese Black. In both breeds, significant LD was observed frequently for marker pairs < 40 cM. Generally, significant LD was observed more frequently in Japanese brown cattle as compared to black. Approximately 5.5% of marker pairs studied in black cattle and 10.8% of marker pairs studied in brown cattle showed significant amounts of nonsyntenic LD.

Thus far, this is the only comparative LD study that utilized more than one breed of cattle. Even though these two breeds originated from the same indigenous cattle, their selective breeding pressures have deviated and Japanese Black cattle have been predominantly bred for marbling while the Japanese Brown cattle are known for their larger mature size and faster growth rate. The variation in average measures of LD seen between these two beef cattle breeds is attributed to difference in selective pressures. No study to date has compared LD between dairy and beef breeds of cattle. Indeed given the large differences in effective population sizes expected, not only between dairy and beef cattle, but between the more popular and rarer breeds within beef cattle such a study would be most informative.

2.5 Linkage disequilibrium in other animal species

2.5.1 Linkage Disequilibrium in Sheep

Two data sets were used to explore linkage disequilibrium in sheep chromosomes 1-10 (McRae *et al.*, 2002). In data set one, 2 sires were crossed with 186 dams resulting in 276 progeny and genotyping was performed with 90 microsatellites. The average D′ of syntenic markers was 0.211 while average D′ of nonsyntenic markers was 0.196. Data set two was used in an attempt to explore LD between tightly linked markers and was comprised of 14 sires crossed with approximately 400 dams resulting in 482 offspring. These were typed with 13 microsatellites and 13 RFLPs on sheep chromosome 6. On average D′ was greater for syntenic markers less than 60 cM in data set 2 compared to data set 1.

McRae *et al*., (2002) aimed at assessing the extent of LD in sheep. Linkage disequilibrium was found to extend over tens of cM. Concerns regarding their approach to haplotype reconstruction and limited sample size were raised as explanations for a potential upward bias in their measures of LD. However, excessive LD was attributed to recent admixture, small effective population size and intense selective pressures. These findings are similar to those found in cattle (Farnir *et al.*, 2000). This work confirmed that long range LD in farm animal species might be quite extensive compared to long range LD findings in humans.

2.5.2 Linkage Disequilibrium in Pigs

Characterization of linkage disequilibrium in the pig genome originated with five populations of pigs and up to 15 microsatellites per population from regions of pig chromosomes (SSC) 4 and 7 (Nsengimana *et al.*, 2004). Mean D′ for all 5 populations was reported as 0.290 for SSC4 and 0.409 for SSC7. Furthermore, the decline of LD as a function of distance occurred over shorter genetic distances on SSC4 compared to SSC7. The authors suggest that the difference in mean LD between chromosomes could be attributed to a selection effect. Unlike cattle (Farnir *et al.*, 2000), significant levels of LD were not found between nonsyntenic marker. The authors suggest that the variation in effective population size, number of haplotypes and sample size could result in dissimilar results concerning use of nonsyntenic markers.

Nsengimana (2004) mentions effective population size as a possible cause for dissimilar findings in LD studies between species. They cite that the estimated effective population size of Holstein-Friesian cattle is less than 50 (Boichard *et al.*, 1996) while the effective population size for their five populations of pigs range from 60 to 300. All of these studies have used limited numbers of microsatellite markers and used D′ to characterize LD. Today with the availability of high density SNPs arrays and the use of r^2 to measure LD, we can more accurately estimate effective population size (*Ne).* Given the recombination fraction (θ) and the number of haplotypes (*n*), we can use r^2 to estimate the effective population size (Ne):

$$r^2=1/(1+4Ne\theta)+1/n$$

However, estimating effective population size using r^2 in cattle, will require the use of tens of thousands of SNPs. McKay et al (in review) estimates that the approximate number of SNPs required for a whole genome LD map in cattle and subsequently a more effective measure of effective population size will require some 30,000-50,000 SNPs.

Further characterization of LD in pigs was studied using 33 and 44 unrelated individuals from commercial pig populations designated A and B (Harmegnies *et al.*, 2006). SSC15 was studied with 29 microsatellites with inter-marker distances ranging from 0.1 to 10.5 cM. SSC2 was included in the study so that nonsyntenic loci could be tested for LD. Five microsatellites on SSC2p all located within a 1 cM range near IGF2 were genotyped. Linkage disequilibrium was seen between syntenic markers up to 40 cM in lines A and B. In line B, LD was detected up to 60 cM. However, non-syntenic LD was not found in either line. r^2 averaged 0.15 and 0.19 for markers less than 1 cM apart in lines A and B on SSC15. For distances greater than 10cM, r^2 averaged approximately 0.05 in both lines. SSC2 r^2 values averaged 0.48 and 0.35 in lines A and B. However, inter-marker distances on SSC2 were much smaller, compared to SSC15. Effective population size was estimated using r^2 values on SSC15. Results suggest a bottleneck occurred some 20 generations ago that reduced the effective population size from thousands to hundreds.

Harmegnies *et al.* (Harmegnies *et al.*, 2006) is one of the first studies to measure LD using r^2 and to observe low r^2 values in farm animal species. In order to test the idea that the low measures could be due to the higher mutation rate in microsatellite markers, the authors reexamined available data from BTA14. At least one microsatellite and multiple SNPs all located within the same bacterial artificial chromosome (BAC) were used to measure LD. Separate measures of LD were obtained from microsatellites and SNPs. They found that the inter-marker r^2 values from the microsatellites did not appear to be lower than the inter-marker r^2 value of the SNPs. Therefore, they concluded that the low r^2 values found on SSC15 and 2 were not attributed to use of microsatellites versus SNP markers. Their finding an average r^2 value of 0.1 suggests that a sample size 10X larger will be needed for association studies.

2.5.3 Linkage Disequilibrium in Dogs

Dogs are a particular species of interest because they are from the same clade as cattle and the dog genome has been sequenced. Before the dog genome had been sequenced, five 5 Mb regions of the canine genome were resequenced for SNPs and LD was measured (Sutter *et al.*, 2004). A total of 97 dogs representing five breeds from unique phylogenetic origins had 5 Mb intervals of canine chromosomes 1, 2, 3, 34 and 37 resequenced to identify SNPs with minor allele frequencies of at least 0.2. Five regions of each chromosome were resequenced, averaging 2.8 sequence reads per region and spanned 775bp to 96 kb. LD was then characterized in these regions and D' was found to fall at 0.5 between 0.4 - 3.2 Mb. The upward bias of the D' measure that was expected due to limited sample size in each breed was reduced by removing rare alleles. Long range LD was expected to be found in dogs for many of the same reasons that we expected to find it in cattle, admixture, use of popular sires and population bottlenecks.

Furthermore, in an attempt to characterize haplotype diversity, haplotypes were inferred for each region of each breed and an average of 12.4 haplotypes was found per region. Additional analysis reported that an average of 2.7 haplotypes accounted for 80% of chromosomes when averaging over all breeds and regions. A high degree of haplotype sharing between breeds is reported. An average of 63% of the chromosomes within each breed carried haplotypes shared between both breeds being compared. However, of the total 124 haplotypes identified, only 10 were found in all 5 breeds. The long range LD reported here suggests that a limited number of markers would be needed for a whole genome association study. The high degree of haplotype sharing between breeds indicates that it may be possible to identify a common set of SNPs that could be studied in many different breeds.

The recent release of the canine genome sequence (Lindblad-Toh *et al.*, 2005) provides access to comprehensive information from an animal within the same clade as cattle. Like cattle, dogs have been selectively bred resulting in periodic population bottlenecks. The canine genome sequencing initiative has yielded a 7.5X draft sequence with 99% coverage of the euchromatic genome and a SNP map comprised of some 2.5 million SNPs resulting in an average density of approximately 1 SNP per kb. Linkage disequilibrium was examined in ten randomly selected 15Mb regions of the canine genome. As expected, the findings showed LD to extend over several Mb within breed. However, LD across breeds extends only over tens of kb. Likewise, haplotypes are shown to extend over long distances. Haplotypes as long as 100 kb are shown to be shared across breeds but with variable frequency. The linkage disequilibrium patterns are an indication of the major bottlenecks in dog history, early domestication and recent breed creation. While the haplotype patterns are further indication that genetic risk factors may be shared across breeds.

2.5.4 Linkage Disequilibrium in Humans

Linkage disequilibrium has been analyzed throughout the human genome (Reich *et al.*, 2001). The authors identified and genotyped 272 high frequency coding SNPs with minor allele frequencies $\geq .35$ in a small panel of 44 unrelated European descendents from Utah. These SNPs were derived from resequencing finished genomic sequence from 19 regions of the genome. Significant p values for LD occurred in greater than 50% of cases at distances less than or equal to 80 kb. Long range LD was reported in Reich *et al.* (2001) to be less than or equal to 80 kb. However, a panel of 96 Nigerians showed LD extending to lengths less than 5 kb (Reich *et al.* 2001). These findings were reported to be typical of sub-Saharan African populations. In order to further investigate these findings, an additional 48 southern Swedish European samples were analyzed. Linkage disequilibrium findings in the Swedish samples were almost identical to those found in the Utah population. The vast difference of LD distance between the European and African populations indicates a possible bottleneck or founder effect within the European descendents after the divergence from the Nigerians some 100,000 years ago. This bottleneck or founder effect would explain the long ranges of LD previously described.

Prior to Reich *et al.* (2001), LD in humans had been studied with very few loci in only a few populations. The author's goal was to obtain a more accurate representation of LD in the human genome. Generally, investigators found that the extent of LD varies throughout the genome and in different populations. Using a minor allele frequency of at least .35 facilitated cross population studies because high frequency SNPs tend to be high frequency in all populations. In fact, it was later discovered that variations in allele frequencies between populations skews D′ and r^2 measures of LD (Boyles *et al.*, 2005).

A genetically isolated human population from the Netherlands, originally founded in the 18th century with 150, people that experienced exponential growth was utilized in a linkage disequilibrium study (Aulchenko *et al.*, 2004). Given that the population was isolated, large segments of their DNA should be identical by descent making this population ideal to study certain diseases. The amount of decay of LD was studied genome wide with 734 autosomal short tandem repeats (STRs) and 47 X-linked STRs across 58 individuals. The mean permutation-corrected measure of LD (D'_{cp}) for autosomal markers, was calculated to be 0.0054 between syntenic markers. The mean permutation-corrected measure of LD for X-linked markers, was calculated to be 0.0114 between syntenic markers. Linkage disequilibrium for nonsyntenic markers was essentially zero, indicating that admixture and drift were not generating detectable LD between unlinked loci.

One of the most comprehensive chromosomal LD maps to date is a first-generation linkage disequilibrium map of human chromosome 22 (HSA22) (Dawson *et al.*, 2002). A linkage disequilibrium report was generated based upon the complete sequence of HSA22. Initially 77 members of the Centre d'Étude du Polymorphisme Humain (CEPH) reference families were genotyped with 1504 SNP markers evenly disbursed approximately every 15 kb along the entire length of the chromosome. The CEPH families were used because they aided in identification of genotyping errors and construction of long haplotypes. Unrelated individuals from the United Kingdom and Estonia were genotyped as well. With use of the CEPH family and unrelated population samples, conserved haplotypes along the chromosome between different populations were identified.

Linkage disequilibrium values were reported as D' and r^2 measurements. Average D' values ranged from 0.7 for markers relatively close in distance to 0.11 for unlinked markers while r^2 measurements declined from 0.38 to 0.01. These measurements were calculated within a sliding window of 1.7 Mb segments along HSA22. Confirmation of areas exhibiting high amounts of LD was confirmed with significance tests. These and other results suggested a variable pattern of LD along the chromosome with some areas displaying almost complete LD adjacent to areas showing approximately no linkage disequilibrium. As was expected, LD was found to decay with increasing distance. A positive correlation was reported between LD and G+C rich regions of the chromosome and a negative correlation was identified between G+C rich DNA regions of HSA22 and genetic distance. The correlation results agree with conclusions previously stated regarding regulation of meiotic recombination.

The previously mentioned studies have vast amounts of variability in population size and structure as well as marker density. The impact of marker density and population size on the measurement of LD can be profound. Variability in LD with respect to marker density and allele frequency was analyzed within a 10 Mb continuous region of HSA20q12-13.2 (Ke *et al.*, 2004a). Initially, 5000 SNPs were genotyped in 12 CEPH families at a density of 1 SNP every 2 kb. They then preceded to genotype a subset of these SNPs at a density of one SNP every 3, 4, 5, 7.5 and 10 kb. Analysis showed marker density of 10 kb density to be similar to that of the 2 kb density. Their results indicate that LD patterns do not change with the marker density that they studied. However, localized patterns of LD are highly dependent on marker density. Adding more SNPs within a close region appears to disrupt specific LD patterns. Additionally boundaries of LD appear variable with marker density.

2.6 Conclusion

Only two genome wide linkage disequilibrium studies have been performed in cattle (Farnir *et al.*, 2000; Odani *et al.*, 2006). Both of these studies reported long range LD in cattle to extend over tens of cM. The similarities between the Farnir and Odani studies are their use of microsatellite markers and D′ to characterize and measure LD. Comparatively, the canine genomics community utilized SNPs and r^2 to measure LD (Lindblad-Toh *et al.*, 2005; Sutter *et al.*, 2004). Whole genome shot gun sequencing of genomes has led to the discovery of tens of thousands of single nucleotide polymorphic markers. Therefore, SNPs are now more prevalent in species whose genomes have been sequenced compared to microsatellite markers. The advantages of using r^2 over D′ as a measure of LD are that r^2 is dependent on genealogy where D′ is dependent on recombination and physical distance (Daly *et al.*, 2001). Furthermore, r^2 quantified the amount of information about one locus provided by another locus (Pritchard & Przeworski, 2001; Zhao *et al.*, 2005). The differences between the Farnir and Odani studies lie in the breeds of cattle used. Farnir's study measured LD in a population of Dutch Black and White Dairy cattle while Odani measured LD in Japanese Black and Japanese Brown beef cattle. Until now there has not been a comparative whole genome linkage disequilibrium map with more than two breeds of cattle.

Generally, only in those species with finished genome sequence, such as the human genome, has it been possible to compare LD between populations (Reich *et al.*, 2001). A study of this magnitude generally requires access to vast amounts of resources, technology and polymorphic markers. In the cattle community, we have been able to take advantage of the release of tens of thousands of bovine SNPs resulting from the bovine genome sequencing initiative, and the generation of new high throughput technology to successfully develop the tools necessary for completion of a comparative genome wide LD map in cattle.

In order to generate a genome wide LD map in cattle, we first had to ensure that we had accurate ordering of all loci used to measure LD. To do so a high throughput approach was employed to generate RH map locations for 4671 SNPs and linkage mapped 2071 SNPs. Genotypes were then produced for 3072 SNPs in eight breeds of cattle. These genotypes along with the map locations were utilized in production of phased haplotypes. Finally, linkage disequilibrium maps were generated and population structure was analyzed both within and between breeds of cattle. The overall goal was to produce tools that could aid in more effective association studies. In pursuit of that goal we have developed a new high throughput mapping method and discovered that the optimal number of markers for a whole genome association study is 30,000-50,000.

2.7 Literature Cited

Ardlie K. G., Kruglyak L., and Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3:** 299-309.

Ashwell M. S., and Van Tassell C. P. (1999). Detection of putative loci affecting milk, health, and type traits in a US Holstein population using 70 microsatellite markers in a genome scan. *J Dairy Sci* **82:** 2497-502.

Aulchenko Y. S., Heutink P., Mackay I., Bertoli-Avella A. M., Pullen J., Vaessen N., Rademaker T. A., Sandkuijl L. A., Cardon L., Oostra B., and van Duijn C. M. (2004). Linkage disequilibrium in young genetically isolated Dutch population. *Eur J Hum Genet* **12:** 527-34.

Boichard D., Maignel L., and Verrier E. (1996). Analyse genealogique des races bovines laitieres francaises. *INRA Prod. Anim* **9:** 323-335.

Boyles A. L., Scott W. K., Martin E. R., Schmidt S., Li Y. J., Ashley-Koch A., Bass M. P., Schmidt M., Pericak-Vance M. A., Speer M. C., and Hauser E. R. (2005).

Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* **59:** 220-7.

Daly M. J., Rioux J. D., Schaffner S. F., Hudson T. J., and Lander E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29:** 229-32.

Dawson E., Abecasis G. R., Bumpstead S., Chen Y., Hunt S., Beare D. M., Pabial J., Dibling T., Tinsley E., Kirby S., Carter D., Papaspyridonos M., Livingstone S., Ganske R., Lohmussaar E., Zernant J., Tonisson N., Remm M., Magi R., Puurand T., Vilo J., Kurg A., Rice K., Deloukas P., Mott R., Metspalu A., Bentley D. R., Cardon L. R., and Dunham I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418:** 544-8. Epub 2002 Jul 10.

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., and Georges M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10:** 220-7.

Funk D. A. (2006). Major advances in globalization and consolidation of the artificial insemination industry. *J Dairy Sci* **89:** 1362-8.

Goode E. L., and Jarvik G. P. (2005). Assessment and implications of linkage disequilibrium in genome-wide single-nucleotide polymorphism and microsatellite panels. *Genet Epidemiol* **29 Suppl 1:** S72-6.

