University of Alberta

Genome-wide Gametic and Zygotic Linkage Disequilibrium in a Composite Beef Population

by

Qi Jiang

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

Master of Science
in
Animal Science

Department of Agricultural, Food & Nutritional Science

Spring 2012
Edmonton, Alberta

Abstract

During 1960 to 1989, the University of Alberta Kinsella Research Ranch had established synthetic beef breeds as a cost-effective crossbreeding system. Animals from these synthetics were subsequently pooled to form a composite population. Despite many breeding and genomic studies on this population, little is known about its genetic structure. This thesis provided the first genome-wide survey of linkage disequilibrium (LD) at both gametic and zygotic levels for the Kinsella population. The survey was based on the genomic data consisting of 1,023 animals genotyped for 50K SNP markers. Similar genomic structures in gametic and zygotic LD were observed, with zygotic LD decaying faster than gametic LD over marker distance. The high-order trigenic and quadrigenic disequilibria were insignificant and decayed rapidly within a very short marker distance. These results support the current intensive focus on use of high-density markers for fine-scale mapping and genomic selection in the Kinsella population.

Acknowledgements

I convey sincere thanks to my supervisors, Dr. Rong-Cai Yang and Dr. Zhiquan Wang. Without their supports, advices, patience, and encouragement throughout these years, this thesis would not have been possible.

A special thanks to Dr. Stephen Moore in my thesis committee for the past two years. I am greatly indebted for his precious suggestions and comments that have benefited this thesis a lot. I thank Prof. Graham Plastow for stepping in to replace Dr. Moore during my thesis defence and for his editorial assistance.

This thesis research has been supported by grants from the Alberta Livestock and Meat Agency (ALMA 2008 F175R, ALMA 2007 F142R) to Dr. Zhiquan Wang, and the Natural Sciences and Engineering Research Council of Canada (NSERC OGP0183983) to Dr. Rong-Cai Yang.

The Department of Agricultural, Food and Nutrition Science (AFNS) provided the facilities and an incredible atmosphere of intellectual excitement and rigorous inquiry that will always stay with me. It was a privilege and fun to work with its great staff members.

I would like to extend thanks to my mother for her great love and supports. She always encouraged me to challenge myself and not to be afraid of adventure. My gratitude also goes to my sisters and their families. They were always my strongest backups.

At last, I would like to give my special thanks to my dearest daughter Alicia for her unwavering love and pleasure. The white coffee she made always kept me energetic during the period of the thesis writing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

cM	Centimorgan
CLD	Composite linkage disequilibrium
GLD	Gametic linkage disequilibrium
GS	Genomic selection
GWAS	Genome-wide association studies
HWD	Hardy-Weinberg disequilibrium
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAS	Marker-assisted selection
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
ZLD	Zygotic linkage disequilibrium

Chapter 1. Introduction and Literature Review

1.1 Introduction

Linkage disequilibrium (LD) is defined as a non-random association of alleles at two or more loci [1], and it is a sensitive indicator of the genetic and demographic forces that influence the genome structure [2]. LD has been extensively studied in the field of population genetics for description of demographic change, construction of evolutionary history and prediction of effective population size [2]. Earlier LD studies in livestock populations were limited to theoretical analysis and computer simulation because genome-wide genomic data was not available until the last decade [3]. Recently, LD has received a considerable amount of attention as it has become a major tool for fine-scale mapping of quantitative trait loci (QTL) [4] and for genomic selection [5].

The recent advancement in molecular biology has enabled the rapid development of single nucleotide polymorphism (SNP) genotyping technology, thereby making the genotyping of cheap and abundant SNP markers possible in many livestock species [6]. The advancements of high throughput genotyping technology using high density SNP panels have provided an opportunity to conduct LD studies in livestock populations [7, 8]. Recently, there is an increasing interest in exploring LD in populations of different livestock species for understanding their past evolutionary and demographic events, for mapping and fine-mapping genomic regions that are associated with economically important traits and for selecting genetically desirable animals [9] in livestock breeding programs.

LD is important to evolutionary biologists and geneticists because there are many factors that affect LD or are affected by LD [2, 10]. The pattern and extent of LD throughout the genome provide information about the population history [11], the breeding system and constraints or potential to responses of selection [2]. The success of fine-scale QTL mapping and genomic selection depends mainly on the strength of LD between markers and QTL [9, 12, 13]. Therefore, quantifying

and characterizing of LD between loci across the entire genome become the important first step towards fine-scale QTL mapping [14], genome-wide association studies (GWAS) [15], marker-assisted selection (MAS) [16], genomic selection (GS) [12], and enhanced understanding of genomic architecture and population structure [17].

MAS and GS have become increasingly important in livestock genetic improvement programs to further accelerate the genetic gain through increasing the selection accuracy and reducing the generation interval in many domestic animals including dairy cattle [18, 19], beef cattle [13, 20-21], and swine [22-24].

1.2 Linkage disequilibrium

In this section, we will first provide an overview of commonly used LD measures (see Figure 1.1 for pictorial description of concepts of these LD measures). We then give a brief discussion on different causes of LD and their effects on genome structure. The discussion will be made with a special reference to livestock populations.

1.2.1 Common LD measurements

The term linkage disequilibrium (LD) was first introduced by Lewontin and Kojima [1] to describe a nonrandom association of alleles at two or more loci. Unfortunately, the use of this term has caused a great deal of confusion because the non-random association may occur between unlinked loci as well. Subsequently, a different term gametic phase disequilibrium or simply gametic disequilibrium [25] is suggested to reflect the true meaning of the term: the presence of gametic disequilibrium indicates a non-random association between alleles at linked as well as independent loci within a gamete. The use of gametic disequilibrium also clarifies the need for several other disequilibria because all these disequilibria are required for a complete characterization of multilocus associations in a general diploid (zygote) population [26,27]. However, since the

term LD remains widely used, it will be included as part of the name for each disequilibrium measure. For example, gametic disequilibrium will be called as gametic LD instead.

1.2.1.1 Gametic LD

Consider two SNP loci (say loci *A* and *B*), each with two alleles, *A* and *a* at locus *A*, and *B* and *b* at locus *B*. In this system, there are four possible gamete types, *AB*, *Ab*, *aB*, and *ab*, with relative frequencies p_{AB}, p_{Ab}, p_{aB}, and p_{ab}. If the two loci are completely independent, the gametic frequencies can be calculated by the products of the allele frequencies. For example, for gamete *AB*, this would mean that $p_{AB} = p_A p_B$, where $p_A = p_{AB} + p_{Ab}$, and $p_B = p_{AB} + p_{aB}$ are the frequencies of alleles *A* at locus *A* and allele *B* at locus *B*, respectively. This is the case where the population is in linkage equilibrium. Otherwise, the population is in gametic LD (i.e., $p_{AB} \neq p_A p_B$) and the gametic LD is simply measured by

$$D_{AB} = p_{AB} - p_A p_B. \quad (1.1)$$

This definition of gametic LD is applicable to each of four gametes:

$$D_{AB} = p_{AB} - p_A p_B = p_{ab} - p_a p_b = -(p_{Ab} - p_A p_b) = -(p_{aB} - p_a p_B). \quad (1.2)$$

where the change in signs for the last two expressions indicates that the deviation between actual and expected gametic frequencies in the coupling phase must be equal but opposite in sign to those in the repulsion phase [28]. Thus, the gametic LD is also defined as half the difference in frequency between coupling and repulsion heterozygotes [29],

$$D_{AB} = p_{AB} p_{ab} - p_{Ab} p_{aB}. \quad (1.3)$$

Clearly, there is the dependence of the gametic LD on gene frequencies. Several methods are available to partially remove such dependence (see ref. [19] for review). One commonly used approach is to normalize the gametic LD measure by recognizing that D_{AB} measures the covariance of four types of gametes in a 2×2 contingency table [30] and that a normalized measure of D_{AB} is

simply the correlation calculated from the contingency table. Thus, the squared correlation is given by:

$$r^2_{D_{AB}} = \frac{D^2_{AB}}{p_A p_a p_B p_b}. \quad (1.4)$$

The gametic LD is a straightforward measure of LD if the haplotype data is available. Complications for the calculation of gametic LD coefficients arise when only genotype data in diploid individuals are available and the gametic phase of individuals that are heterozygous at two or more loci cannot be directly observed or specified. For example, a heterozygous individual at locus *A* is *A/a* and the same individual at locus *B* is *B/b*, but it is often impossible to distinguish between the two double heterozygotes *AB/ab* and *Ab/aB*, so that the phases and frequencies of gametes cannot be inferred. In other words, it is uncertain whether the double heterozygote is made up from an *AB* haplotype and an *ab* haplotype (double heterozygote in coupling phase) or instead is made up from an *Ab* haplotype and an *aB* haplotype (double heterozygote in repulsion phase). Thus, while the gametic LD has the advantages of being simple to compute and easy to understand, it is a useful measure only for the analysis of diploid data from a Hardy-Weinberg Equilibrium (HWE) population in which there is no deed to distinguish coupling-phase from repulsion-phase double heterozygotes. Thus the gametic LD may not be appropriate for a nonequilibrium population where a complete characterization of two-locus association also requires other disequilibria [26, 31].

1.2.1.2 Composite LD

The composite LD was developed by Peter Burrows in an unpublished manuscript but was first introduced by Cockerham and Weir [32] in an attempt to avoid the problem arising from the uncertainty in the phase of the double heterozygotes. It is a combined measure of the gametic (D_{AB}) and non-gametic ($D_{A/B}$) disequilibria simultaneously based on the allelic and genotypic frequencies obtainable directly from data. The non-gametic LD is defined as

$D_{A/B}=P_{A/B} - p_A p_B$,

with $P_{A/B}$ being the joint frequency of allele *A* present in one gamete and allele *B* present in a different gamete. Thus, the composite LD is simply the sum of gametic and non-gametic disequilibria [33] as:

$$\Delta_{AB} = D_{AB} + D_{A/B}. \quad (1.5)$$

The scaled statistics r^2 is formulated as follows:

$$r^2_{\Delta_{AB}} = \frac{\Delta^2_{AB}}{(p_A p_a + D_A)(p_B p_b + D_B)}, \quad (1.6)$$

where $D_A = P_{Aa} - 2p_A p_a$ and $D_B = P_{Bb} - 2p_B p_b$ measure the departures from HWE, with P_{Aa} and P_{Bb} being the frequencies of genotypes *Aa* and *Bb* at loci *A* and *B*, respectively. $r^2_{\Delta_{AB}}$ would be the same as $r^2_{D_{AB}}$ if it was estimated in a HWE population [33]. Composite LD has been used to study multilocus structures in different species including Drosophila [33, 34] and Canine [35].

As mentioned above, the composite LD measures the disequilibria between alleles at two loci within and between gametes regardless of whether the population is in HWE or not. The advantage of the composite LD is to relax the HWE assumption with the population, but it is still limited to the association between two non-allelic genes. For a diploid individual, there are four genes at two loci. Thus, to further include three-gene (trigenic) and four-gene (quadrigenic) disequilibria, a different two-locus association measure at the zygote level is needed.

1.2.1.3 Zygotic LD

The concept of zygotic LD (zygotic associations) was first introduced by Haldane [36] who showed that partial inbreeding may either increase or decrease the double heterozygosity in a population. Later, Bennet & Binet [37] investigated the effect of inbreeding on the gametic and zygotic imbalance and obtained mathematical expressions of zygotic associations for a pair of loci in a population

undergoing mixed self and random mating. More recently, Yang [31] developed a set of summary statistics for the measure of multilocus zygotic LD for both equilibrium and non-equilibrium population. For locus *A* and *B*, this can be defined as the deviation between the joint heterozygote frequencies:

$$\omega_{AB} = H_{AB} - H_A H_B, \tag{1.7}$$

where $H_{AB} = P_{ab}^{AB} + P_{aB}^{Ab}$ is the joint heterozygosity between loci *A* and *B*, with P_{ab}^{AB} and P_{aB}^{Ab} being the observed frequencies of coupling-phase and repulsion-phase double heterozygotes, respectively; $H_A = P_{a.}^{A.}$ and $H_B = P_{.b}^{.B}$ are the heterozygosities at loci *A* and *B*, respectively. The diploid (zygotic) analogy to the normalized gametic LD is calculated as the squared correlation ($r_{\omega_{AB}}^2$)

$$r_{\omega_{AB}}^2 = \frac{\omega_{AB}^2}{H_A(1-H_A)H_B(1-H_B)} . \tag{1.8}$$

This zygotic LD includes all two-, three- and four-gene disequilibria at two loci in a zygote that would arise from union of two gametes [31]. As zygotic LD is calculated directly from counting zygotes, there is no need for the HWE assumption. This LD measure has provided a method to examine and characterize populations with complex genetic structures such as the Kinsella composite beef population from the University of Alberta.

1.2.2 Causes of linkage disequilibrium in livestock populations

Shown in Figure 1.2 are the genetic and demographic factors that can cause LD [38]. In many crop plants and laboratory animals, LD can also be created by deliberately crossing between two inbred lines of contrast genotypes. The populations from such crosses are often used for QTL mapping. However, this type of LD will not be discussed here.

In livestock populations, finite population size is generally implicated as a major force that can cause the LD [39]. The effective population sizes for most

livestock populations are relatively small. For example, in *Bos taurus* cattle, the effective population size was large (>50,000) before domestication, but was drastically declined to 1,000-2,000 after domestication and, in many breeds, was further declined to approximately 100 after recent breed formation [40, 41]. The pattern of LD observed in a population depends on the history of the population, especially the history of its effective population size [42]. A small effective population size means that alleles in the current population coalesce in a common ancestor in the pedigree in only a few generations [9]. Such a few coalescences can cause genome-wide LD and can have significant effects on a genome structure.

Mixing and crossbreeding (migration) between two or more populations with distinct genes frequencies can cause a large amount of LD in the crossbred population. However, the LD would be small if the breeds do not differ markedly in gene frequencies, and the LDs can only last for a limited number of generations [39].

Selection is another important force that can cause LD in livestock populations. Directional (truncation) selection is known to reduce the genetic variance among animals observed before selection. This reduction is due to the presence of negative gametic LD that is induced by the selection [29, 42]. Most livestock populations undergo some degrees of truncation selection. The size of the impact of selection is depends on the selection intensity. The accumulative impact also depends on the length of the generation interval of the specific species. Generally, the higher selection intensity with a shorter generation interval can accumulate a relatively large amount of LD. However, it should be recognized that the effect of selection on LD must be trait specific and thus involves genes or genomic regions localized at certain parts of the genome.

There are many mutations that cause monogenic genetic defects in cattle and other animal species (see Table 1 of ref. [9] for a list). When these mutations occur, there are LDs that are generated between the mutated allele and alleles at neighbouring loci. For polygenic complex traits, the effects of individual

mutations are probably small and thus the amount of LD induced by such mutations is likely too small as well. For practical livestock improvement, mutation is probably a rare event and thus will likely have a little impact. LDs that are generated due to mutation are localized over the genome.

Animals of consanguineous mating are more likely to be homozygous by descent for a large chromosome segment containing a causative gene than those of random mating [44]. In other words, LDs arise between the causative gene and neighbouring loci in inbred populations. The intensity of LD (i.e., homozygosity by descent) increases with the level of inbreeding [45]. For example, LD would be stronger for siblings than for second-cousin mating.

1.3 LD studies in livestock populations

LD has been examined extensively in livestock populations, especially from those economically important domestic species, such as cattle, pig, sheep and dogs. Most of the LD studies have focused on gametic LD with the HWE assumption.Gardard and Hayes [9] provided a brief review on gametic LD levels among different farmed animals.

1.3.1 LD in pigs

The reported extent of gametic LD in pigs is considerably larger than that in cattle. Du et al. [46] evaluated the extent of gametic LD in six commercial pig lines using 4,500 SNP markers, and reported the average value of r^2 around 0.2 for SNP markers that are 1cM apart. Amaral et al. [47] reported that the average r^2 value of 0.30 was observed for markers spanning around 0.1 cM in European pig breeds but the same r^2 value was observed at a much shorter distance of 0.005 cM for Chinese breeds. Based on the average r^2 of 0.3 with markers being 0.1 cM apart, they recommended a marker spacing of 0.1 cM for a whole genome association study in European breed pig populations with an assay of 30,000 evenly distributed SNPs to cover the entire genome [47]. The required marker

density for GWAS in pigs is smaller than what would be needed for GWAS in cattle.

1.3.2 LD in sheep

Meadow et al. [48] evaluated the extent of gametic LD in five domestic sheep populations. They reported short range gametic LD for the distance of 0 – 5 cM in all five populations. However the persistence with increasing distance and magnitude of LD varied considerably among the populations. The LD decayed faster within the crossbreeds than within purebreds. This confirmed that LD is likely to be breed-specific; LD information is important for the design of successful genome scans in sheep.

1.3.3 LD in dog

Liu et al. [35] investigated the extent and distribution of different LD measures across the canine genome, using 247 microsatellite markers genotyped for a total of 148 dogs. They evaluated LD using Weir's [33] composite LD, and trigenic and quadrigenic genic disequilibria between two loci. The study found that the composite digenic was stronger than the trigenic and quadrigenic disequilibria. They also observed considerable variation in individual genic disequilibria among different chromosomes. Perhaps, the large variation of the LD patterns in dog genome was at least partially due to the low density marker panel and marker distributions on the genome that they used.

1.3.4 LD in cattle

Cattle may be the most extensively studied of livestock species because the sequencing of the Bovine genome was the first to complete in livestock species. The first whole-genome gametic LD study in Dutch Holstein cattle was based on a few hundreds of microsatellite markers [4]. Several subsequent studies have confirmed the extensive gametic LD in cattle, described the LD patterns. These

studies revealed that different normalized measures of gametic LD such as r^2 and D' yielded different conclusions in terms of the strength of LD [49-54]. The LD studies based on the Affymetrix 10K SNP array in Holstein population of North America found a lower level of gametic LD for SNP pairs than previously reported [55,56]. More recently, several studies have been published, on using the Illumina Bovine SNP50K Beadchip to analyze gametic LD structure on the dairy populations. Qanbari et al [57] found lower levels of gametic LD at marker distance ≤100kb, and estimated about 26% of useful LD (r^2>0.25) on average adjacent marker distance of 50 kb-75 kb in Germen Holstein cattle,.

Only recently was the extent of gametic LD examined in beef cattle population. A large, extended gametic LD in Japanese brown and Japanese black cattle based on a few hundreds of microsatellite markers [50]. Mckay et al. [51] compared gametic LD patterns in eight cattle breeds, and found that the gametic LD for marker pairs spanning no more than 500 kb.

Almost all of the previous LD studies in cattle population has been mainly focused pure breed Holstein [40, 55, 57, 58], Jersey [40], Angus [40] and Japanese Black and Brown beef cattle [50] populations. However, commercial stock production often involves crossing of several breeds or lines to generate crossbreds or hybrids [17]. There are more factors shaping the pattern and distribution of LD in an admixture population and no study of LD in crossbred populations have been reported yet.

1.4 The Kinsella composite beef population

The animal population studied in this thesis is the beef composite population at the Kinsella Research Ranch of the University of Alberta. This population is the progenies of three synthetic lines that had been maintained at the Kinsella Research Ranch during 1960 to 1989 [59]. The beef synthetic 1 was established in 1960, mainly composed of Charolais, Angus, and Galloway. The beef synthetic 2 was established in 1982, made up of approximately 60% Hereford and 40% other beef breeds. The dairy beef synthetic was composed of

approximately 60% dairy cattle and 40% of other breeds [59]. The three synthetic lines were subsequently pooled to form the current composite beef population. Obviously this composite population arose from recent mixing of the synthetic lines that were under a moderate selection for growth traits and reproduction abilities. The population may not be in HWE.

The Kinsella beef composite population has been the subject for at least 10 research projects from QTL mapping, candidate genes identification to MAS and GS studies (eg. 60, 61). However, the extent and patterns of gametic LD or other LD measures in this population has never been examined. It would be desirable to fully investigate the extent of LD and their distribution patterns of the whole genome for this population to provide the baseline information about multilocus structures for helping future genomic research. In addition, the beef composite is an excellent population to test the different LD measures (gametic, composite, and zygotic LDs) and to examine relative contributions of the different genic disequilibrium components to the zygotic LD measurements.

1.5 Objectives

This thesis is designed to investigate the extent and distribution of different LD measures at both gametic and zygotic levels in the Kinsella composite beef population. The objectives of this thesis are:

1. To conduct a comparative assessment of extent and patterns of gametic, composite and zygotic LDs;
2. To determine the significance of different components in the zygotic LD.

1.6 References

1. Lewontin RC, Kojima K: **The evolutionary dynamics of complex polymorphisms.** *Evolution* 1960, **14**:458-72.
2. Slatkin M: **Linkage disequilibrium-understanding the evolutionary past and mapping the medical future.** *Nat Rev Genet* 2008, **9**:477-485.
3. Terwilliger JD, Weiss KM: **Linkage disequilibrium mapping of complex disease: fantasy or reality?** *Cur Opin Biot* 1998, **9**:578-594.
4. Farnir, F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M: **Extensive genome-wide linkage disequilibrium in cattle.** *Genome Res* 2000, **10**:220-227.
5. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
6. Hayes BJ, Chamberlain AJ, McPartlan H, MacLeod I, Sethuraman L, Goddard ME: **Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle.** *Genet Res Camb* 2007, **89**:215-220.
7. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, C'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP: **Development and characterization of a high density SNP genotyping assay for cattle.** *PLoS ONE* 2009, **4**:e5350.
8. Ramos AM, Crooijmans RPMA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Dehais P, Affara NA, Hansen MS, Hedegaard J, Hu Z-L, Kerstens HH, Law AS, Megens HJ, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Clark R, Churcher C, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS ONE* 2009, **4**:e6524.

9. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programs.** *Nat Rev Genet* 2009, **10**:381-391.
10. Hu XS, Yeh FC, Wang Z: **Structural genomics: correlation blocks, population structure, and genome architecture.** *Current genomics* 2011, **12**:55-70.
11. Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S, Matukumalli LK, McEwan JC, Nazareth LV, Schnabel RD, Weinstock GM, Wheeler DA, Ajmone-Marsan P, Boettcher PJ, Caetano AR, Garcia JF, Hanotte O, Mariani P, Skow LC, Sonstegard TS, Williams JL, Diallo B, Hailemariam L, Martinez ML, Morris CA, Silva LO, Spelman RJ, Mulatu W, Zhao K, Abbey CA, Agaba M, Araujo FR, Bunch RJ, Burton J, Gorni C, Olivier H, Harrison BE, Luff B, Machado MA, Mwakaya J, Plastow G, Sim W, Smith T, Thomas MB, Valentini A, Williams P, Womack J, Woolliams JA, Liu Y, Qin X, Worley KC, Gao C, Jiang H, Moore SS, Ren Y, Song XZ, Bustamante CD, Hernandez RD, Muzny DM, Patil S, San Lucas A, Fu Q, Kent MP, Vega R, Matukumalli A, McWilliam S, Sclep G, Bryc K, Choi J, Gao H, Grefenstette JJ, Murdoch B, Stella A, Villa-Angulo R, Wright M, Aerts J, Jann O, Negrini R, Goddard ME, Hayes BJ, Bradley DG, Barbosa da Silva M, Lau LP, Liu GE, Lynn DJ, Panzitta F, Dodds KG: **Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds.** *Science* 2009, **324**:528-532.
12. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
13. Garrick DJ: **The nature, scope and impact of genomic prediction in beef cattle in the United States.** *Genet Sel Evol* 2011, **43**:17.
14. Meuwissen TH, Goddard ME**: Fine mapping of quantitative trait loci using linkage disequilibria with closely linked markers.** *Genetics* 2000,

155:421-430.

15. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516-1517.
16. Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR: **Extent of linkage disequilibrium in Holstein cattle in North America.** *J Dairy Sci* 2008, **91**:2106-2117.
17. Dekkers JCM, Hospital F: **The use of molecular genetics in the improvement of agricultural populations.** *Nat Rev Genet* 2002, **3**:22-32.
18. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: Reliability of genomic predictions for North American Holstein bulls.** *J Dairy Sci* 2009, **92**:16-24.
19. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME: **Accuracy of genomic breeding values in multi-breed dairy cattle populations.** *Genet Sel Evol* 2009, **24**:41-51.
20. MacNeil MD, Nkrumah JD, Woodward BW, Northcutt SL: **Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators.** *J Anim Sci* 2010, **88**:517-522.
21. Tang G, Li X, Plastow G, Moore SS, Wang Z: **Developing marker-assisted models for evaluating growth traits in Canadian beef cattle genetic improvement.** *Livestock Science* 2011, **138**:62-68.
22. Ibanez-Escriche N, Fernando RL, Toosi A, Dekkers JC: **Genomic selection of purebreds for crossbred performance.** *Genet Sel Evol* 2009, **41**:12.
23. Toosi A, Fernando RL, Dekkers JCM: **Genomic selection in admixed and crossbred populations.** *J Anim Sci* 2010, **88**:32-46.
24. Sun XC, Habier D, Fernando RL, Garrick DJ, Dekkers JCM: **Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods**. *BMC Proceedings* 2011, **5**:S13.
25. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117**:331-341.

26. Cockerham CC, Weir BS: **Descent measures for two loci with some applications.** *Theor Popul Biol* 1973, **4**:300-330.
27. Weir BS, Cockerham CC: **Complete characterization of linkage disequilibrium at two loci.** In *Mathematical evolutionary theory* Edited by: Feldman MW. Princeton, NJ: Princeton University Press; 1989:86-110.
28. Mueller JC: **Linkage disequilibrium for different scales and applications.** *Briefings in Bioinformatics* 2004, **5**:355-364.
29. Falconer DS, MacKay TFC: *Introduction to Quantitative Genetics.* Harlow: Prentice Hall; 1996.
30. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38**:226-231.
31. Yang RC: **Zygotic associations and multilocus statistics in a nonequilibrium diploid population.** *Genetics* 2000, **155**:1449-1458.
32. Cockerham CC, Weir BS: **Digenic decent measures for finite populations.** *Genet Res Camb* 1977, **30**:121-147.
33. Weir BS: **Inferences about linkage disequilibrium.** *Biometrics* 1979, **35**:235-254.
34. Laurie-Ahlberg CC, Weir BS: **Allozymic variation and linkage disequilibrium in some laboratory populations of Drosophila Melanogaster.** *Genetics* 1979, **92**:1295-1314.
35. Liu T, Todhunter RJ, Lu Q, Schoettinger L, Li H, Littell RC, Burton-Wurster N, Acland GM, Lust G, Wu R: **Modelling extent and distribution of zygotic disequilibrium: implications for a multigenerational canine pedigree**. *Genetics* 2006, **174**:439-453.
36. Haldane JBS: **The association of characters as a result of inbreeding and linkage.** *Ann Eugen* 1949, **15**:15-23.
37. Bennett JH, Binet FE: **Association between Mendelian factors with mixed selfing and random mating.** *Heredity* 1956, **10**:51–55.
38. Lander ES, Schork NJ: **Genetic dissection of complex traits.** *Science* 1994, **265**:2037-2048.

39. Goddard ME: **Mapping genes for quantitative traits using linkage disequilibrium.** *Genet Sel Evol* 1991, **23**:131s-134s.
40. De Roos APW, Hayes BJ, Spelman R, Goddard ME, **Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle.** *Genetics* 2008, **179**:1503-1512.
41. MacEachern S, Hayes B, McEwan J, Goddard M: **An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle.** *BMC Genomic* 2009, **10**:181.
42. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME: **Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size.** *Genome Res* 2003, **13**:635-643.
43. Bulmer MG: **The effect of selection on genetic variability.** *Am Natur* 1971, **105**:201-211.
44. Charlier C, Farnir F, Berzi P, Vanmanshoven P, Brouwers B, Vromans H, Georges M: **Application to map the bovine syndactyly locus to chromosome 15. Identity-by-descent mapping of recessive traits in livestock.** *Genome Res* 1996, **6**:580-589.
45. Lander ES, Botstein D: **Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children.** *Science* 1987, **236**:1567-1570.
46. Du FX, Clutter AC, Lohuis MM: **Characterizing linkage disequilibrium in pig populations.** *Biol Sci* 2007, **3**:166-178.
47. Amaral AJ, Megens HJ, Kerstens HHD, Heuven HCM, Dibbits B, Crooijmans R, den Dunnen JT, Groenen MAM: **Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome.** *BMC Genomics* 2009, **10**:374.
48. Meadow JRS, Chan EKF, Kijas JW: **Linkage disequilibrium compared between five poppulations of domestic sheep.** *BMC Genetics* 2008, **9**:61.

49. Khatkar MS, Collins A, Cavanagh JA, Hawken RJ, Hobbs M, Zenger KR, Barris W, McClintock AE, Thomson PC, Nicholas FW, *et al.*: **A first-generation metric linkage disequilibrium map of bovine chromosome 6.** *Genetics* 2006a, **174**:79-85.
50. Odani M, Narita A, Watanabe T, Yokouchi K, Sugimoto Y, Fujita T, Oguni T, Matsumoto M, Sasaki Y: **Genome-wide linkage disequilibrium in two Japanese beef cattle breeds.** *Anim Genet* 2006, **37**:139-144.
51. McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Dias Neto E, Gill CA, Gao C, *et al.*: **Whole genome linkage disequilibrium maps in cattle.** *BMC Genomics* 2007, **8**:74.
52. Marques E, Schnabel RD, Stothard P, Kolbehdari D, Wang Z, Taylor JF, Moore SS: **High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle.** *BMC Genomics* 2008, **9**:45.
53. Nilsen H, Hayes B, Berg PR, Roseth A, Sundsaasen KK, Nilsen K, Lien S: **Construction of a dense SNP map for bovine chromosome 6 to assist the assembly of the bovine genome sequence.** *Anim Genet* 2008, **39**:97-104.
54. Prasad A, Schnabel RD, McKay SD, Murdoch B, Stothard P, Kolbehdari D, Wang Z, Taylor JF, Moore SS: **Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle.** *Anim Genet* 2008, **39**:597-605.
55. Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR: **Extent of linkage disequilibrium in Holstein cattle in North America.** *J Dairy Sci* 2008, **91**:2106-2117.
56. Kim ES, Kirkpatrick BW: **Linkage disequilibrium in the North American Holstein population.** *Anim Genet* 2009, **40**:279-288.
57. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H: **The pattern of linkage disequilbriom in German Holstein cattle.** *Anim Genet* 2010, **41**:346-356.

58. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, Schnabel RD, Taylor JF, Raadsma HW: **Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel.** *BMC Genomics* 2008, **9**:187.
59. Berg RT, Makarechian M, Arthur PF: **The University of Alberta beef breeding project after 30 years—A review.** *Univ Alberta Annu Feeders' Day Rep* 1990, **69**:65-69.
60. Mujibi FDN, Nkrumah JD, Durunna ON, Stothard P, Mah J, Wang Z, Basarab J, Plastow G, Crews Jr DH, Moore SS: **Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle.** [http://jas.fass.org/content/early/2011/06/03/jas] 2010:3361.
61. Hu XS, Wang Z: **Estimating the correlation of non-allele descents along chromosomes.** *Genet Res Camb* 2011, **93**:23-32.

Figures

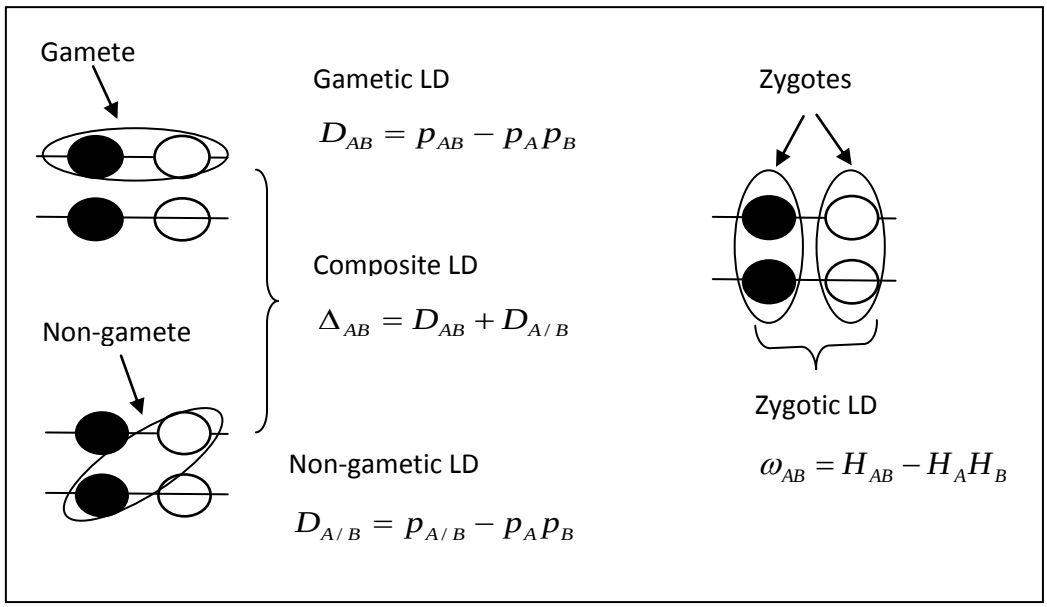

Figure 1.1 Pictorial presentation of different LD measures at two loci (*A* and *B*), each with two alleles, *A* and *a* at locus *A* and *B* and *b* at locus *B*: gametic LD ($D_{AB} = p_{AB} - p_A p_B$), non-gametic LD ($D_{A/B} = p_{A/B} - p_A p_B$), composite LD ($\Delta_{AB} = D_{AB} + D_{A/B}$) and zygotic LD ($\omega_{AB} = H_{AB} - H_A H_B$), where p_{AB} and $p_{A/B}$ are the gametic and non-gametic frequencies involving alleles *A* at locus *A* and allele *B* at locus *B*, p_A and p_B are the frequencies of allele *A* at locus *A* and allele *B* at locus *B*, and H_{AB}, H_A and H_B are the heterozygosities at both loci, locus *A* and locus *B*, respectively.

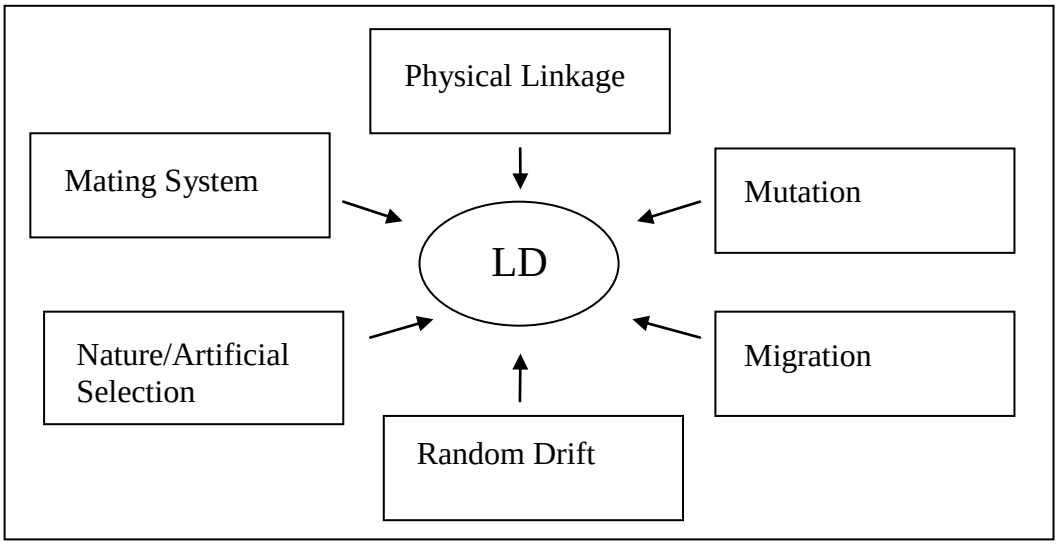

Figure 1.2 Genetic and demographic factors that can cause linkage disequilibrium (LD).

Chapter 2. Genome-wide Assessment of Gametic, Composite and Zygotic Linkage Disequilibria

2.1 Introduction

A cost-effective beef genetic improvement program for use under commercial conditions was carried out at the Kinsella Research Ranch of the University of Alberta during 1960 to 1989 [1]. Instead of the typical pure breeds or two- and three-breed crossbreeding systems that would be difficult to sustain particularly in small beef operations due to the cost of selecting superior pure breeds for breeding from a large genetic base, the Kinsella program had established synthetic (composite) breeds as an alternative system that was generally competitive with the usual crossbreeding systems but easier to manage regardless of herd size. The main breeding objective of the Kinsella program had been selection for growth performance and cow reproduction under commercial management conditions similar to typical beef operations in Alberta. Over the years, there was a clear trend of increased growth even after controlling birth weight in later years, but there was no clear trend of cow reproduction. After continued single-sire crossbreeding for 30 years (M.A. Price, private communication), animals from these synthetic lines were subsequently pooled to form the current composite beef population. This composite population has since been the subject of many breeding and genomic studies (e.g., [2-4]). However, little attention has been paid to the study on the genetic structure of this population.

Recently there are several studies on the use of single nucleotide polymorphism (SNP) markers for uncovering the population genetic structures of dairy and beef cattle [5-9]. In particular, there is a recent intensive focus on characterizing patterns and extent of gametic linkage disequilibrium (LD) in cattle populations due to the predominant role of gametic LD in livestock genomic selection [10]. Most assessments of gametic LD assume that the study populations undergo random mating and thus are in Hardy-Weinberg equilibrium (HWE). In

such a case, gametic LD alone is sufficient to describe nonrandom associations between alleles at different loci. However, the HWE assumption is obviously unwarranted in the Kinsella beef composite population as it arises from repeated mixing of multiple breeds and selection for growth and cow reproduction within and between breeds. Consequently, multilocus associations in this population need to be characterized at both gametic and zygotic levels [11, 12].

It is well known [13] that the extent of gametic LD is related to the genetic distance between loci on the same chromosome: the closer the locus pair, the stronger the gametic LD. This relationship is the basis of current intensive interest in genome-wide association studies (GWAS) in which most of the QTL effects would be picked up by the tightly linked adjacent markers. It is less clear, however, if and how this relationship would hold when zygotic LD is considered. It is long recognized [12, 14-17] that zygotic LD can be generated as a result of partial inbreeding, mixing of two or more distinct gene pools or heterotic selection even in the absence of gametic LD. In addition, chromosome-to-chromosome variation in different LD measures may provide some signals for selection or other locus-specific events in the recent history of the target cattle population.

In this study, we compared and contrasted three LD measures: gametic LD, composite LD, and zygotic LD, based on a 50K SNP data set that was obtained from genotyping the Kinsella beef composite population. Composite LD is similar to gametic LD but it allows for the presence of Hardy-Weinberg disequilibrium (HWD) [11]. In addition, we determined whether or not the patterns and relationships known for gametic or composite LD would also hold for zygotic LD.

2.2 Materials and methods

2.2.1 Description of animals and genotyping data

The blood samples of 1,023 beef steers were collected and genotyped using the Illimina Infinium genotyping system (the BovineSNP50 Beadchip). All steers

were progenies of the University of Alberta Hybrid dam lines that were derived from mixing of the three composite lines, namely Beef Synthetic 1 (SY1), Beef Synthetic 2 (SY2), and Dairy × Beef Synthetic (SD) [1]. SY1 was composed of approximately 33% each of Angus and Charolais, 20% Galloway, 5% Brown Swiss, and small amounts of other breeds. SY2 was composed of approximately 60% Hereford and 40% other beef breeds mainly including Augus, Charolais, and Galloway. SD was composed of approximately 60% dairy cattle (Holstein, Brown Swiss, or Simmental) and approximately 40% of other breeds, mainly including Angus and Charolais [18]. Selection for growth and cow reproduction were practiced during 1960 to 1989 and continued after mixing. There was a clear effect of selection on growth. Annual increase of adjusted 180-day weight prior to birth weight control (1960-1982) was 2.15, 2.05, and 0.93 kg for animals from SY1, SY2, and SD, respectively; such annual increase after birth weight control (1982-1989) was 3.62, 1.19, and 0.16 kg for SY1, SY2, and SD, respectively. Positive genetic trends for birth weight and weaning gain were also observed [2]. However, there was no obvious trend of change in cow reproduction as measured by the number of cows exposed, the percent of calf crop born and the percent of calf crop weaned.

A total of 51,828 SNP markers were originally obtained. These markers were distributed across 29 autosomes and one sex chromosome in the entire bovine genome. For our analyses, we only used 43,124 SNPs after removing those markers (i) with monomorphism, (ii) with unknown genomic position and (iii) on the sex chromosome, (iv) with minor allele frequency (MAF) of ≤ 2%, and (v) with a Chi-square (χ^2) value > 600 for the HWD test.

2.2.2 Two-locus gametic frequency, homozygosity, and heterozygosity

As described in Chapter 1, the frequency of gamete *AB* at loci *A* and *B* was denoted as p_{AB} and the basic measure of gametic LD was defined as, $D_{AB} = p_{AB} - p_A p_B$, with p_A and p_b being the frequencies of allele *A* at locus *A* and allele *B* at locus *B*. Similar expressions could be obtained for the other three

gametes, *Ab*, *aB* and *ab*. It should be noted that gametic LD would also measure the excess or deficiency of gametes in a 2×2 two-way contingency table for the two loci, $D_{AB} = p_{AB}p_{ab} - p_{Ab}p_{aB}$.

SNP genotypes at a given locus (say locus *A*) were classified either homozygous or heterozygous. The sample frequencies of the two classes estimated single-locus homozygosity (1 - H_A) and heterozygosity (H_A), respectively. Similarly, SNP genotypes at a pair of loci (say loci *A* and *B*) were grouped into four classes: (i) double homozygotes, (ii) homozygotes at locus *A* and heterozygotes at locus *B*, (iii) heterozygotes at locus *A* and homozygotes at locus *B*, and (iv) double heterozygotes. The frequencies of the four classes estimated two-locus homozygosity ($1 - H_A - H_B + H_{AB}$), homozygosity at locus *A* and heterozygosity at locus *B* ($H_B - H_{AB}$), heterozygosity at locus *A* and homozygosity at locus *B* ($H_A - H_{AB}$) and two-locus heterozygosity (H_{AB}), respectively. These frequencies were used for defining zygotic disequilibrium (see below).

2.2.3 Measures of linkage disequilibrium (LD)

- **Gametic LD**

The most commonly used statistical measures of gametic LD were reviewed in Chapter 1 (also see [19] and [11] for reviews). In this study, we used the squared correlation (r^2) as a normalized measure of gametic LD (GLD) to minimize its dependence on allele frequencies:

$$r^2_{GLD} = \frac{D^2_{AB}}{p_A p_a p_B p_b} \quad . \qquad (2.1)$$

The maximum likelihood (ML) method [20] was used to estimate the unknown gametic frequencies from unphased SNP data. The estimated gametic frequencies were in turn used for the estimation of gametic LD. The ML estimation was based on the HWE assumption which would not be warranted for

populations with complex genetic structure as in the Kinsella composite beef population.

- **Composite LD**

For the unphased SNP data, we also allowed for the presence of both gametic disequilibrium (D_{AB}) and nongametic disequilibrium ($D_{A/B}$). The composite LD (CLD) was calculated as the sum of the two components, i.e., $\Delta_{AB} = D_{AB} + D_{A/B}$ [21]. The squared correlation for the composite LD was calculated as follows:

$$r^2_{CLD} = \frac{\Delta^2_{AB}}{(p_A p_a + D_A)(p_B p_b + D_B)} \qquad (2.2)$$

where D_A and D_B measure the departures from HWE at loci *A* and *B*, respectively.

Composite LD was estimated directly from the unphased SNP data (indistinguishable coupling and repulsion double heterozygotes) without the need to estimate gametic frequencies. In addition, the estimate did not require the HWE assumption. r^2_{CLD} would be the same as r^2_{GLD} in a HWE population [21].

- **Zygotic LD**

Zygotic LD (ZLD) was calculated as the squared correlation between heterozygosities at loci *A* and *B* [22],

$$r^2_{ZLD} = \frac{\omega^2_{AB}}{H_A(1-H_A)H_B(1-H_B)} \quad . \qquad (2.3)$$

where H_A and H_B are the heterozygosities at loci *A* and *B*, respectively, and ω_{AB} is the zygotic association between loci *A* and *B*, $\omega_{AB} = H_{AB} - H_A H_B$, with H_{AB} being the two-locus heterozygosity as defined above. The zygotic association consists of all non-allelic genic disequilibria at the two loci [12]. This aspect will be further examined in Chapter 3.

Another standardized measure of zygotic association ω'_{AB} [22] was also calculated,

$$\omega'_{AB} = \begin{cases} \dfrac{\omega_{AB}}{\max^{-}}, & \text{if } \omega_{AB} < 0 \\ \dfrac{\omega_{AB}}{\max^{+}}, & \text{if } \omega_{AB} > 0 \end{cases}$$

where $\max^{-} = \max[-H_A H_B, -(1-H_A)(1-H_B)]$ and $\max^{+} = \min[(1-H_A)H_B, H_A(1-H_B)]$ are the limiting values of ω_{AB} when it is negative and positive, respectively.

2.2.4 Bonferroni correction for linkage disequilibrium test

Since the number of chi-square tests for gametic LD, composite LD and zygotic LD between pairs of loci would be very large (e.g., 2,841 × 2,840/2 = 4,034,220 for BTA 1), the Bonferroni correction [11] was used to avoid spurious rejection of the null hypothesis of no LD. Instead of the usual critical value of 3.84 for the chi-square test with one degree of freedom and the significance level of $\alpha = 0.05$, a more stringent chi-square test was obtained using the new significance level of $\beta = \alpha / m$, where m is the number of all possible syntenic marker pairs (e.g., m = 4,034,220 for BTA 1). Thus the critical chi-square values adjusted under the Bonferroni correction range from 27.4 for BTA 28 to 31.6 for BTA 1. The chi-square value for a marker pair was computed as $X^2 = nr_x^2$, where r_x^2 is the LD measure defined above with x indexing *GLD*, *CLD* or *ZLD* and n is the number of animals.

2.2.5 Data analysis

The data analysis was done using SAS 9.2 [23]. The calculation of gametic and composite LD was carried out using PROC ALLELE of SAS/Genetics 9.2. The SNP marker data was read in as columns of genotypes using the GENOCOL and DELIMITER= options in the PROC ALLELE statement. Gametic LD was

calculated if the HAPLO= EST option in the PROC ALLELE statement was invoked, whereas composite LD was calculated if the HAPLO= NONEHWD option was specified. Zygotic LD was calculated through SAS programming.

2. 3 Results

2.3.1 Single locus statistics

2.3.1.1 Marker density

Different data quality control measures filtered out some 17% of SNP markers and a total of 43,124 markers were retained for the single and multi-locus analyses. These SNP markers were distributed across 29 autosomes (BTA 1 to BTA 29) of the bovine genome (Table 2.1). The total physical distance of the genome was estimated to be 2,544 Mb, which would correspond to the total genetic map distance of 2,764 cM [24]. On average, there were approximately 17 markers per 1 Mb, with little variation among individual chromosomes.

An adjacent marker distance was calculated as the absolute physical distance between two immediate neighbour markers. The genome-wide averaged distance of adjacent markers was 58,933 bp, with the standard deviation of 54,650 bp. Although mean adjacent marker distances were similar for all chromosomes, some extremely small and large gaps were observed. The smallest gap was only 3 bp apart and this tiny gap is located on BTA 7 whereas the largest gap was more than 2 million bp apart on BTA 10 (Table 2.1). However, the majorities (87%) of the adjacent markers were spaced within 100 kb with about half of these being located within 25-50kb and only 2.5% being spaced over 200kb apart (Figure 2.1).

2.3.1.2 Minor allele frequency (MAF) & Hardy-Weinberg disequilibrium (HWD)

Genome-wide average of MAF was 0.273 with the standard deviation being 0.136. This estimate of MAF might have been biased upward slightly as a MAF of < 2% was removed prior to the analysis. There was little variation in mean MAF variation across chromosomes, with a narrow range of 0.258 for BTA 26 to 0.281 for BTA 18. The MAF was almost uniformly distributed at 10 equally divided frequency intervals of the range from zero to 0.5, with approximately 400 markers falling into each category, except for category of MAF less than 0.05 where only 93 markers were observed (Figure 2.2).

