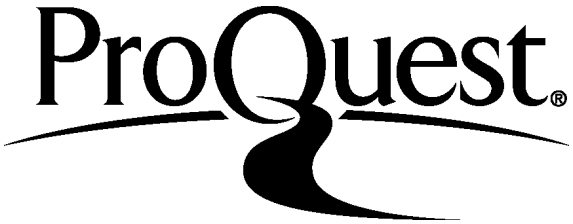
LINKAGE DISEQUILIBRIUM AND RECENT SELECTION SIGNATURES IN COMMERCIAL BROILERS

by

Weixuan Fu

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Animal Science

Spring 2016



LINKAGE DISEQUILIBRIUM AND RECENT SELECTION SIGNATURES IN COMMERCIAL BROILERS

by

Weixuan Fu

Approved: _____
Behnam Abasht, Ph.D.
Assistant Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Limin Kung, Jr., Ph.D.
Chair of the Department of Animal and Food Sciences

Approved: _____
Mark Rieger, Ph.D.
Dean of the College of Agriculture and Natural Resources

Approved: _____
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I deeply appreciate my advisor Dr. Behnam Abasht for allowing me to join his lab and guiding me in this project about commercial broilers. The door to his office was always open whenever I had a question or issue about my researches. I would like to thank Dr. Jack C.M. Dekkers for his helpful advices and guidance on this project. I would like to thank William R. Lee for sharing his experience and knowledge in chicken industry.

I would also like to thank all the people in Dr. Abasht's Lab for their support and assistance. Dr. Joseph Tomlinson helped me in English writing of the thesis. Zhu Zhuo was essential in helping me understanding the protocols for RNA library preparation; Michael Babak helped me gain knowledge about histology. Thank you all for your support!

Finally, I must express my very profound gratitude to my family for providing me with continuous encouragement and maximum support throughout years of study no matter what happened. This thesis would not have been accomplished without them. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Over the past 70 years, modern breeding programs have significantly improved traits of economical importance in commercial broilers. Due to recent strong selection and large reduction in effective population size in the breeding program, linkage disequilibrium (LD) in broiler population has changed over generations. Characterizing LD is of fundamental importance for implementation of both genome-wide association studies and genomic selection as well as identification of recent selection signatures.

Therefore, in this study, using a 60K single nucleotide polymorphism (SNP) panel, we firstly estimated LD and haplotype structure in crossbred broiler chickens and their component pure lines (one sire (male) and two dam (female) lines) and calculated the consistency of LD between these populations. The average level of LD (measured by r^2) between adjacent SNPs across the chicken autosomes studied here ranged from 0.34 to 0.40 in the pure lines but was only 0.24 in the crossbred populations, with 28.4% of adjacent SNP pairs having an r^2 higher than 0.3. Compared with the pure lines, the crossbred populations consistently showed a lower level of LD, smaller haploblock sizes and lower haplotype homozygosity on macro-, intermediate and micro-chromosomes. Furthermore, correlations of LD between markers at short distances (0 to 10 kb) were high between crossbred and pure lines (0.83 to 0.94). Our results suggest that using crossbred populations instead of pure lines can be advantageous for high-resolution QTL (quantitative trait loci) mapping in genome-wide association (GWA) studies and to achieve good persistence of accuracy

of genomic breeding values over generations in genomic selection. These results also provide useful information for the design and implementation of GWA studies and genomic selection using crossbred populations.

Secondly, with better understanding of the extent of LD in commercial broiler populations, we applied two methods, cross-population extended haplotype homozygosity (XP-EHH) and cross-population composite likelihood ratio (XP-CLR), to detect signatures of positive selection in five elite lines of commercial broiler chickens, including three broiler sire (male) lines and two broiler dam (female) lines. A total of 321 candidate selection regions were detected by both methods, 42 of which were shared by 2 or more purebred lines. Our results provide a genome-wide scan of recent selection signatures in five purebred lines of commercial broiler chickens. We found several candidate genes for recent selection in multiple lines, such as *SOX6* (*Sex Determining Region Y-Box 6*) and *cTR* (*Thyroid hormone receptor beta*), which may have been under selection due to their essential roles in growth, development and reproduction in chickens. Furthermore, our results suggest that in some candidate selection regions, the same or opposite alleles have been under recent selection in multiple lines. Most of the candidate genes in selection regions are novel, and as such they should be of great interest for future research into the genetic architecture of traits relevant to modern broiler breeding.