Harmegnies N., Farnir F., Davin F., Buys N., Georges M., and Coppieters W. (2006). Measuring the extent of linkage disequilibrium in commercial pig populations. *Anim Genet* **37:** 225-31.

Ihara N., Takasuga A., Mizoshita K., Takeda H., Sugimoto M., Mizoguchi Y., Hirano T., Itoh T., Watanabe T., Reed K. M., Snelling W. M., Kappes S. M., Beattie C. W., Bennett G. L., and Sugimoto Y. (2004). A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res* **14:** 1987-98.

Jorde L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10:** 1435-44.

Ke X., Durrant C., Morris A. P., Hunt S., Bentley D. R., Deloukas P., and Cardon L. R. (2004). Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum Mol Genet* **13:** 2557-65.

Lewontin R. C. (1988). On measures of gametic disequilibrium. *Genetics* **120:** 849-52.

Lindblad-Toh K., Wade C. M., Mikkelsen T. S., Karlsson E. K., Jaffe D. B., Kamal M., Clamp M., Chang J. L., Kulbokas E. J., 3rd, Zody M. C., Mauceli E., Xie X., Breen M., Wayne R. K., Ostrander E. A., Ponting C. P., Galibert F., Smith D. R., DeJong P. J., Kirkness E., Alvarez P., Biagi T., Brockman W., Butler J., Chin C. W., Cook A., Cuff J., Daly M. J., DeCaprio D., Gnerre S., Grabherr M., Kellis M., Kleber M., Bardeleben C., Goodstadt L., Heger A., Hitte C., Kim L., Koepfli K. P., Parker H. G., Pollinger J. P., Searle S. M., Sutter N. B., Thomas R., Webber C., Baldwin J., Abebe A., Abouelleil A., Aftuck L., Ait-Zahra M., Aldredge T., Allen N., An P., Anderson S., Antoine C., Arachchi H., Aslam A., Ayotte L., Bachantsang P., Barry A., Bayul T., Benamara M., Berlin A., Bessette D., Blitshteyn B., Bloom T., Blye J., Boguslavskiy L., Bonnet C., Boukhgalter B., Brown A., Cahill P., Calixte N., Camarata J., Cheshatsang Y., Chu J., Citroen M., Collymore A., Cooke P., Dawoe T., Daza R., Decktor K., DeGray S., Dhargay N., Dooley K., Dooley K., Dorje P., Dorjee K., Dorris L., Duffey N., Dupes A., Egbiremolen O., Elong R., Falk J., Farina A., Faro S., Ferguson D., Ferreira P., Fisher S., FitzGerald M., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803-19.

McRae A. F., McEwan J. C., Dodds K. G., Wilson T., Crawford A. M., and Slate J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160:** 1113-22.

McVean G. A., Myers S. R., Hunt S., Deloukas P., Bentley D. R., and Donnelly P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581-4.

Myers S., Spencer C. C., Auton A., Bottolo L., Freeman C., Donnelly P., and McVean G. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans* **34:** 526-30.

Nsengimana J., Baret P., Haley C. S., and Visscher P. M. (2004). Linkage disequilibrium in the domesticated pig. *Genetics* **166:** 1395-404.

Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., and Sasaki Y. (2006). Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Anim Genet* **37:** 139-44.

Petes T. D. (2001). Meiotic recombination hot spots and cold spots. *Nat Rev Genet* **2:** 360-9.

Pritchard J. K., and Przeworski M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69:** 1-14.

Reich D. E., Cargill M., Bolk S., Ireland J., Sabeti P. C., Richter D. J., Lavery T., Kouyoumjian R., Farhadian S. F., Ward R., and Lander E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Sandor C., Farnir F. P., Meuwissen T. H., Coppieters W., and Georges M. (2006). Linkage disequilibrium on the bovine X chromosome: characterization and use in QTL mapping. *Genetics*.

Schaffner S. F. (2004). The X chromosome in population genetics. *Nat Rev Genet* **5:** 43-51.

Sutter N. B., Eberle M. A., Parker H. G., Pullar B. J., Kirkness E. F., Kruglyak L., and Ostrander E. A. (2004). Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Res* **14:** 2388-96. Epub 2004 Nov 15.

Tenesa A., Knott S. A., Ward D., Smith D., Williams J. L., and Visscher P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *J Anim Sci* **81:** 617-23.

Vallejo R. L., Li Y. L., Rogers G. W., and Ashwell M. S. (2003). Genetic diversity and background linkage disequilibrium in the North American Holstein cattle population. *J Dairy Sci* **86:** 4137-47.

Zhao H., Nettleton D., Soller M., and Dekkers J. C. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res* **86:** 77-87.

Figures

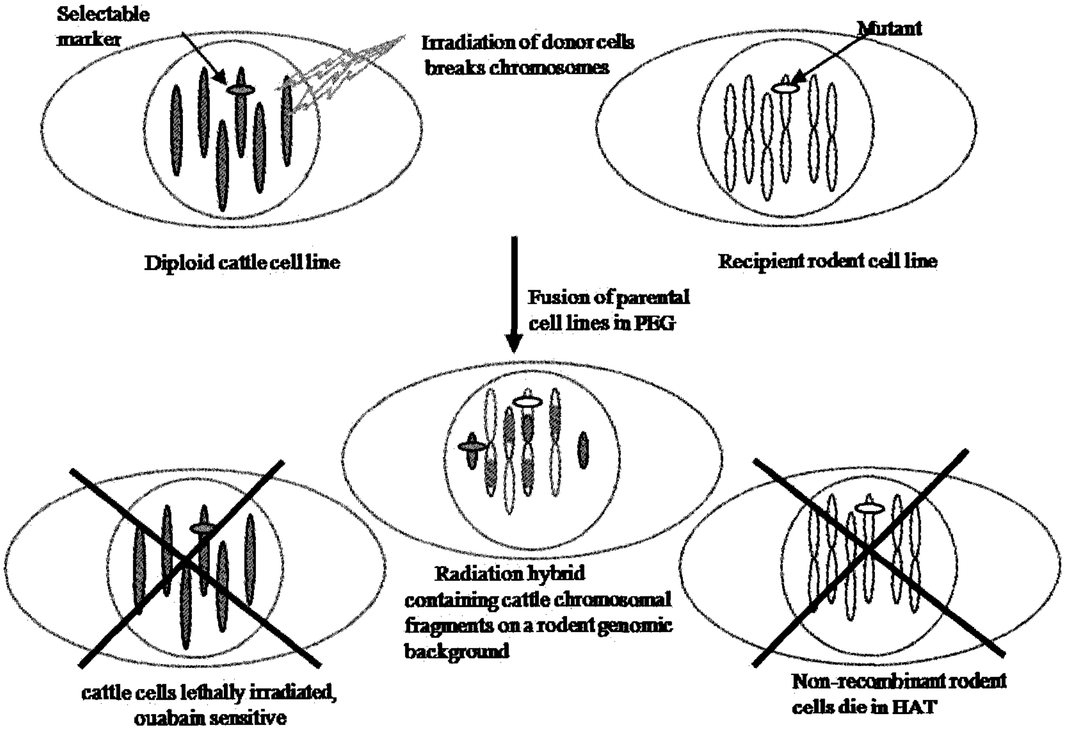
Selectable marker
Irradiation of donor cells breaks chromosomes
Mutant
Diploid cattle cell line
Recipient rodent cell line
Fusion of parental cell lines in PEG
Radiation hybrid containing cattle chromosomal fragments on a rodent genomic background
cattle cells lethally irradiated, ouabain sensitive
Non-recombinant rodent cells die in HAT

Figure 2.1 How to make a radiation hybrid panel.

CHAPTER 3

Construction of whole genome radiation hybrid and linkage maps for bovine using high-throughput genotyping

3.1 Introduction

Radiation hybrid (RH) mapping is a powerful tool that can be utilized for the production of in-depth comparative maps of single chromosomes and whole genomes. RH mapping in cattle was resurrected in the late 1990's with the creation of the 5000 rad panel (Womack *et al.*, 1997). Since then, bovine RH panels of various resolution have been created and utilized, including 3000 (Williams *et al.*, 2002), 7000 (Itoh *et al.*, 2005) and 12,000 (Rexroad *et al.*, 2000) rad panels. Radiation hybrid mapping relies upon scoring the presence or absence of markers in a hybrid cell panel constructed by fusing irradiated donor cells with recipient rodent cells. Conventional methods for typing markers on RH panels rely on individual or low complexity multiplex PCR assays for all typed markers on DNA derived from each of the cell lines in the panel, followed by agarose gel electrophoresis to detect the presence or absence of the marker in individual RH cell lines. The proportion of the DNA fragments from the donor genome that harbor any particular marker will vary between the cell lines and hence signal intensity varies among the positive cells. This problem as well as PCR artifacts and differentiation of rodent versus target species amplification products usually necessitates running assays in duplicate and on occasion in triplicate to confirm results. To increase marker throughput, the Illumina® BeadStation 500G was used to type a large number of markers which were distributed throughout the bovine genome on a whole genome radiation hybrid panel. The accuracy of typing was confirmed by

Chapter 3 accepted for publication in Animal Genetics
http://www.blackwellpublishing.com/journal.asp?ref=0268-9146

building the markers into radiation hybrid maps together with 1125 markers that had been typed conventionally on the RH panel. With this high throughput approach to RH mapping 4690 loci were rapidly mapped. Next, paternally related registered Angus artificially inseminated (AI) sires were genotyped and linkage maps were constructed. Of the 4690 loci that were RH mapped, 2701 of the same loci were utilized in construction of linkage maps. Thus resulting in human-cattle comparative maps with direct comparison to the bovine whole genome sequence assembly (Btau 2.0).

3.2 Materials and Methods

3.2.1 Marker selection

Information for sequences containing SNPs was obtained from public databases (http://www.ncbi.nlm.nih.gov/projects/SNP/, ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp) and oligonucleotides were designed, synthesized and assembled into oligo pooled assays (OPA) by Illumina Inc. (San Diego, CA).

3.2.2 Oligo pooled assays

GoldenGate SNP assays were designed at Illumina using a proprietary assay design tool that selects oligonucleotides with balanced melting temperature characteristics at each locus and compares these oligonucleotide sequences to an internal database to ensure sequence uniqueness. Assay designs were based on build Btau 2.0 of the cattle genome sequence. Sequences of up to 200 bases in length, with at least 50 bases on each side of the SNP of interest with ambiguities or additional SNPs indicated as "N" or IUPAC codes for mixed base, were used as input for the design program. The portion of the oligos that bind to the genomic sequence are all between 20 and 23 bases long, with a pair of oligos being utilized for each locus being analyzed. Therefore, the footprint of each assay on the genomic DNA is between 45 and 55 bases generally, including a gap of 1 to 20 bases between the two oligos. Of the 10,631 candidate SNPs submitted for scoring, 6738 were included as part of the final assay pool. The remaining 3893 SNPs were not utilized for this experiment.

3.2.3 Typing of the RH panel

I typed DNA from the 94 cell lines of the Roslin-Cambridge 3000 rad bovine-hamster whole genome radiation hybrid ($WGRH_{3000}$) panel plus hamster and bovine control DNA samples (Williams *et al.*, 2002) following the manufacturer's protocol for the Illumina® BeadStation 500G (Oliphant *et al.*, 2002); www.illumina.com). Typing was carried out in five multiplex reactions which included all 6738 loci which passed the Illumina design protocol. I used Illumina®Gencall software to manually score the presence or absence of loci across the 94 clones of the $WGRH_{3000}$ rad RH panel. For markers for which the donor bull used to construct the RH panel was homozygous (Figure 3.1a), positive and negative hybrids formed distinct clusters and a clear boundary could be defined between the clusters allowing simple scoring of the positive and negative controls. All hybrids within the negative control cluster were marked using the "exclude selected sample" option of the Gencall software resulting in these samples being identified as "No Calls" which produced "U genotypes." RH vectors were generated for these loci by combining all types and scoring the cells harboring the screened locus as present (1) and the U genotypes as absent (0). However, for loci for which the donor animal was heterozygous, four clusters were produced: cells harboring the A allele, the B allele, both alleles (H) and the negative cells (U) (Figure 3.1b). For these heterozygous markers, RH vectors were generated by combining the A, B, and H genotypes and scoring these cells as present (1) and scoring the U types as absent (0).

In conventional RH mapping studies, it is reasonably common to find markers that amplify with low intensity in a hybrid DNA and for which the signal can be difficult to replicate. These low intensity positives can be attributed to mispriming, to loci for which a mutation lies within a primer site for the animal used to construct the panel or to low target copy number of a hybrid cell culture and these cells are scored as ambiguous (2) for the marker. The high throughput approach described here does not eliminate this issue. As is evident in Figure 3.1c, one hybrid is labeled as ambiguous. Ambiguous genotypes were visually identified as outliers from the clusters, and vectors for these loci were manually changed to "2".

Markers were assigned to chromosomes by two-point analysis between the loci scored on the Illumina and 1125 markers previously included in an earlier version of the WGRH$_{3000}$ map (Williams *et al.*, 2002) using RHmap (Lange *et al.*, 1995). To minimize the number of false positive assignments, a minimum LOD score of 4.5 was used as evidence for linkage. I constructed whole chromosome maps using the default algorithm of Carthagene (Schiex & Gaspin, 1997) as described by (Williams *et al.*, 2002). A total of 1423 typed SNPs were physically separated by approximately 1-2 kilobases (Supplementary Table 3.1). For these loci RH map locations were determined for only one locus and inferred the map position of the remaining loci based upon their order in the Btau_2.0 assembly. Homologous synteny blocks (HSB) and inversions were defined according to criteria of (Murphy *et al.*, 2005) and the accepted threshold for BLASTn hits was E < 0.00001.

3.2.4 Genotypes for linkage map

I genotyped DNA from 80 paternally related registered Angus AI sires with 3,072 of the 6738 SNPs typed on the RH panel following the manufacturer's protocol for the Illumina® BeadStation 500G (Oliphant *et al.*, 2002) www.illumina.com). These animals are a subset of a larger 14 generation pedigree comprising 1697 registered Angus sires (Morsci *et al.*, 2006).

Homologous bovine sequence coordinates were obtained by BLAST (Altschul *et al.*, 1990) analysis (blastn) of the 500 bp sequences harboring each target SNP against the bovine sequence assembly (Btau 2.0). Orthologous human sequence coordinates were derived from UCSC alignments of bovine scaffolds (Btau 1.0) to the human assembly (build 35). Only SNP markers with un-ambiguous match positions to the human and cow genomes were retained. Approximately 40% of these SNPs were not assigned to a bovine chromosome in the Btau_2.0 assembly and the best alignment to build 35 of the human assembly and the human-bovine comparative map of Itoh *et al.* (2005) were used to order these loci relative to the SNPs contained within the bovine sequence assembly.

3.2.5 Map construction

A framework map was constructed using loci for which the best BLAST hit to the bovine assembly agreed with its map location predicted in the RH analysis. We then integrated into the map those loci for which the best human BLAST hit mapped to an orthologous bovine location in agreement with the RH map position. Any locus within a scaffold which could not be assigned by these processes but for which another locus in the same scaffold had previously been assigned using this method was then integrated into the map. Linkage analysis of the 80 paternally related registered Angus AI sires was then performed using Crimap (Green *et al.* 1990) and loci that were significantly linked (LOD > 3.0) to SNPs already included in the map were integrated into the map only if an available RH location, bovine BLAST hit or human BLAST hit localized the SNP to the same chromosomal region. Finally the FLIPS5 option of Crimap was utilized to identify ordering errors for each chromosome. This approach was primarily used to detect small inversions of groups of loci which were not identified in the bovine-human comparative map or to correct erroneous inversions suggested by markers that were incorrectly ordered in the comparative map (Itoh *et al.* 2005).

3.3 Results

Of the 6738 loci included in the Illumina oligo pooled assays, 5815 (86.3%) were successfully typed and 4690 of the 5815 (80.7%) were successfully typed and RH mapped. These rates exceeded the 62.3% (data not shown) success rate in our laboratory using conventional RH typing methods. Of the 4690 markers whose RH map locations were determined, 4334 were SNPs identified as a result of the bovine whole genome sequencing initiative. Furthermore, 2701 of these markers were selected to construct a high-density linkage map to estimate recombination distance among loci and to confirm locus order produced in the RH map (Supplementary Table 3.2). Due to the limited number of informative meioses within the mapping pedigree, we required at least two independent measures of support for integrating loci into the linkage map which resulted in 2701 of the 3072 markers (87.9%) being included in the map. Only 1801 (66.7%) of these loci have been assigned chromosomal coordinates in Btau_2.0 (Table 3.1). Disagreement in marker order between the linkage map and Btau_2.0 was observed on chromosomes 3, 7, 10, 15, 18, 27 and 29 with inversions on BTA13 and 19. Comparisons between the linkage map and Btau_2.0 identified 133 of 2701 (4.9%) markers with chromosomal assignment discordancies. Discordant markers are labeled in red on the Btau_2.0 maps contained in the Supplementary Figures 3.1. Btau_2.0 chromosome 11 had 14 discordant markers. However, Btau_2.0 chromosomes 4, 10, 16 and 28 had no discordant markers. Human sequence orthologs which were supported by either RH map or Btau_2.0 positions were identified for 1495 of the 2701 (55.3%) linkage mapped SNPs (Table 3.2). These coordinates indicate 185 homologous synteny blocks and 28 putative major inversions (Table 3.2 and Supplementary Table 3.3).