A total of 4,024 (9.3%) markers genome-wide showed significant HWD as the chi-square values for these markers exceeded 3.84, the critical value of chi-square test with 5% significance level and one degree of freedom. Small chromosome-to-chromosome variation was observed with the proportions of HWD markers ranging from 6.3% for BTA 19 to 9.9% for BTA 6 (data not shown).

2.3.1.3 Single-locus heterozygosity

The mean heterozygosity over all SNP markers was 0.357 with the standard deviation 0.136. The heterozygosities varied considerably across individual SNP markers, ranging from 0.015 for a locus on BTA 15 to 0.599 for a locus on BTA 11. The heterozygosities were very similar among chromosomes (Figure 2.3) with the mean heterozygosities ranging from 0.346 for BTA 1 to 0.379 for BTA 25.

The mean heterozygosity for the markers with a significant HWD was 0.381 with the standard deviation 0.131 (Table 2.2). This heterozygosity was slightly larger than 0.357, the mean heterozygosity over all SNP markers. The mean heterozygosities at the HWD loci on the individual chromosomes have a slightly wider range (0.352-0.414) than the mean heterozygosities at all SNP loci.

The genome-wide average of fixation indexes was -0.008, with the range of -0.202 to 0.715. The mean fixation indices were similar among individual

chromosomes with a range of -0.003 for BTA 25 to -0.014 for BTA 3. Figure 2.4 showed the distribution of the fixation indices at all loci on 29 autosomes. There were a total of 25 markers with a fixation index of > 0.25; further analysis found significant HWD at all these loci (data not shown).

2.3.2 Multilocus statistics

There would be a total 36,131,636 syntenic pairs of SNP markers across all 29 autosomes. We identified those marker pairs with a genetic distance of ≤ 50 cM, beyond which free recombination would occur and a syntenic marker pair with a distance of > 50 cM would behave just like a non-syntenic pair.

2.3.2.1 Two-locus homozygosity

The mean chromosome-wide two-locus homozygosities ranged from 0.393 for BTA 23 to 0.429 for BTA 1 (Table 2.3). BTA 28 had the narrowest range of the two-locus homozygosities (0.2-0.931) whereas BTA 11 had the widest range (0.168-0.945).

When the two-locus homozygosities were averaged over only those marker pairs with significant zygotic LD, they ranged from 0.526 for BTA 25 to 0.605 for BTA 9, which were significantly (P<0.0001) higher than the mean two-locus homozygosities for all syntenic pairs.

The mean two-locus homozygosity for a given chromosome could be very accurately predicted from the squared mean single-locus homozygosity of the same chromosome with a prediction error of ≤1.6%.

2.3.2.2 Two-locus heterozygosity

The mean chromosome-wide two-locus heterozygosities ranged from 0.120 for BTA 1 to 0.145 for BTA 25 (Table 2.4). BTA 20 had the narrowest range of

the two-locus heterozygosities (0-0.499) whereas BTA 13 had the widest range (0-0.551).

When the two-locus heterozygosities were averaged over only those marker pairs with significant zygotic LD, they ranged from 0.118 for BTA 9 to 0.161 for BTA 14, which were significantly (P<0.0001) higher than the mean two-locus heterozygosities for all syntenic pairs.

The mean two-locus heterozygosity for a given chromosome could be very accurately predicted from the squared mean single-locus heterozygosity of the same chromosome with a prediction error of ≤0.1%.

2.3.2.3 Gametic LD

The 95% empirical intervals constructed from gametic LD values between all syntenic pairs on individual chromosomes ranged from (0, 0.047) for BTA 1 to (0, 0.073) for BTA25 (Table 2.5). These intervals had a significant negative correlation (r = -0.832, P < 0.0001) with the lengths of chromosomes.

Genome-wide, 7.14 % of all gametic LDs between syntenic marker pairs were significant (Table 2.5). The percentages of significant gametic LD varied among individual chromosomes, ranging 4.82% for BTA1to 11.7% for BTA 25. Strong negative correlation was found between the proportion of significant gametic LD and chromosome length (r = 0.94; P < 0.0001). When the 95% empirical intervals were constructed using only those marker pairs with significant gametic LD, these intervals range from (0.023, 0.180) for BTA 28 to (0.026, 0.23) found on BTA 7. The genome-wide interval was (0.026, 0.207).

Of 36,131,636 syntenic marker pairs genome-wide, 47,659 marker pairs showed strong gametic LD (r^2_{GLD}≥0.25) (Table 2.5). These marker pairs were identified over individual chromosomes, ranging from 433 marker pairs on BTA 28 to 3,467 marker pairs on BTA 1.

2.3.2.4 Composite LD

The 95% empirical intervals constructed from composite LD values between all syntenic pairs on each chromosome ranged from (0, 0.048) for BTA 1 to (0, 0.074) for BTA25 (Table 2.6). These intervals had a significant negative correlation (r = -0.856, P < 0.0001) with the chromosome length.

Genome-wide, 7.09 % of all composite LDs between syntenic marker pairs were significant (Table 2.6). The percentages of significant composite LD varied among individual chromosomes, ranging 4.95% for BTA1to 11.94% for BTA 25. Strong negative correlation was found between the proportion of significant composite LD and chromosome length (r = -0.94; P < 0.0001). When the 95% empirical intervals were constructed using only those marker pairs with significant gametic LD, these intervals ranged from (0.023, 0.180) for BTA 28 to (0.031, 0.24) for BTA 7 with the genome-wide interval being (0.030, 0.214).

Of 36,131,636 syntenic marker pairs genome-wide, 46,497 marker pairs showed strong composite LD (r^2_{CLD}≥0.25) (Table 2.6). These marker pairs were distributed over individual chromosomes, ranging from 410 marker pairs on BTA 28 to 3,531 marker pairs on BTA 1.

2.3.2.5 Zygotic LD

The 95% empirical intervals constructed from zygotic LD values between all syntenic pairs on each chromosome ranged from (0, 0.012) for BTA 1 to (0, 0.018) for BTA26 (Table 2.7). These intervals had a significant negative correlation (r = -0.731, P < 0.0001) with the chromosome length.

Genome-wide, 0.85 % of all gametic LDs between syntenic marker pairs were significant (Table 2.7). The percentages of significant zygotic LD varied among individual chromosomes, ranging 0.59% for BTA1to 1.47% for BTA 26. Strong negative correlation was found between the proportion of significant zygotic LD and chromosome length (r = -0.834; P < 0.0001). When the 95% empirical intervals were constructed using only those marker pairs with

significant zygotic LD, these intervals ranged from (0.027, 0.363) for BTA 28 to (0.03, 0.533) found on BTA 16 with the genome-wide interval being (0.03, 0.478).

Of 36,131,636 syntenic marker pairs genome-wide, 18,039 marker pairs show strong zygotic LD (r^2_{ZLD}≥0.25) (Table 2.7). These marker pairs were identified on individual chromosomes, ranging from 146 marker pairs on BTA 28 to 1,448 marker pairs on BTA 1.

2.3.3 Comparisons between gametic, composite and zygotic LD

2.3.3.1 Correlations between different LD measures

Significant gametic LD (r^2_{GLD}) was observed in 2,578,526 syntenic marker pairs across all 29 autosomes. There was nearly the same number of marker pairs with significant composite LD (r^2_{CLD}). Of the total marker pairs, 90% syntenic marker pairs (2,327,726) showed significant gametic and composite LD (Table 2.8).

A total of 306,391 syntenic marker pairs over all autosomes showed significant zygotic LD (r^2_{ZLD}). Of these, 303,860 pairs showed significant LD at both gametic and zygotic levels and they were distributed genome-wide, ranging from 3,287 marker pairs on BTA 28 to 23,652 marker pairs on BTA1. Similarly, 299,979 marker pairs showed significant LD at both composite and zygotic level, ranging from 3,254 marker pairs on BTA 28 to 23,192 marker pairs on BTA1.

The correlations between gametic LD, composite LD and zygotic LD were strong and significant ($r_{GLD,CLD}$ = 0.991, P < 0.001; $r_{GLD,ZLD}$ = 0.782, P < 0.001and $r_{CLD,ZLD}$ = 0.771, P < 0.001). When the marker pairs with non-significant LD were excluded these correlations were slightly higher ($r_{GLD,CLD}$ = 0.987, $r_{GLD,ZLD}$ = 0.863, and $r_{CLD,ZLD}$ = 0.847). These correlations were depicted in Figure 2.5 for all 29 autosomes.

2.3.3.2 Comparative analysis of gametic and zygotic LD

In the absence of gametic LD (r^2_{GLD}≤0.001), standardized zygotic LDs decayed exponentially with two-locus heterozygosities on each of the 29 autosomes (Figure 2.6). Such decay pattern would have been masked if all gametic LDs were included regardless of their magnitudes and significance levels. Thus, the 95% empirical intervals of zygotic LD were much wider at low two-locus heterozygosities (<0.05) than those at the intermediate heterozygosities (~0.3).

2.4 Discussion

This study is the first systematic comparison between gametic and zygotic disequilibrium measures for the Kinsella composite beef population that was established from recent mixing of multiple breeds and was under selection for growth and cow reproduction. Single-locus heterozygosities fell approximately in the range of 0.3 to 0.4 over different chromosomes (Table 2.2). There was little change when only those markers with significant HWD were kept. Two-locus heterozygosities had lower values but displayed similar patterns (Table 2.4) and they were almost perfectly predicted by the products of single-locus heterozygosities. In contrast, two-locus homozygosities were significantly higher when only those marker pairs with significant zygotic LD were kept than when all pairs were included. The two-locus homozygosities (all pairs) were almost predicted by the products of single-locus homozygosities (Table 2.3). The levels of significant gametic and composite LD (Tables 2.5 and 2.6) were much higher than those of significant zygotic LD (Table 2.7). Similar patterns between gametic and zygotic LD were observed when only those marker pairs with extremely high gametic LD values (r^2_{GLD}> 0.25) were retained. Such tightly linked marker pairs were chosen because a similar stringent criterion was used for detecting marker effects in many livestock genomic selection programs [10].

Our estimates of gametic LD are similar to those reported for Holstein and other cattle in the recent literature [5-9] though our estimates are generally slightly lower based on different comparisons. Khatkar et al. [5] observed the genome-wide r^2_{GLD} average and median of 0.016 and 0.003, respectively, in comparison to our estimates of 0.0105 and 0.0036. Khatkar et al. [5] used a smaller set of markers (15,036 SNPs) covering all 29 autosomes but a larger number of animals (1,546). Sargolzaei et al. [6] presented r^2_{GLD} values between adjacent markers with a genome-wide mean of 0.31, comparing to our mean of 0.195 (detailed data not shown). However, an updated study by the same group [8] with more markers (38,590 SNPs) had a genome-wide r^2_{GLD} average of 0.20 which is very close to our value. Villa-Angulo et al. [7] calculated r^2_{GLD} values using 101 targeted high-density regions (non-overlapping genomic windows of 100kb containing 10 or more markers and a maximum gap between markers was 20 kb) on QTL-rich chromosomes 6, 14 and 25 to calculate values for 19 beef and dairy breeds; the mean values of r^2_{GLD} ranged from 0.204 for Nelore to 0.397 for Hereford with an overall average of 0.294. Qanbari et al. [9] obtained a genome-wide r^2_{GLD} average of 0.30 for SNP pairs with a distance of < 25 kb for German Holstein, which is very comparable to our estimate of 0.285 for the same distance range (<25 kb) (Table 2.9). Our r^2_{CLD} and r^2_{GLD} estimates are very similar and thus the above comparisons of gametic LD estimates with other studies would be applicable to composite LD estimates as well. However, our zygotic LD (r^2_{ZLD}) estimates are much lower than gametic LD estimates in our and other studies for the reason that will be discussed below. Thus, useful LD most likely occurs between those markers that are adjacent or tightly linked.

In our composite beef population that has been maintained through continued mixing and selection for growth with inflow of new Augus and Charolais bulls every generation, the observed high single-locus heterozygosities (~0.3-0.4) are probably expected. A somewhat surprising result was that there were only 8.2% of total SNP loci with significant HWD despite mixing of animals with diverse breed (genetic) backgrounds every generation. One possible reason could be the removal of ~17% of SNP loci especially with MAF <2.0% and HWD

chi-square values of >600. It could also be that after many generations of breed mixing and crossbreeding, certain level of genetic homogeneity might have been achieved (i.e., genetic integrity of distinct breeds is no longer clear) and thus unplanned mixing of different animals in recent generations resembles largely to the situation of random mating. It is reasonable to speculate that selection may have occurred at or around those SNP loci with significant HWD, but it would be desirable to examine whether or not QTLs for growth and cow reproduction are present in those genomic regions when such data become available.

The observation that gametic or composite LD was higher than zygotic LD is expected because the latter may be viewed as a weighted average of gametic LD and other higher-order genic disequilibria. Simulation results [25] showed that gametic or composite LD was predominant and the high-order disequilibria involving three or more genes are generally smaller. In Chapter 3, we will provide a detailed examination of high-order genic disequilibria relative to gametic LD and their importance to zygotic LD in the same composite population.

Faster decay of zygotic LD than gametic or composite LD with physical distances on individual chromosomes observed in this study is consistent with the predictions from population genetic theory. It is well known [13] that the gametic LD decays every generation of random mating by a factor of $(1-c)$, where c is the recombination frequency between a pair of loci. In other words, the amount of gametic LD is halved by each generation when the loci are unlinked (free recombination; $c = ½$) but the decay is slower when the loci are linked ($c < ½$). While the exact relationship of zygotic LD with physical distance is quite complicated, Sabatti and Risch (equation 6) [26] showed that the decay of zygotic LD in a random mating population is proportional to the factor of $(1-c)^2$, thereby explaining why zygotic LD decays faster than gametic LD. In non-random mating populations such as our composite population, the rate of decay of zygotic LD depends on if inbreeding is preferred or avoided. The amount of zygotic LD should be reduced every generation by a factor lying between $(1-c)^2$ and $(1-c)$ if there is inbreeding or by a factor of $<(1-c)^2$ if there is avoidance of inbreeding.

There were many marker pairs where zygotic LD were strong but with little or no gametic LD in our composite population. Such genomic structure could arise from selection against double homozygotes in a large population where gametic LD may be negligible [27]. Based on a limited amount of isozyme data from a few animal and plant species, Mitton [27] did not find the desired genetic structure. There are two potential issues with Mitton's analysis. First, his data sets were generally very small with a limited number of loci and selection signals might have been located at genomic regions far from the tested loci. Second, most populations he examined were natural populations and natural selection would generally be weaker than truncated (directional) selection that would generally be practiced in breeding or production populations. These two issues should not be a major concern in our study. We used a set of 50K SNP markers that are densely distributed over 29 autosomes for the current analysis. Additionally, our composite population would have been under a "stronger" directional selection for growth and cow reproduction. In particular, to improve cow reproduction, heterozygous cows would be preferred because they usually have a higher level of fertility than homozygous cows [13]. Coupled with selection for growth, the desired cows would be those that are capable of producing large numbers of rapidly growing calves.

Our study has important implications for genetic improvement in beef composites. First of all, the genomic regions marked by those SNP loci with significant gametic and zygotic LD should be targets for QTL identification or candidate gene search because strong LD may arise from selection for growth or cow reproduction in these genomic regions. This may form a basis for a new gene discovery strategy which would otherwise be difficult to be implemented. For example, in the future study searching for candidate genes corresponding for growth or reproduction traits, the researchers may want to increase maker density in those regions. Second, for genomic selection to be successful (achieving a prediction accuracy of 0.85), Meuwissen et al. [28] showed from their simulation study that the required level of gametic LD (r^2_{GLD}) should be >0.2. On the other hand, Ardlie et al. [29] suggested the use of r^2_{GLD}> 1/3 for genome-wide

association studies (GWAS) in human. When the threshold of useful gametic LD is set to be 0.25 to ensure the SNP spacing ~ 35 kb as suggested by Qanbari et al. [9], the GWAS approach would require the use of more than 75,000 SNPs per individual, assuming that all SNPs are informative (with a MAF $\geq$ 0.05). If all of the current 50K data set were usable, we would have to use less extreme frequencies (MAF $\geq$ 0.15) to achieve the improved accuracy and magnitude of estimated LD between pairs of SNP markers. However, the removal of many rare alleles may lose the opportunity to capture potentially novel casual mutations in the population. In addition, our zygotic LD estimates serve as a reminder that non-random mating may be an important cause of LD even when the loci are very tightly linked.

In this chapter, we quantified the zygotic disequilibria for all syntenic marker pairs. The zygotic LD is a measure of the overall zygotic LD to which the individual genic LD contribute. Thus a logical next step is to decompose the overall zygotic disequilibrium into individual genic LDs to investigate the patterns and extent of these genic LDs. Such decomposition may provide more complete information on the possible selection pattern and/or effect of mating system. Thus, the next chapter is intended to investigate the individual components of the zygotic disequilibrium. The genetic forces effecting the LD pattern and distribution are further investigated at individual and combined genic levels.

2.5 Conclusion

Our study suggests that: a) traditional gametic LD is incapable of capturing nongametic LD and other non-allelic genic LDs; b) moderate zygotic associations likely exist in non-equilibrium populations such as our beef composite; c) the use of zygotic LD allows us to assess the relative importance of gametic LD vs. all other non-allelic genic LDs even when HWE is not seriously violated. No previous LD studies have evaluated linkage disequilibrium at both gametic and zygotic levels in cattle and other livestock populations. Our study demonstrates

the need to investigate both gametic and zygotic LD for a complete evaluation of gene associations at multiple loci.

2.6 References

1. Berg RT, Makarechian M, Arthur PF: **The University of Alberta beef breeding project after 30 years—A review.** *Univ Alberta Annu Feeders' Day Rep* 1990, **69**:65-69.
2. Wang Z, Goonewardene LA, Yang RC, Price MA, Makarechian M, Knapp J, Okine EK, Berg TR: **Estimation of Genetic Parameters and Trends in Pre-Weaning Traits of Beef Lines Subject to Phenotypic Selection.** *Journal of Animal and Veterinary Advances* 2005, **4**:202-209.
3. Durunna ON, Mujibi FDN, Goonewardene L, Okine EK, Basarab JA, Wang Z, Moore SS: **Feed efficiency differences and reranking in beef steers fed grower and finisher diets.** *J Anim Sci* 2011, **89**:158-167.
4. Mujibi FDN, Nkrumah JD, Durunna ON, Stothard P, Mah J, Wang Z, Basarab J, Plastow G, Crews Jr DH, Moore SS: **Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle.** [http://jas.fass.org/content/early/2011/06/03/jas] 2010:3361.
5. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, Schnabel RD, Taylor JF, Raadsma HW: **Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel.** *BMC Genomics* 2008, **9**:187.
6. Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR: **Extent of linkage disequilibrium in Holstein cattle in North America.** *J Dairy Sci* 2008, **91**:2106-2117.
7. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ: **High-resolution haplotype block structure in the cattle genome.** *BMC Genetics* 2009, **10**:19.
8. Bohmanova J, Sargolzaei M, Schenkel FS: **Characteristics of linkage disequilibrium in North American Holsteins.** *BMC Genomics* 2010, **11**:421.
9. Qanbari S, Pimentel ECG, Teten J, Thaller G, Lichtner P, Sharifi AR, Simianer H: **The pattern of linkage disequilibrium in German Holstein**

cattle. *Anim Genet* 2010, **41**:346-356.

10. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programs.** *Nat Rev Genet* 2009, **10**:381-391.
11. Weir BS: *Genetic Data Analysis II.* MA Sunderland: Sinauer Associates; 1996.
12. Yang RC: **Zygotic associations and multilocus statistics in a nonequilibrium diploid population.** *Genetics* 2000, **155**:1449-1458.
13. Falconer DS, MacKay TFC: *Introduction to Quantitative Genetics.* Harlow: Prentice Hall; 1996.
14. Haldane JBS: **The association of characters as a result of inbreeding and linkage.** *Ann Eugen* 1949, **15**:15-23.
15. Bennett JH, Binet FE: **Association between Mendelian factors with mixed selfing and random mating.** *Heredity* 1956, **10**:51-55.
16. Allard RW, Jain SK, Workman PL: **The genetics of inbreeding populations.** *Adv Genet* 1968, **14**:55-131.
17. Barton NH, Gale KS: **Genetic analysis of hybrid zones.** In *Hybrid Zones and the Evolutionary Process*. Edited by Harrison RG. New York: Oxford University Press; 1993:13-45.
18. Goonewardene LA, Wang Z, Price MA, Yang RC, Berg RT, Makarechian M: **Effect of udder type and calving assistance on weaning traits of beef and dairy x beef calves.** *Livestock Production Science* 2003, **81**:47-56.
19. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117**:331-341.
20. Hill WG: **Estimation of linkage disequilibrium in randomly mating populations.** *Heredity* 1974, **33**:229-239.
21. Weir BS: **Inferences about linkage disequilibrium.** *Biometrics* 1979, **35**:235-254.
22. Yang RC: **Analysis of multilocus zygotic associations.** *Genetics* 2002, **161**:435-445.

23. SAS Institute Inc.: *SAS/Genetics User's Guide Version 9.2.* Cary, NC, USA: SAS Institute Inc.; 2008.
24. Li W, Freudenberg J: **Two-parameter characterization of chromosome-scale recombination rate.** *Genome Res* 2009, **19**:2300-2307.
25. Weir BS, Cockerham CC: **Complete characterization of disequilibrium at two loci.** In *Mathematical Evolutionary Theory*. Edited by Feldman MW. Princeton New Jersey: Princeton University Press; 1989:86-110.
26. Sabatti C, Risch N: **Homozygosity and linkage disequilibrium.** *Genetics* 2002, **160:** 1707–1719.
27. Mitton JB: *Selection in Natural Populations.* Oxford: Oxford University Press; 1997.
28. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:** 1819-1829.
29. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3**:299-309.

Tables

Table 2.1 Number of Single Nucleotide Polymorphism (SNP) markers (m) and chromosome length (mega base pairs, Mb) for 29 bovine autosomes (BTA 1 to BTA 29) in the Kinsella composite beef population. Mean, standard deviation (SD), minimum and maximum distances (in base pairs) between all pairs of adjacent markers are also presented.

BTA	SNPs (m)	Length (Mb)	Adjacent Marker Distance (bp)			
			Mean	SD	Minimum	Maximum
1	2841	161	56658	46568	131	470043
2	2320	141	60578	59780	75	661839
3	2180	128	58700	57515	108	807545
4	2126	124	58388	47620	3406	430390
5	1815	126	69347	73232	2444	1115633
6	2183	123	56112	51997	2660	826193
7	1895	112	58974	55986	3	950817
8	2026	117	57742	47799	3416	489257
9	1747	108	61883	57633	449	729114
10	1844	106	57363	68139	284	2081464
11	1892	110	58194	52300	886	890683
12	1393	85	61205	60291	237	760907
13	1491	84	56427	47106	382	592177
14	1440	81	56479	45280	108	575964
15	1400	85	60446	54116	3701	683257
16	1326	78	58774	59748	178	1051359
17	1331	77	57408	49438	333	725528
18	1109	66	59597	57330	5603	867228
19	1130	65	57604	45262	3937	553067
20	1321	76	57195	50727	9201	837059
21	1129	69	61297	56769	903	742465
22	1061	62	58135	42967	2432	360641
23	900	52	59256	50555	1568	476317
24	1073	65	60570	51851	95	531092
25	813	44	53478	42015	1342	350281
26	884	52	57759	41705	281	373781
27	811	49	60128	78931	151	1889396
28	785	46	58642	47409	675	363454
29	858	52	60327	58828	1860	806694
Overall	43124	2544	58933	54650	3	2081464

Table 2.2 Distribution of single marker heterozygosity for 29 bovine autosomes (BTA 1-BTA 29) in the Kinsella composite beef population. Mean, standard deviation (SD), minimum and maximum of heterozygosity over all markers and over only those markers with significant Hardy-Weinberg disequilibrium (HWD) are presented.