Chapter 1

BACKGROUND AND LITERATURE REVIEWS

In conventional breeding program, most economically important traits in livestock are selected in elite purebreds, which are mainly based on their phenotypes or estimated breeding values (EBVs). During the last century, this breeding strategy has proved to be very successful in commercial broilers. Intensive artificial selection has largely improved traits of economic importance for farmers, such as growth rate, feed efficiency and body composition. For example, when comparing a modern broiler chicken strain like Ross 308 with a broiler population that had not been subjected to artificial selection since 1957 [i.e., Athens-Canadian Random-bred Control (ACRBC) strain], Havenstein et al. (2003) found that the average body weight at 42 days of age increased from 539 g in 1957 for the ACRBC strain to 2,672 g in 2001 for the Ross 308 strain and that the feed conversion ratio decreased from 2.34 to 1.43 over the same time period [1]. The authors indicated that genetic selection contributed 85–90% of the improvement in growth rate over the past 45 years. These dramatic phenotypic changes imply that the underlying causal polymorphisms themselves have been under strong positive selection by choosing limited numbers of individuals with the best performance to reproduce next generation. Both reduction in the effective population size and strong selection in commercial broilers may largely influence the linkage disequilibrium (LD) in their populations. LD is a term in population genetics which descripts that alleles at two or more loci are non-randomly associated and are descended from a common ancestor [2] , and it is related to the fundamental

assumptions for many population-level genomics studies, such as GWAS (genome-wide association study), GS (genomic selection) and detection of recent selection signatures. Thus, gaining knowledge about the extent of LD in commercial broiler populations is important for application of these LD-based studies in commercial broilers.

1.1 Linkage Disequilibrium in Livestock Species

In an idealized population under neutral model, two factors affecting LD in population are mutation and recombination. Since the mutation rate is very low (~1.1×10^{-8} per site per generation [3]), recombination has more effect on decreasing LD by exchanging corresponding chromosomal segments from both parents during meiosis. Because the likelihood of recombination between two loci increases with genetic distance between them, the extent of LD usually decays with the distance. For these alleles in a short genetic distance (e.g. 0.01cM) with each others, the extent of LD among them can still decay over thousands of generations. For this reason, in molecular evolution, LD between a novel neutral mutation and alleles in a short genetic distance to the loci will take many generations to decay by the recombination events, and it also takes many generations for the mutation to reach a high allele frequency by random (genetic) drift [4].

However, unlike the situations in the neutral model, a beneficial mutation or polymorphism under strong positive selection in livestock species can reach a higher frequency in fewer generations, and the LD around it will be more extensive because the haplotype carrying the mutation/polymorphism becomes more frequent due to selection in population. Besides selection, another main factor influencing LD in livestock species is effective population size, which is the size of an idealized

population under neutral model which has the same rate of losing heterozygosity as the observed population [5]. In comparison with human populations, livestock species generally have much smaller effective population size due to animal breeding programs [2, 6, 7]. In an animal breeding program, only a limited number of elite individuals have the chance to reproduce thousands or millions of progenies in the following generation. For example, the effective population size in Holstein dairy cattle ranges from 100 to 150 based on estimations in previous studies [6, 8, 9] although more than 90% of dairy cattle on earth belong to this breed. Also, Qanbari et al. (2010) have reported that the recent effective population size for commercial chickens is around 70 [10]. Due to the low effective population size, genetic drift has a large effect on LD in livestock species. Genetic drift can randomly generate or increase LD between loci by making some allele combinations become more or less common during reproduction. Thus, livestock species are thought to have longer extent of LD within each breeds and larger persistence of LD between breeds within the same species [8, 11–14].