The overall length of the autosomal RH map was 80086.8 cR_{3000} (Supplementary Table 3.1) with an average intermarker distance of 13.77 cR_{3000}. *Bos taurus* autosome (BTA) 5 was the most densely mapped chromosome with 366 markers and a total length of 3798.5 cR_{3000} compared to 177 markers that were mapped by linkage with a total length of 133.9 cM. Conversely, BTA27 was the most sparsely mapped chromosome with 92 markers RH mapped and a total chromosomal length of 1298.5 cR_{3000} compared to 40 markers linkage mapped with a total chromosomal length of 51.6 cM.

There are 1125 common markers between the RH maps and the USDA-MARC linkage map. Comparison of the RH and linkage maps described here to the USDA-MARC linkage map (Supplementary Figures 3.1) indicates general agreement. However, marker orders on chromosomes 2, 9, 10, 19, 26 and 27 indicate inversions between the RH and MARC maps. Major order disagreements were found on BTA19, where a substantial marker gap at the proximal end of the chromosome prevented alignment of the RH and linkage maps.

3.4 Discussion

RH map locations were determined for 4690 SNP loci and 1125 previously typed markers. Of these, 4334 SNPs were identified in the bovine genome sequencing project (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/bovine-snp.txt). Linkage maps containing 2701 loci were constructed for all autosomes and human-cattle comparative maps were built. We report cR_{3000} locations for 4690 markers, cM location for 2701 markers and Mbp positions for 1801 of the 4690 loci,. This approach for high throughput RH mapping has allowed us to rapidly assemble comprehensive maps of the bovine genome.

Generally, linkage maps are conducive to the high-level ordering of loci along a chromosome with a resolution that is generally determined by the number of informative meioses. On the other hand, RH maps are often superior for fine mapping closely linked loci (Mellersh *et al.*, 2000). Alignment of the high density linkage map with Btau_2.0 confirmed the overall quality of the Btau_2.0 sequence assembly but indicated that there are 133 incorrectly assigned loci and localized inversions in scaffold ordering within the assembly. The marker density of the maps presented here is higher than that used to assign and order the sequence scaffolds in the current Btau_2.0 genome assembly. Therefore additional scaffolds were able to be positioned within the sequence assembly. These data were contributed to the bovine composite map (Snelling *et al.*, personal communication) that will be used for the next assembly of the bovine genome sequence.

In comparison to PCR-based typing methods used for RH mapping, this approach is rapid and cost effective. The use of high density SNP assays such as the Illumina GoldenGate or Infinium platforms harboring as many as 64,000 loci per assay will allow typing of a whole genome RH panel in as little as one day. Thus, the generation of *de novo* RH maps based only upon the knowledge of short sequences can now be accomplished extremely rapidly and cost effectively. This strategy appears to be particularly useful for the development of high-content maps necessary for the assembly of whole genome shotgun sequences. On the other hand, for those genomes where a sequence assembly is not likely in the near future, the approach described here may provide an effective means of generating comprehensive comparative maps which extend the utility of existing whole genome sequences to closely related species with low sequence information content. High throughput RH mapping may contribute towards low cost genome sequencing by coupling the methodology described here with 454 sequencing technology (Margulies *et al.*, 2005).

3.5 Literature Cited

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Itoh T., Watanabe T., Ihara N., Mariani P., Beattie C. W., Sugimoto Y., and Takasuga A. (2005). A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics* **85:** 413-24.

Lange K., Boehnke M., Cox D. R., and Lunetta K. L. (1995). Statistical methods for polyploid radiation hybrid mapping. *Genome Res* **5:** 136-50.

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y. J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L., Jarvie T. P., Jirage K. B., Kim J. B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., and

Rothberg J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-80.

Mellersh C. S., Hitte C., Richman M., Vignaux F., Priat C., Jouquand S., Werner P., Andre C., DeRose S., Patterson D. F., Ostrander E. A., and Galibert F. (2000). An integrated linkage-radiation hybrid map of the canine genome. *Mamm Genome* **11:** 120-30.

Morsci N. S., Schnabel R. D., and Taylor J. F. (2006). Association analysis of adiponectin and somatostatin polymorphisms on BTA1 with growth and carcass traits in Angus cattle. *Anim Genet* **37:** 554-62.

Murphy W. J., Larkin D. M., Everts-van der Wind A., Bourque G., Tesler G., Auvil L., Beever J. E., Chowdhary B. P., Galibert F., Gatzke L., Hitte C., Meyers S. N., Milan D., Ostrander E. A., Pape G., Parker H. G., Raudsepp T., Rogatcheva M. B., Schook L. B., Skow L. C., Welge M., Womack J. E., O'Brien S J., Pevzner P. A., and Lewin H. A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309:** 613-7.

Oliphant A., Barker D. L., Stuelpnagel J. R., and Chee M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **Suppl:** 56-8, 60-1.

Rexroad C. E., 3rd, Owens E. K., Johnson J. S., and Womack J. E. (2000). A 12,000 rad whole genome radiation hybrid panel for high resolution mapping in cattle: characterization of the centromeric end of chromosome 1. *Anim Genet* **31:** 262-5.

Schiex T., and Gaspin C. (1997). CARTHAGENE: constructing and joining maximum likelihood genetic maps. *Proc Int Conf Intell Syst Mol Biol* **5:** 258-67.

Williams J. L., Eggen A., Ferretti L., Farr C. J., Gautier M., Amati G., Ball G., Caramorr T., Critcher R., Costa S., Hextall P., Hills D., Jeulin A., Kiguwa S. L., Ross O., Smith A. L., Saunier K., Urquhart B., and Waddington D. (2002). A bovine whole-genome radiation hybrid panel and outline map. *Mamm Genome* **13:** 469-74.

Womack J. E., Johnson J. S., Owens E. K., Rexroad C. E., 3rd, Schlapfer J., and Yang Y. P. (1997). A whole-genome radiation hybrid panel for bovine gene mapping. *Mamm Genome* **8:** 854-6.

Figures

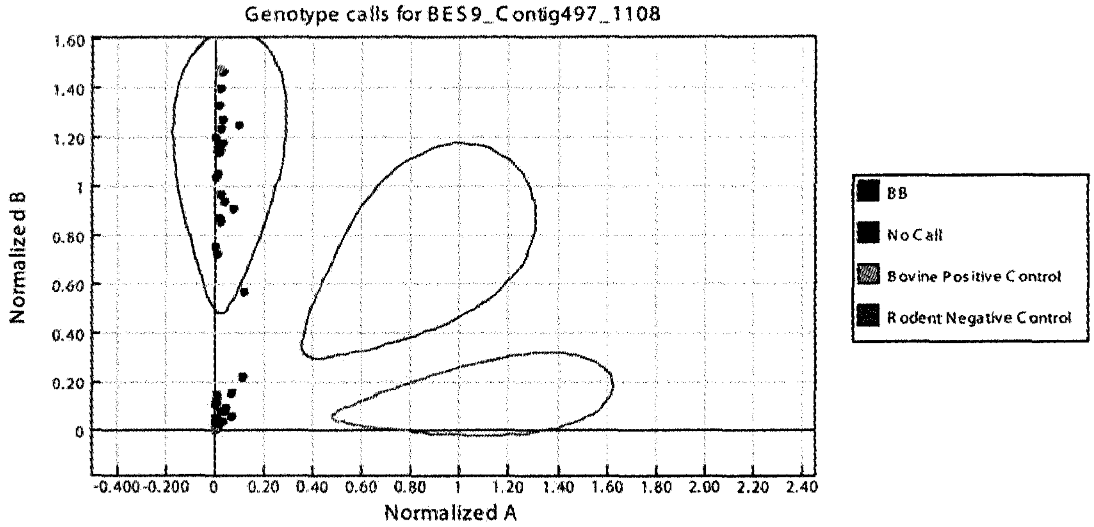

Figure 3.1a

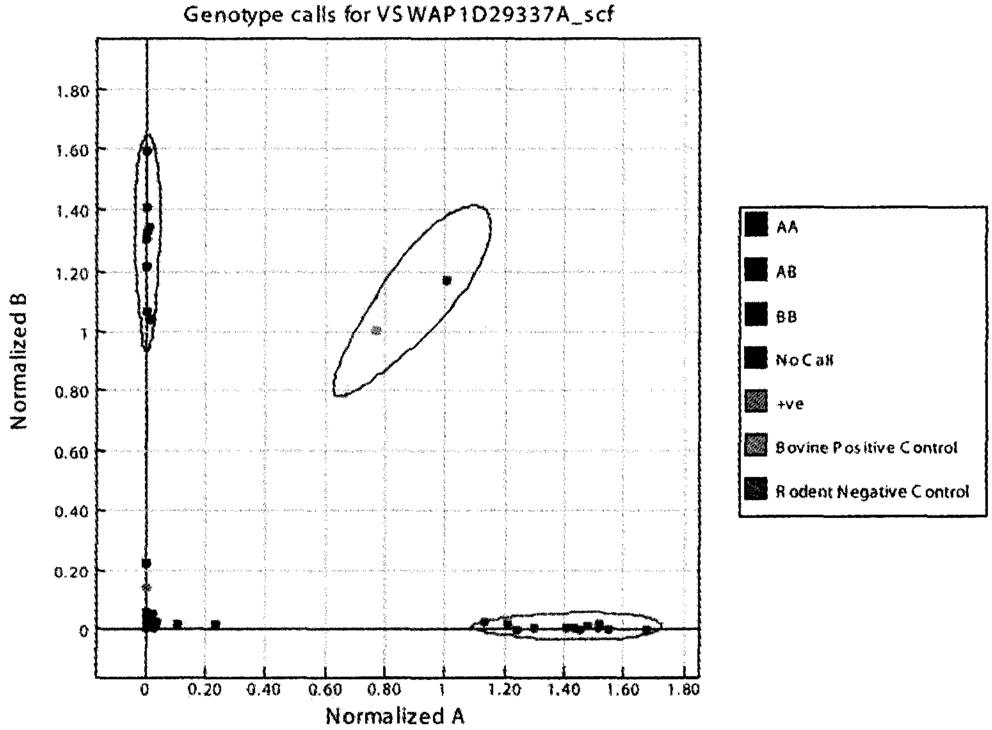

Figure 3.1b

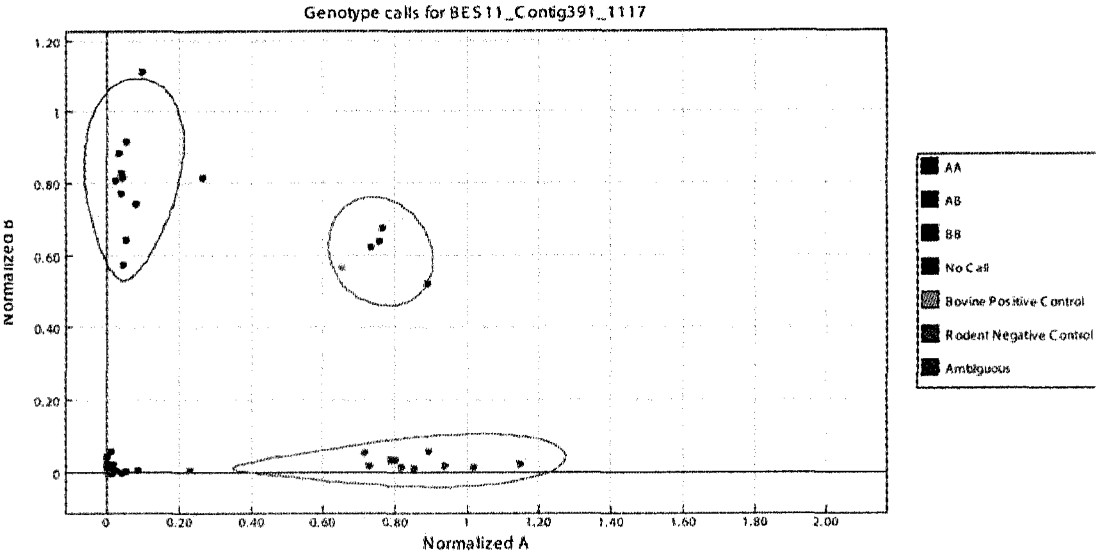

Figure 3.1c

Figure 3. 1: A Cartesian plot depicting radiation hybrid typing generated by the Illumina® BeadStation. a) Cell lines located under the sweeping arc from 0.30 on the Y axis to 0.30 on the X axis were identified as negatives and labeled no calls (U). Once selected, the negative cluster was excluded and the calling algorithm was rerun. Cells containing BB genotypes were scored as present. Cells containing U genotypes were scored as absent. b) Cell lines located under the sweeping arc from 0.40 on the Y axis to 0.40 on the X axis were identified as negatives and labeled no calls (U). Once selected, the negative cluster was excluded and the calling algorithm was rerun. Cells containing A, H and B genotypes were scored as present. Cells containing U genotypes were scored as absent. c) Cell lines located under the sweeping arc from 0.10 on the Y axis to 0.10 on the X axis were identified as negatives and labeled no calls (U). Once selected, the negative cluster was excluded and the calling algorithm was rerun. One clone were labeled as ambiguous and marked as "2". It is located at approximately 0 on the Y axis and just to the right of 0.2 on the X axis respectively. Cells containing A, H and B genotypes were scored as present. Cells containing U genotypes were scored as absent. Cells marked as "2" were scored as ambiguous.

Table 3.1: Summary statistics when comparing the second version of the bovine sequence assembly, Btau_2.0, and the linkage map.

Btau_2.0	Btau_2.0 Length Mbp	No.BLASTN hits	Avg. Marker Spacing Mbp	No. dicordant markers on Btau_2.0
1	97.2	71	1.37	6
2	83.1	89	0.93	11
3	81.5	121	0.67	10
4	66.4	67	0.99	0
5	75.4	107	0.70	5
6	65.6	67	0.98	6
7	67.8	74	0.92	4
8	60.4	53	1.14	3
9	62.6	64	0.98	11
10	66.4	67	0.99	0
11	86.5	89	0.97	14
12	46.8	44	1.06	1
13	61.2	65	0.94	3
14	46.2	42	1.10	4
15	53.5	74	0.72	1
16	56.5	69	0.82	0
17	44	41	1.07	2
18	55.5	58	0.96	3
19	54.6	53	1.03	9
20	40.5	44	0.92	1
21	46.3	51	0.91	6
22	45.7	41	1.11	2
23	41.5	75	0.55	12
24	44.7	47	0.95	2
25	39.6	62	0.64	9
26	34.9	47	0.74	3
27	24.6	25	0.98	2
28	32.6	49	0.67	0
29	43.6	45	0.97	3
Total	1625.2	1801	0.90	133

Table 3.2: Summary statistics of human-cattle comparative maps.

Human Chromosome (HSA)	HSA length Mbp	No. BLASTN hits	Average marker spacing Mbp	No. Homologous synteny blocks	No. putative inversions
1	240.2	171	1.40	20	6
2	222.2	140	1.59	15	3
3	192.7	115	1.68	17	4
4	174.4	90	1.94	15	2
5	172	98	1.76	12	3
6	162	85	1.91	6	2
7	155.1	73	2.12	8	1
8	129.1	51	2.53	6	3
9	126.9	67	1.89	10	1
10	128.8	100	1.29	10	1
11	132.6	99	1.34	11	0
12	129	82	1.57	13	0
13	89.7	40	2.24	8	0
14	81.5	40	2.04	6	0
15	72.1	43	1.68	6	0
16	81.9	47	1.74	5	0
17	63.5	30	2.12	4	0
18	66.2	40	1.66	5	1
19	51.2	19	2.69	2	0
20	60.1	31	1.94	4	1
21	31.3	19	1.65	1	0
22	23	15	1.53	1	0
Total	2585.5	1495	1.73	185	28

CHAPTER 4

Whole genome linkage disequilibrium maps in eight breeds of cattle

4.1 Background

Linkage disequilibrium (LD) maps are fundamental tools for exploring the genetic basis of economically important traits in cattle. Likewise, comparative LD maps enable us to explore the degree of diversity between breeds of cattle and to detect genomic regions that have been subject to selective sweeps within the different dairy and beef breeds which represent different biological types (e.g. Continental European vs. British). The currently available information regarding LD in cattle is primarily based on microsatellite studies performed in dairy cattle. The most extensive of these studies used 284 genome-wide microsatellites in a population of Dutch Black and White Dairy cattle (Farnir *et al.*, 2000) to show that syntenic LD extended up to several tens of centimorgans (cM). Haplotypes for 581 maternally inherited gametes were used to estimate LD using Lewontin's normalized D'. The results indicated high levels of LD not only between closely linked markers but for markers located as much as 40 cM (~40 Mb) apart. Two subsequent studies examined the extent of LD in cattle although both used fewer animals and microsatellites (Tenesa *et al.*, 2003; Vallejo *et al.*, 2003). Vallejo et al. (2003) selected distantly related animals to quantify the level of genetic diversity in United States Holstein cattle. While only 23 Holstein bulls were genotyped with 54 microsatellite loci that spanned most of the autosomal genome, extensive LD was detected in the United States Holstein population in agreement with the findings of Farnir et al. (2000). Tenesa et al. (2003) genotyped 50 Holstein bulls for 13

Chapter 4 in review at BMC Genetics

microsatellites spanning *Bos taurus* autosome (BTA)2 and BTA6, to determine the extent of LD in the United Kingdom Holstein population. The average *D'* value was 44% with significant LD reported only for distances less than 10.3 cM. Linkage disequilibrium among non syntenic loci was not significant.