BTA	Single locus heterozygosity at all loci				Single locus heterozygosity at HWD loci			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
1	0.346	0.141	0.038	0.562	0.353	0.139	0.040	0.562
2	0.359	0.136	0.042	0.563	0.396	0.124	0.058	0.563
3	0.356	0.136	0.027	0.559	0.383	0.128	0.027	0.559
4	0.355	0.136	0.038	0.549	0.360	0.134	0.056	0.549
5	0.356	0.139	0.038	0.554	0.396	0.130	0.038	0.554
6	0.359	0.138	0.040	0.555	0.389	0.133	0.066	0.555
7	0.355	0.139	0.039	0.549	0.382	0.134	0.039	0.549
8	0.354	0.135	0.037	0.544	0.377	0.137	0.037	0.544
9	0.349	0.14	0.037	0.551	0.360	0.134	0.037	0.551
10	0.355	0.135	0.039	0.561	0.376	0.129	0.053	0.561
11	0.354	0.138	0.041	0.599	0.375	0.130	0.049	0.599
12	0.353	0.138	0.038	0.554	0.394	0.123	0.052	0.554
13	0.363	0.136	0.040	0.553	0.381	0.130	0.049	0.553
14	0.366	0.132	0.039	0.546	0.386	0.129	0.053	0.546
15	0.357	0.134	0.015	0.560	0.391	0.124	0.015	0.560
16	0.352	0.138	0.038	0.550	0.352	0.143	0.051	0.550
17	0.351	0.134	0.026	0.544	0.373	0.127	0.026	0.544
18	0.370	0.131	0.040	0.559	0.360	0.144	0.069	0.559
19	0.368	0.133	0.041	0.569	0.408	0.117	0.054	0.569
20	0.355	0.141	0.040	0.561	0.388	0.136	0.045	0.561
21	0.358	0.136	0.041	0.558	0.372	0.130	0.055	0.558
22	0.353	0.134	0.040	0.542	0.357	0.130	0.040	0.542
23	0.374	0.129	0.041	0.553	0.395	0.128	0.061	0.553
24	0.367	0.134	0.038	0.552	0.379	0.130	0.057	0.552
25	0.379	0.127	0.041	0.551	0.394	0.121	0.141	0.551
26	0.352	0.135	0.040	0.543	0.400	0.112	0.067	0.543
27	0.358	0.134	0.040	0.557	0.383	0.133	0.040	0.557
28	0.366	0.131	0.040	0.543	0.406	0.12	0.075	0.543
29	0.363	0.133	0.040	0.569	0.414	0.118	0.074	0.569
Overall	0.357	0.136	0.015	0.599	0.381	0.131	0.015	0.599

Table 2.3 Distribution of two-locus homozygosity across 29 bovine autosomes (BTA 1-BTA 29) in the Kinsella composite beef population. Mean, standard deviation (SD), minimum and maximum values over all syntenic pairs and over those pairs with significant zygotic linkage disequilibrium (ZLD) are presented. The predicted two-locus homozygosities based on single-locus homozygosity are given as well.

BTA	Two-locus homozygosity (all pairs)				Two-locus homozygosity (pairs with ZLD)				Predicted two-locus homozygosity
	Mean	SD	Min	Max	Mean	SD	Min	Max	
1	0.429	0.132	0.186	0.959	0.602	0.156	0.263	0.959	0.428
2	0.411	0.125	0.185	0.957	0.590	0.143	0.261	0.957	0.410
3	0.416	0.125	0.189	0.948	0.576	0.147	0.269	0.948	0.415
4	0.417	0.125	0.196	0.959	0.566	0.151	0.245	0.959	0.416
5	0.416	0.128	0.193	0.960	0.571	0.155	0.255	0.960	0.415
6	0.413	0.127	0.182	0.947	0.577	0.151	0.261	0.947	0.411
7	0.417	0.128	0.193	0.960	0.573	0.158	0.213	0.960	0.415
8	0.418	0.124	0.196	0.952	0.571	0.141	0.263	0.952	0.417
9	0.424	0.130	0.193	0.959	0.605	0.144	0.277	0.959	0.423
10	0.418	0.124	0.184	0.943	0.581	0.142	0.267	0.943	0.416
11	0.419	0.128	0.168	0.945	0.577	0.150	0.259	0.945	0.417
12	0.419	0.128	0.195	0.945	0.580	0.147	0.274	0.945	0.418
13	0.408	0.124	0.186	0.960	0.540	0.147	0.267	0.960	0.406
14	0.403	0.119	0.195	0.950	0.529	0.143	0.254	0.950	0.402
15	0.415	0.123	0.192	0.953	0.578	0.142	0.252	0.953	0.414
16	0.421	0.128	0.194	0.955	0.594	0.154	0.253	0.955	0.420
17	0.423	0.124	0.197	0.938	0.578	0.144	0.257	0.938	0.421
18	0.398	0.118	0.186	0.951	0.551	0.138	0.261	0.951	0.397
19	0.401	0.120	0.188	0.955	0.548	0.145	0.272	0.955	0.400
20	0.417	0.130	0.191	0.946	0.585	0.153	0.260	0.946	0.416
21	0.413	0.125	0.189	0.946	0.568	0.150	0.263	0.946	0.412
22	0.421	0.123	0.206	0.944	0.571	0.139	0.272	0.944	0.419
23	0.393	0.115	0.183	0.959	0.542	0.139	0.268	0.959	0.391
24	0.402	0.122	0.185	0.959	0.552	0.143	0.238	0.959	0.401
25	0.402	0.112	0.195	0.936	0.526	0.143	0.276	0.936	0.386
26	0.422	0.125	0.201	0.945	0.568	0.139	0.258	0.945	0.420
27	0.413	0.123	0.193	0.946	0.564	0.139	0.266	0.946	0.412
28	0.403	0.119	0.200	0.931	0.558	0.137	0.276	0.931	0.402
29	0.407	0.121	0.184	0.958	0.577	0.142	0.258	0.958	0.406

Table 2.4 Distribution of two-locus heterozygosity across 29 autosomes (BTA 1-BTA 29) in the Kinsella composite beef population. Mean, standard deviation (SD), minimum and maximum values over all syntenic pairs and over pairs with significant zygotic linkage disequilibrium (ZLD) are presented. The predicted two-locus heterozygosities based on single locus heterozygosity are given as well.

BTA	Two-locus heterozygosity (all pairs)				Two-locus heterozygosity (pairs with ZLD)				Predicted two-locus heterozygosity
	Mean	SD	Min	Max	Mean	SD	Min	Max	
1	0.120	0.072	0	0.521	0.123	0.090	0	0.521	0.120
2	0.129	0.072	0	0.521	0.126	0.085	0	0.521	0.129
3	0.127	0.072	0	0.546	0.135	0.088	0	0.546	0.127
4	0.127	0.072	0	0.530	0.143	0.093	0	0.530	0.126
5	0.127	0.073	0	0.541	0.138	0.094	0	0.541	0.127
6	0.129	0.073	0	0.526	0.134	0.089	0	0.526	0.129
7	0.127	0.073	0	0.545	0.140	0.096	0.006	0.545	0.126
8	0.126	0.071	0	0.507	0.137	0.086	0	0.507	0.126
9	0.122	0.072	0	0.510	0.118	0.083	0.006	0.510	0.122
10	0.126	0.071	0	0.514	0.129	0.083	0	0.514	0.126
11	0.126	0.073	0	0.517	0.134	0.091	0	0.517	0.125
12	0.126	0.073	0	0.522	0.129	0.086	0	0.522	0.125
13	0.132	0.074	0	0.551	0.155	0.093	0.004	0.551	0.132
14	0.135	0.072	0	0.533	0.161	0.094	0	0.533	0.134
15	0.128	0.071	0	0.531	0.127	0.083	0.004	0.531	0.127
16	0.124	0.072	0	0.522	0.127	0.094	0	0.522	0.124
17	0.124	0.070	0	0.513	0.129	0.084	0.003	0.513	0.123
18	0.137	0.071	0	0.508	0.145	0.087	0.002	0.508	0.137
19	0.136	0.072	0	0.521	0.145	0.091	0.005	0.521	0.135
20	0.126	0.074	0	0.499	0.126	0.088	0.001	0.499	0.126
21	0.129	0.072	0	0.510	0.135	0.094	0	0.510	0.128
22	0.125	0.070	0	0.503	0.133	0.083	0.004	0.503	0.124
23	0.141	0.071	0	0.518	0.147	0.092	0.001	0.518	0.140
24	0.135	0.073	0	0.518	0.144	0.089	0	0.518	0.135
25	0.145	0.071	0	0.550	0.158	0.095	0.003	0.550	0.144
26	0.124	0.071	0	0.526	0.130	0.083	0.006	0.526	0.124
27	0.129	0.072	0	0.517	0.134	0.084	0.009	0.517	0.128
28	0.134	0.071	0	0.526	0.133	0.085	0.007	0.526	0.134
29	0.132	0.072	0	0.518	0.127	0.086	0	0.518	0.132

Table 2.5 Distribution of two-locus gametic linkage disequilibrium (r^2_{GLD}) in the Kinsella composite beef population. Total number of syntenic pairs, % of syntenic pairs with distance ≤ 50 cM*, % of pairs with significant r^2_{GLD} and the proportion of marker pairs with $r^2_{GLD} \geq 0.25$ are presented. The 95% empirical intervals are given for all syntenic pairs, pairs with significant r^2_{GLD} and pairs with $r^2_{GLD} \geq 0.25$ on 29 autosomes.

BTA	Total pairs	% of pairs with distance ≤50 cM	% of pairs with significant GLD	# of pairs with $r^2_{GLD}>0.25$	All syntenic pairs		Pairs with significant LD		Pairs with $r^2_{GLD} \geq 0.25$	
					2.50%	97.50%	2.50%	97.50%	2.50%	97.50%
1	4034220	51.93	4.82	3467	0	0.047	0.027	0.201	0.254	1.000
2	2690040	56.12	5.70	2337	0	0.050	0.027	0.186	0.253	0.996
3	2375110	63.12	6.17	2911	0	0.054	0.027	0.218	0.253	0.990
4	2258875	65.14	6.82	3415	0	0.058	0.027	0.226	0.253	0.991
5	1646205	60.47	6.65	2297	0	0.056	0.026	0.220	0.254	1.000
6	2381653	63.51	6.49	2835	0	0.056	0.027	0.210	0.253	0.993
7	1794565	69.07	6.73	2864	0	0.056	0.026	0.230	0.253	0.987
8	2051325	67.93	6.72	2885	0	0.056	0.026	0.220	0.253	0.993
9	1525131	70.81	6.35	1697	0	0.052	0.026	0.200	0.252	0.992
10	1699246	71.47	7.31	2297	0	0.059	0.026	0.206	0.253	0.984
11	1788886	67.24	6.78	2342	0	0.056	0.026	0.210	0.254	0.992
12	969528	81.67	7.59	1107	0	0.057	0.025	0.188	0.254	0.988
13	1110795	82.85	8.42	2006	0	0.066	0.026	0.227	0.253	0.986
14	1036080	83.76	8.65	1865	0	0.064	0.026	0.219	0.253	0.988
15	979300	85.24	8.08	1210	0	0.059	0.025	0.194	0.252	0.989
16	878475	86.53	7.83	1444	0	0.059	0.025	0.216	0.253	0.998
17	885115	85.60	8.12	1115	0	0.059	0.025	0.194	0.252	0.996
18	614386	94.60	9.34	939	0	0.063	0.025	0.199	0.253	1.000
19	637885	95.86	8.85	954	0	0.061	0.025	0.200	0.255	1.000
20	871860	86.27	8.90	1148	0	0.062	0.025	0.188	0.253	0.990
21	636756	92.59	8.20	850	0	0.057	0.024	0.194	0.253	0.981
22	562330	95.71	9.47	1005	0	0.066	0.024	0.212	0.253	0.918
23	404550	99.53	9.67	590	0	0.063	0.024	0.193	0.252	0.986
24	575128	94.06	10.77	1161	0	0.072	0.025	0.213	0.253	0.994
25	330078	100	11.70	622	0	0.073	0.024	0.197	0.253	0.955
26	390286	99.95	11.22	683	0	0.071	0.024	0.194	0.253	0.985
27	328455	100	11.55	599	0	0.070	0.024	0.194	0.253	0.997
28	307720	100	10.38	433	0	0.063	0.023	0.180	0.253	0.984
29	367653	99.91	10.37	581	0	0.063	0.024	0.183	0.253	0.954
Overall	36131636	71.64	7.14	47659	0	0.057	0.026	0.207	0.253	0.992

* The genetic distance was obtained from the physical distance by a simple conversion of 1 cM = 1 Mb.

Table 2.6 Distribution of two-locus composite linkage disequilibrium (r^2_{CLD}) in the Kinsella composite beef population. Total number of syntenic pairs, % of syntenic pairs with distance ≤ 50 cM*, % of pairs with significant r^2_{CLD} and the proportion of marker pairs with $r^2_{CLD} \geq 0.25$ are presented. The 95% empirical intervals are given for all syntenic pairs, pairs with significant r^2_{CLD} and pairs with $r^2_{CLD} \geq 0.25$ on 29 autosomes.

BTA	Total pairs	% of pairs with distance ≤50cM	% of pairs with significant CLD	# of pairs with r^2_{CLD}>0.25	All syntenic pairs		Pairs with significant LD		Pairs with r^2_{CLD} ≥0.25	
					2.50%	97.50%	2.50%	97.50%	2.50%	97.50%
1	4034220	51.93	4.95	3531	0	0.048	0.032	0.212	0.253	0.985
2	2690040	56.12	5.63	2256	0	0.049	0.031	0.193	0.253	0.995
3	2375110	63.12	5.93	2731	0	0.052	0.031	0.223	0.254	0.942
4	2258875	65.14	6.79	3370	0	0.057	0.031	0.235	0.253	0.961
5	1646205	60.47	6.59	2202	0	0.055	0.031	0.226	0.253	0.975
6	2381653	63.51	6.21	2652	0	0.054	0.031	0.212	0.253	0.949
7	1794565	69.07	6.64	2746	0	0.056	0.031	0.240	0.254	0.959
8	2051325	67.93	6.71	2861	0	0.055	0.031	0.229	0.253	0.978
9	1525131	70.81	6.33	1661	0	0.052	0.030	0.208	0.254	0.979
10	1699246	71.47	7.39	2311	0	0.059	0.031	0.216	0.253	0.956
11	1788886	67.24	6.83	2263	0	0.056	0.031	0.215	0.253	0.958
12	969528	81.67	7.63	1116	0	0.057	0.030	0.195	0.253	1.002
13	1110795	82.85	8.17	1854	0	0.064	0.030	0.229	0.253	0.945
14	1036080	83.76	8.54	1851	0	0.064	0.030	0.229	0.253	0.993
15	979300	85.24	8.02	1194	0	0.059	0.030	0.199	0.253	0.950
16	878475	86.53	7.89	1415	0	0.059	0.029	0.227	0.253	0.957
17	885115	85.60	8.26	1120	0	0.060	0.029	0.201	0.254	0.968
18	614386	94.60	9.26	922	0	0.063	0.029	0.205	0.253	0.981
19	637885	95.86	8.91	955	0	0.062	0.029	0.208	0.253	1.014
20	871860	86.27	8.78	1121	0	0.061	0.029	0.195	0.252	0.970
21	636756	92.59	8.02	829	0	0.056	0.029	0.200	0.254	0.958
22	562330	95.71	9.63	1032	0	0.066	0.029	0.220	0.253	0.933
23	404550	99.53	9.49	565	0	0.062	0.028	0.197	0.253	0.930
24	575128	94.06	10.42	1064	0	0.070	0.029	0.212	0.253	0.959
25	330078	100.00	11.94	636	0	0.074	0.028	0.204	0.252	0.922
26	390286	99.95	10.88	675	0	0.070	0.028	0.200	0.252	0.989
27	328455	100.00	11.52	583	0	0.070	0.028	0.202	0.253	0.989
28	307720	100.00	10.29	410	0	0.062	0.027	0.185	0.255	0.988
29	367653	99.91	10.35	571	0	0.063	0.028	0.192	0.253	0.924
Overall	36131636	71.64	7.09	46497	0	0.056	0.030	0.214	0.253	0.968

* The genetic distance was obtained from the physical distance by a simple conversion of 1 cM = 1 Mb.

Table 2.7 Distribution of two-locus zygotic linkage disequilibrium (r^2_{ZLD}) in the Kinsella composite beef population. Total number of syntenic pairs, % of syntenic pairs with distance ≤ 50 cM*, % of pairs with significant r^2_{ZLD} and the proportion of marker pairs with $r^2_{ZLD} \geq 0.25$ are presented. The 95% empirical intervals are given for all syntenic pairs, pairs with significant r^2_{ZLD} and pairs with $r^2_{ZLD} \geq 0.25$ on 29 autosomes.

BTA	Total pairs	% of pairs with distance ≤50cM	% of pairs with significant ZLD	# of pairs with $r^2_{ZLD}>0.25$	All syntenic pairs		Pairs with significant LD		Pairs with r^2_{ZLD} ≥0.25	
					2.50%	97.50%	2.50%	97.50%	2.50%	97.50 %
1	4034220	51.93	0.59	1448	0	0.012	0.032	0.509	0.256	1.000
2	2690040	56.12	0.62	995	0	0.013	0.032	0.485	0.255	0.996
3	2375110	63.12	0.75	1054	0	0.014	0.031	0.492	0.255	1.000
4	2258875	65.14	0.81	1148	0	0.014	0.031	0.501	0.254	1.000
5	1646205	60.47	0.77	856	0	0.014	0.031	0.507	0.255	1.000
6	2381653	63.51	0.81	1111	0	0.014	0.031	0.474	0.255	0.996
7	1794565	69.07	0.88	1120	0	0.014	0.031	0.532	0.258	1.000
8	2051325	67.93	0.84	1127	0	0.014	0.031	0.481	0.259	1.000
9	1525131	70.81	0.84	767	0	0.014	0.031	0.500	0.255	0.996
10	1699246	71.47	0.92	907	0	0.015	0.031	0.454	0.256	1.000
11	1788886	67.24	0.84	847	0	0.014	0.031	0.479	0.255	1.000
12	969528	81.67	0.86	443	0	0.014	0.030	0.444	0.253	1.000
13	1110795	82.85	0.98	669	0	0.015	0.030	0.475	0.255	1.000
14	1036080	83.76	0.93	654	0	0.014	0.030	0.530	0.260	0.996
15	979300	85.24	0.99	445	0	0.015	0.030	0.377	0.255	1.000
16	878475	86.53	0.99	585	0	0.015	0.030	0.533	0.255	1.000
17	885115	85.60	1.00	415	0	0.015	0.029	0.376	0.254	1.000
18	614386	94.60	0.95	332	0	0.014	0.029	0.490	0.259	1.000
19	637885	95.86	0.91	331	0	0.014	0.029	0.531	0.263	1.000
20	871860	86.27	1.00	410	0	0.015	0.029	0.392	0.258	1.000
21	636756	92.59	0.93	313	0	0.015	0.029	0.478	0.255	0.992
22	562330	95.71	1.21	378	0	0.016	0.029	0.410	0.254	0.991
23	404550	99.53	0.97	204	0	0.014	0.028	0.450	0.257	0.996
24	575128	94.06	1.26	434	0	0.017	0.029	0.471	0.257	1.000
25	330078	100.00	1.08	205	0	0.015	0.028	0.444	0.256	0.992
26	390286	99.95	1.47	250	0	0.018	0.028	0.377	0.255	0.996
27	328455	100.00	1.29	207	0	0.017	0.028	0.419	0.254	1.000
28	307720	100.00	1.08	146	0	0.016	0.027	0.363	0.253	1.000
29	367653	99.91	1.12	238	0	0.016	0.028	0.452	0.252	1.000
Overall	36131636	71.64	0.85	18039	0	0.014	0.030	0.478	0.255	1.000

* The genetic distance was obtained from the physical distance by a simple conversion of 1 cM = 1 Mb.

Table 2.8 Numbers of pairs with significant gametic LD (N_{GLD}), composite LD (N_{CLD}) and zygotic LD (N_{ZLD}); numbers of pairs shared by gametic and composite LD ($N_{GLD,CLD}$), gametic and zygotic LD ($N_{GLD,ZLD}$) and composite and zygotic LD ($N_{CLD,ZLD}$) and the correlations between pairs of the three LD measures ($r_{GLD,CLD}$, $r_{GLD,ZLD}$, and $r_{CLD,ZLD}$).

BTA	N_{GLD}	N_{CLD}	N_{ZLD}	$N_{GLD,CLD}$	$N_{GLD,ZLD}$	$N_{CLD,ZLD}$	$r_{GLD,CLD}$	$r_{GLD,ZLD}$	$r_{CLD,ZLD}$
1	194587	199707	23812	176906	23652	23192	0.988	0.914	0.899
2	153231	151516	16648	136598	16481	16236	0.987	0.919	0.905
3	146561	140779	17878	129748	17758	17446	0.988	0.906	0.891
4	154021	153420	18303	139550	18189	18004	0.987	0.901	0.886
5	109540	108539	12725	98752	12592	12433	0.987	0.903	0.888
6	154685	147840	19225	136410	19109	18755	0.986	0.904	0.886
7	120788	119102	15853	108495	15726	15554	0.988	0.911	0.900
8	137888	137572	17175	124151	17019	16857	0.988	0.911	0.902
9	96817	96502	12799	87444	12698	12546	0.990	0.920	0.906
10	124176	125582	15664	113030	15534	15352	0.987	0.914	0.900
11	121357	122197	14950	110291	14855	14635	0.988	0.902	0.888
12	73585	74013	8331	66861	8267	8165	0.988	0.911	0.902
13	93583	90798	10917	83983	10815	10727	0.988	0.901	0.889
14	89575	88447	9626	81504	9557	9468	0.988	0.904	0.890
15	79088	78532	9705	71266	9625	9496	0.987	0.899	0.884
16	68813	69285	8687	62799	8606	8512	0.990	0.914	0.901
17	71903	73123	8895	66030	8831	8689	0.987	0.899	0.882
18	57354	56909	5860	52275	5799	5759	0.988	0.909	0.897
19	56465	56851	5780	51663	5732	5680	0.989	0.910	0.900
20	77597	76583	8725	69994	8632	8512	0.986	0.900	0.887
21	52196	51095	5919	46630	5847	5773	0.987	0.902	0.890
22	53248	54135	6794	48940	6746	6677	0.987	0.901	0.885
23	39131	38395	3928	35323	3881	3837	0.987	0.896	0.882
24	61926	59907	7223	55574	7161	7096	0.986	0.907	0.891
25	38617	39423	3578	35966	3546	3521	0.989	0.896	0.887
26	43786	42481	5724	39321	5657	5620	0.983	0.906	0.891
27	37933	37841	4225	34760	4185	4148	0.988	0.901	0.886
28	31943	31665	3325	28896	3287	3254	0.986	0.893	0.884
29	38132	38050	4117	34566	4073	4035	0.987	0.908	0.890

Table 2.9 Frequency and mean values of gametic, composite and zygotic LD between syntenic SNP pairs for different ranges of distance between markers at a close vicinity (≤ 5 Mb) in the Kinsella composite beef population.