The advantages of the long-range extensive LD in most livestock species include: 1) it may benefit for some initial studies of identifying QTL (quantitative trait locus) regions controlling the traits of economical importance [15]. And, with LD markers for causal mutations identified in these QTL studies, the long-range extensive LD can subsequently contribute to some initial applications of marker-assisted selection (MAS) in livestock breeding [16]; 2) it can largely reduce costs of population-level SNP genotyping by application of imputation methods for inferring the missing genotypes on low-density SNP chip, because markers were linked with each other in a long-range region [17]. However, it can also be a problem in fine

mapping of QTL, such as genome-wide association studies (GWAS), and then identifying the causal mutations, because long-range extensive LD around causal mutations can cause long-range QTL regions which harbor too many positional candidate genes to discover causal gene(s) or mutation(s) subsequently [18]. Also, like GWAS, genomic selection (GS) is currently widely applied in studies and breeding programs of livestock animals [19–22]. But, instead of directly detecting QTLs with large effects on traits of interests, GS focuses on utilizing thousands of SNP markers to predict genomic breeding values (GEBV) of genotyped individuals for breeding [23]. Even with differences in application strategy, the basic assumption of GWAS and GS is these markers are in detectable LD with nearby causal mutations. Long-range extensive LD may benefit initial GS application in breeding programs for selecting elite purebreds with high accuracy of GEBV; however, the accuracy can decline over generations due to decays of LD mainly by recombination [24]. Thus, a resource population with relatively short-range extensive LD can benefit both fine mapping of QTL in GWAS and persistence of GEBV accuracy in GS.

In comparison with purebreds, LD decays more quickly in a multi-breed or a crossbred population compared with their parental populations [14, 25–27], because crossbreeding of multiple breeds may increase haplotype diversity and effective population size in crossbreds. In reality, excluding dairy cattle, the majority of livestock industries widely utilize crossbreeding of purebred animals to generate commercial populations for final animal products. So, it will be highly practical to utilize these crossbreds as training (reference) population to predict breeding values of purebred lines as an alternative way to utilize GS in livestock species. Two simulation studies [24, 28] have shown training in the crossbred population provide comparable

accurate prediction of true breeding values of purebred candidates when compared to training in a single purebred. Moreover, short-range LD in crossbreds can captured markers with high consistent association with QTL across the training and validation population for GEBV prediction, which could keep the persistence of accuracy in GS over generations, in comparison with training in purebreds [24].

Another important factor in GS is consistency of LD phase between training population and validation population. This factor is quite critical for accuracies of estimated GEBV for a desired trait in GS [29]. As mentioned before, recombination events can decrease LD over generations, especially for alleles in a long genetic distance. Besides that, the differences in genetic background and breeding history can also cause the low consistency of LD phase among purebreds even in the same agricultural species. Thus, the performance of GS is not ideal when training in one purebred but validating in another one. For example, Hayes et al. (2009) found that the accuracies of GEBV among 5 milk production traits were only -0.01 to 0.17 when training in Jersey cattle and validating in Holstein cattle. However, the accuracies of Holstein's GEBV dramatically increase to 0.44-0.70 when training in both Holstein and Jersey population [30]. On the contrary, alleles in a short genetic distance may maintain high consistency of LD phase among breeds due to low effect from recombination, especially for the agricultural species domesticated from a few common ancestor (s). Using combination of multiple breeds (like crossbreeds in animal industry) as reference population, the statistical methods in GS can utilize the high consistency of short-range LD phases among purebreds to estimate GEBV and improve the performance of GS in all component breeds. Thus, crossbreds could be a better choice than purebreds for being reference populations in GS of multiple related

purebreds. Therefore, gaining the knowledge of LD in crossbred and its parental purebred populations would benefit future genome-wide studies, such as GWAS and GS, in commercial chickens.

Previous studies [10, 15, 31–35] have investigated extent of linkage disequilibrium in chickens. However, because of biotechnical limitations and incomplete chicken reference sequence, most previous studies LD on a few chicken chromosomes with low-density markers. With the developments in SNP array technology, Qanbari et al. (2010) and Wragg et al. (2012) used the Illumina 60K genotyping chips for investigating whole genome LD pattern in layer (egg-type) chicken populations and non-commercial chickens. But the LD patterns in commercial broiler populations using a high-density of SNP chip were uncovered. Thus, to explore the possibility of using crossbreds in GWAS and GS, the study on chapter 2 of the thesis investigated the extent and consistency of LD using high-density SNP chip in commercial broiler populations, including 3 pure lines and their crossbreds. The comparison of extent and structure of LD and haplotype structure among purebreds and their crossbreds has been also performed in this study.