More recently Khatkar et al. (2006) scored 220 BTA6 single nucleotide polymorphisms in a sample of 433 Australian dairy bulls and estimated LD between marker pairs using D'. While they found that LD decayed with increasing distance between markers, D' did not reach background until an average distance of 20 Mb separated the markers. They also found that there was extensive variability in the magnitude of D' at any one distance. The rate of decay of LD estimated using SNPs (Khatkar *et al.*, 2006) was much greater than that estimated using microsatellites (Farnir *et al.*, 2000), which is consistent with the findings of Varilo et al. (2003) that more informative marker systems are able to detect LD over greater physical distances.

Only recently has the extent of LD been examined in beef cattle populations. A sample of 162 half-sib progeny from a Japanese black sire and 406 half-sib Japanese brown cattle were genotyped with 246 and 156 autosomal microsatellite markers, respectively (Odani *et al.*, 2006). For syntenic markers, the mean *D'* was 16.3% for Japanese Brown and 25.1% for Japanese Black. Characteristic of D' as a measure of LD, significant LD was observed for marker pairs separated by as much as 40 cM in both breeds.

Quantifying the extent of LD in the bovine genome is a necessary first step for determining the number of markers that will be sufficient for quantitative trait loci (QTL) mapping by linkage disequilibrium. The previous studies which used microsatellite markers were either too narrowly focused on particular chromosomes, or were of insufficient resolution to precisely estimate genome-wide LD and almost certainly were unable to precisely estimate short-range LD. The high density and low inherent rates of mutation of SNPs relative to microsatellites within mammalian genomes allows for the identification of ancestral haplotype blocks and the estimation of identity by descent probabilities which are crucial for haplotype-based association studies (Vignal *et al.*, 2002). In this study, LD was estimated in 8 breeds of cattle utilizing 2670 single nucleotide polymorphism markers that were derived from the bovine genome sequence and were aligned to the Btau_3.1 genome sequence assembly.

4.2 Results and Discussion

4.2.1 Haplotype Estimation

The program GENOPROB 2.0 (Thallman *et al.*, 2001a; Thallman *et al.*, 2001b), which utilizes multi-generation pedigrees including both genotyped and non-genotyped animals, was used for the estimation of phased haplotypes. For all breeds, greater than 97% of the scored genotypes were determined by GENOPROB 2.0 to have a probability of at least 95% of being correct conditional on the pedigree and marker map (McKay *et al.*, 2007) (Figure 4.1a, Additional Table 4.1). While the overall level of genotype accuracy was high, the level of genotype certainty was clearly dependent on pedigree structure. Holstein, Limousin and Angus samples were obtained from the most complex pedigrees and produced the most accurately estimated genotypes and phased chromosomes (Figure 4.1b, Additional Table 4.1). The depth of the pedigree as well as the location of the genotyped individuals within the pedigree (generation), had the largest influence on the estimation of phase probabilities (oGmx). Figure 4.1b clearly demonstrates that the breeds with the greatest pedigree complexity produced the highest probabilities of correctly phased genotypes. The cumulative proportion of heterozygous genotypes which could be phased by GENOPROB 2.0 with order probabilities >0.99 was 87.6%, 76.9% and 69.2% for Holstein, Limousin and Angus, respectively. The Brahman sample comprises several independent three generation pedigrees consisting of grandparent – parent and multiple offspring in which only one parent and grandparent was genotyped and there were no additional close pedigree relationships between individuals within or between families. This is in contrast to the Nelore sample which represents a two generation pedigree and which, in most cases, both parents of each animal were genotyped. For these pedigree structures, the three generation Brahman pedigree produced 5.3% (BR 63.4%, NEL 58.1%) more heterozygous genotypes with a phase order probability of >0.99. However, using a three generation pedigree structure with larger numbers of individuals per generation and including complete pedigree relationships among ungenotyped and genotyped animals such as in the Holstein, Limousin and Angus samples, produced a significant increase in the proportion of heterozygous genotypes which could be accurately phased.

4.2.2 General LD Findings

Comparative linkage disequilibrium maps were generated for eight breeds of cattle for the 29 bovine autosomes. The majority of the SNPs used in this study were chosen because they had previously been putatively identified as being variable within *Bos* *taurus*. This ascertainment bias resulted in the SNP minor allele frequencies being substantially lower in the two *Bos indicus* breeds than in the *Bos taurus* breeds (Figure 4.2). It also resulted in a set of SNPs in which common SNPs within the *Bos taurus* genome were over-represented. However, even though these loci were identified from *Bos taurus* derived sequences, more than 50% of the loci were polymorphic in *Bos indicus* and had a minor allele frequency >0.05 (Figure 4.2). This indicates that a substantial fraction of the loci identified by the Bovine Genome Sequencing Project will have utility for QTL mapping within *Bos indicus* breeds.

The current estimate of the size of the bovine genome is 2.87 Gb (http://www.hgsc.bcm.tmc.edu/projects/bovine/) and with equal spacing, the 2,670 SNP loci used in this study would have an inter-marker distance of approximately 1 Mb. However, the loci were selected according to genomic location, likely assay conversion rate and minor allele frequency in *Bos taurus.* Consequently they were not uniformly distributed (Figure 4.3) with 30% of the loci having inter-marker distances less than 0.5 Mb, and 13% separated by more than 3 Mb. The non-uniform distribution of marker locations allows for the estimation of LD across several orders of magnitude of differences in physical distance.

The r^2 values for pairs of loci were binned according to the physical distance separating the loci and were averaged within each breed (Figure 4.4). As has previously been observed, there is an inverse relationship between LD and physical or genetic distance (Sved, 1971) and r^2 is essentially at long-range background levels in all eight breeds by a locus separation of approximately 500 kb. A similar study performed in pigs found that average r^2 values had fallen to 0.1 for SNPs with an inter-marker distance of 3 cM (Du *et al.*, 2007); similarly, LD in dog breeds extends across several Mb (Lindblad-Toh *et al.*, 2005). However LD in humans extends for only tens of kb (Hinds *et al.*, 2005) which is consistent with the large effective size and rapid recent expansion of human populations.

Our findings indicate a substantially shorter range of LD than has previously been reported in cattle (Farnir *et al.*, 2000; Odani *et al.*, 2006). The differences between previous reports and our findings are attributed to the differences in measures used to report LD, namely D' versus r^2. The Dutch Black and White Dairy cattle used in this experiment are a subset of the animals used in the original publication (Farnir *et al.*, 2000) and the Holstein cattle are a subset of the animals previously used to fine map milk production traits on BTA6 (Schnabel *et al.*, 2005). To provide a direct comparison between the approaches, we estimated genome-wide average measures of LD using both r^2 and D' in these two breeds (Figure 4.5). Both estimates of LD show an inverse relationship between LD and distance, however, in general, D' overestimates the extent of LD (Ardlie *et al.*, 2002; Ke *et al.*, 2004b). The use of D' suggests that LD extends for several tens of centimorgans (or Mb), consistent with the earlier reports (Figure 4.5). However, the use of r^2 indicates that LD is at background levels by approximately 0.5 Mb. Similar discrepancies between measures of LD have recently been reported in cattle (Barendse *et al.*, 2007; Khatkar *et al.*, 2007).

It has been suggested that when large discrepancies exist between marker allele frequencies, due to the presence of a rare allele, these two measures of LD are divergent (Boyles *et al.*, 2005). D′ estimates historical recombination through allelic association whereas r^2 measures the squared correlation coefficient between locus allele frequencies and is strongly influenced by the order in which the mutations arose (genealogy) and not necessarily the physical distance between loci (Daly *et al.*, 2001). In the context of QTL mapping, r^2 is the preferred measure of LD, because it quantifies the amount of information that can be inferred about one (perhaps nonobservable quantitative trait or disease) locus from another (Pritchard & Przeworski, 2001; Zhao *et al.*, 2005), and can therefore be used to estimate the number of loci needed for association studies (Kruglyak, 1999; Pritchard & Przeworski, 2001). For this reason r^2 was used as the primary measure of LD in this study.

Variation in average r^2 values between breeds is evident in Figures 4.4 and 4.6. Considering the similarity between the Holstein and Dutch Black and White Dairy breeds, we expected comparable average r^2 values between these breeds (Figure 4.4). In fact, the extent of LD is quite similar within all of the *Bos taurus* and within the *Bos indicus* breeds, however, the *Bos indicus* appear to have substantially less short-range LD than do the *Bos taurus.* This could be the result of ascertainment bias, as the SNPs used in this study were detected because they were common SNPs within *Bos taurus* and their average minor allele frequency was much lower in *Bos indicus* (Weiss & Clark, 2002). Alternatively, the lower levels of short-range LD could also reflect historically larger effective population sizes, (Tenesa *et al.*, 2007), which seems particularly appropriate for the Nelore. On the other hand, long range LD in Brahman appears to be greater than for the Nelore and other *Bos taurus* breeds which suggests a smaller current effective population size, which is consistent with the relatively recent formation of the breed as an admixture between extant *Bos taurus* and several imported *Bos indicus* breeds imported into the U.S. between 1854 and 1926 (Sanders, 1980).

4.2.3 r^2 Findings by Chromosome

Variation in LD between chromosomes and breeds was examined using the 18.7% of all possible syntenic locus pairs (Additional Table 4.2 and Figure 4.2) that were separated by less than 1 Kb. The average r^2 values by chromosome and for each breed are shown in Figure 4.6 (Additional Table 4.3). In Figure 4.6, breeds are grouped according to subspecies and the primary agricultural purpose of each breed. The first six *Bos taurus* breeds include Angus, Charolais, Limousin, and Japanese Black representing meat breeds, followed by Dutch Black and White Dairy and Holstein which are dairy breeds. While the Brahman is used primarily for meat in the U.S. and Australia, the Nelore is a *Bos indicus* breed used for both milk and meat production in South America. With the exception of BTA7, 12 and 21, the average r^2 across the *Bos taurus* breeds was 0.5603 with minimum and maximum r^2 values obtained on BTA29 in Limousin (0.12) and BTA14 in Holstein (0.91), respectively. The average r^2 values across the *Bos indicus* breeds was 0.37 with minimum and maximum r^2 values on BTA20 in Nelore (0.06) and BTA22 in Nelore (0.69). The relatively low level of short-range LD contrasts with previously published reports in cattle (Farnir *et al.*, 2000; Odani *et al.*, 2006). Comparable results were found for pairs of syntenic loci separated by approximately 100 kb and 500 kb (Additional tables 4.2 and 4.3), respectively, which further supports our contention that usable LD (Kruglyak, 1999) does not extend beyond 0.5 Mb and that LD reaches background levels by 1 Mb (Figure 4.4). These findings have a profound impact on the number of loci and the number of individuals that will need to be tested in association-based QTL scans.

4.2.4 BTA 7, 12 and 21

Figure 4.6 indicates that the average r^2 values are low for all breeds on BTA 7, 12 and 21 when compared to all other autosomes. The average r^2 values on these chromosomes does not appear to be a sampling artifact, since the number of loci used to calculate the average r^2 values was 19, 21 and 7 locus pairs on BTA 7, 12 and 21, respectively, which is not significantly different than for the other autosomes (Additional Table 4.3). Additionally, each of the loci on these chromosomes had similar allele frequencies to those loci on the other autosomes (data not shown). This suggests that the loci on BTA 7, 12 and 21 may have been clustered around one or more recombination hotspots on each of these chromosomes. To determine if the loci were clustered, the location of the SNP pairs along each chromosome were plotted (Figure 4.7). Figure 4.7 demonstrates that the SNP pairs used to examine the extent of short range LD are distributed along the length of these three chromosomes and we therefore conclude that BTA 7, 12 and 21 have intrinsically lower levels of LD than do the other autosomes.

We have two theories as to why lower than average LD may exist on BTA 7, 12 and 21. First, because cattle have been selected for production traits for at least 50 generations, there is the possibility that selection on QTL distributed throughout the genome has generated different patterns of LD on individual chromosomes. However, compared to chromosomes of similar size, there does not appear to be fewer QTLs on BTA7, 12 and 21 (Hu & Reecy, 2007; Polineni *et al.*, 2006) and selection should have resulted in similar patterns of LD on these chromosomes as on all others. Second, it is possible that chromosomes 7, 12 and 21 have higher than average rates of recombination than do the other autosomes. A comparison between the physical (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9913) and genetic (http://www.marc.usda.gov/genome/genome.html) maps of BTA7, 12 and 21 as well as chromosomes of similar physical size, indicates that the physical to genetic size relationship is similar for BTA7, 12 and 21 and other chromosomes of similar physical size. However, regions of increased recombination have been detected on human chromosomes 14 and 15 (Yu *et al.*, 2001), which are partially orthologous to BTA21. A complete exploration of these chromosomes in order to study aspects of genome organization that potentially affect recombination rate will require additional markers and animals.

4.3 Conclusions

While loci were included in this analysis for which order was consistent between the Btau_3.1 assembly and our radiation hybrid map (Thallman *et al.* 2001), the genome coordinates for each locus were obtained from the Btau_3.1 assembly. This assembly spans only 2.43 Gb and while an additional 319 Mb of sequence exists as contigs which are unassigned to chromosomes, the final sequence is expected to be much closer to 2.8 Gb. The unassembled contigs are likely to be biased towards centromeric and telomeric sequences and duplications which are difficult to assemble, but some are no doubt interstitial to chromosomes. The fact that these are unassembled would likely cause a systematic bias towards the underestimation of the physical distance between loci. There also appear to be a significant number of problems with the ordering and orientation of scaffolds within the assembly and these errors are likely to produce random effects on the estimation of distance between syntenic loci. Thus overall, we suspect that the incomplete nature of the assembly results in about a 10% underestimate of the distance between loci. This has only a minor affect on our conclusions and the extent of LD available for association analysis still does not significantly exceed 500 kb. At a physical distance of 100 kb separating flanking SNP loci, the average r^2 is 0.15-0.2 and the average r^2 between these markers and a QTL located at mid-interval is about 0.3 (Figure. 4.4). This would appear to be the lowest desirable resolution for whole genome association mapping in bovine and assuming a 2.87 Gb genome, it would require 28,700 fully informative SNPs to saturate the genome at an average resolution of 100 kb. Since the number of validated bovine SNPs is currently insufficient to achieve an even spacing and because many SNPs are likely to have low minor allele frequencies leading to their being uninformative in many populations, we believe that 50,000 SNPs will be the minimum required for whole genome association studies in cattle. Furthermore, the extent of LD on BTA 7, 12 and 21 appears to be much lower than for the autosomal genome as a whole and suggests that SNP density may need to be enhanced on these chromosomes. The construction of a high resolution LD map of the bovine genome will provide further insight into the effects of selection and evolutionary forces upon the genomes of breeds which have been selected for different agricultural purposes.

4.4 Methods

4.4.1 DNA Collection

I collected DNA from 70 Angus (USA), 20 Canadian Angus, 40 Charolais (Canada), 40 Brahman (USA), 97 Dutch Black and White Dairy cattle (Belgium), 48 Holstein (USA), 65 Japanese Black (Japan), 43 Limousin (USA) and 97 Nelore (Brazil) cattle. In order to phase the chromosomes using linkage information, small families were selected where members within the families were closely related but the families themselves were not closely related. Family structure and the number of individuals per family varied between the breeds but the general family structure consisted of a male grandparent, male parent and three or more progeny (Additional figure 4.1). This three generation family structure allowed for the efficient estimation of marker phase relationships in the progeny and also produced the most likely phase relationships in each of the parents/grandparents.

4.4.2 Marker Selection and Genotyping

Sequence information for SNPs was obtained from public databases (http://www.ncbi.nlm.nih.gov/projects/SNP/, ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp). Loci included in this study met the following criteria; minor allele frequency (MAF) ≥ 0.05 in Angus based on previous screens (data not shown) and concordant order determined by radiation hybrid (RH) mapping (McKay *et al.*, 2007) and genomic sequence location. Oligonucleotides were designed, synthesized and assembled into oligo pooled assays (OPA) by Illumina Inc. (San Diego, CA). Genotyping was performed using the manufacturer's protocol for the Illumina® BeadStation 500G (Oliphant *et al.*, 2002) (www.illumina.com).

4.4.3 Locus locations within the bovine genome sequence assembly

Chromosomal coordinates for each SNP were obtained by aligning approximately 250 bp flanking each SNP by BLAST to the latest release of the bovine genome sequence assembly, Btau_3.1. These physical coordinates were compared to the linkage and RH maps of McKay *et al.* (2007). Thirty four markers were excluded from the analysis because their assignment in the sequence assembly was to a chromosome that differed to their linkage or RH map assignment or because they had no chromosomal assignment in Btau_3.1. Marker information can be found in Additional Table 4.4.