Distance Range (Mb)	Pairs (n)	Mean r^2 ±SD (gld)	Mean r^2 ± SD (cld)	Mean r^2 ± SD (zld)	$r^2 \geq 0.25$ (gld) (%)	$r^2 \geq 0.25$ (cld) (%)	$r^2 \geq 0.25$ (zld) (%)
<0.025	5692	0.2853 ±0.3071	0.2857 ±0.3070	0.1812 ±0.3011	38.3	38.5	22.9
0.025-0.05	20032	0.2175 ±0.2638	0.2178 ±0.2638	0.1230 ±0.2425	29.5	29.4	15.8
0.05-0.075	18041	0.1615 ±0.2191	0.1619 ±0.2192	0.0799 ±0.1884	20.8	20.9	10.0
0.075-0.1	17838	0.1277 ±0.1832	0.1278 ±0.1832	0.0561 ±0.1479	15.6	15.5	6.7
0.1-0.2	69905	0.0912 ±0.1420	0.0914 ±0.1422	0.0340 ±0.1059	9.4	9.5	3.6
0.2-0.5	205321	0.0569 ±0.0885	0.0571 ±0.0887	0.0159 ±0.0559	3.7	3.7	1.0
0.5-1.5	665397	0.0394 ±0.0580	0.0395 ±0.0581	0.0091 ±0.0300	1.3	1.3	0.3
1.5-3	968768	0.0286 ±0.0424	0.0286 ±0.0424	0.0062 ±0.0195	0.5	0.5	0.1
3-5	1252994	0.0214 ±0.0315	0.0214 ±0.0316	0.0045 ±0.0131	0.1	0.1	0.0

Figures

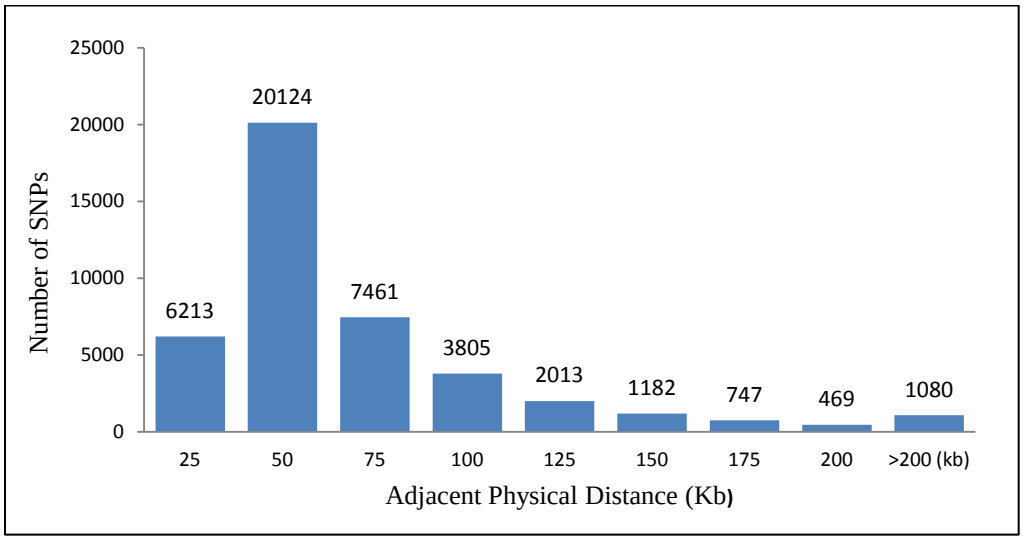

Figure 2.1 Distribution of all adjacent marker distances at 9 distance intervals (0-25, 25-50, 50-75, 75-100, 100-125, 125-150, 150-175, 175-200 and >200 kb) for all 43,124 single nucleotide polymorphic markers.

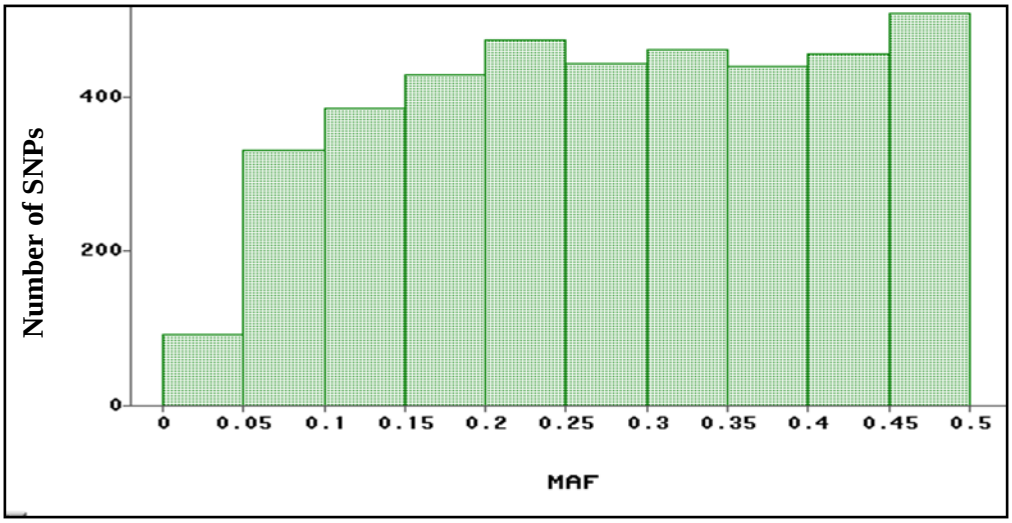

Figure 2.2 Distribution of 4,024 SNP markers with significant Hardy-Weinberg disequilibrium over different minor allele frequencies (MAF) in the Kinsella composite beef population.

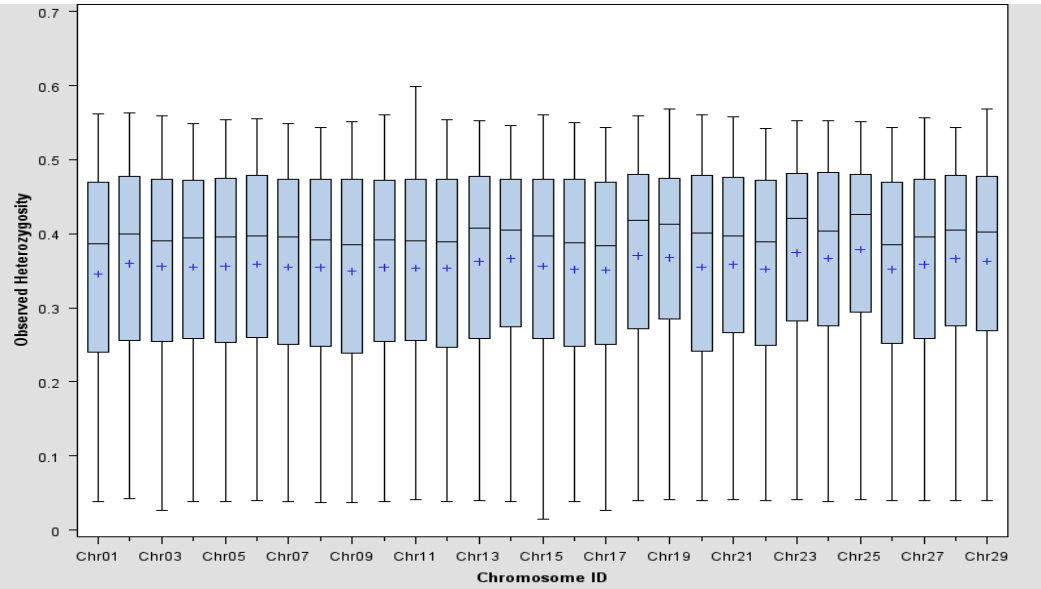

Figure 2.3 Boxplots for describing observed heterozygosities over 29 autosomes in the Kinsella composite beef population.

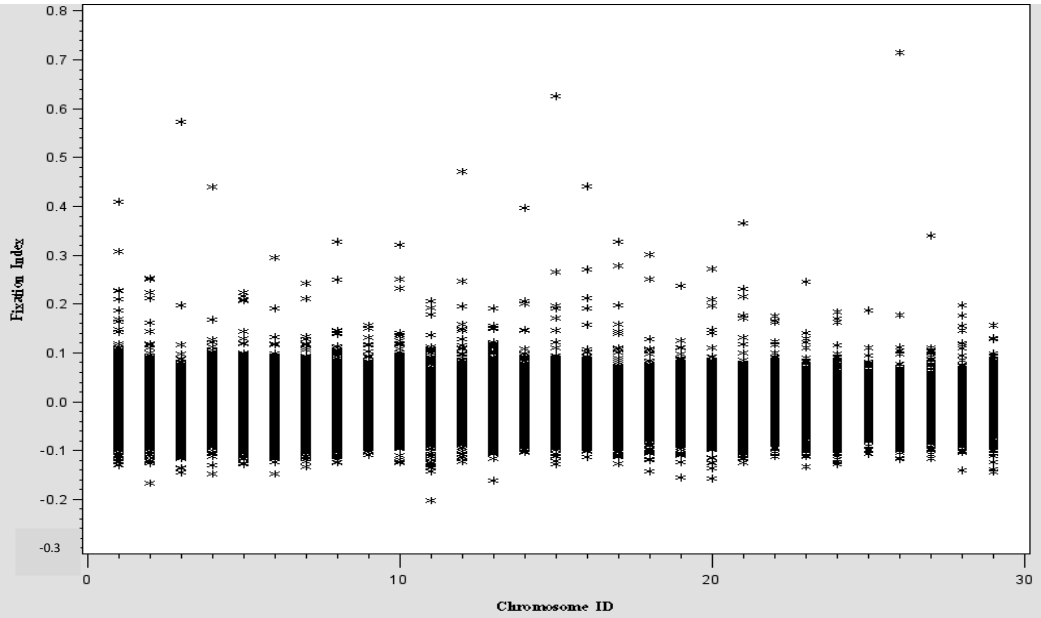

Figure 2.4 Fixation indices of 43,124 markers with significant HWD in the Kinsella composite beef population.

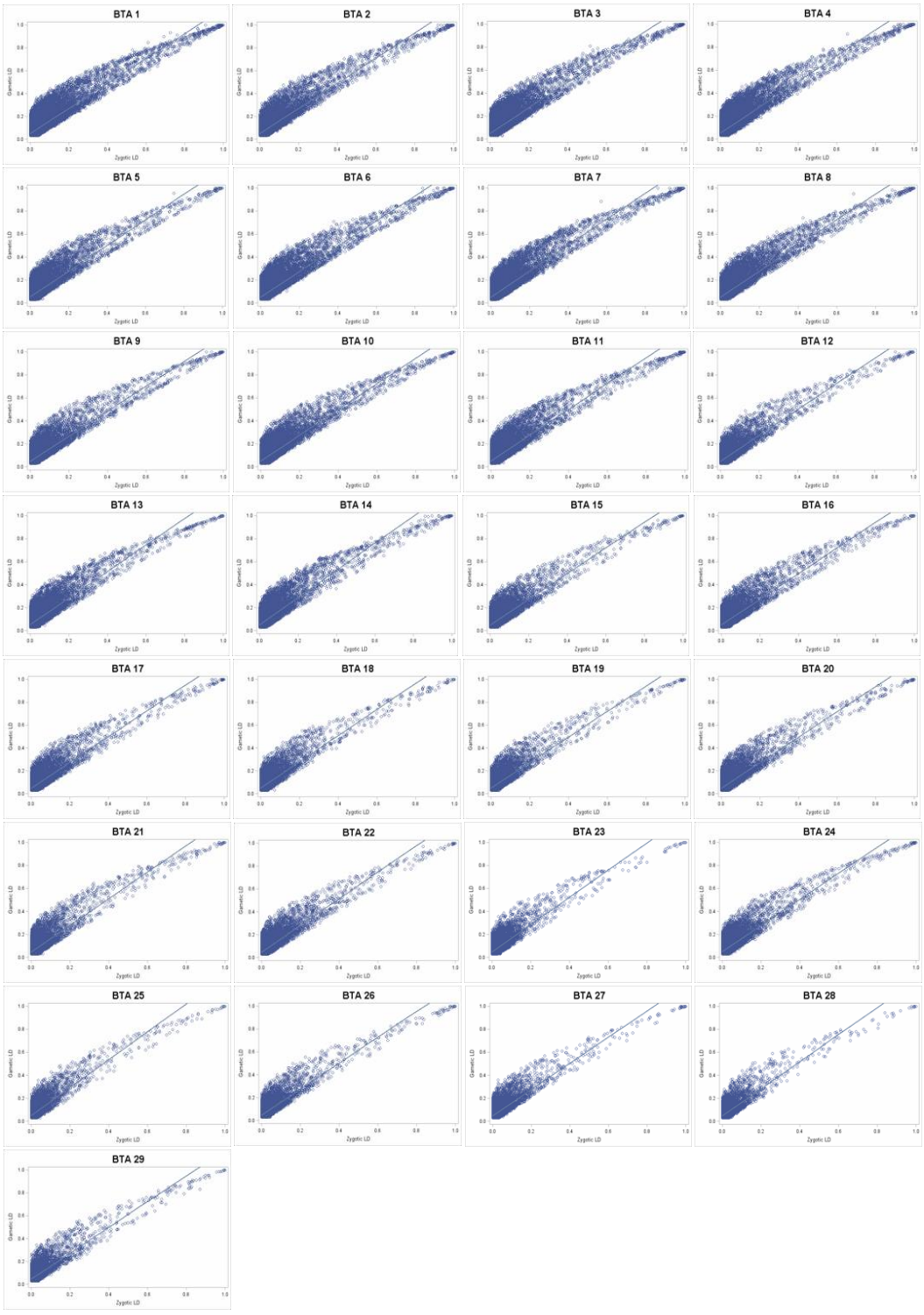

Figure 2.5 Correlation between gametic and zygotic LD for marker pairs with significant gametic and zygotic LD. The straight lines are the fitted regression lines in the Kinsella composite beef population.

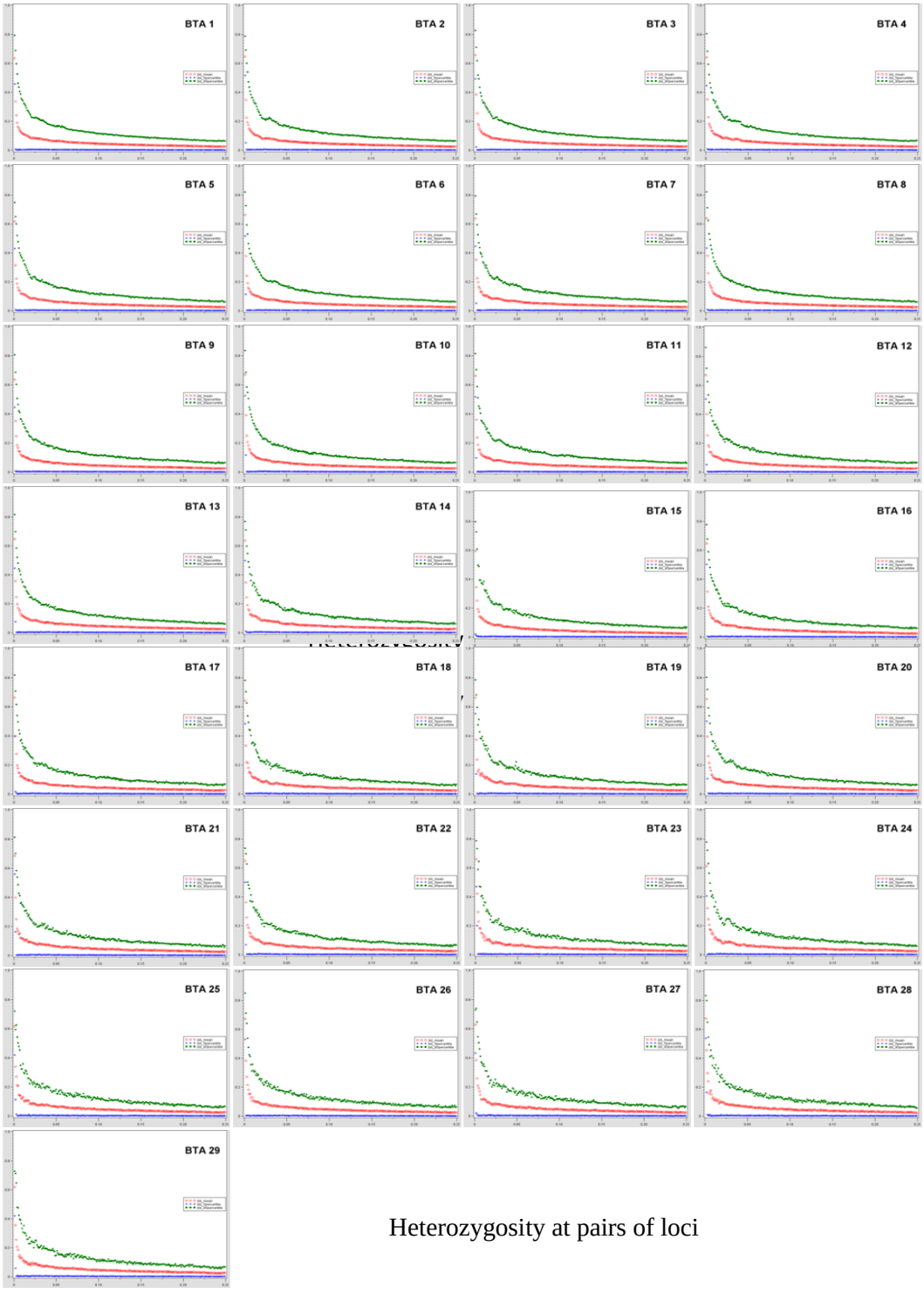

Figure 2.6 Distribution of the standardized zygotic linkage disequilibrium with little gametic LD (r^2<0.001) on all 29 autosomes. The 95 percentile, the mean and the 5 percentile were calculated at 1000 two-locus heterozygosity intervals (at 0.001 increments). The horizontal axis is represented by the level of two-locus heterozygosity, and the vertical axis is represented by the standardized zygotic LD.

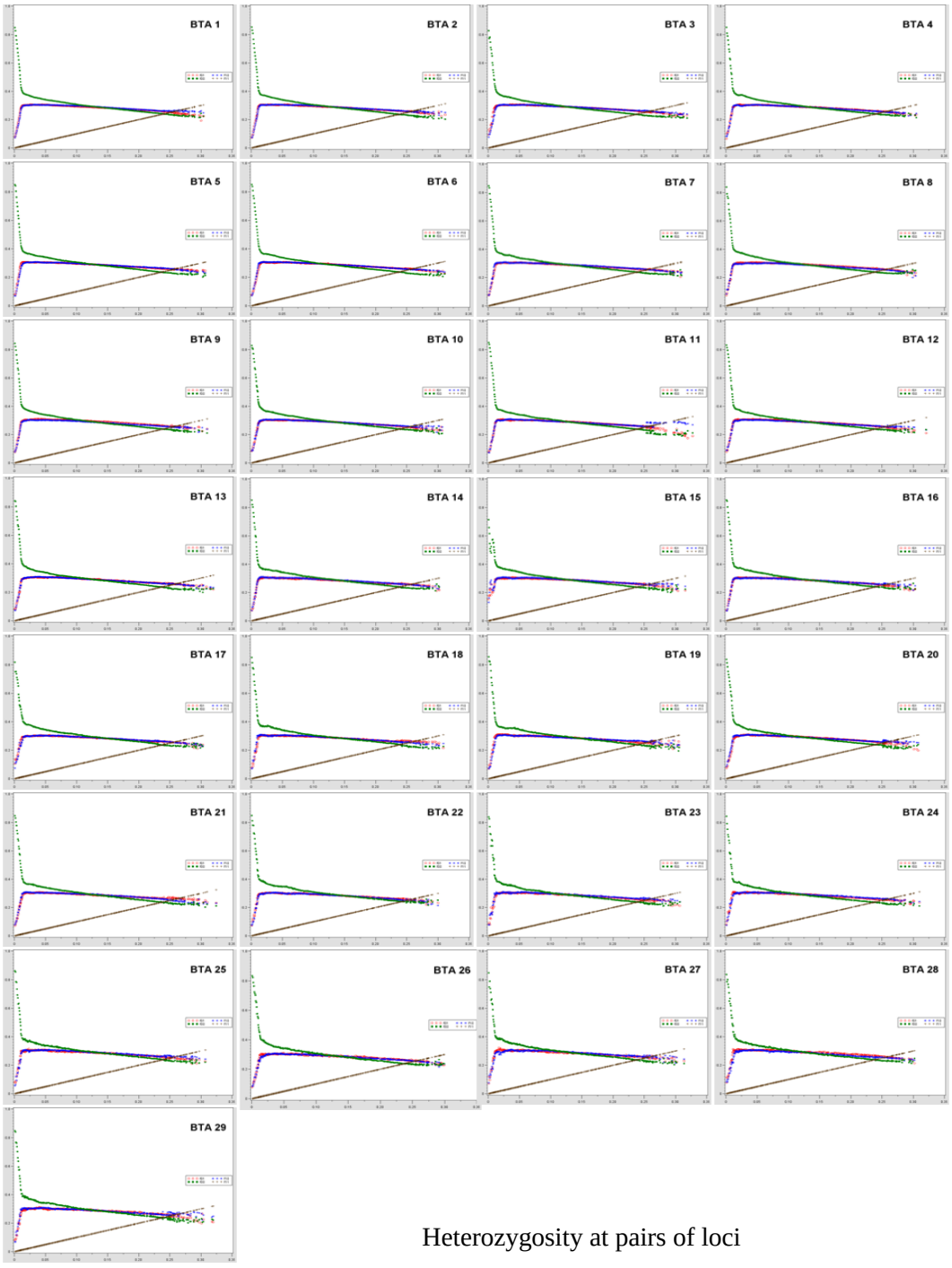

Figure 2.7 Distribution of four classes of two-locus genotypic frequencies for all marker pairs with significant zygotic linkage disequilibrium but with little gametic linkage disequilibrium. The horizontal axis represents the two-locus heterozygosity, and the vertical axis represents the two-locus genotype frequencies. The four zygote classes are: (i) the two-locus homozygosity (green); (ii) homozygote at locus A but heterozygote at locus B (red); (iii) heterozygote at locus A but homoztgote at locus B(blue) and (iv) two-locus zygotic heterozygosity (yellow).

Chapter 3. Genome-wide Analysis of Components of Zygotic Linkage Disequilibrium

3.1 Introduction

In Chapter 2, we provided a comparative assessment of zygotic, composite and gametic LD in the Kinsella beef composite population. Such comparison allowed for characterizing the pattern and distribution of the genome-wide non-allelic associations at both gametic and zygotic levels. However, the zygotic LD is a summary statistic that consists of digenic (gametic or composite LD), trigenic and quadrigenic components [1, 2]. There is little knowledge on the significance of individual genic disequilibria in this and other livestock populations. We know only one study by Liu et al. [3] who attempted to examine high-order trigenic and quadrigenic disequilibria in a canine population but with a small number of dogs and a limited number of markers. Such information would certainly help explain the pattern observed in Chapter 2 that zygotic LD was smaller than gametic or composite LD. In addition, the magnitudes and patterns of individual genic disequilibria across different chromosomes may be compared and contrasted to infer about the effects of cross breeding and selection on genomic structure of the Kinsella population.

It was also confirmed in Chapter 2 that the relationship between the zygotic LD and marker distance was just like that between the gametic or composite LD and marker distance. However the zygotic LD decayed faster than gametic or composite LD over the physical distance. It remains to be investigated if such relationship would be held for individual components of the zygotic LD and physical distance.

Since all LD measures (zygotic LD and its components) are just functions of gene frequencies at different loci, their patterns and extent are largely dependent on gene frequencies. For example, gametic LD is generally smaller if gene frequencies are towards extreme values than if gene frequencies are

intermediate. Thus there is a need to examine the dependence of zygotic LD and its components on gene frequencies.

The objectives of this chapter are (i) to determine if individual genic disequilibria are significant; (ii) to investigate the relationship between individual components of zygotic LD and physical distance; and (iii) to examine effects of changing gene frequencies on different LD measures.

3.2 Materials & methods

3.2.1 Genomic data

The same genomic data set as used in Chapter 2 was employed here again for estimating and testing for two-, three- and four-gene disequilibria. The data set for the analysis consisted of 1,023 animals each with 43,124 SNP markers over 29 bovine autosomes.