1.2 Detection of Genome-wide Recent Selection Signatures

As motioned before, strong artificial selection in modern breeding program dramatically improved the traits of economical importance in commercial broilers, and it simultaneously generated selection signatures, such as long extensive LD and high allele frequencies, around the underlying causal polymorphisms controlling these traits in chicken genome. Thus, detecting these signatures of artificial selection should help researchers to identify the causal polymorphisms underlying phenotypic changes and to better understand the biological and genetic mechanism controlling these traits.

Before the advent of genome-wide sequences and polymorphisms data, the most practical way to detect positive selection was to examine functional candidate genes that usually were chosen due to their important biological function [36]. However, this single locus approach is very time-consuming for genotyping and choosing functional candidate genes is very subjective. Even though the candidate gene may be very important in term of its biological function, the causal polymorphisms under positive selection may be located in a neighboring region and could regulate the behavior of this functional candidate gene. Moreover, most traits, like disease resistance and body weight, are quantitative (complex) traits that are controlled by numerous genes, and recent positive selection on the complex traits may have worked on these genes simultaneously. Moreover, a large proportion of functional genes controlling a complex traits may be unknown due to incomplete pathway information in agricultural species. Therefore, we need an objective and high throughput method to screen positive selection signatures in the whole genome.

With the availability of large-scale SNP data and high throughput SNP genotyping tools (such as SNP array and next generation sequence (NGS)) in human and recently in many agricultural species, it has become possible to measure thousands and even millions of polymorphisms in each individual genome. These newly developed tools allow us to systematically interrogate an entire species genome to detect the positional candidate genes under positive selection while comparing with the genetic variation in the genome as a whole. Many genetic methods have been developed to detect genome-wide recent selection signatures based on the extent of linkage disequilibrium and haplotype structure.

1.3 Statistical Methods for Detecting Recent Selection Signatures

As prior mentioned, recent selection can cause long-range extensive LD and high allele frequency in a specific chromosomal region harboring the causal polymorphisms/mutations. Thus, the methods based on extended haplotype homozygosity (EHH) [37] or change in allele frequency spectrum have both been suggested to be useful for detecting signatures of recent selection [38, 39]. Both classifications of methods are described below.

1.3.1 Methods Based on Extended Haplotype Homozygosity

1.3.1.1 EHH and Relative EHH

EHH (extended haplotype homozygosity) is defined as the probability that two randomly chosen haplotypes (including the core haplotype or SNP) are homozygous at all loci in a given interval from the candidate core haplotype or SNP [37]. The EHH of a tested core haplotype t in a population is:

$$EHH_t = \frac{\sum_{i=1}^{g} \binom{n_{ti}}{2}}{\binom{N_t}{2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1)$$

where N_t is the number of chromosomes of a particular core haplotype t, n_{ti} is the number of chromosomes of a particular extended haplotype i, and the g is the number of unique extended haplotypes. The extended haplotype here means that the markers used to construct the core haplotype *t* as well as other markers around the core haplotype to reconstruct a set of haplotypes, and all extended haplotypes will include the whole core haplotype t. Thus, it is obvious that: $\sum_{i=1}^{g} n_{ti} = N_t$.

This test is relatively robust to different genetic markers (like microsatellite and SNP) in practice, and can also help to narrow down the candidate regions carrying

the causal polymorphisms under strong recent positive selection [36]. In the application of this test in genome-wide studies, it is assumed that the recombination occurs at the same rate in each chromosomal region across the whole genome. But actually, it has been observed that both recombination hot spots and cold spots do exist in many organisms' genome [40–42]. The recombination cold spots may cause false positive selection signals in these regions while hot spots may cause false negatives instead. Moreover, in chicken, the differences in LD between small- and large-size autosomes have been attributed mainly to the differences in recombination rates against the same physical distance, with micro-chromosomes showing the largest rate of recombination (6.4 cM/Mb), in comparison with lowest one in macro-chromosomes (2.8 cM/Mb) [43]. Since differences in recombination rate is a very critical issue with this kind of selection signature, relative EHH test therefore has been developed to deal with it.