4.4.4 Haplotypes and LD Analysis

GENOPROB V2.0 (Thallman *et al.*, 2001a; Thallman *et al.*, 2001b) was used to assess genotype score quality and produce whole chromosome phased haplotypes based on the pedigree and physical map locations of the loci. Briefly, GENOPROB uses an allelic peeling algorithm to estimate both the probability that a genotype is correct, denoted as pGmx, and the probability that the order (phase) of the alleles are correct, denoted as oGmx. Only genotypes with a pGmx ≥ 0.95 were used for LD analysis but no restriction was placed on order probability, oGmx. This produced a set of whole chromosome haplotypes comprised of accurately scored genotypes that were in the most likely phase configuration. I assessed LD by generating r^2 values using GOLD (Abecasis & Cookson, 2000) independently for the maternally- and paternally-inherited haplotypes. LD data presented here is based only on the maternally inherited haplotypes which avoids the overrepresentation of paternally inherited haplotypes within the primarily male pedigrees.

4.5 Literature Cited

Abecasis G. R., and Cookson W. O. (2000). GOLD--graphical overview of linkage disequilibrium. *Bioinformatics* **16:** 182-3.

Ardlie K. G., Kruglyak L., and Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3:** 299-309.

Barendse W., Bunch R. J., Kijas J. W., and Thomas M. B. (2007). The effect of genetic variation of the retinoic acid receptor-related orphan receptor C gene on fatness in cattle. *Genetics* **175:** 843-53.

Boyles A. L., Scott W. K., Martin E. R., Schmidt S., Li Y. J., Ashley-Koch A., Bass M. P., Schmidt M., Pericak-Vance M. A., Speer M. C., and Hauser E. R. (2005). Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* **59:** 220-7.

Daly M. J., Rioux J. D., Schaffner S. F., Hudson T. J., and Lander E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29:** 229-32.

Du F. X., Clutter A. C., and Lohuis M. M. (2007). Characterizing linkage disequilibrium in pig populations. *Int J Biol Sci* **3:** 166-78.

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., and Georges M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10:** 220-7.

Hinds D. A., Stuve L. L., Nilsen G. B., Halperin E., Eskin E., Ballinger D. G., Frazer K. A., and Cox D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072-9.

Hu Z. L., and Reecy J. M. (2007). Animal QTLdb: beyond a repository : A public platform for QTL comparisons and integration with diverse types of structural genomic information. *Mamm Genome* **18:** 1-4.

Ke X., Hunt S., Tapper W., Lawrence R., Stavrides G., Ghori J., Whittaker P., Collins A., Morris A. P., Bentley D., Cardon L. R., and Deloukas P. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* **13:** 577-88.

Khatkar M. S., Collins A., Cavanagh J. A., Hawken R. J., Hobbs M., Zenger K. R., Barris W., McClintock A. E., Thomson P. C., Nicholas F. W., and Raadsma H. W. (2006). A first generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics.*

Khatkar M. S., Zenger K. R., Hobbs M., Hawken R. J., Cavanagh J. A., Barris W., McClintock A. E., McClintock S., Thomson P. C., Tier B., Nicholas F. W., and Raadsma H. W. (2007). A primary assembly of a bovine haplotype block map based on a 15k SNP panel genotyped in Holstein-Friesian cattle. *Genetics.*

Kruglyak L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22:** 139-44.

Lindblad-Toh K., Wade C. M., Mikkelsen T. S., Karlsson E. K., Jaffe D. B., Kamal M., Clamp M., Chang J. L., Kulbokas E. J., 3rd, Zody M. C., Mauceli E., Xie X., Breen M., Wayne R. K., Ostrander E. A., Ponting C. P., Galibert F., Smith D. R., DeJong P. J., Kirkness E., Alvarez P., Biagi T., Brockman W., Butler J., Chin C. W., Cook A., Cuff J., Daly M. J., DeCaprio D., Gnerre S., Grabherr M., Kellis M., Kleber M., Bardeleben C., Goodstadt L., Heger A., Hitte C., Kim L., Koepfli K. P., Parker H. G., Pollinger J. P., Searle S. M., Sutter N. B., Thomas R., Webber C., Baldwin J., Abebe A., Abouelleil A., Aftuck L., Ait-Zahra M., Aldredge T., Allen N., An P., Anderson S., Antoine C., Arachchi H., Aslam A., Ayotte L., Bachantsang P., Barry A., Bayul T., Benamara M., Berlin A., Bessette D., Blitshteyn B., Bloom T., Blye J., Boguslavskiy L., Bonnet C., Boukhgalter B., Brown A., Cahill P., Calixte N., Camarata J., Cheshatsang Y., Chu J., Citroen M., Collymore A., Cooke P., Dawoe T., Daza R., Decktor K., DeGray S., Dhargay N., Dooley K., Dooley K., Dorje P., Dorjee K., Dorris L., Duffey N., Dupes A., Egbiremolen O., Elong R., Falk J., Farina A., Faro S., Ferguson D., Ferreira P., Fisher S., FitzGerald M., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803-19.

McKay S. D., Schnabel R. D., Murdoch B. M., Aerts J., Gill C. A., Gao C., Li C., Matukumalli L. K., Stothard P., Wang Z., Van Tassell C. P., Williams J. L., Taylor J. F., and Moore S. S. (2007). Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Anim Genet* **38:** 120-5.

Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., and Sasaki Y. (2006). Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Anim Genet* **37:** 139-44.

Oliphant A., Barker D. L., Stuelpnagel J. R., and Chee M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **Suppl:** 56-8, 60-1.

Polineni P., Aragonda P., Xavier S. R., Furuta R., and Adelson D. L. (2006). The bovine QTL viewer: a web accessible database of bovine Quantitative Trait Loci. *BMC Bioinformatics* **7:** 283.

Pritchard J. K., and Przeworski M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69:** 1-14.

Sanders J. O. (1980). History and development of Zebu cattle in the United States. *J. Anim. Sci.* **50:** 1188-1200.

Schnabel R. D., Kim J. J., Ashwell M. S., Sonstegard T. S., Van Tassell C. P., Connor E. E., and Taylor J. F. (2005). Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proc Natl Acad Sci U S A* **102:** 6896-901.

Sved J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2:** 125-41.

Tenesa A., Knott S. A., Ward D., Smith D., Williams J. L., and Visscher P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *J Anim Sci* **81:** 617-23.

Tenesa A., Navarro P., Hayes B. J., Duffy D. L., Clarke G. M., Goddard M. E., and Visscher P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17:** 520-6.

Thallman R. M., Bennett G. L., Keele J. W., and Kappes S. M. (2001a). Efficient computation of genotype probabilities for loci with many alleles: I. Allelic peeling. *J Anim Sci* **79:** 26-33.

Thallman R. M., Bennett G. L., Keele J. W., and Kappes S. M. (2001b). Efficient computation of genotype probabilities for loci with many alleles: II. Iterative method for large, complex pedigrees. *J Anim Sci* **79:** 34-44.

Vallejo R. L., Li Y. L., Rogers G. W., and Ashwell M. S. (2003). Genetic diversity and background linkage disequilibrium in the North American Holstein cattle population. *J Dairy Sci* **86:** 4137-47.

Vignal A., Milan D., SanCristobal M., and Eggen A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol* **34:** 275-305.

Weiss K. M., and Clark A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18:** 19-24.

Yu A., Zhao C., Fan Y., Jang W., Mungall A. J., Deloukas P., Olsen A., Doggett N. A., Ghebranious N., Broman K. W., and Weber J. L. (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409:** 951-3.

Zhao H., Nettleton D., Soller M., and Dekkers J. C. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res* **86:** 77-87.

Figures

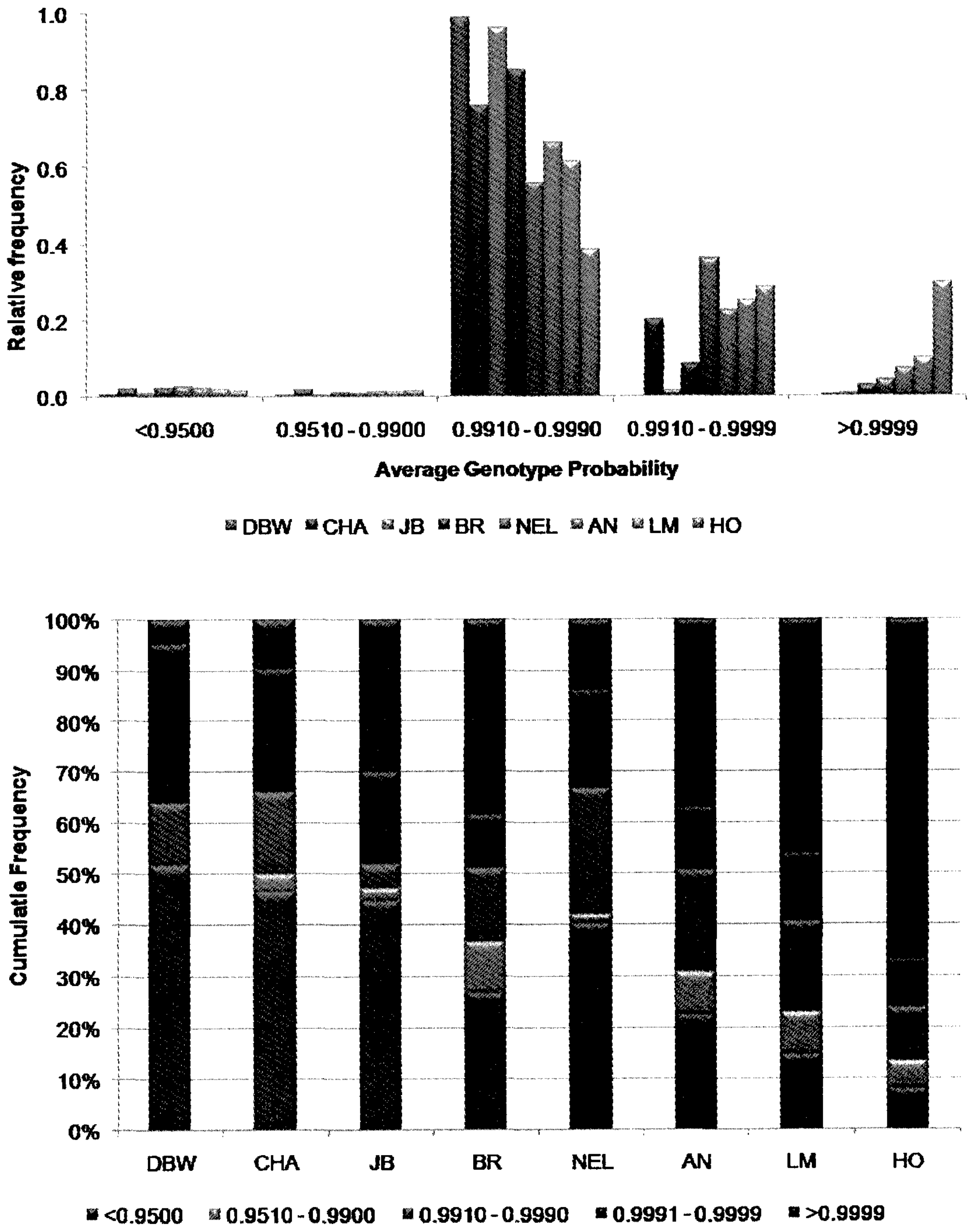

Figure 4.1 a and b. Summary of genotype (pGmx) and phase (oGmx) probabilities for each breed based on GENOPROB 2.0 results. Only genotyped progeny with at least one genotyped parent were used. Shown are the following breeds of cattle: Dutch Black and White dairy (DBW), Charolais (CHA), Japanese Black (JB), Brahman (BR), Nelore (NEL), Angus (AN), Limousin (LM) and Holstein (HO).

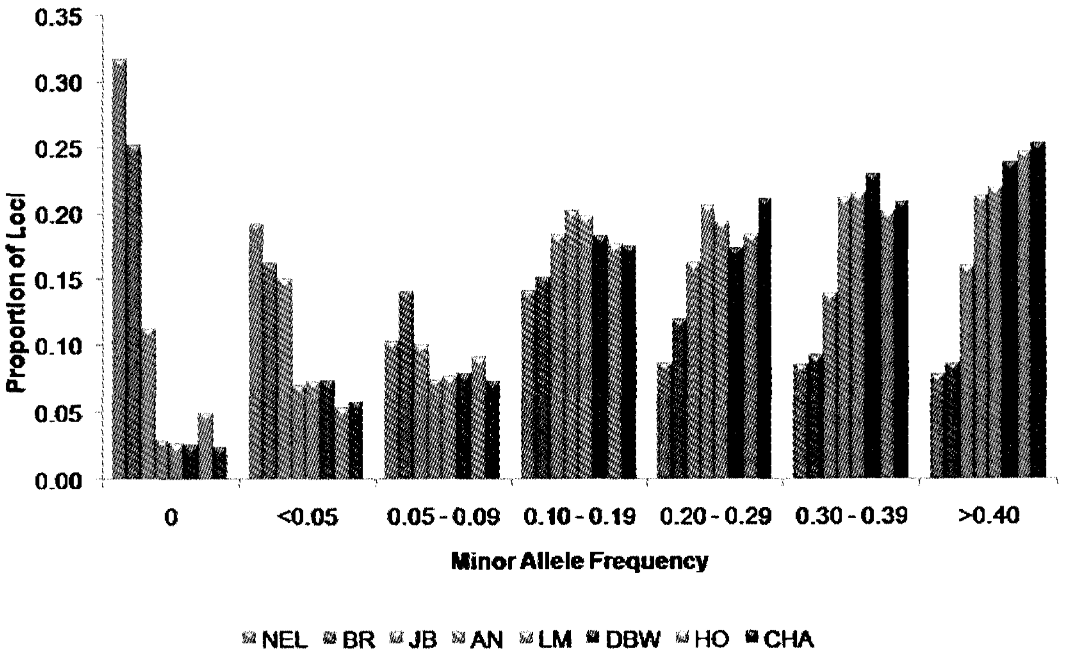

Figure 4.2. Minor allele frequencies (MAF) for all breeds.

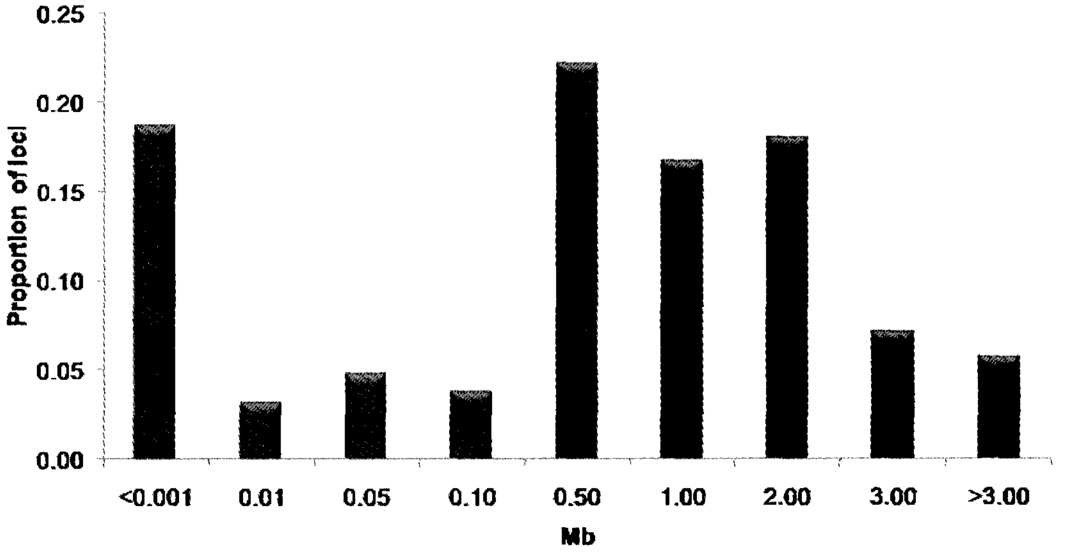

Figure 4.3. Percentage of loci within each bin. Proportion of loci for each bin are shown based on inter-marker distances measured in Mb. The horizontal axis depicts the maximum inter-marker distance for each bin. Here, a majority of the markers fall within the 0.50 Mb bin, This bin includes all markers with an inter-marker distance greater than 0.10 Mb but less than or equal to 0.50 Mb.