3.2.2 Components of zygotic linkage disequilibrium

Following Yang [4], definitions and notations of allele frequencies, single- and two-locus genotypic frequencies and zygotic LD between loci *A* and *B* are given in Appendix 3.1. It was established [5,2] that the total zygotic LD between loci *A* and *B* as given in Chapter 2 could equivalently be defined as a sum of individual zygotic LDs for double homozygotes, for example,

$$\omega_{AB} = \omega_{AB}^{AB} + \omega_{Ab}^{Ab} + \omega_{aB}^{aB} + \omega_{ab}^{ab} \ ,$$

with each zygotic LD being a complex function of digenic, trigenic and quadrigenic disequilibria. For example, the zygotic LD for double homozygote *AABB* (ω_{AB}^{AB}) could be written as,

$$\begin{aligned}\omega_{AB}^{AB} &= P_{AB}^{AB} - P_{A\cdot}^{A\cdot} P_{\cdot B}^{\cdot B} \\ &= 2p_A D_{\cdot B}^{AB} + 2p_B D_{A\cdot}^{AB} + 2p_A p_B D_{\cdot\cdot}^{AB} + 2p_A p_B D_{\cdot B}^{A\cdot} + (D_{\cdot\cdot}^{AB})^2 + (D_{\cdot B}^{A\cdot})^2 + D_{AB}^{AB}\end{aligned}$$

(3.1)

where each genic disequilibrium (D) is the deviation of a frequency from that based on random association of genes and accounting for any lower order disequilibria. The usual gametic LD ($D_{\cdot\cdot}^{AB}$) would be the deviation of frequency of gamete *AB* from the product of frequencies of allele *A* at locus *A* and allele *B* at locus *B*, $D_{\cdot\cdot}^{AB} = P_{\cdot\cdot}^{AB} - p_A p_B$ with

$$P_{\cdot\cdot}^{AB} = P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}.$$

With zygotes arising from random union of gametes as often assumed in other LD studies, all inter-gametic disequilibria including Hardy-Weinberg disequilibrium (HWD) would disappear (*e.g.*, $D_{A\cdot}^{A\cdot} = D_{\cdot B}^{A\cdot} = D_{\cdot B}^{AB} = D_{AB}^{AB} = 0$). In this case, the zygotic LD for genotype *AABB* (ω_{AB}^{AB}) would reduce to,

$$\omega_{AB}^{AB} = 2p_A p_B D_{\cdot\cdot}^{AB} + (D_{\cdot\cdot}^{AB})^2$$

This formula was the basis for possible use of double homozygosity to measure gametic LD [6, 7].

Since the two types of double heterozygote (AB/ab and Ab/aB) in our SNP data could not be distinguished, we used the composite LD (Δ_{AB}) and a composite quadrigenic component (Δ_{AABB}) in place of gametic and quadrigenic disequilibria. Thus, the zygotic LD for genotype *AABB* (ω_{AB}^{AB}) in equation (3.1) was rewritten as

$$\begin{aligned}\omega_{AB}^{AB} &= P_{AB}^{AB} - P_{A\cdot}^{A\cdot} P_{\cdot B}^{\cdot B} \\ &= 2p_A D_{ABB} + 2p_B D_{AAB} + 2p_A p_B \Delta_{AB} + \Delta_{AB}^2 + \Delta_{AABB}\end{aligned} \qquad (3.2)$$

where

$$\begin{aligned}\Delta_{AB} &= P_{\cdot\cdot}^{AB} + P_{\cdot B}^{A\cdot} - 2p_A p_B \\ &= D_{\cdot\cdot}^{AB} + D_{\cdot B}^{A\cdot}\end{aligned}$$

and

$$\Delta_{AABB} = D_{AB}^{AB} - 2D_{\cdot\cdot}^{AB} D_{\cdot B}^{A\cdot}.$$

It should be noted from equations (3.1) and (3.2) that the two trigenic disequilibria in (3.2) were rewritten without superscripts for notational simplicity.

3.2.3 Maximum likelihood estimation

Following Weir and Cockerham [8] and Weir [9], we used the procedure of statistical inference based on the assumption of multinomial sampling of individual diploids from a population. The observed frequencies and disequilibria with tildes (~) were maximum likelihood (ML) estimates of corresponding parametric values. Since the additive models described in Section 3.2.2 allowed for defining the same number of parameters as there would be degrees of freedom, the ML estimates were simply replacing all parametric values of frequencies and disequilibria with corresponding observed values. For example, the ML estimates of composite LD was simply given by,

$$\begin{aligned} \tilde{\Delta}_{AB} &= \tilde{P}_{\cdot\cdot}^{AB} + \tilde{P}_{\cdot B}^{A\cdot} - 2\tilde{p}_A\tilde{p}_B \\ &= \tilde{D}_{\cdot\cdot}^{AB} + \tilde{D}_{\cdot B}^{A\cdot} \end{aligned}$$

However, the ML estimates might be biased because they would involve quadratic terms of multinomial variables. For example, the expectation of the squared gene frequency of allele *A* over replicate samples of size *n* would be,

$$E(\tilde{p}_A^2) = p_A^2 + [p_A(1-p_A) + D_A]/2n,$$

where D_A is the HWD measure at locus *A* [9]. With the sufficiently large sample (n = 1023 animals) in our data set, we invoked large-sample theory for statistical inference about genic disequilibria. Thus, we ignored the possible biases of order $1/n$ and stayed with the ML estimates for hypothesis testing.

3.2.4 Hypothesis testing

With a ML estimate ($\tilde{D}$ or $\tilde{\Delta}$) of a given genic disequilibrium D or Δ, along with its sampling variance, [Var($\tilde{D}$) or Var($\tilde{\Delta}$)], we constructed a test statistic,

$$X^2 = \tilde{D}^2 / Var(\tilde{D}) \quad \text{or} \quad X^2 = \tilde{\Delta}^2 / Var(\tilde{\Delta})$$

to test the hypothesis of zero disequilibrium (i.e., H_0: $D = 0$ or H_0: $\Delta = 0$). Assuming the asymptotic normality of the ML estimate, X^2 under the hypothesis of zero disequilibrium would be distributed as chi-square with one degree of freedom.

The sampling variances of individual genic disequilibrium estimates [Var($\tilde{D}$) and Var($\tilde{\Delta}$)] were derived by Weir and Cockerham [8] using Fisher's [10] expression for the approximate variance of a quadratic function (T) of multinomial variables. For easier reference, they were reproduced in Appendix 3.2. Since the parametric values of frequencies and/or disequilibria in the variance expressions were unknown, they were substituted by respective observed or estimated values. When testing a given genic disequilibrium, we did not follow Weir and Cockerham's [8] suggestion of setting that particular disequilibrium to zero in its variance expression. For example, to test for composite LD, Weir and Cockerham [8] suggested the test statistic after setting $\tilde{\Delta}_{AB} = 0$,

$$X^2_{AB} = n\tilde{\Delta}^2_{AB} / [(\tilde{\pi}_A + \tilde{D}_A)(\tilde{\pi}_B + \tilde{D}_B) + \tilde{\tau}_A \tilde{D}_{ABB} + \tilde{\tau}_B \tilde{D}_{AAB} + \tilde{\Delta}_{AABB}] \qquad (3.3a)$$

where $\pi_i = p_i(1\text{-}p_i)$ and $\tau_i = 1\text{-}2p_i$. However, we did not set $\tilde{\Delta}_{AB} = 0$ in the variance expression because our view was that there would be no basis for the existence of three- and four-gene disequilibria in the absence of composite LD. Thus, our test statistic for composite LD was,

$$X^2_{AB} = n\tilde{\Delta}^2_{AB} / [(\tilde{\pi}_A + \tilde{D}_A)(\tilde{\pi}_B + \tilde{D}_B) + \tilde{\tau}_A \tilde{\tau}_B \tilde{\Delta}_{AB} / 2 + \tilde{\tau}_A \tilde{D}_{ABB} + \tilde{\tau}_B \tilde{D}_{AAB} + \tilde{\Delta}_{AABB}]$$

(3.3b)

3.2.5 Chi-square statistic and correlation

In Chapter 2, we routinely used the squared correlation (r^2) as a measure of gametic LD (r^2_{GLD}), composite LD (r^2_{CLD}), or zygotic LD (r^2_{ZLD}). We used a chi-square statistic ($X_i^2 = nr_i^2$, *i* = *GLD*, *CLD* or *ZLG*) to test for the significance of the LD estimate. We knew from the literature [11] that the relationship of $X_i^2 = nr_i^2$ would hold exactly only for a 2×2 contingency table. This was the case for GLD and ZLD, but not for CLD. When dropping out three- and four-gene disequilibria in testing for zero composite LD ($\tilde{\Delta}_{AB} = 0$), we would obtain a simpler version of the chi-square statistic from (3.3a) or (3.3b),

$$X_{AB}^2 = n\tilde{\Delta}_{AB}^2 / [(\tilde{\pi}_A + \tilde{D}_A)(\tilde{\pi}_B + \tilde{D}_B)]$$

which would be equal to nr^2_{CLD} as given in Chapter 2. Similar approximations or restrictions would be needed if the relationship of $X_i^2 = nr_i^2$ were desired for three- and four-gene disequilibria. Thus, to avoid such approximations or restrictions, we used a generalized measure of square correlation $\phi^2 = X^2/n$ [11] in place of r^2 as a standardized measure of genic disequilibria. As pointed out above, the relationship of $\phi^2 = r^2$ would hold only for a 2×2 contingency table.

3.3 Results

3.3.1 Zygotic LD and its components

The estimated powers of chi-square test statistic for gametic, composite and zygotic LDs are presented for two groups of SNP pairs, those with a distance of ≤ 50 cM (Linked Group) and those with a distance of > 50 cM (Unlinked Group) (Table 3.1). The genetic distance was obtained from the physical distance through the simple conversion of 1cM = 1 Mb. The threshold of 50 cM would be an indicator of whether or not a pair of markers is freely recombined. Within each group, the chi-square tests for gametic LD and composite LD had similar powers,

but they both were more powerful than the chi-square tests for zygotic LD. Genome-wide, the chi-square tests for gametic and composite LD were ~15% more powerful in the Linked Group than in the Unlinked Group, but the chi-square tests for zygotic LD were only ~7% more powerful in the Linked Group than in the Unlinked Group. For the Linked Group, the range of mean powers for gametic LD was from 50.6% on BTA 19 to 53.9% on BTA 25; the range of mean powers for composite LD was from 50.6% on BTA 19 to 54.0% on BTA 25; and the range of mean powers for zygotic LD was from 16.9% on BTA 24 to 19.7% on BTA 11. The corresponding ranges of mean powers for the Unlinked Group were 0.348 (BTA12) – 0.401 (BTA 29) for gametic LD, 0.344 (BTA 12) – 0.392 (BTA 29) for composite LD and 0.105 (BTA 23) – 0.126 (BTA 11) for zygotic LD. It should be noted that chromosomes 25, 26, 27 and 28 are shorter than 50 cM.

The estimated powers of the chi-square tests for trigenic and quadrigenic components of the zygotic LD are given for the Linked (≤ 50 cM) and Unlinked (> 50 cM) Groups (Table 3.2). For a given trigenic or quadrigenic disequilibrium, the powers of chi-square tests were similar regardless of the distance between marker pairs. The ranges of the powers for each of the two trigenic disequilibria were 0.103-0.145 in the Linked Group and 0.081-0.184 in the Unlinked Group. Such ranges for the quadrigenic disequilibrium were 0.072-0.092 in the Linked Group and 0.052-0.066 in the Unlinked Group. It should be noted that the estimated powers were based on the number of marker pairs left after removing those with the generalized squared correlations (ϕ^2) being outside the acceptable range of 0 to 1. We recorded separately the frequencies of the two out-of-bound situations ($\phi^2 < 0$ and $\phi^2 > 1$) in Appendix 3.3. First, for $\phi^2 < 0$, the sampling variances of estimated trigenic disequilibria were negative for about 69% of the genome-wide syntenic marker pairs (36,131,636); in contrast, the sampling variances of estimated quadrigenic disequilibrium were positive for all the syntenic marker pairs. Second, for $\phi^2 > 1$, there was 0.02% of the genome-wide syntenic marker pairs for both trigenic disequilibria, but only 0.001% for quadrigenic disequilibrium.

The plot of the estimated power of chi-square tests for individual genic disequilibria against the distance between markers in the Linked Group (≤ 50 cM) (Figure 3.1) showed the power decreased with the increasing marker distance. The pace of the power decay varied with individual genic disequilibria with the composite LD being the slowest but the two trigenic disequilibria being the fastest.

Presented in Table 3.3 are the generalized squared correlations (ϕ^2) of gametic, composite, trigenic and quadrigenic disequilibria averaged over all syntenic marker pairs. The genome-wide ϕ^2 values for digenic disequilibria (gametic and composite LD) were about three times those for trigenic and quadrigenic disequilibria. There was variation among chromosomes in terms of individual genic disequilibria. For example, the gametic LD averaged over all pairs on chromosomes ranged from 0.0082 on BTA 1 to 0.0126 on BTA 25 whereas the quadrigenic LD ranged from 0.0012 on BTA1 to 0.0016 on BTA 25. When looking at the percentages of SNP pairs with $\phi^2 \geq 0.2$ (Table 3.4), the values for digenic disequilibria were also two- to three-fold higher than for trigenic and quadrigenic disequilibria.

The mean values of LD and estimated powers for chi-square tests for gametic, composite, trigenic, quadrigenic and zygotic LD were summarized for all syntenic marker pairs (intra-chromosome pairs) and all non-syntenic pairs (inter-chromosome pairs) (Table 3.5). The mean values of individual genic disequilibria and test powers were greater for intra-chromosome pairs than for inter-chromosome pairs though such difference between intra- vs. inter-chromosome pairs were more pronounced for the digenic disequilibria than for trigenic and quadrigenic disequilibria. In particular, the two trigenic disequilibria and their test powers were almost the same for intra- and inter-chromosome pairs. The magnitudes of LD values and test powers decreased with the number of genes in the LD measures for both intra- and inter-chromosome pairs with the order of digenic LD > trigenic LD > quadrigenic LD. It should be noted that despite the same number of possible intra- or inter-chromosome marker pairs for all individual genic disequilibria, only those pairs whose generalized squared

correlations fell within the acceptable range of $0 \leq \phi^2 \leq 1$ were retained for calculating the mean LD values and estimating the test powers.

3.3.2 Effects of gene frequencies

The mean, minimum, and maximum values of estimated powers of chi-square tests for individual genic disequilibria between marker pairs belonging to nine classes of minor allele frequency (MAF) with three MAF intervals (<0.1, 0.1-0.3 and 0.3-0.5) at each of the two loci are given separately for the syntenic marker pairs (intra-chromosome pairs) and for non-syntenic pairs (inter-chromosome pairs) (Table 3.6). In all nine MAF classes, the digenic disequilibria (composite LD) and zygotic LD were greater for intra-chromosome pairs than for inter-chromosome pairs but trigenic and quadrigenic disequilibria were similar for both intra- and inter-chromosome pairs.

The estimated powers of chi-square tests for digenic, trigenic, and quadrigenic disequilibria were increased with the increasing MAF at both loci, whereas those for zygotic LD were decreased with the increasing MAF (Table 3.5). For example, for intra-chromosome pairs, the powers for composite LD were increased from 0.389 when MAF at both loci were less than 0.1 to 0.512 when MAF at both loci were above 0.3; in contrast, the powers for zygotic LD were decreased from around 0.3 when MAF at both loci were below 0.30 to about 0.1 when MAF at either locus was above 0.3. In all nine MAF classes, the powers for trigenic and quadrigenic disequilibria were much smaller than those for digenic disequilibria. In particular, the powers for trigenic and quadrigenic disequilibria in most MAF classes were below 0.05, confirming the hypothesis of zero trigenic and quadrigenic disequilibria. Similar patterns of changes in the estimated powers for individual genic disequilibria with gene frequency were observed for inter-chromosome pairs though the powers were generally smaller for inter-chromosome pairs than for intra-chromosome pairs throughout all MAF classes.

3.4 Discussion

This study represents the first major genome-wide survey of high-order genic disequilibria between three or four genes at pairs of loci in a farmed animal species. We chose the Kinsella composite beef population for such a survey because continued crossbreeding and selection for growth and cow reproduction compelled the population to constantly stay in a HWD condition, thereby providing an excellent opportunity for uncovering the high-order genic disequilibria. The survey showed that the trigenic and quadrigenic disequilibria were generally two- to three- fold smaller than the usual digenic disequilibria (gametic or composite LD) (Table 3.4). Correspondingly, there was less power of testing for these high-order genic disequilibria than for the digenic disequilibria (Tables 3.1 and 3.2). The powers decreased with the distance between markers though the decay is more obvious for the digenic disequilibria than for high-order disequilibria (Figure 3.1).

To the best of our knowledge, the only other survey of high-order genic disequilibria was made by Liu et al. [3] for a canine population with an outbred multigenerational pedigree that was initiated with a limited number of unrelated founders (seven greyhounds and six Labrador retrievers). A total of 148 dogs were sampled from this pedigree for genotyping at 247 microsatellite markers over 39 chromosomes (38 autosomes and one sex chromosome). Using the essentially same statistical analysis as in our study, Liu et al. [3] observed that the genome-wide powers of tests for composite LD, two trigenic disequilibria and quadrigenic disequilibrium were 61%, 23%, 19%, and 22%, respectively. These power estimates are clearly higher than those observed in our study (Tables 3.1 and 3.2). There are at least two possible reasons for the different powers observed in the two studies. First, the microsatellite markers used by Liu et al. [3] showed a high level of allelic diversity with the number of alleles at a marker ranging from 2 to 11 [12]. In order to fit the simple biallelic model for detecting individual genic disequilibria, Liu et al. [3] used the most frequent allele and a new synthetic allele consisting of all other alleles for the markers with more than two alleles.

This pooling of all less frequent alleles into a single new allele certainly reduces the likelihood of detecting low MAF. In contrast, the biallelic SNP markers in our study were directly used without any need for modifying allelic states. Additionally our less stringent threshold of MAF $\geq$ 2% would allow for the presence of low MAF. As evident in Table 3.6 individual genic disequilibria increased with increasing MAF. Second, 148 dogs sampled from the pedigree by Liu et al. [3] were closely inbred relatives. Moreover, as indicated above, the pedigree was established from a limited number of founders. Thus, strong founder effect coupled with high level of inbreeding would have caused large LD in the dog population; this is in comparison to our composite beef population which should have only limited founder and inbreeding effects on LD.

Liu et al. [3] observed a much greater chromosome-to-chromosome variation in individual genic disequilibria than did our study. For example, the range of test powers for composite LD in Liu et al. [3] was from 20% to 100% whereas the range of test powers for composite LD in our study was from 44.49% to 54% (Table 3.4). Clearly there is a huge difference in marker density between the two studies, with a genome-wide average marker distance being 9.3 cM in Liu et al. [3] but with a genome-wide average distance being < 1 cM (62.3 kb) in our study. In particular, there were only 2 to 13 markers on individual chromosomes of the canine genome in Liu et al. [3] but 785 to 2,841 markers on individual chromosomes of the bovine genome in our study (cf. Table 2.1). The limited number of markers sampled from individual chromosomes across the canine genome would make the study by Liu et al. [3] more likely to suffer from biased sampling of the genome. In addition, with a small sample size (148 dogs) in Liu et al. [3], the tests for individual genic disequilibria must be based on a two-way contingency table with many empty cells. Due to unpredictable distributions of these empty cells in the two-way table, the genic disequilibria might have been under- or over-emphasized, thereby resulting in a much wider range of the power values.

Our study and Liu et al. [3] have both observed that the tests for digenic disequilibria (gametic or composite LD) are much more powerful than those for trigenic and quadrigenic disequilibria. It is also shown in our study and implied in Liu et al. [3] that the digenic disequilibria are two- to three-fold larger than the high-order genic disequilibria. Liu et al. [3] did not explain this observation. In developing the statistical analysis used by our study and by Liu et al. [3], Weir and Cockerham [8] assumed the random union of gametes taken from an infinite large founder population so that all initial disequilibrium would be digenic (gametic LD). It is evident from Table 6.4 of Weir and Cockerham [8] that individual genic disequilibria in subsequent inbred generations decay by a rate gauged in terms of two-locus descent measures. Further numerical results (Table 6.5 of Weir and Cockerham [8]) showed that relative to the digenic disequilibria, the trigenic and quadrigenic disequilibria would be always small in a given inbred generation and they could take a long time to reach the equilibrium values. Thus, since the composite beef population in our study and the canine population in Liu et al. [3] are obviously not in an equilibrium condition, it is expected that the digenic LDs overpower the high-order disequilibria.

Our study is the first empirical evaluation of the effect of allele frequency on different genic disequilibria. The powers of chi-square tests for digenic, trigenic and quadrigenic disequilibria increased with the increasing gene frequency at both loci, but those for zygotic LD decreased with the increasing gene frequency. Weir and Cockerham [8] used extensive computer simulations to show the similar trend for individual genic disequilibria but these authors did not consider zygotic LD. The simulation results also confirmed that the allowance for digenic disequilibria not only led to nonzero gametic or composite LD as expected, but also to nonzero quadrigenic disequilibrium as implied by the power of the chi-square test being more than 5%. It is difficult to explain exactly the trend for zygotic LD. Since zygotic LD is a complex function of individual genic disequilibria weighted by gene frequency [2], the combinations of gene frequencies and individual disequilibria are too numerous to identify the exact

combinations of genic disequilibria at different gene frequencies for the observed trend of zygotic LD. this will certainly be an area for more research in the future.

In our study, we proposed an *ad hoc* measure of non-allelic associations (ϕ^2) based on chi-square statistics for individual genic disequilibria. It is equal to the squared correlation (r^2) only when the chi-square statistics are calculated from a 2×2 contingency table [11]. We used this *ad hoc* measure for gauging the degree of association. More importantly the measure was also used for detecting outlier chi-square statistics by setting the range of the observed ϕ^2 values as $0 \leq \phi^2 \leq 1$ for individual genic disequilibria. Before the removal of outliers, we observed extremely large standard deviations of chi-square values over marker pairs in many frequency classes for the trigenic and quadrigenic disequilibria. Weir and Cockerham [8] noted a similar problem of unusually large standard deviations of chi-square values but offered no explanation about it. After the removal of outliers, we noted that all standard deviations of chi-square values fell within the normal range. Of course, we used a pragmatic and conservative approach to trim off the outlier chi-square statistics. When the ϕ^2 value is calculated from a general $I \times J$ contingency table, its allowable range is given by $0 \leq \phi^2 \leq \min[(I\text{-}1), (J\text{-}1)]$ [11]. It is already known (e.g., [2], [9]) that the gametic and zygotic LD are calculated from a 2×2 contingency table and thus their acceptable range should be $0 \leq \phi^2 \leq 1$. However, it remains unclear of the size of the contingency tables for other genic disequilibria. Thus, more research is needed to determine appropriate contingency tables for digenic, trigenic, and quadrigenic disequilibria.

Our study has practical implications. Our results of predominant digenic disequilibria coupled with insignificant high-order disequilibria suggest that current intensive focus on the use of gametic LD for GWAS and genomic selection in cattle and other animal species is reasonable. It has been demonstrated (e.g., [13]) that gametic LD in domestic animals may occur at long distances (> 1 cM) due to the rapid and sudden decline of population size (bottleneck effect) during and after breed formation, in comparison to the

situation in human where there is no gametic LD at long distances. However, it has also been suspected that such LD may be false positive associations due to a mixture of multiple breeds. If the breeds can be identified through pedigree information, Goddard and Hayes [13] suggested the use of breeds as a covariate in the statistical model to minimize such false positive associations. However, animals in our composite beef population were long pooled and their breed identity is no longer available. So the suggestion by Goddard and Hayes is not feasible for the composite beef population. Perhaps, our demonstration of insignificant trigenic and quadrigenic disequilibria may be the only way of indicating that the admixture effect is a negligible cause of the false positive associations in the composite population.

3.5 Conclusions

This study is the first major genome-wide survey of high-order genic disequilibria between three or four genes at pairs of 50K SNP markers in a farmed animal species. The survey showed that the trigenic and quadrigenic disequilibria were generally insignificant and two- to three-fold smaller than the usual digenic disequilibria (gametic or composite LD). The powers of tests for these high-order genic disequilibria dropped rapidly even at a very short distance between SNPs. These results support the current intensive focus on the use of gametic LD for GWAS and genomic selection activities in the Kinsella composite beef population.