Due to different recombination rates on chromosomes in many species genome, the relative EHH test has been proposed to correct the EHH for each core haplotype [37]. Relative EHH is defined as the ratio of the EHH on the tested core haplotype compared with the EHH of the grouped set of all other core haplotypes at the region. It can be simply calculated by $EHH_t/\overline{EHH}$.

$$\overline{EHH} = \frac{\sum_{j=1,j\neq t}^{h}\left[\sum_{i=1}^{g_j}\binom{n_{ji}}{2}\right]}{\sum_{j=1,j\neq t}^{h}\binom{N_j}{2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.2)$$

where h is the number of different core haplotypes, g_j is the number of unique extended haplotypes carrying core haplotype j and the rest of components in the equation above is the same as the equation (1.1). Thus, relative EHH is on a scale of 0 to infinity.

1.3.1.2 Integrated Haplotype Score

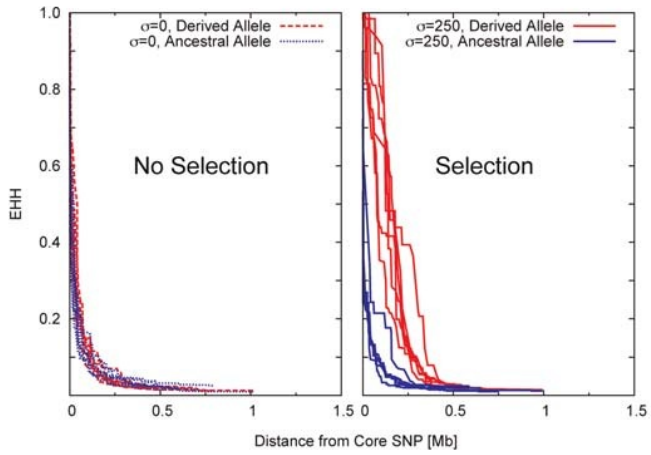


Figure 1.1 Decay of EHH in Simulated Data for an Allele Starting at Frequency 0.5

Note: without selection (s = 0, left side), the haplotype homozygosity decays at very similar pattern for both ancestral (blue) and derived (red) alleles. However, with positive selection on derived alleles (s = 2Ns = 250; right side) than ancestral allele, the haplotype homozygosity decays much slower for the derived alleles in comparison with the ancestral alleles. Reproduced from [44], an open-access article under the terms of the Creative Commos Attribution License.

The integrated haplotype score (iHS) test have been developed by Voight et al. (2006) based on EHH test, and it is more suitable for genome-wide SNP data because the SNP array usually is designed to capture only two alleles on each locus. The EHH in their simulation study was defined as homozygosity of haplotypes carrying a specified 'core' allele rather than core haplotype, and was supported to decay from 1 to 0 with increasing distance from the core site (Figure 1.1). Utilizing simulation data they found that, in comparison with the core allele under a neutral model, EHH for the core allele under selection was higher and could expand much farther as shown in

Figure 1.1. To quantitate the difference in EHH between these two core alleles, the unstandardized iHS test [44] can be computed as:

$$\text{unstandardized } iHS = \ln\left(\frac{iHH_A}{iHH_D}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.3)$$

where iHH_A and iHH_D are the integrated EHH (*iHH*) for ancestral allele and derived allele respectively, and they are computed by the integral of the observed decay of EHH until EHH reaches to 0.05.

From equation (1.3), it is obvious that the unstandardized iHS will be a negative value if derived allele is under positive selection, otherwise it will be approximately equal to 0 under neutral model. As mentioned above, the allele under recent positive selection has high allele frequency and is associated with long-range haplotype. But if a neutral allele was generated more recently, it could have a low allele frequency due to random drift and be also associated with a long-range haplotype because of not enough recombination events, so this low frequency neutral allele might have a larger *iHH* than the high frequency one. Hence, iHS test is adjusted as [44]:

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.4)$$

where $E_p[\ln(iHH_A/iHH_D)]$ and $SD_p[\ln(iHH_A/iHH_D)]$ are the expectation and the standard deviation of $\ln(iHH_A/iHH_D)$ estimated from the empirical distribution at SNPs whose derived allele frequency is p. The iHS statistics follow an approximately standard normal distribution, which is convenient for researchers to do significant test. Moreover, due to the strong LD in the region under positive selection, this method

showed more power for screening windows of consecutive SNPs that contain a lot of high iHS scores[44].