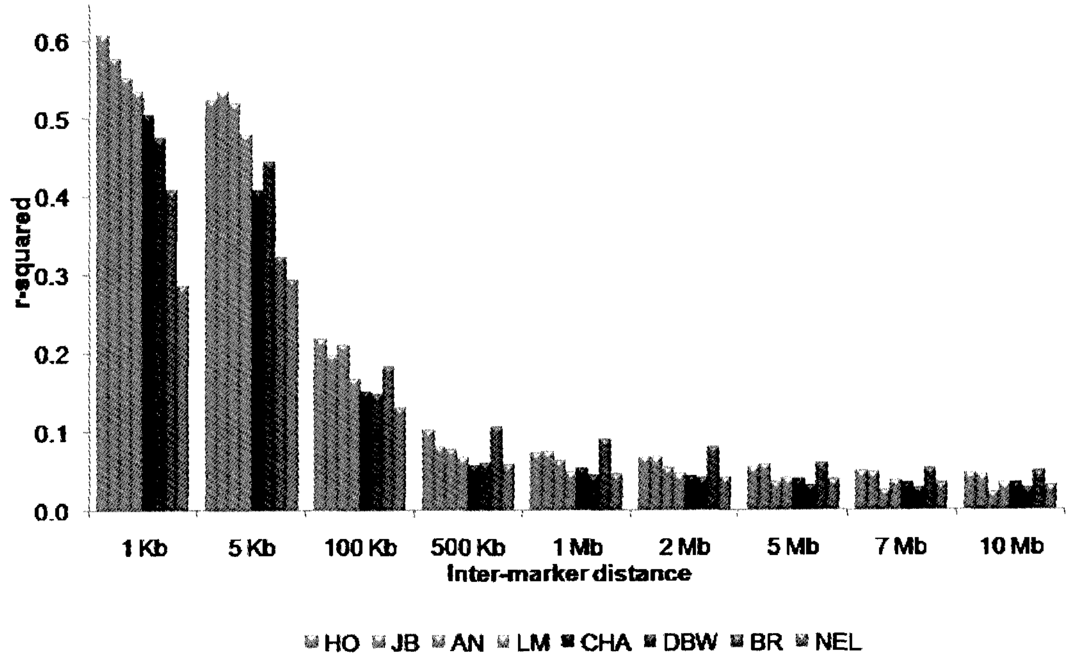

Figure 4.4. Overall average r^2 values.

Average r^2 values are shown for each breed. The maximum value for each bin is shown on the horizontal axis.

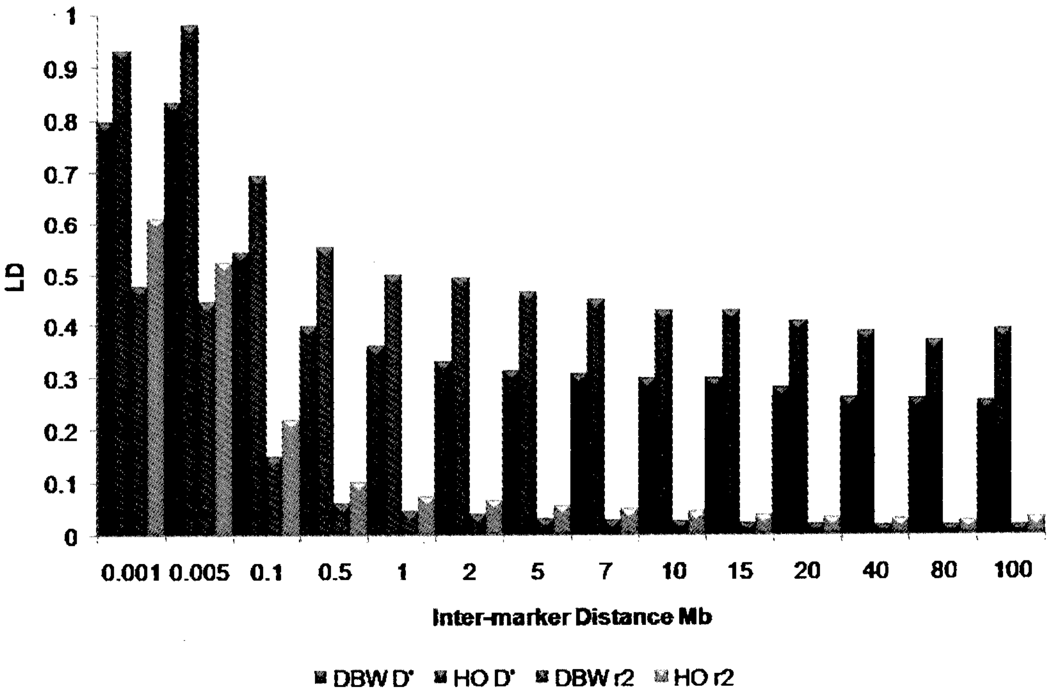

Figure 4.5. Estimates of D′ and r^2 for Holstein and Dutch Black and White Dairy cattle. Average LD is shown for each breed in each bin. The maximum inter-marker value for each bin is shown on the horizontal axis.

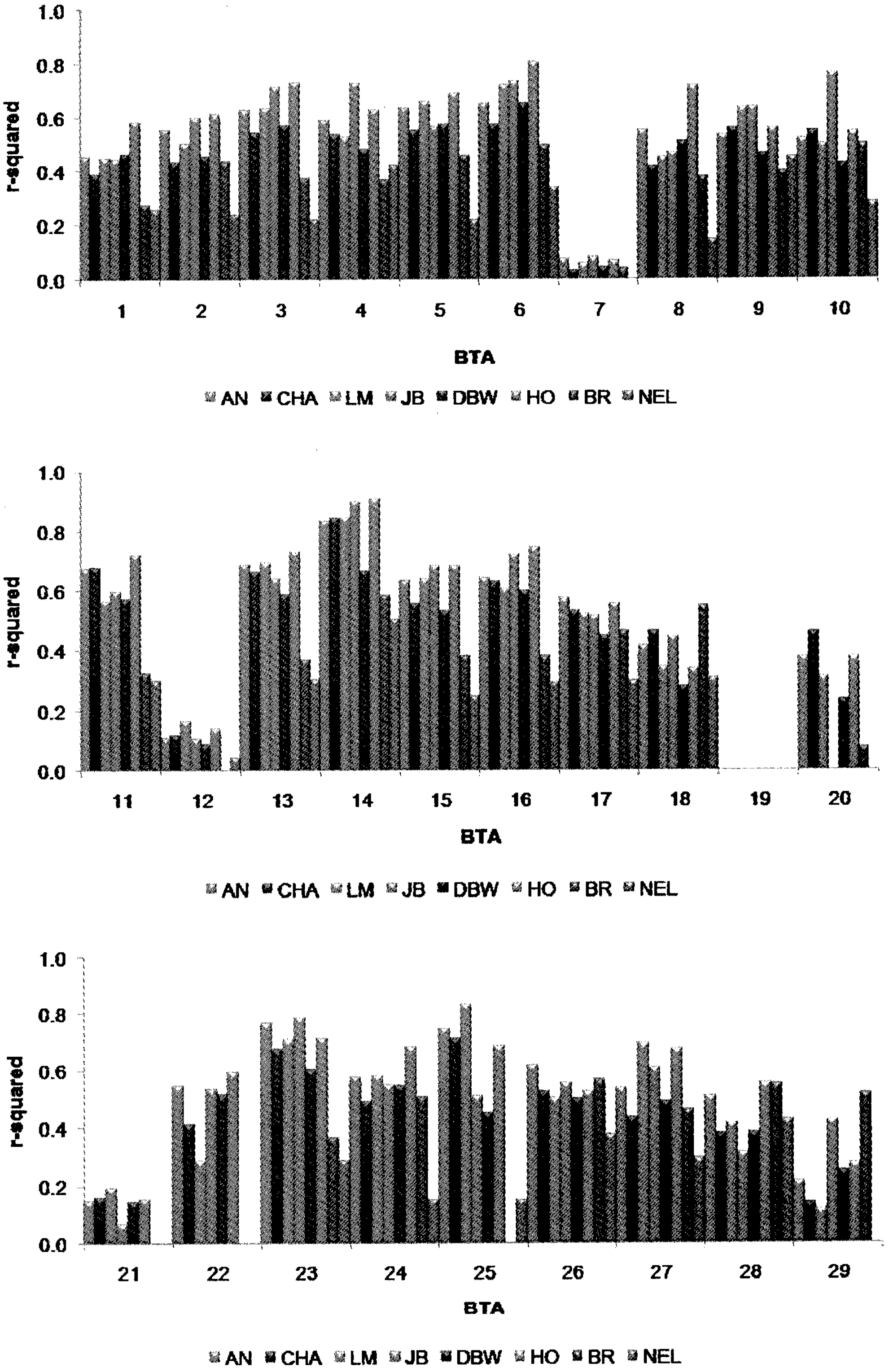

Figure 4.6. Average r^2 values for inter-marker distances <1 kb for each breed and chromosome. Data are not presented for breed by chromosome combinations for which there were less than 5 informative locus pairs (BTA19).

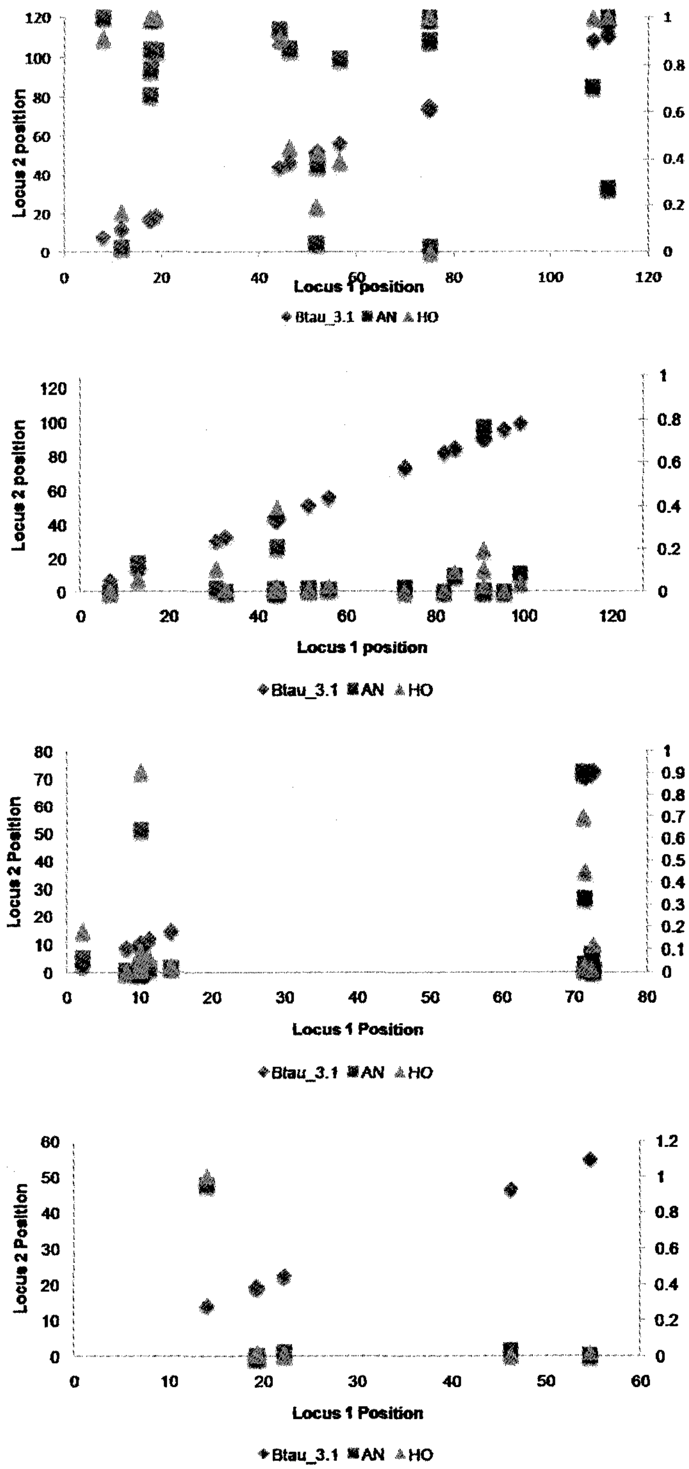

Figure 4.7. Distribution of SNPs along BTA6, 7, 12 and 21 respectively. Measures of LD, r^2, are shown for Angus and Holstein. Measures of LD for Angus are represented by red squares and Holstein LD measures are represented by green boxes. BTA6 was included for comparison purposes.

CHAPTER 5

An Assessment of Population Structure in Eight Breeds of Cattle Using a Whole Genome SNP Panel

5.1 Background

Population structure and diversity within and between breeds of cattle has been studied in order to learn more about the origin, history and evolution of cattle (Troy *et al.*, 2001). Admixture analysis has also been used to characterize the genetic relationships between breeds (Kumar *et al.*, 2003; Macneil *et al.*, 2007; Negrini *et al.*, 2007; Wiener *et al.*, 2004). Kumar *et al.* (2003) used 20 microsatellite markers to estimate the extent of admixture among breeds of cattle from India, Europe and the Near East. Assuming two ancestral populations, the mean admixture coefficients ranged from 0.0 to 0.1 in Indian *Bos indicus* breeds, 0.9 to 1.0 in European *Bos taurus* breeds and from 0.1 to 0.9 in hybrid breeds from the Near East. This variation in admixture coefficients clearly demonstrates the ancestral divergence between the *Bos taurus* and *Bos indicus* subspecies. Similarly, Wiener *et al.* (2004) characterized the diversity within and between eight British breeds of cattle using 30 microsatellite markers. A majority of the allelic variation, 87%, was found to exist within breeds and the breeds of cattle did not cluster according to their geographic source of origin. In a study of the origin of Chirikof Island cattle, MacNeil *et al.* (2007) also found that 86% of the genetic variation in 34 microsatellite loci was found within *Bos taurus* breeds. However, the indigenous Chirikof Island cattle were strongly differentiated from the European *Bos taurus* cattle suggesting that a comparison to Asian *Bos taurus* breeds might next be appropriate. On the other hand, no significant divergence appears to exist between

Chapter 5 is in review at BMC Genetics

geographically separated populations of Holstein cattle probably due to historic occurrences of gene flow between populations and selection for similar traits (Zenger *et al.*, 2007). The extent of nuclear diversity among cattle breeds has yet to be extensively explored. Here, evidence is examined for population substructure and interbreed diversity among eight breeds of cattle using 2,641 autosomal genome-wide SNPs.

5.2 Results and Discussion

Preliminary STRUCTURE analyses were performed to determine the appropriate numbers of iterations for the initial burn-in and estimation phases of the analysis. These preliminary analyses indicated that the probability of the number of ancestral population (the K parameter from STRUCTURE) being greater than five was very small and therefore the analyses of all datasets was restricted to $K \leq 5$ to limit computation time. Analyses were performed on three datasets which used the full complement of markers but varied according to breed representation. The first analysis included data for all eight breeds, the second dataset included only the six taurine breeds and the third analysis included all *Bos taurus* breeds excluding the Japanese Black. The number of ancestral populations (K) that were subsequently admixed to form these breeds was estimated using the method described by Evanno *et al.* (2005) and was found to be no greater than two for each data set (Figure 5.1). The ΔK method of Evanno *et al.* (2005) is unable to identify the optimal K to be one; however, the log-likelihood of the data, logP(X|K) in Figure 5.1, indicates that K=1 can be excluded for all three of the analyzed datasets. The average estimated admixture coefficients (the Q parameters from

STRUCTURE) of individuals from each breed are summarized in Figure 5.2 assuming two ancestral populations in each case.

The results presented in Figures 5.1a and 5.2a demonstrate that a considerable source of variation among cattle is the partitioning of breeds into the *Bos indicus* and *Bos taurus* subspecies. The variance between these groups (Table 5.1) accounted for 18.8% of the total variation ($F_{CT} = 0.19$) which was significant ($P \leq 0.036$). However, 71.06% of the genetic variation was found within populations. However, the SNP loci used within this study were detected because they were common SNPs within *Bos taurus* and their average minor allele frequency was much lower in *Bos indicus*. This ascertainment bias may have resulted in underestimated F_{ST} values and caused the breeds to appear more similar than they really are. The clustering of *Bos taurus* and *Bos indicus* breeds within subspecies using STRUCTURE has previously been reported in an analysis using 20 microsatellites. Kumar *et al.* (2004) found in a model with $K = 2$, that the mean admixture coefficient of taurine breeds ranged from 0.9 to 1.0 while that for the indicine breeds ranged from 0 to 0.1. Our mean admixture coefficients ranged from 0.03 to 0.08 in *Bos indicus,* and but only from 0.54 to 0.67 in *Bos taurus* breeds. While our findings are similar for the *Bos indicus* breeds, results for the *Bos taurus* breeds are substantially different and may be due either to the difference in number of markers examined or, perhaps more likely, due to the vastly different mutation rates between microsatellite and SNP loci. The increased rate of mutation for microsatellite loci would likely lead to greater estimates of divergence as measured by admixture.

A second STRUCTURE analysis was performed using data only from the six taurine breeds to determine if population substructure could be detected among the *Bos taurus* breeds. This analysis partitioned Japanese Black cattle from the cluster comprising the remaining five taurine breeds (Figures 5.1b, 5.2b). However, the partitioning was not strongly supported by the analysis of molecular variance ($F_{CT} = 0.09$; $P < 0.17$; Table 5.1). The mean admixture coefficients for the European taurine breeds ranged from 0.43 to 0.60 while for the Japanese Black ranged from 0.1 to 0.29. The upper and lower quartile range of the admixture coefficients for the individual Japanese Black animals are not as symmetric as found for the European taurine breeds (Figure 5.2b) and are skewed in the direction of the European taurine breeds, suggesting a recent influence of European *Bos taurus* breeds within Japanese Black. Previously published reports describe the use of European breeds to upgrade Japanese Black cattle (Mannen *et al.*, 1998) and support this conclusion. However, it has also been suggested that multiple domestication events occurred among the different strains of aurochs including an independent taurine domestication event in Asia (Mannen *et al.*, 2004; Mannen *et al.*, 1998). This suggestion was based on mitochondrial DNA variation between Japanese Black and European taurine breeds. Our results show that there is as much variation between Japanese Black and European *Bos taurus* breeds as exists among the European breeds (Table 5.1) which may support the hypothesis of an independent taurine domestication event in Asia.