3.6 References

1. Cockerham CC, Weir BS: **Descent measures for two loci with some applications.** *Theor Popul Biol* 1973, **4**:300-330.
2. Yang RC: **Analysis of multilocus zygotic associations.** *Genetics* 2002, **161**:435-445.
3. Liu T, Todhunter RJ, Lu Q, Schoettinger L, Li H, Littell RC, Burton-Wurster N, Acland GM, Lust G, Wu R: **Modelling extent and distribution of zygotic disequilibrium: implications for a multigenerational canine pedigree.** *Genetics* 2006, **174**:439-453.
4. Yang RC: **Epistasis of quantitative trait loci under different gene action models.** *Genetics* 2004, **167**:1493-1505.
5. Yang RC: **Zygotic associations and multilocus statistics in a nonequilibrium diploid population.** *Genetics* 2000, **155**:1449-1458.
6. Yang RC: **Gametic and zygotic associations.** *Genetics* 2003, **165:**447-450.
7. Sabatti C, Risch N: **Response to the letter "gametic and zygotic associations" by Rong-Cai Yang.** *Genetics* 2003, **165**:451-452.
8. Weir BS, Cockerham CC: **Complete characterization of disequilibrium at two loci.** In *Mathematical Evolutionary Theory*. Edited by Feldman MW. Princeton New Jersey: Princeton University Press; 1989:86-110.
9. Weir BS: *Genetic Data Analysis II.* MA Sunderland: Sinauer Associates; 1996.
10. Fisher RA: *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd; 1925.
11. Bishop Y, Fienberg SE, Holland PW: *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press; 1975.
12. Todhunter RJ, Bliss SP, Casella G, Wu R, Lust G, Burton-Wurster NI, Williams AJ, Gilbert RO, Acland GM: **Genetic structure of susceptibility traits for hip dysplasia and microsatellite informativeness of an outcrossed canine pedigree.** J *Hered* 2003, **94**:39-48.

13. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programs.** *Nat Rev Genet* 2009, **10**: 381-391.

Tables

Table 3.1 The estimated powers* of test statistics for gametic LD, composite LD and zygotic LD for marker pairs with a distance ≤50 cM and >50cM on 29 autosomes in the Kinsella composite beef population.

BTA	$X^2_{D_{AB}}$		$X^2_{\Delta_{AB}}$		$X^2_{\omega_{AB}}$	
	≤50 cM	>50 cM	≤50 cM	>50 cM	≤50 cM	>50 cM
1	0.522	0.358	0.525	0.359	0.193	0.119
2	0.536	0.374	0.536	0.372	0.189	0.119
3	0.524	0.364	0.523	0.358	0.193	0.122
4	0.531	0.379	0.531	0.378	0.190	0.120
5	0.531	0.375	0.532	0.373	0.194	0.124
6	0.527	0.359	0.524	0.353	0.191	0.115
7	0.517	0.375	0.515	0.371	0.189	0.119
8	0.528	0.374	0.530	0.372	0.192	0.121
9	0.511	0.362	0.512	0.361	0.188	0.122
10	0.535	0.373	0.538	0.372	0.197	0.125
11	0.529	0.387	0.529	0.385	0.189	0.126
12	0.513	0.348	0.514	0.344	0.182	0.114
13	0.517	0.352	0.515	0.349	0.185	0.114
14	0.524	0.375	0.522	0.370	0.177	0.112
15	0.521	0.367	0.521	0.364	0.191	0.121
16	0.511	0.365	0.514	0.369	0.184	0.121
17	0.518	0.360	0.519	0.361	0.191	0.120
18	0.517	0.386	0.514	0.383	0.172	0.107
19	0.506	0.368	0.506	0.369	0.169	0.116
20	0.530	0.378	0.531	0.372	0.191	0.118
21	0.508	0.375	0.507	0.372	0.182	0.119
22	0.517	0.374	0.518	0.369	0.189	0.123
23	0.513	0.359	0.510	0.347	0.169	0.105
24	0.536	0.376	0.535	0.372	0.188	0.113
25	0.539	-[a]	0.540	-	0.170	-
26	0.518	0.000[b]	0.530	0.000	0.170	0.000
27	0.537	-	0.538	-	0.189	-
28	0.521	-	0.518	-	0.186	-
29	0.531	0.401	0.531	0.392	0.185	0.123
Overall	0.523	0.371	0.523	0.367	0.185	0.118

* Proportion of marker pairs that X^2 exceeded 3.84, the 5% critical value of $\chi^2_{(1)}$.

[a] The chromosome length was short than 50 cM and thus no marker pairs with distance exceeded 50 cM.

[b] Chromosome 26 was 52 cM long but no marker pair with a distance >50 cM exceeded 3.84.

Table 3.2 The estimated powers* of test statistics for the trigenic and quadrigenic disequilibria for marker pairs with a distance ≤50 cM and >50 cM on 29 autosomes in the Kinsella composite beef population.

	$X^2_{D_{ABB}}$		$X^2_{D_{AAB}}$		$X^2_{\Delta_{AABB}}$	
BTA	≤50 cM	>50 cM	≤50 cM	>50 cM	≤50 cM	>50 cM
1	0.124	0.122	0.120	0.108	0.076	0.052
2	0.140	0.130	0.140	0.134	0.077	0.054
3	0.119	0.113	0.120	0.115	0.076	0.053
4	0.119	0.119	0.117	0.115	0.082	0.055
5	0.131	0.126	0.133	0.124	0.079	0.053
6	0.124	0.120	0.121	0.108	0.079	0.052
7	0.114	0.114	0.115	0.121	0.078	0.056
8	0.122	0.119	0.119	0.114	0.081	0.056
9	0.128	0.119	0.135	0.140	0.072	0.053
10	0.129	0.126	0.126	0.118	0.082	0.055
11	0.132	0.126	0.131	0.129	0.078	0.053
12	0.128	0.120	0.125	0.118	0.076	0.053
13	0.122	0.127	0.122	0.117	0.082	0.055
14	0.120	0.113	0.126	0.122	0.084	0.053
15	0.128	0.137	0.127	0.138	0.075	0.056
16	0.135	0.097	0.135	0.120	0.073	0.053
17	0.103	0.111	0.112	0.134	0.076	0.052
18	0.145	0.165	0.131	0.138	0.078	0.055
19	0.126	0.178	0.117	0.116	0.078	0.053
20	0.134	0.126	0.136	0.133	0.077	0.055
21	0.127	0.116	0.117	0.092	0.072	0.052
22	0.114	0.081	0.127	0.148	0.079	0.055
23	0.141	0.184	0.129	0.117	0.08	0.059
24	0.139	0.116	0.145	0.137	0.083	0.055
25	0.145	-[a]	0.142	-	0.092	-
26	0.115	0.000[b]	0.110	0.000	0.077	0.000
27	0.12	-	0.126	-	0.078	-
28	0.122	-	0.138	-	0.077	-
29	0.141	0.121	0.125	0.107	0.077	0.066
Overall	0.130	0.125	0.130	0.123	0.080	0.055

* Proportion of marker pairs that X^2 exceeded 3.84, the 5% critical value of $\chi^2_{(1)}$.

[a] The chromosome length was short than 50 cM and thus no marker pairs with distance exceeded 50 cM.

[b] Chromosome 26 was 52 cM long but no marker pair with a distance >50 cM exceeded 3.84.

Table 3.3 The estimated digenic (gametic and composite), trigenic and quadrigenic disequilibria averaged over syntenic SNP pairs on 29 autosomes in the Kinsella composite beef population.

BTA	$\phi^2_{D_{AB}}$ [a]	$\phi^2_{\Delta_{AB}}$ [a]	$\phi^2_{\mathrm{D}_{\mathrm{ABB}}}$ [a]	$\phi^2_{\mathrm{D}_{\mathrm{AAB}}}$ [a]	$\phi^2_{\Delta_{\mathrm{AABB}}}$ [a]
1	0.0082	0.0077	0.0035	0.0033	0.0012
2	0.0088	0.0082	0.0036	0.0036	0.0012
3	0.0093	0.0085	0.0030	0.0032	0.0013
4	0.0099	0.0091	0.0034	0.0032	0.0013
5	0.0096	0.0089	0.0037	0.0037	0.0013
6	0.0095	0.0086	0.0031	0.0032	0.0013
7	0.0098	0.0089	0.0031	0.0032	0.0013
8	0.0098	0.0090	0.0033	0.0033	0.0013
9	0.0092	0.0085	0.0036	0.0036	0.0012
10	0.0101	0.0094	0.0034	0.0031	0.0013
11	0.0098	0.0090	0.0036	0.0039	0.0013
12	0.0099	0.0091	0.0036	0.0034	0.0013
13	0.0109	0.0099	0.0034	0.0032	0.0014
14	0.0111	0.0101	0.0031	0.0035	0.0014
15	0.0103	0.0096	0.0033	0.0038	0.0013
16	0.0103	0.0095	0.0041	0.0040	0.0013
17	0.0103	0.0096	0.0032	0.0031	0.0013
18	0.0111	0.0103	0.0039	0.0034	0.0014
19	0.0108	0.0100	0.0034	0.0032	0.0014
20	0.0108	0.0100	0.0037	0.0037	0.0013
21	0.0103	0.0095	0.0035	0.0030	0.0013
22	0.0113	0.0104	0.0028	0.0032	0.0014
23	0.0111	0.0102	0.0045	0.0036	0.0014
24	0.0123	0.0112	0.0039	0.0041	0.0014
25	0.0126	0.0117	0.0044	0.0037	0.0016
26	0.0118	0.0111	0.0032	0.0031	0.0014
27	0.0124	0.0114	0.0035	0.0035	0.0014
28	0.0113	0.0104	0.0033	0.0036	0.0014
29	0.0115	0.0106	0.0041	0.0035	0.0014
Ave.	0.0105	0.0097	0.0035	0.0034	0.0013

[a] The generalized measures of squared correlation for digenic, trigenic and quadrigenic disequilibria: $\phi^2_{\mathrm{D}_{\mathrm{AB}}} = X^2_{\mathrm{D}_{\mathrm{AB}}}/n$, $\phi^2_{\Delta_{\mathrm{AB}}} = X^2_{\Delta_{\mathrm{AB}}}/n$, $\phi^2_{\mathrm{D}_{\mathrm{ABB}}} = X^2_{\mathrm{D}_{\mathrm{ABB}}}/n$, $\phi^2_{\mathrm{D}_{\mathrm{AAB}}} = X^2_{\mathrm{D}_{\mathrm{AAB}}}/n$, $\phi^2_{\Delta_{\mathrm{AABB}}} = X^2_{\Delta_{\mathrm{AABB}}}/n$, where n is the number of animals at individual SNP pairs.

Table 3.4 The proportion of syntenic SNP pairs with the generalized measures of square correlation estimated for gametic, composite, trigenic and quadrigenic disequilibria that exceeded 0.2 on 29 autosomes in the Kinsella composite beef population.

BTA	$\phi^2_{D_{AB}}$ [a]	$\phi^2_{\Delta_{AB}}$ [a]	$\phi^2_{\mathrm{D_{ABB}}}$ [a]	$\phi^2_{\mathrm{D_{AAB}}}$ [a]	$\phi^2_{\Delta_{\mathrm{AABB}}}$ [a]
1	0.1366	0.1440	0.0711	0.0666	0.0050
2	0.1345	0.1279	0.0626	0.0619	0.0061
3	0.1971	0.1778	0.0487	0.0559	0.0068
4	0.2368	0.2326	0.0676	0.0672	0.0109
5	0.2212	0.2132	0.0745	0.0775	0.0077
6	0.1934	0.1690	0.0615	0.0664	0.0074
7	0.2418	0.2298	0.0650	0.0617	0.0104
8	0.2230	0.2137	0.0559	0.0622	0.0105
9	0.1744	0.1692	0.0681	0.0646	0.0081
10	0.2147	0.2162	0.0623	0.0476	0.0074
11	0.2066	0.1984	0.0770	0.0816	0.0084
12	0.1782	0.1768	0.0744	0.0669	0.0081
13	0.2926	0.2607	0.0620	0.0584	0.0113
14	0.2836	0.2851	0.0645	0.0750	0.0080
15	0.2052	0.1972	0.0587	0.0804	0.0056
16	0.2507	0.2466	0.0834	0.0850	0.0130
17	0.2079	0.2109	0.0738	0.0600	0.0075
18	0.2494	0.2349	0.0760	0.0596	0.0094
19	0.2417	0.2262	0.0610	0.0655	0.0102
20	0.2130	0.2055	0.0759	0.0681	0.0073
21	0.2108	0.1957	0.0726	0.0589	0.0074
22	0.2943	0.3005	0.0462	0.0475	0.0085
23	0.2408	0.2230	0.1060	0.0704	0.0131
24	0.3267	0.2822	0.0795	0.0911	0.0099
25	0.3060	0.3160	0.1130	0.0848	0.0155
26	0.2806	0.2806	0.0623	0.0635	0.0108
27	0.2947	0.2962	0.0746	0.0725	0.0091
28	0.2213	0.2096	0.0497	0.0569	0.0075
29	0.2396	0.2358	0.0849	0.0756	0.0103
Ave.	0.2316	0.2233	0.0701	0.0674	0.0090

[a] The generalized measures of squared correlation for digenic, trigenic and quadrigenic disequilibria: $\phi^2_{\mathrm{D_{AB}}} = X^2_{\mathrm{D_{AB}}}/n$, $\phi^2_{\Delta_{\mathrm{AB}}} = X^2_{\Delta_{\mathrm{AB}}}/n$, $\phi^2_{\mathrm{D_{ABB}}} = X^2_{\mathrm{D_{ABB}}}/n$, $\phi^2_{\mathrm{D_{AAB}}} = X^2_{\mathrm{D_{AAB}}}/n$, $\phi^2_{\Delta_{\mathrm{AABB}}} = X^2_{\Delta_{\mathrm{AABB}}}/n$, where n is the number of animals at individual SNP pairs.

Table 3.5 Descriptive statistics for generalized measures of squared correlation for digenic (gametic and composite), trigenic, quadrigenic and zygotic disequilibria averaged over all syntenic (intra-chromosome) SNP pairs and over all non-syntenic (inter-chromosome) SNP pairs across the composite beef genome.

LD	Intra-chromosome					Inter-chromosome				
	# of pairs	Strength		Power*		# of pairs	Strength		Power*	
		mean	range	mean	range		mean	range	mean	range
$\phi^2_{D_{AB}}(r^2_{\mathrm{D_{AB}}})$[a]	36131611	0.0105	0.0082-0.0126	0.4946	0.4432-0.5385	83686490	0.0044	0.0041-0.0047	0.3526	0.3338-0.3740
$\phi^2_{\Delta_{AB}}$ [a]	36131636	0.0097	0.0077-0.0117	0.4945	0.4449-0.5400	893686490	0.0044	0.0041-0.0047	0.3526	0.3338-0.3740
$\phi^2_{\mathrm{D_{ABB}}}$ [a]	11158132	0.0035	0.0028-0.0045	0.1264	0.1043-0.1459	272879855	0.0033	0.0027-0.0042	0.1198	0.1025-0.1404
$\phi^2_{\mathrm{D_{AAB}}}$ [a]	11224541	0.0034	0.0030-0.0041	0.1257	0.1098-0.1450	273790442	0.0034	0.0027-0.0042	0.1222	0.1030-0.1428
$\phi^2_{\Delta_{\mathrm{AABB}}}$ [a]	36128429	0.0013	0.0012-0.0016	0.0740	0.0643-0.0921	893685393	0.0010	0.0010-0.0011	0.0540	0.0513-0.0579
$\phi^2_{\omega_{AB}}(r^2_{\omega_{\mathrm{AB}}})$ [a]	36071428	0.0029	0.0025-0.0034	0.1726	0.1577-0.1887	893012960	0.0016	0.0015-0.0016	0.1145	0.0105-0.1205

* Proportion of SNP pairs that X^2 exceeded 3.84, the 5% critical value of $\chi^2_{(1)}$.

[a] The generalized measures of squared correlation for digenic, trigenic and quadrigenic disequilibria: $\phi^2_{\mathrm{D_{AB}}} = X^2_{\mathrm{D_{AB}}}/n$, $\phi^2_{\Delta_{\mathrm{AB}}} = X^2_{\Delta_{\mathrm{AB}}}/n$, $\phi^2_{\mathrm{D_{ABB}}} = X^2_{\mathrm{D_{ABB}}}/n$, $\phi^2_{\mathrm{D_{AAB}}} = X^2_{\mathrm{D_{AAB}}}/n$, $\phi^2_{\Delta_{\mathrm{AABB}}} = X^2_{\Delta_{\mathrm{AABB}}}/n$, where n is the number of animals at individual SNP pairs.

Table 3.6 The mean, minimum and maximum of powers* of the test statistics for digenic, trigenic and quadrigenic disequilibria obtained for nine combinations of minor allele frequency (MAF) categories at each of the two loci (MAF_A and MAF_B) for all syntenic (intra-chromosome) SNP pairs and for all non-syntenic (inter-chromosome) SNP pairs in the Kinsella composite beef population.

MAF_A	MAF_B		Intra-chromosome					Inter-chromosome				
			$X^2_{\Delta_{AB}}$	$X^2_{D_{ABB}}$	$X^2_{D_{AAB}}$	$X^2_{\Delta_{AABB}}$	$X^2_{\omega_{AB}}$	$X^2_{\Delta_{AB}}$	$X^2_{D_{ABB}}$	$X^2_{D_{AAB}}$	$X^2_{\Delta_{AABB}}$	$X^2_{\omega_{AB}}$
		Mean	0.389	0.031	0.032	0.018	0.290	0.282	0.014	0.014	0.022	0.212
< 0.1	< 0.1	Min	0.342	0.016	0.008	0.013	0.251	0.243	0.002	0.004	0.017	0.183
		Max	0.444	0.049	0.084	0.024	0.328	0.317	0.037	0.052	0.027	0.241
		Mean	0.464	0.000	0.056	0.029	0.308	0.323	0.000	0.035	0.029	0.202
	0.1-0.3	Min	0.396	0.000	0.034	0.024	0.261	0.286	0.000	0.022	0.026	0.180
		Max	0.544	0.002	0.104	0.036	0.364	0.361	0.000	0.058	0.034	0.225
		Mean	0.477	0.000	0.187	0.051	0.136	0.323	0.000	0.183	0.041	0.094
	0.3-0.5	Min	0.420	0.000	0.157	0.039	0.120	0.286	0.000	0.154	0.034	0.086
		Max	0.565	0.000	0.216	0.072	0.160	0.361	0.000	0.219	0.049	0.104
		Mean	0.463	0.053	0.000	0.030	0.310	0.320	0.034	0.000	0.030	0.204
0.1-0.3	< 0.1	Min	0.403	0.042	0.000	0.025	0.271	0.283	0.022	0.000	0.025	0.181
		Max	0.514	0.077	0.001	0.035	0.356	0.354	0.057	0.000	0.034	0.232
		Mean	0.501	0.004	0.004	0.071	0.275	0.357	0.002	0.002	0.056	0.171
	0.1-0.3	Min	0.448	0.003	0.003	0.065	0.236	0.336	0.001	0.001	0.053	0.161
		Max	0.544	0.006	0.008	0.085	0.318	0.377	0.003	0.003	0.059	0.188
		Mean	0.505	0.003	0.201	0.090	0.127	0.363	0.002	0.194	0.062	0.086
	0.3-0.5	Min	0.464	0.002	0.182	0.082	0.117	0.340	0.002	0.168	0.060	0.081
		Max	0.552	0.004	0.221	0.108	0.139	0.389	0.003	0.222	0.065	0.091
		Mean	0.476	0.188	0.000	0.051	0.137	0.326	0.182	0.000	0.042	0.095
0.3-0.5	< 0.1	Min	0.424	0.158	0.000	0.042	0.125	0.288	0.152	0.000	0.034	0.083
		Max	0.537	0.209	0.000	0.062	0.150	0.365	0.215	0.000	0.049	0.105
		Mean	0.506	0.201	0.003	0.090	0.128	0.365	0.194	0.002	0.062	0.086
	0.1-0.3	Min	0.460	0.170	0.002	0.080	0.116	0.347	0.170	0.002	0.059	0.081
		Max	0.548	0.223	0.005	0.101	0.140	0.388	0.221	0.004	0.065	0.092
		Mean	0.512	0.200	0.199	0.095	0.101	0.373	0.193	0.193	0.063	0.067
	0.3-0.5	Min	0.479	0.158	0.178	0.083	0.089	0.350	0.167	0.167	0.061	0.065
		Max	0.559	0.226	0.220	0.107	0.113	0.397	0.220	0.221	0.066	0.071

* Proportion of marker pairs that X^2 exceeded 3.84, the 5% critical value of $\chi^2_{(1)}$.

Figure

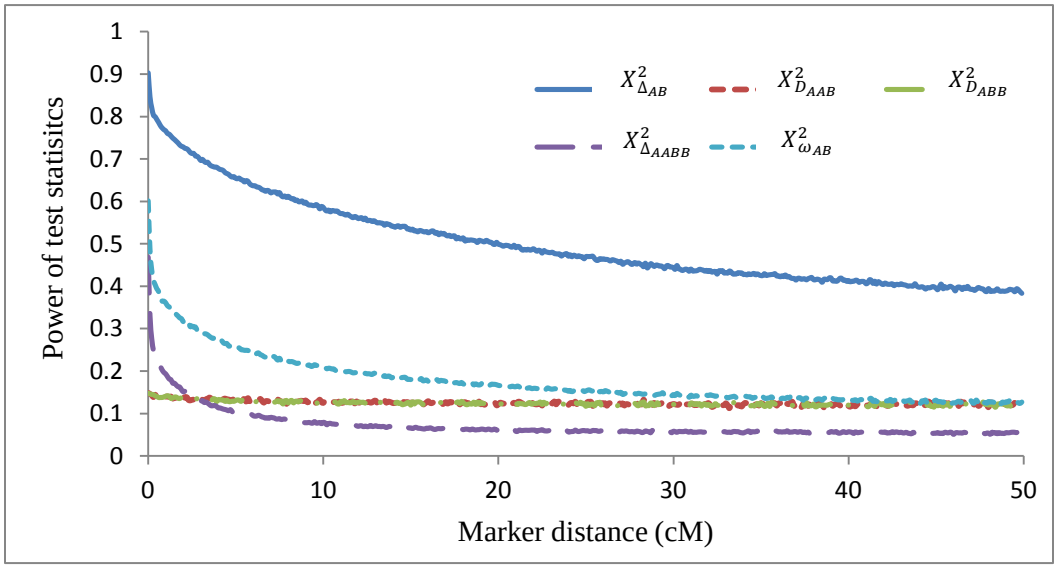

Figure 3.1 The relationship between the estimated powers of chi-square tests for zygotic LD and its individual genic components and marker distance for SNP markers that were apart within 50cM on 29 autosomes in the Kinsella beef composite population. Note that the powers of the tests for the two trigenic components were very similar over the whole range of marker distance as indicated by the lines for the two disequilibria being overlapped to each other.

Appendix 3.1 Definitions and notations of gene and genotypic frequencies and zygotic LD

Yang (2004) constructed a table for joint frequencies of the nine genotypes for two loci, each with two alleles, *A* and *a* at locus *A* and *B* and *b* at locus *B*. In the table, these genotypic frequencies are expressed in terms of their single-locus genotypic frequencies and zygotic associations. The Table is reproduced here for easier reference to definitions and notations of gene and genotypic frequencies and zygotic LD.