1.3.1.3 Cross Population Extended Haplotype Homozogysity

All the prior mentioned methods are developed to detect selection sweeps in single population, but they have a common limitation. The prior methods can be applied to detect alleles that have undergone recent selection, but do not to detect alleles that have swept to near-fixation or fixation within a population. For example, in the iHS test, only SNP with minor allele frequency (MAF) larger than 5% can be estimated the iHS score because we cannot tell the difference between a SNP with small MAF are caused by alleles under positive selection or just some neutral alleles that have reached fixation by random drift in the common ancestral population. In animal and plant breeding, some mutations with high fitness for economically important traits can reach fixation under strong selection pressure due to artificially increasing the number of progeny of best selection candidates, particularly in MAS system where only the selection candidates carrying the favored allele can have progenies in breeding programs. Thus, the limitation caused by these methods may have low power to identify the alleles with high fitness that means more biological importance in the population. To overcome this limitation, the XP-EHH test (cross population extended haplotype homozogysity) have been developed by Sabeti et al. (2007) to detect selection sweeps in which the favored allele has reached near-fixation or fixation in two different populations.

Very similar to the development of iHS test, XP-EHH test also uses the integrated EHH (*iHH*) of a core SNP but in two populations, A and B rather than two

alleles in a single population. The unnormalized *XP-EHH* statistic can be calculated as [45]:

$$\text{unnormalized XP} - EHH = \ln\left(\frac{iHH_A}{iHH_B}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.5)$$

where iHH_A and iHH_B are the integrated EHH of a given core SNP in population A and B respectively. The unstandardized XP-EHH statistics were adjusted using their means and variances. Thus, a big positive value of XP-EHH suggests selection in population A, or a negative value in B. Because of differences in evolution history and geographic environment, two populations may not share identical long-range haplotype. Thus, in each chromosomal region, each population could be utilized as a control population under neutral model for another population, which makes it possible to detect the strong selection sweeps where alleles have a very small MAF. Obviously, the method requires both populations have different genetic background with separated selection direction.

1.3.2 Methods Based on Allele Frequency Spectrum

1.3.2.1 Composite Likelihood Ratio

The composite likelihood ratio (CLR [46, 47]) is a statistical method that can identify regions with differential allele frequency pattern from overall pattern in the genome of a single breed. Kim & Stephan (2002) first developed the method for detecting selection signatures in subgenome data, Nielsen et al. (2005) improved it for genome-wide data. The basic assumption in the methods by Nielsen et al. (2005) is that most loci are not under selection and can be used to estimate the distribution of allele frequencies in a neutral model. On the contrary, the selected loci and their nearby loci assumed to be presenting a different distribution of allele frequencies in

comparison with neutral model. Thus, the composite likelihood of a fix window (a chromosomal region) can be calculated as:

$$CL_{kA} = \prod_{i=1}^{k} \int_0^1 f\left(p^i\right)_k \binom{n}{m} \left(p^i\right)^m \times \left(1-p^i\right)^{n-m} dp^i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.6)$$

$$CL_{k0} = \prod_{i=1}^{k} \int_0^1 f\left(p^i\right)_K \binom{n}{m} \left(p^i\right)^m \times \left(1-p^i\right)^{n-m} dp^i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.7)$$

where k is the number of SNPs in fixed size of sliding window, and p^i is the allele frequency of a derived allele on locus i, n is the sample size and m is the count of the derived allele; K is the total number of SNPs; Thus, $f\left(p^i\right)_k$ and $f\left(p^i\right)_K$ are empirical background distribution of allele frequencies estimated from k SNPs and all SNPs, respectively, which is the only different between equation (1.6) AND (1.7). The CLR statistic is given by:

$$CLR = 2(\log CL_{kA} - \log CL_{k0}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.8)$$