The third STRUCTURE analysis considered the remaining *Bos taurus* breeds after excluding the Japanese Black and resulted in a clustering of the meat and dairy breeds

(Figures 5.1c, 5.2c). The mean admixture coefficients demonstrate considerably less variation among the Continental European and Holstein breeds which is consistent with the small number of Continental breed animals introduced to North America and the history of strong selection for milk production among Holstein cattle. Quite surprisingly, the Dutch Black and White Dairy cattle possessed the greatest variation of all of these breeds suggesting that selection for milk production has been less intense in this breed than in Holsteins. Interestingly, 4.65% of the variation was found among the beef and dairy groups ($F_{CT} = 0.04$) (Table 5.1) with a p value of 0.10 that was suggestive, but not significant. This variation suggests that artificial selection within cattle for alternate agricultural purposes has led to a genome wide divergence among the beef and dairy breeds. Additional analyses in which the genomic regions at which divergence between the types is greatest are overlaid with detected meat and milk QTL would be of considerable interest.

All of the STRUCTURE analyses using the three datasets supported the existence of two ancestral populations partitioning the breed types. Assuming that these represent the true number of ancestral populations, we sought to answer the question, how many loci would be required to precisely estimate the number of ancestral populations? The dataset of 2,641 loci was sampled without replacement to produce 10 datasets with 25, 50 or 100 loci and repeated each of the previous analyses. The results presented in Figure 5.3 show the results of each of the replicate analyses. At the subspecies level, Figure 5.3a, the correct number of ancestral populations was accurately inferred with as few as 25 loci. This is clearly due to the large divergence between the *Bos indicus* and

Bos taurus subspecies which is demonstrated by the difference in the mean admixture coefficients in Figure 5.2a. The second analysis, which included only taurine breeds, demonstrates that using as many as 100 randomly chosen loci only yields the correct number of clusters in 50% of the instances (Figure 5.3b). This is most likely a result of the closer relationship between the taurine breeds(Figure 5.2b), and the presence of two levels of substratification among these breeds (Asian vs. beef vs. dairy). The third analysis, which excluded the Japanese Black breed, more frequently detected two ancestral populations (Figure 5.3c), which primarily detected the remaining beef vs. dairy strata in the data. Not surprisingly, it is evident from this set of analyses that the number of random SNP loci needed to accurately infer population structure is dependent on the divergence between populations. Seldin *et al.* (2006) had similar results when trying to differentiate Northern and Southern European human populations using 400 randomly selected SNP loci and suggested the limited number of SNPs used as the potential problem. Clearly, future studies which seek to evaluate the relationships among closely related cattle breeds will require a substantial number of SNP loci to accurately infer breed relationships.

Finally, pairwise population F_{ST} estimates were generated using the complete dataset of 2,641 loci (Table 5.1) and used these to construct an unrooted Neighbor-Joining tree (Figure 5.4). These estimates of genetic distance and the tree topology support the findings of the STRUCTURE analyses.

5.3 Conclusions

The recent completion of a draft bovine genome sequence assembly has provided sufficient numbers of SNP loci to replace microsatellite loci and augment mitochondrial DNA (mtDNA) sequences for population genetic analyses in cattle. We have shown that SNP loci can be used to identify population substructure among cattle breeds. However, we have demonstrated that a large number of SNP loci must be used to obtain precise population inferences due to the lower information content of individual SNP loci.

5.4 Methods

5.4.1 DNA Collection

I collected DNA from 70 Angus (USA), 20 Canadian Angus, 40 Charolais (Canada), 40 Brahman (USA), 97 Dutch Black and White Dairy cattle (Belgium), 48 Holstein (USA), 65 Japanese Black (Japan), 43 Limousin (USA) and 97 Nelore (Brazil) cattle. Family structure and the number of individuals per family varied between breeds but the general family structure consisted of a grandparent, parent and three or more progeny. In order to phase the chromosomes using linkage information, small families were selected where members within the families were closely related but the families themselves were as unrelated as possible. This three generation family structure allowed for the efficient estimation of marker phase relationships in the progeny and also produced the most likely phase relationships in each of the parents/grandparents.

5.4.2 Marker Selection and Genotyping

A detailed description of the SNP loci used in this study and of the genotyping methods was presented in Chapter 4 of this thesis. Briefly, sequence information for SNPs (Additional Table 5.1) was obtained from public databases and SNPs were genotyped as a GoldenGate® assay using an Illumina BeadStation 500G (Oliphant *et al.*, 2002) (www.illumina.com). The software GENOPROB V2.0 (Thallman *et al.*, 2001a; Thallman *et al.*, 2001b) was used to assess genotype score quality and to produce whole chromosome phased haplotypes based on the pedigree and map locations of the loci.

5.4.3 Population Structure Analysis

I ran the program STRUCTURE and the linkage model of (Falush *et al.*, 2003) to evaluate the extent of substructure among contemporary breeds of European and Asian *Bos taurus* and *Bos indicus* cattle. Exploratory STRUCTURE runs were performed to determine the optimum number of iterations to use for the initial burn-in and estimation phases of the analysis to ensure that the Gibbs sampler had explored a sufficiently large sample space to provide reliable posterior probabilities. From these preliminary analyses, it was determined that an initial burn-in of 10,000 iterations followed by 10,000 iterations for parameter estimation was sufficient to ensure convergence of parameter estimates (data not shown). A series of analyses (runs) were performed based on inclusion of differing combinations of cattle breeds in an attempt to determine the minimum number of ancestral populations that were admixed to best explain the genomic architecture of the current breeds. The first run used all of the animals from all 8 breeds. The second run used the 6 taurine breeds (Angus, Charolais, Limousin, Dutch Black and White Dairy, Holstein and Japanese Black) while the third run used the taurine breeds without the Japanese Black. To estimate the number of populations (the K parameter of STRUCTURE), I analyzed each of these three data sets allowing the value of K to vary from 1 to 5 and repeated each run five times to produce a total of 75 STRUCTURE runs. I used the method of Evanno *et al.* (2005) and calculated ΔK which is an *ad hoc* quantity related to the second order rate of change of the log probability (likelihood) of the data Pr(X|K) (equation 12 in (Pritchard *et al.*, 2000)) with respect to the number of population clusters K.

Assuming the full dataset of 2,641 loci would yield the most accurate estimate of the true number of ancestral populations, we sought to determine the effect the number of loci analyzed would have on inferences of K. Three new data sets, each with 10 replicates, were created by randomly sampling without replacement 25, 50 or 100 loci from the 2,641 markers. Each replicate was analyzed using STRUCTURE as previously described, except the admixture model was used rather than the linkage model, since linkage among the sampled loci was assumed to be lost due to randomly sampling loci throughout the genome.

Finally, I used the program ARLEQUIN (Schneider *et al.*, 2000) to calculate F_{ST} values, population corrected average pairwise differences and analyses of molecular variance (AMOVA) using the phased genotypes produced by GENOPROB V2.0. Significance levels for variance components and F statistics were estimated using 10,000 permutations. MEGA (Kumar *et al.*, 2004) was used to construct a Neighbor-Joining tree from the pairwise Fst values (Table 5.2).

5.5 Literature Cited

Falush D., Stephens M., and Pritchard J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164:** 1567-87.

Kumar P., Freeman A. R., Loftus R. T., Gaillard C., Fuller D. Q., and Bradley D. G. (2003). Admixture analysis of South Asian cattle. *Heredity* **91:** 43-50.

Kumar S., Tamura K., and Nei M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5:** 150-63.

Macneil M. D., Cronin M. A., Blackburn H. D., Richards C. M., Lockwood D. R., and Alexander L. J. (2007). Genetic relationships between feral cattle from Chirikof Island, Alaska and other breeds. *Anim Genet* **38:** 193-7.

Mannen H., Kohno M., Nagata Y., Tsuji S., Bradley D. G., Yeo J. S., Nyamsamba D., Zagdsuren Y., Yokohama M., Nomura K., and Amano T. (2004). Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Mol Phylogenet Evol* **32:** 539-44.

Mannen H., Tsuji S., Loftus R. T., and Bradley D. G. (1998). Mitochondrial DNA variation and evolution of Japanese black cattle (Bos taurus). *Genetics* **150:** 1169-75.

Negrini R., Nijman I. J., Milanesi E., Moazami-Goudarzi K., Williams J. L., Erhardt G., Dunner S., Rodellar C., Valentini A., Bradley D. G., Olsaker I., Kantanen J., Ajmone-Marsan P., and Lenstra J. A. (2007). Differentiation of European cattle by AFLP fingerprinting. *Anim Genet* **38:** 60-6.

Oliphant A., Barker D. L., Stuelpnagel J. R., and Chee M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **Suppl:** 56-8, 60-1.

Pritchard J. K., Stephens M., and Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155:** 945-59.

Schneider S., Roessli D., and Excoffier L. (2000). Arlequin: A software for population genetics data analysis, Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Thallman R. M., Bennett G. L., Keele J. W., and Kappes S. M. (2001a). Efficient computation of genotype probabilities for loci with many alleles: I. Allelic peeling. *J Anim Sci* **79:** 26-33.

Thallman R. M., Bennett G. L., Keele J. W., and Kappes S. M. (2001b). Efficient computation of genotype probabilities for loci with many alleles: II. Iterative method for large, complex pedigrees. *J Anim Sci* **79:** 34-44.

Troy C. S., MacHugh D. E., Bailey J. F., Magee D. A., Loftus R. T., Cunningham P., Chamberlain A. T., Sykes B. C., and Bradley D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410:** 1088-91.

Wiener P., Burton D., and Williams J. L. (2004). Breed relationships and definition in British cattle: a genetic analysis. *Heredity* **93:** 597-602.

Zenger K. R., Khatkar M. S., Cavanagh J. A., Hawken R. J., and Raadsma H. W. (2007). Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. *Anim Genet* **38:** 7-14.

Figures

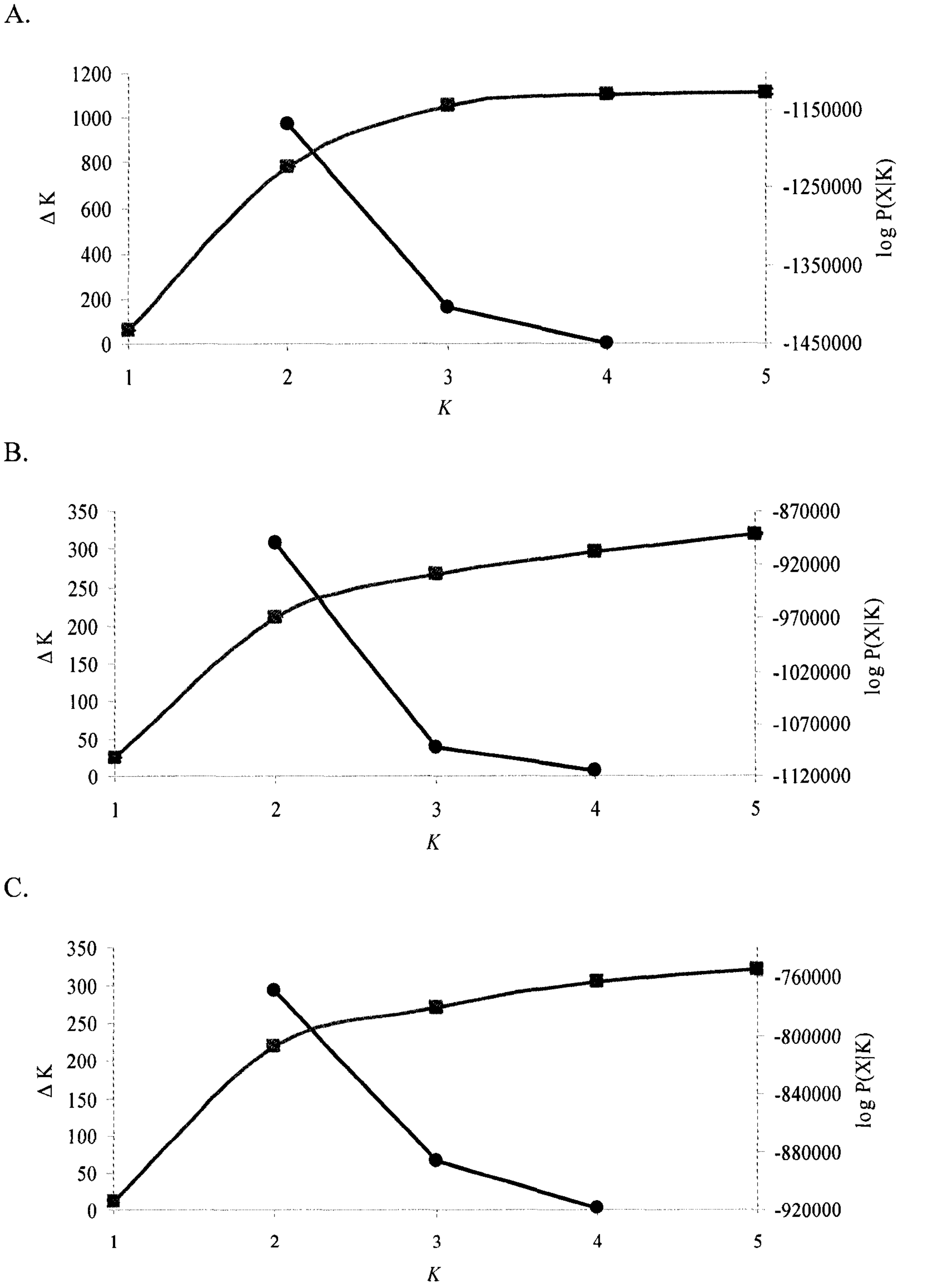

Figure 5.1 (A-C). Δ*K* (-○-) and the mean log *P*(*X*|*K*) (-□-) based on the 5 replicate STRUCTURE runs indicate that *K*=2 is optimal for each dataset. The highest point on the blue line depicts the optimal *K* value. The red line depcits the mean log *P*(*X*|*K*) (-□-) for each K value. (A) All eight breeds included, (B) Only *Bos taurus* and (C) *Bos taurus* without Japanese Black.

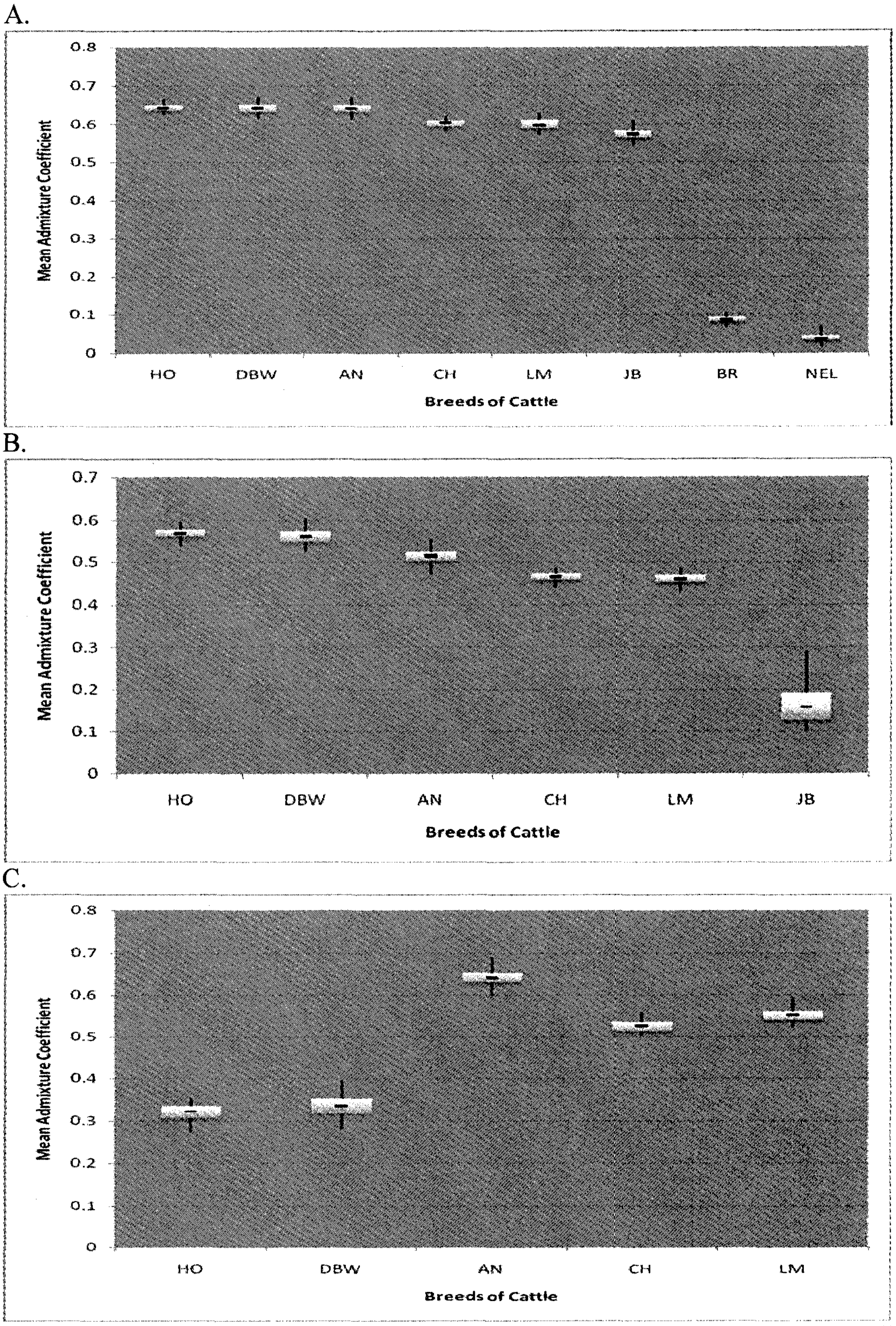

Figure 5.2 (A-C): Box plot of mean individual admixture coefficient for the 5 replicate STRUCTURE runs using *K*=2 for each dataset. (A) All eight breeds included, (B) Only *Bos taurus* and (C) *Bos taurus* without Japanese Black. Breed abbreviations are defined in the text.