	Locus *B*			
Locus *A*	*BB*	*Bb*	*bb*	Total
AA	$P_{AB}^{AB} = P_{A\cdot}^{A\cdot} P_{\cdot B}^{\cdot B} + \omega_{AB}^{AB}$	$P_{Ab}^{AB} = P_{A\cdot}^{A\cdot} P_{\cdot b}^{\cdot B} + \omega_{Ab}^{AB}$	$P_{Ab}^{Ab} = P_{A\cdot}^{A\cdot} P_{\cdot b}^{\cdot b} + \omega_{Ab}^{Ab}$	$P_{A\cdot}^{A\cdot} = p_A^2 + D_A$
Aa	$P_{aB}^{AB} = P_{a\cdot}^{A\cdot} P_{\cdot B}^{\cdot B} + \omega_{aB}^{AB}$	$P_{ab}^{AB} + P_{Ab}^{aB} = P_{a\cdot}^{A\cdot} P_{\cdot b}^{\cdot B} + \omega_{ab}^{AB}$	$P_{ab}^{Ab} = P_{a\cdot}^{A\cdot} P_{\cdot b}^{\cdot b} + \omega_{ab}^{Ab}$	$P_{a\cdot}^{A\cdot} = 2p_A p_a - 2D_A$
Aa	$P_{aB}^{aB} = P_{Aa}^{a\cdot} P_{\cdot B}^{\cdot B} + \omega_{aB}^{aB}$	$P_{ab}^{aB} = P_{a\cdot}^{a\cdot} P_{\cdot b}^{\cdot B} + \omega_{ab}^{aB}$	$P_{ab}^{ab} = P_{a\cdot}^{a\cdot} P_{\cdot b}^{\cdot b} + \omega_{ab}^{ab}$	$P_{a\cdot}^{a\cdot} = p_a^2 + D_A$
Total	$P_{\cdot B}^{\cdot B} = p_B^2 + D_B$	$P_{\cdot b}^{\cdot B} = 2p_B p_b - 2D_B$	$P_{\cdot b}^{\cdot b} = p_b^2 + D_B$	1

The possible values of zygotic associations are constrained by the single-locus frequencies such that only four of the nine zygotic associations need to be defined and the remaining five are entirely expressed in terms of the four defined zygotic associations. For example, if the zygotic associations for four double homozygotes, ω_{AB}^{AB}, ω_{Ab}^{Ab}, ω_{aB}^{aB} and ω_{ab}^{ab}, are defined, then the zygotic associations for the remaining five genotypes are expressed as follows: $\omega_{Ab}^{AB} = -(\omega_{AB}^{AB} + \omega_{Ab}^{Ab})$, $\omega_{aB}^{AB} = -(\omega_{AB}^{AB} + \omega_{aB}^{aB})$, $\omega_{ab}^{aB} = -(\omega_{aB}^{aB} + \omega_{ab}^{ab})$, $\omega_{ab}^{Ab} = -(\omega_{Ab}^{Ab} + \omega_{ab}^{ab})$, and $\omega_{ab}^{AB} = \omega_{AB}^{AB} + \omega_{Ab}^{Ab} + \omega_{aB}^{aB} + \omega_{ab}^{ab}$.

Appendix 3.2 Sampling variances of individual genic disequilibria in zygotic LD

Weir and Cockerham (1989) provided the formulas for the large-sample variances of different components of zygotic LD. These formulas are reproduced here, with minor modifications, for easier reference. Definitions and notations of different genic disequilibria were detailed in the Materials and Methods section.

Gametic disequilibrium $(\widetilde{D}_{AB})$

$$\mathrm{Var}(\widetilde{D}_{AB}) = [\pi_A\pi_B + \tau_A\tau_B D_{AB} + D_A D_B - D_{AB}^2 + D_{A/B}^2 + D_{AB}^{AB}]/2n.$$

Non-gametic disequilibrium $(\widetilde{D}_{A/B})$

$$\mathrm{Var}(\widetilde{D}_{A/B}) = [\pi_A\pi_B + \tau_A\tau_B D_{A/B} + D_A D_B + D_{AB}^2 - D_{A/B}^2 + D_{AB}^{AB}]/2n.$$

Composite disequilibrium $(\tilde{\Delta}_{AB})$

$$\mathrm{Var}(\tilde{\Delta}_{AB}) = [(\pi_A + D_A)(\pi_B + D_B) + \tau_A\tau_B\Delta_{AB}/2 + \tau_A D_{ABB} + \tau_B D_{AAB} + \Delta_{AABB}]/n.$$

Trigenic disequilibria $(\widetilde{D}_{AAB}$ and $\widetilde{D}_{ABB})$

$$\begin{aligned}\mathrm{Var}(\widetilde{D}_{AAB}) = \{&(\pi_A^2 + \tau_A^2 D_A - D_A^2)(\pi_B + D_B) + \pi_A\tau_A\tau_B\Delta_{AB} \\ &+(1 - 5\pi_A + D_A)\Delta_{AB}^2 + 2\pi_A\tau_A D_{ABB} \\ &+(\tau_A^2\tau_B - 2D_A\tau_B - 4\tau_A\Delta_{AB})D_{AAB} - 2D_{AAB}^2 \\ &+(\tau_A^2 - 2D_A)\left(D_{AB}^{AB} - 2D_{AB}D_{A/B}\right)\}/2n.\end{aligned}$$

$$\mathrm{Var}\left(\widetilde{D}_{ABB}\right) = \{(\pi_B^2 + \tau_B^2 D_B - D_B^2)(\pi_A + D_A) + \pi_B \tau_B \tau_A \Delta_{AB}$$
$$+(1 - 5\pi_B + D_B)\Delta_{AB}^2 + 2\pi_B \tau_B D_{AAB}$$
$$+(\tau_B^2 \tau_A - 2D_B \tau_A - 4\tau_B \Delta_{AB}) D_{ABB} - 2D_{ABB}^2$$
$$+(\tau_B^2 - 2D_B)\left(D_{AB}^{AB} - 2D_{AB} D_{A/B}\right)\}/2n.$$

Quadrigenic disequilibrium $\left(\tilde{\Delta}_{AABB}\right)$

$$\mathrm{Var}\left(\tilde{\Delta}_{AABB}\right) = \{(\pi_A^2 + \tau_A^2 D_A - D_A^2)(\pi_B^2 + \tau_B^2 D_B - D_B^2)$$
$$+2\tau_A \tau_B (\pi_A \pi_B - 4D_A D_B)\Delta_{AB}$$
$$+(\tau_A^2 \tau_B^2 - 4\tau_B^2 D_A - 4\tau_A^2 D_B + 4D_A D_B + 2D_A + 2D_B)\Delta_{AB}^2$$
$$-6\tau_A \tau_B \Delta_{AB}^3 + 3\Delta_{AB}^4$$
$$+[2\pi_B \tau_B (\tau_A^2 - 2D_A) + 4\Delta_{AB}(2\tau_A D_B - 2\tau_A \pi_B +$$
$\tau_B \Delta_{AB})] D_{AAB}$
$$+[2\pi_A \tau_A (\tau_B^2 - 2D_B) + 4\Delta_{AB}(2\tau_B D_A - 2\tau_B \pi_A + \tau_A \Delta_{AB})] D_{ABB}$$
$$+2(3D_B - \pi_B) D_{AAB}^2 + 2(3D_A - \pi_A) D_{ABB}^2 + 20\Delta_{AB} D_{AAB} D_{ABB}$$
$$+[(\tau_A^2 - 2D_A)(\tau_B^2 - 2D_B) - 8\tau_A \tau_B \Delta_{AB} + 6\Delta_{AB}^2$$
$$-4\tau_A D_{ABB} - 4\tau_B D_{AAB}]\Delta_{AABB} - \Delta_{AABB}^2\}/n.$$

Appendix 3.3 Counts of out-of-bound estimates high-order genic disequilibria

The proportion of syntenic SNP pairs with out-of-bound estimates of generalized measures of squared correlation for trigenic and quadrigenic disequilibria in the Kinsella composite population.

BTA	$\phi^2_{D_{ABB}}$ [a]		$\phi^2_{D_{AAB}}$ [a]		$\phi^2_{\Delta_{AABB}}$ [a]	
	< 0	>1	< 0	>1	< 0	>1
1	70.18	0.019	70.34	0.017	0	0.001
2	70.47	0.018	69.16	0.016	0	0.001
3	69.12	0.012	68.90	0.013	0	0.001
4	68.51	0.018	67.22	0.016	0	0.001
5	70.75	0.016	69.00	0.018	0	0.002
6	66.62	0.015	66.81	0.016	0	0.001
7	66.68	0.019	68.32	0.015	0	0.002
8	70.32	0.012	69.64	0.016	0	0.001
9	70.42	0.017	70.16	0.014	0	0.001
10	70.32	0.013	70.01	0.011	0	0.002
11	69.00	0.020	69.38	0.021	0	0.001
12	67.45	0.018	67.96	0.018	0	0.001
13	69.69	0.014	67.33	0.016	0	0.003
14	64.77	0.014	66.79	0.018	0	0.001
15	67.48	0.018	69.55	0.019	0	0.002
16	73.03	0.020	71.29	0.022	0	0.001
17	69.73	0.019	68.11	0.017	0	0.001
18	68.56	0.023	69.22	0.016	0	0.001
19	68.02	0.014	66.28	0.016	0	0.001
20	68.28	0.017	69.99	0.019	0	0.001
21	66.70	0.019	66.55	0.014	0	0.002
22	70.78	0.013	73.45	0.014	0	0.002
23	67.99	0.028	70.61	0.014	0	0.000
24	68.40	0.020	67.50	0.020	0	0.002
25	66.58	0.027	64.77	0.020	0	0.002
26	68.87	0.014	68.61	0.011	0	0.002
27	69.31	0.017	69.51	0.022	0	0.002
28	72.49	0.013	73.72	0.019	0	0.001
29	70.16	0.019	68.31	0.018	0	0.002
Overall	68.99	0.020	68.91	0.020	0	0.001

[a] The generalized measures of squared correlation for trigenic and quadrigenic disequilibria: $\phi^2_{\mathrm{D_{ABB}}} = X^2_{\mathrm{D_{ABB}}}/n$, $\phi^2_{\mathrm{D_{AAB}}} = X^2_{\mathrm{D_{AAB}}}/n$, $\phi^2_{\Delta_{\mathrm{AABB}}} = X^2_{\Delta_{\mathrm{AABB}}}/n$, where n is the number of animals at individual SNP pairs.Chapter 4. General Discussion and Conclusions

Chapter 4. General Discussion and Conclusions

4.1 Introduction

The recent advancement in molecular biology has enabled the rapid development of single nucleotide polymorphism (SNP) genotyping technology, thereby making the genotyping of cheap and abundant SNP markers possible in many livestock species [1]. Commercial SNP chips (e.g., Illumina BovineSNP50 beadchip) are now available for cattle. Such large panels of SNPs have allowed animal geneticists to search for the quantitative trait nucleotides (QTNs) underlying variation in complex traits through the use of genome-wide association studies (GWAS) or to predict animal's performance through genomic selection [2, 3]. The success of GWAS and genomic selection depends crucially on the extent of gametic LD between SNPs and QTNs on chromosomes. It is shown (e.g., [3]) that strong gametic LD is found at long distances (> 1 cM) in domestic animals (e.g., cattle and dog) but not in human. The strong gametic LD in cattle is likely due to drastic reduction of population sizes that occurs during animal domestication and breed formation.

With the availability of high-density SNPs, many studies have also conducted population genetic analysis of gametic LD in cattle and other animal species (e.g., [4-8]). However, the gametic LD cannot be calculated directly for unphased SNP markers because the gametic phase of animals that are heterozygous at two or more loci cannot be directly observed or specified. Thus, these studies have followed the classic approach of Hill [9] to estimating gametic LD for unphased data, but such estimation was carried out under the HWE assumption. For pure breed populations as often in the above studies, the HWE assumption may be reasonable.

In this study, we set out to study the genome-wide extent and patterns of LD in the Kinsella composite beef population. Since this population arose from repeated mixing of multiple breeds and selection for growth and cow reproduction,

it might have stayed constantly in the HWD condition. Furthermore, in a general diploid nonequilibrium population, the assessment of gametic LD alone would miss an opportunity to explore other types of non-allelic associations within and between uniting gametes that provide additional insightful information about genomic structure.

4.2 Summary of results and significance

Our study is the first major genome-wide survey of genotypic disequilibrium in the Kinsella composite beef population where the HWE assumption is obviously unjustifiable. We first compared and contrasted gametic LD, composite LD, and zygotic LD based on a 50K SNP data set that was obtained from genotyping the beef composite population (Chapter 2). We then determined the extent and significance of individual digenic, trigenic and quadrigenic disequilibria as components of the overall zygotic LD (Chapter 3).

In Chapter 2, a genome-wide comparison was made between gametic and zygotic disequilibrium measures for the Kinsella composite beef population arising from mixing of multiple breeds that was under selection for growth and cow reproduction. Single-locus heterozygosities fell approximately in the range of 0.3 to 0.4 over different chromosomes (Table 2.2). There was little change in these heterozygosity estimates when only those markers with significant HWD were kept. Two-locus heterozygosities had lower values but displayed similar patterns (Table 2.4) and they were almost perfectly predicted by the products of single-locus heterozygosities. In contrast, two-locus homozygosities were significantly higher when only those marker pairs with significant zygotic LD were kept than when all pairs were included, but again the two-locus homozygosities were almost predicted by the products of single-locus homozygosities (Table 2.3). The levels of significant gametic and composite LD (Tables 2.5 and 2.6) were much higher than those of significant zygotic LD (Table 2.7). Similar patterns between gametic and zygotic LD were observed when only those marker pairs with extremely high gametic LD values (r^2_{GLD} >

0.25) were retained. Such tightly linked marker pairs were chosen because a similar stringent criterion was used for detecting marker effects in many livestock genomic selection programs [3]. No such comparison has been made in any other studies with domestic animal species.

In Chapter 3, we examined the extent and significance of high-order genic disequilibria between three or four genes at pairs of loci in the Kinsella composite beef population. This population was chosen because continued crossbreeding and selection for growth and cow reproduction would allow the population to constantly stay in a HWD condition, thereby providing an excellent opportunity for uncovering the high-order genic disequilibria. The results showed that the trigenic and quadrigenic disequilibria were generally two- to three-fold smaller than the usual digenic disequilibria (gametic or composite LD) (Table 3.4). Correspondingly, there was less power of testing for these high-order genic disequilibria than for the digenic disequilibria (Tables 3.1 and 3.2). The powers decreased with the distance between markers though the decay is more obvious for the digenic disequilibria than for high-order disequilibria (Figure 3.1). We also provided an empirical evaluation of the effect of allele frequency on different genic disequilibria. The powers of chi-square tests for digenic, trigenic and quadrigenic disequilibria increased with the increasing gene frequency at both loci, but those for zygotic LD decreased with the increasing gene frequency.

Instead of focusing just on gametic LD as in other LD assessments reported recently for cattle and other animal species, our study went a step further to assess the genome-wide extent and patterns of zygotic LD and its components including high-order trigenic and quadrigenic disequilibria. Such a comparative assessment of genome-wide gametic and zygotic LD allowed for inference about selection, crossbreeding and other historical breeding events in a composite cattle population. The only other attempt of this kind was made by Mitton [10] who provided a similar comparison between gametic and zygotic LD for natural populations of some marine animals and conifer plants, but based on a very limited number of isozyme loci and thus his results would be less conclusive. The

only other study on high-order genic disequilibria in a dog population by Liu at el. [11] revealed more power of detecting genic disequilibria, but again the study by Liu et al. [11] would be less conclusive because it was based on only < 250 markers typed for 148 dogs. We anticipate that there will be a growing emphasis on comparative assessment of gametic and zygotic LD as SNP markers are now ubiquitously available in cattle and other farmed animals.

4.3 Implications for genetic improvement in cattle

This study has several important implications for genetic improvement in beef composites. First, it may be advisable that the genomic regions marked by those SNP loci with significant gametic and zygotic LD are targeted for QTL identification or candidate gene search because strong LD between SNPs may arise from the hitchhiking effect of neighbouring genes under selection for growth or cow reproduction. These LD 'hot spots' may form a basis for a new strategy for gene discovery particularly with genes of minor effects which would otherwise be difficult to be detected by the traditional QTL mapping methods.

Second, the success of fine-scale QTL mapping and genomic selection depends largely on the optimal balance between the level of LD and SNP density. For example, Meuwissen et al. [2] suggested that the required level of gametic LD (r^2_{GLD}) should be >0.2 in order for genomic selection to be successful (achieving a prediction accuracy of $\geq$0.85). On the other hand, Ardlie et al. [12] suggested the use of r^2_{GLD} > 1/3 for GWAS in human. When the threshold of useful gametic LD is set to be 0.25 to ensure the SNP spacing ~ 35 kb as suggested by Qanbari et al. [8], the GWAS approach would require the use of more than 75,000 SNPs per individual, assuming that all SNPs are informative (with a MAF $\geq$ 0.05). If all of the current 50K data set were usable, we would have to use less extreme frequencies (MAF $\geq$ 0.15) to achieve the improved accuracy and magnitude of estimated LD between pairs of SNP markers. However, the removal of many rare alleles may lose the opportunity to capture potentially novel causal mutations in the population. In addition, our zygotic LD estimates serve as a reminder that

non-random mating may be an important cause of LD even when the loci are very tightly linked.

Third, the main result from our study is the predominance of digenic disequilibria coupled with insignificant high-order disequilibria. This result supports the current intensive effort of using gametic LD for GWAS and genomic selection in cattle and other animal species. It has been demonstrated (e.g., [3]) that gametic LD in domestic animals may occur at a long distance (> 1 cM) due to the rapid and sudden decline of population size (bottleneck effect) during domestication and at breed formation, in comparison to the situation in human where there is no gametic LD at long distances. However, it has also been suspected that such LD may be in part due to false positive associations due to a mixture of multiple breeds or other causes. If the breeds can be identified through pedigree information, Goddard and Hayes [3] suggested the use of breeds as a covariate in the statistical model to minimize such false positive associations. However, animals in our composite beef population were long mixed and their breed identity is no longer available. So the suggestion by Goddard and Hayes [3] is not feasible for the composite beef population. Thus, our analysis of high-order trigenic and quadrigenic disequilibria may provide a quick, practical means of assessing the significance of the false positive associations in admixed populations.

4.4 Future directions

In this study, we quantified and determined the significance of gametic, composite and zygotic linkage disequilibria. The zygotic LD is a summary statistic to which the individual genic LD contributes, and gametic or composite LD is a predominant component of zygotic LD in comparison to trigenic and quadrigenic disequilibria. This is an indirect assessment in terms of the power of chi-square tests. A direct assessment is needed to assess the relative importance of individual genic disequilibria in terms of percentages of contributions by individual genic disequilibria to the overall zygotic LD. Furthermore, our

observed patterns of zygotic LD and its components are a net result of combined effects of several demographic factors including those that are known and unknown to us. We believe that a computer simulation study will be able to substantiate our findings here by singling out and then combining the effects of individual factors

Our study examined the genome-wide LD patterns for only one composite beef population as it would give a better opportunity to capture the effects of nonrandom mating and selection, etc. Our analytic procedures can certainly be applicable to other animal populations. In particular, it is desired to have a comparative assessment of genome-wide LD patterns between different pure breeds to infer about other genetic and demographic determinants. For example, a comparison between a dairy breed and a beef breed will be of great interest as it allows for inference about the effect of random drift and selection at and after breed formation. Archaeological and genetic data have suggested that the domestication and artificial selection of cattle occurred approximately 8,000 to 10,000 years ago in the Near East [13-15]. Prior to modern cattle breeding, there was no or little distinction between dairy and beef cattle, with the same animals often being used for both meat and milk production. Only in the past many decades have dairy cows been specialized and bred to produce large quantities of milk with little or no regard for their production of meat. The opposite is probably true for beef cattle. Thus, the comparison of LD patterns between dairy and beef cattle will help determine if divergent selection for milk in dairy cattle and meat yields in beef cattle has produced different patterns of multilocus structure.

When the data from multiple populations are available, both within- and among-population inferences about the cause of LD patterns can be made. Current theory and statistical methods are limited to gametic LD [16-18]. Ohta [16, 17] partitioned the variance of gametic LD into a set of D^2-statistics, analogous to *F*-statistics [19]. Various components of gametic LD can be defined in a mixed pool of multiple populations, but the expected values of these LD components are all zero under the assumption of selective neutrality. For this reason, Ohta [16, 17]

suggested the use of five different gametic LD variances: D_{IT}^2 is the variance of gametic LD in the total population; D_{ST}^2 is the variance of the expected allelic associations among populations; $D_{IS}^{'2}$ is the variance of the expected allelic associations within populations; D_{IS}^2 is the variance of gametic LD within individual populations and $D_{ST}^{'2}$ is the variance of gametic LD within the total population. The use of subscripts, *IS*, *ST* and *IT*, in these variances is analogous to that in the well-known *F*-statistics [19], but the meanings of Ohta's D^2-statistics may be quite different. For example, $D_{IS}^{'2}$ is actually a between-population component even though its subscript *IS* might have suggested that it is a within-population component. These D^2-statistics were the basis for our inference about the relative importance of natural selection, random drift and gene flow as causes of gametic LD [16, 17]. Specifically, under an equilibrium island model, if the ratios, D_{IS}^2 / D_{ST}^2 and $D_{ST}^{'2} / D_{IS}^{'2}$, are less than unity, then gene flow between populations is limited and the observed gametic LD are more likely due to random drift; if, on the other hand, these ratios are greater than unity, then epistatic natural selection is more likely responsible for the observed gametic LD. There is an obvious need for an extension of this theory for gametic LD to zygotic LD and other genic disequilibria.

4.5 Conclusions

This study is the first major genome-wide survey of all non-allelic associations between pairs of SNPs in cattle. Such analysis allows us to assess the relative importance of gametic LD vs. all other non-allelic genic LDs regardless of whether or not the HWE assumption holds. Our study shows that: (i) both gametic and zygotic LD are significant but the zygotic decays more rapidly than the gametic LD over the whole range of marker distances; (ii) digenic LD (gametic or composite LD) remains predominant in comparison to high-order trigenic and quadrigenic disequilibria; and (iii) the powers of chi-square tests for

digenic, trigenic and quadrigenic disequilibria increase with the increasing gene frequency at both loci, but those for zygotic LD decrease with the increasing gene frequency. These results support the current intensive focus on the use of high-density SNP markers for GWAS and genomic selection activities in the Kinsella composite beef population.

4.6 References

1. Hayes BJ, Chamberlain AJ, McPartlan H, MacLeod I, Sethuraman L, Goddard ME: **Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle.** *Genet Res Camb* 2007, **89**:215-220.
2. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
3. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programs.** *Nat Rev Genet* 2009, **10**:381-391.
4. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, Schnabel RD, Taylor JF, Raadsma HW: **Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel.** *BMC Genomics* 2008, **9**:187.
5. Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR: **Extent of linkage disequilibrium in Holstein cattle in North America.** *J Dairy Sci* 2008, **91**:2106-2117.
6. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ: **High-resolution haplotype block structure in the cattle genome.** *BMC Genetics* 2009, **10**:19.
7. Bohmanova J, Sargolzaei M, Schenkel FS: **Characteristics of linkage disequilibrium in North American Holsteins.** *BMC Genomics* 2010, **11**:421.
8. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H: **The pattern of linkage disequilbriom in German Holstein cattle.** *Anim Genet* 2010, **41**:346-356.
9. Hill WG: **Estimation of linkage disequilibrium in randomly mating populations.** *Heredity* 1974, **33**:229-239.
10. Mitton JB: *Selection in Natural Populations.* Oxford: Oxford University

Press; 1997.

11. Liu T, Todhunter RJ, Lu Q, Schoettinger L, Li H, Littell RC, Burton-Wurster N, Acland GM, Lust G, Wu R: **Modelling extent and distribution of zygotic disequilibrium: implications for a multigenerational canine pedigree**. *Genetics* 2006, **174**:439-453.
12. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3**:299-309.
13. Bradley DG, Loftus RT, Cunningham P, MacHugh DE: **Genetics and domestic cattle origins.** *Evolutionary Anthropology* 1998, **6**:79-86.
14. Diamond J: **Evolution, consequences and future of plant and animal domestication.** *Nature* 2002, **418**:700-707.
15. Bruford MW, Bradley DG, Luikart G: **DNA markers reveal the complexity of livestock domestication.** *Nat Rev Genet* 2003, **4**:900-910.
16. Ohta T: **Linkage disequilibrium due to random genetic drift in finite subdivided populations.** *Proc Natl Acad Sci USA* 1982, **79**:1940-1944.
17. Ohta T: **Linkage disequilibrium with the island model.** *Genetics* 1982, **101**: 139-155.
18. Yang RC, Yeh FC, Yeh TZ: **Multilocus Structure in *Pinus contorta* - *Pinus banksiana* Complex.** *Canadian Journal of Botany* 2007, **85**:774-784.
19. Wright S: **The interpretation of population structure by *F*-statistics with special regard to systems of mating.** *Evolution* 1965, **19**:395-420.