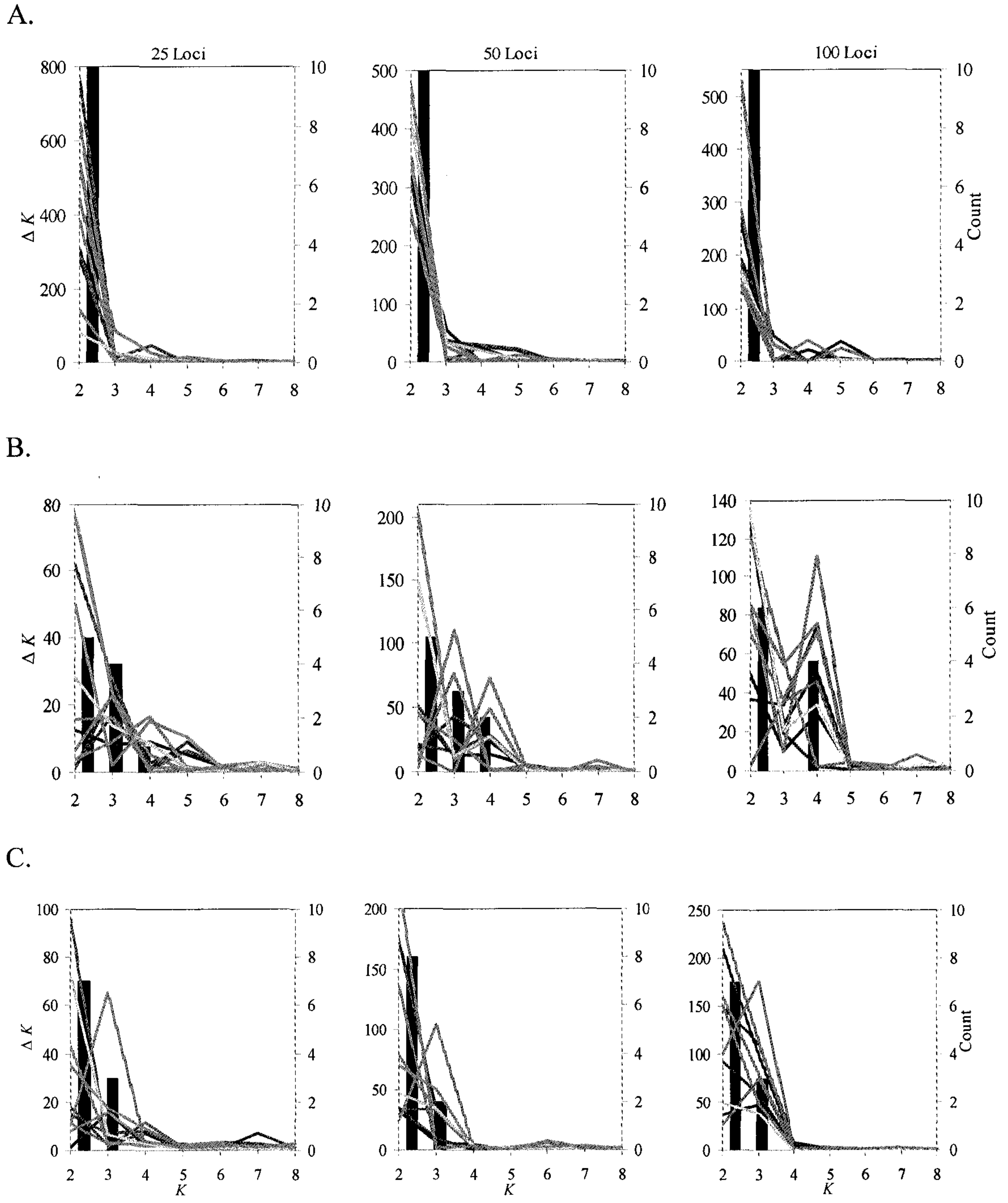

Figure 5.3. Estimation of the number of ancestral populations based on samples of 25, 50 or 100 loci randomly sampled from the complete dataset. In each panel, the solid bars represent the number of times each K was found to be optimal and the colored lines represent the ΔK values for each of the 10 replicate runs. (A) All eight breeds included, (B) All *Bos taurus* breeds and (C) *Bos taurus* without Japanese Black.

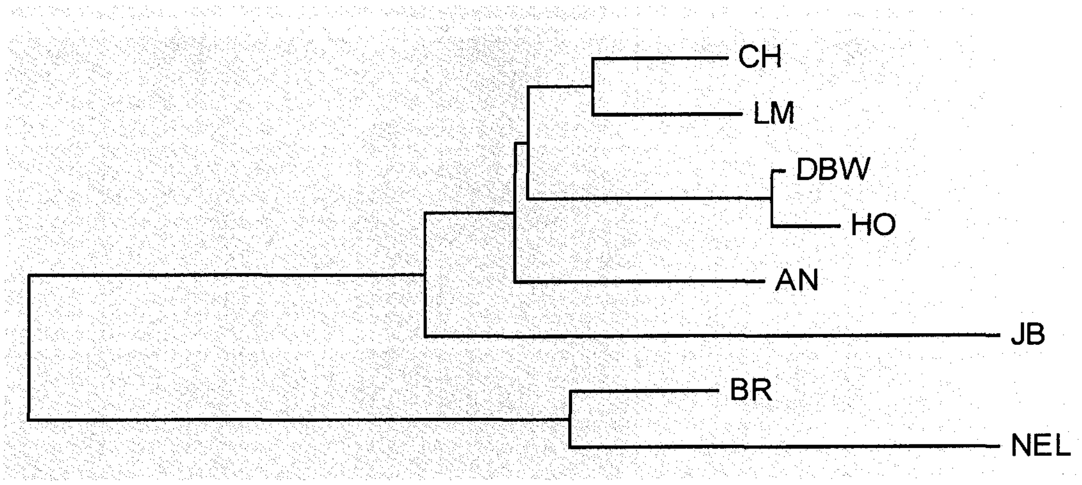

Figure 5.4: Neighbor-Joining Tree based on pairwise Fst values calculated using 2,641 SNP loci showing phylogenetic relationships between cattle breeds. Breed abbreviations are defined in the text.

Tables

Table 5.1 Analysis of Molecular Variance. F_{CT} is the correlation of random haplotypes within a group of populations, relative to that of random pairs of haplotypes drawn from the whole species. F_{SC} is the correlation of the molecular diversity of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the region. F_{ST} is the correlation of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the whole species.

		Variance Components (%)			Fixation indices			
Data Set	# groups (*K*)	Among groups	Among populations within groups	Within populations	F_{CT}	p value	F_{SC}	F_{ST}
All 8 breeds	2	18.79	10.15	71.06	0.19	0.036±0.002	0.12	0.29
only *Bos taurus*	2	8.64	8.3	83.06	0.09	0.168±0.003	0.09	0.17
Bos taurus without Japanese Black	2	4.65	5.59	89.76	0.05	0.101±0.003	0.06	0.10

Table 5.2: Pairwise Fst values based on 2,641 SNP loci. Shown are the following breeds of cattle: Angus (AN), Brahman (BR), Charolais (CHA), Dutch Black and White (DBW), Holstein (HO), Japanese Black (JB), Limousin (LM) and Nelore (NEL). Breed abbreviations are defined in the text.

	AN	BR	CH	DBW	HO	JB	LM
BR	0.271						
CH	0.090	0.273					
DBW	0.105	0.269	0.089				
HO	0.111	0.290	0.094	0.015			
JB	0.185	0.324	0.165	0.183	0.195		
LM	0.088	0.276	0.055	0.097	0.101	0.168	
NEL	0.320	0.112	0.335	0.316	0.350	0.373	0.336

CHAPTER 6

General Discussion and Conclusions

6.1 General Conclusion

Comprehensive radiation hybrid, linkage and linkage disequilibrium maps were successfully generated for all bovine autosomes in eight breeds of cattle. In doing so, a new methodology for high throughput construction of radiation hybrid and linkage maps was developed. Human cattle comparative maps were constructed and the bovine genome sequence assembly version 2.0 was compared to the linkage map generated in Chapter 3. These maps aided in construction of linkage disequilibrium maps for all bovine autosomes.

6.2 Radiation Hybrid and Linkage Map Conclusions

A new high throughput approach to radiation hybrid and linkage mapping was successfully implemented for generation of radiation hybrid maps comprised 4690 SNP markers and 1125 previously typed microsatellite markers. Inclusion of the previously typed microsatellite markers allowed for comparison to the USDA linkage map (http://www.marc.usda.gov/genome/genome.html). Furthermore, linkage maps and subsequent human-cattle comparative maps were generated with 2701 markers that had been RH mapped. Therefore, comprehensive mapping resources have been developed for 2701 SNP markers through generation of RH map locations, linkage map locations, comparative human genome information and comparative information from the bovine genome sequence assembly.

This new high throughput approach has allowed rapid and cost effective assembly of comprehensive maps.

6.3 Linkage Disequilibrium Map Conclusions

Approximately 2670 markers were utilized for construction of whole genome linkage disequilibrium maps in eight breed of cattle. Given that the current size of the bovine genome is estimated to be 2.87 Gb (http://www.hgsc.bcm.tmc.edu/projects/bovine/), it is assumed that a uniform marker distribution would result in an average inter-marker distance of approximately 1 Mb. However, markers are not uniformly distributed. The beneficial aspect of non uniform distribution is that estimation of LD across several orders of magnitude of distance was produced.

Markers across the genome were binned and averaged by breed. Similar to Farnir (2000) and Odani (2006), a decline in LD was found as inter-marker distance increased, however, r^2 reached baseline levels at approximately 500 kb. This is different than previously published findings of several tens of cM (Farnir *et al.*, 2000; Odani *et al.*, 2006). The variation in these findings is attributed to the different measures used to characterize LD. Fortunately, the Dutch Black and White Dairy and the Holstein animals used in this study were a subset of animals used in previous reports (Farnir *et al.*, 2000; Schnabel *et al.*, 2005). Linkage disequilibrium was measured for each of these breeds in Farnir (2000) and Schnabel (2005) and D′ was used as a measure of LD. Therefore, for comparative purposes, D′ measures of LD were generated for all breeds and chromosomes. It was determined that D′ decays with distance and levels of LD agree with previously published reports. However, the D′ values appeared to be overestimated. Consequently, given the current resources available for analysis, r^2 should be used as a measure of linkage disequilibrium gives a better estimation of LD.

Variation in LD was examined between breeds for all chromosomes. With the exception of BTA 7, 12 and 21, r^2 averaged 0.56 across *Bos taurus* breeds. Minimum r^2 values were obtained on BTA29 in Limousin while maximum r^2 values were obtained in Holstein on BTA14. However, BTA7, 12 and 21 displayed average r^2 values that were significantly lower when compared to all other autosomes. There are two possible explanations to the lower than average LD on BTA7, 12 and 21. The first is that a lack of quantitative traits on these chromosomes has meant that selection has not occurred resulting in lower than average measures of LD. Alternatively, the possibility that chromosome structure has influenced the extent of LD was suggested. However, after closer examination, neither of these theories appeared to explain the lower than average LD on BTA7, 12 and 21.

Our results indicate that long range LD in cattle is not as extensive as previously reported and that usable LD may extend only 100 – 500 Kb. Furthermore, bovine chromosomes 7, 12 and 21 have been identified as anomalies with lower than expected LD at very small distances which merits a more in depth analysis. Based on findings in Chapter 4, 30,000 – 50,000 markers will be needed to obtain an optimal LD map. Construction of a higher resolution LD map has the potential to provide additional insight into what role selection and evolutionary forces may play upon quantitative traits and how those may vary between breeds.

6.4 Population Structure Conclusions

Previous work involving population structure between breeds of cattle have utilized mtDNA sequence or microsatellites (Kumar *et al.*, 2003; Mannen *et al.*, 1998; Troy *et al.*, 2001). The recent completion of a draft bovine genome sequence assembly has provided sufficient numbers of SNP loci to replace traditional mtDNA sequence or microsatellite loci as markers for population genetic analyses in cattle. Two thousand six hundred forty one SNPs and eight breeds of cattle were utilized in three types of analysis in order to analyze population structure. These include analysis with the linkage model (Falush *et al.*, 2003) in the program STRUCTURE (Pritchard *et al.*, 2000) as well as generation of Fst values and construction of a Neighbor-Joining tree. Initially, STRUCTURE differentiated between *Bos taurus* and *Bos indicus* breeds. However, the divergence between taurine and indicine breed was so great that it eclipsed population structure within taurine breeds. With the indicine breeds removed and the analysis repeated, Japanese Black were determined to be most divergent among all other taurine breeds. With the *Bos indicus* and Japanese Black removed and the analysis repeated, the meat breeds clustered independently of the dairy breeds. Pairwise Fst measurement and an unrooted Neighbor-Joining tree support the STRUCTURE findings. The successful use of SNP markers to detect population structure between breeds of cattle was demonstrated in Chapter 5.

6.5 Future Direction

Results from the genome wide linkage disequilibrium map reported in this thesis have provided us with the optimal number of markers required for implementing LD as a tool for whole genome association studies. Fortunately, two high throughput arrays containing bovine SNPs are in development. Affymextrix has a GeneChip® Bovine Mapping 10K SNP kit containing 10,000 bovine SNPs and reportedly has a 35,000 SNP kit in development (www.affymetrix.com). Likewise, Illumina has a Bovine BeadChip containing 58,000 bovine SNP markers has recently been produced (www.illumina.com). High resolution SNP platforms such as these will provide us with the optimal number of markers needed for creating linkage disequilibrium and haplotype resources that will expand our abilities to detect causative mutations associated to quantitative traits. Furthermore, additional knowledge can be gained concerning the evolution of a genome under intense selective breeding pressures. An enhanced understanding of quantitative traits and selection have the potential to provide insight into the how the genome functions.

6.6 Literature Cited

Casas E., Shackelford S. D., Keele J. W., Koohmaraie M., Smith T. P., and Stone R. T. (2003). Detection of quantitative trait loci for growth and carcass composition in cattle. *J Anim Sci* **81:** 2976-83.

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., and Georges M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10:** 220-7.

Kumar P., Freeman A. R., Loftus R. T., Gaillard C., Fuller D. Q., and Bradley D. G. (2003). Admixture analysis of South Asian cattle. *Heredity* **91:** 43-50.

Li C., Basarab J., Snelling W. M., Benkel B., Kneeland J., Murdoch B., Hansen C., and Moore S. S. (2004). Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21, and 23 in a commercial line of Bos taurus. *J Anim Sci* **82:** 967-72.

Mannen H., Tsuji S., Loftus R. T., and Bradley D. G. (1998). Mitochondrial DNA variation and evolution of Japanese black cattle (Bos taurus). *Genetics* **150:** 1169-75.

McKay S. D., Schnabel R. D., Murdoch B. M., Aerts J., Gill C. A., Gao C., Li C., Matukumalli L. K., Stothard P., Wang Z., Van Tassell C. P., Williams J. L., Taylor J. F., and Moore S. S. (2007). Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Anim Genet* **38:** 120-5.

Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M., and Sasaki Y. (2006). Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Anim Genet* **37:** 139-44.

Pritchard J. K., Stephens M., and Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155:** 945-59.

Schnabel R. D., Kim J. J., Ashwell M. S., Sonstegard T. S., Van Tassell C. P., Connor E. E., and Taylor J. F. (2005). Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proc Natl Acad Sci U S A* **102:** 6896-901.

Troy C. S., MacHugh D. E., Bailey J. F., Magee D. A., Loftus R. T., Cunningham P., Chamberlain A. T., Sykes B. C., and Bradley D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410:** 1088-91.