1.3.2.2 Cross Population Composite Likelihood Ratio

One essential assumption of CLR method is that the empirical background distribution of allele frequencies estimated using genome-wide markers should be consistent with the distribution of allele frequencies under neutral model, so this method assumes that only a small proportion of polymorphisms in genome are influenced by selection in the population. However, as mentioned before, selection for complex traits in agricultural species can work simultaneously on multiple causal polymorphisms/mutations as well as nearby polymorphisms in strong LD with them across the whole genome, which can limit the usage of this method in commercial broilers. Another obvious limitation of the CLR method is that it is very sensitive to SNP ascertainment bias. This bias is a systematic deviations from an expected

distribution of allele frequencies due to the sampling processes for SNP detection, especially sampling limited number of non-random individuals [48]. For example, for the Illumina 60K chicken SNP array, only four commercial breeding lines (two broilers and two layer lines) were used for SNP detection, and SNPs were identified by sequencing the DNA pooled from 25 individuals from each of these commercial breeding lines [49]. Furthermore, only SNPs with medium to high MAF among these purelines were selected. Each of these limitations can be a potential factor contributing to SNP ascertainment bias in estimating distribution of allele frequencies using the Illumina 60K chicken SNP array. Also, the single-population CLR method does not take advantage of larger differences in allele frequencies between two breeds. Thus, similar to strategy of XP-EHH, Chen et al. (2010) developed the XP-CLR (cross population composite likelihood ratio) test to overcome these limitations. The composite likelihood of the k SNPs in a fixed size of sliding window is calculated as [50]:

$$CL(\mathbf{r}, w, s) = \prod_{i=1}^{k} \int_0^1 f\left(p_1^i \middle| p_2^i, w, s, r^i\right) \binom{n}{m_i} \left(p_1^i\right)^{m_i} \times \left(1 - p_1^i\right)^{n-m_i} dp_1^i \quad \ldots\ldots (1.9)$$

where **r** is the vector of recombination rate $\{r^1, r^2, \ldots, r^k\}$, *w* is a weight factor based on linkage disequilibrium, *s* is the selection coefficient, which can be estimated from , *k* is the number of SNPs in fixed size of sliding window, and *p* is the allele frequency, n is the sample size and m_i is the count of neutral allele at locus *i* with assumption that the nearby chromosomal region in reference population is not under selection; $f\left(p_1^i \middle| p_2^i, w, s, r^i\right)$ is the distribution of allele frequencies estimated after a selective sweep. And then a *XP-CLR* statistic is given by:

$$XP - CLR = 2\{\max[\log CL(\mathbf{r}, w, s)] - \log CL(\mathbf{r}, w, s = 0)\} \ldots\ldots\ldots\ldots (1.10)$$

where max refers to the maximization of $\log CL(\mathbf{r}, w, s)$ with respect to the parameter *s.* Another advantage of XP-CLR method is that the distribution of allele frequencies are estimated with modeling genetic drift in reference population by Brownian motion, for identifying the selection signatures where differences in derived allele frequency between objective and reference populations were too large to be due to genetics drift [50]. It makes the methods more suitable for detecting selection signatures in livestock species that usually has low effective population size as mentioned above.

A famous statistician, George E. P. Box, once said, "Essentially, all models are wrong, but some are useful" [51]. This is absolutely true because the methods mentioned above cannot work in every situation. Because they rely on the neighboring LD which surrounds the specific allele of interests, each method is applicable for detecting recent selection sweeps under the correct situation.

1.4 Application of Methods for Detecting Selection Sweeps in Livestock Species

In dairy cattle, Qanbari et al. (2010) adopted the REHH method to detect positive selection signature in Holstein–Friesian cattle using a 50k SNP array [38]. In the study, they first examined ten well-known functional candidate genes that were associated to economically important traits (like milk yield and protein percentage). They found that *DGAT1 (*Diacylglycerol O-Acyltransferase Homolog 1*)* and casein cluster gene showed very high relative EHH in their core region, and haplotype homozygosity extended extremely long from the core region (Figure 1.2). Both genes have been proved to have major effect on milk production traits in dairy cattle [52, 53]. Besides that, another two genes, *SST* (*Somatostatin*) and *LPR* (*Lymphoproliferation*), also showed significant relative EHH. Even though not all of these functional

candidate genes were detected to have recent positive selection, it still showed the method could be used to detect the positive selection in the Holstein population. They then applied the relative EHH method in the whole genome screen for selection signatures.

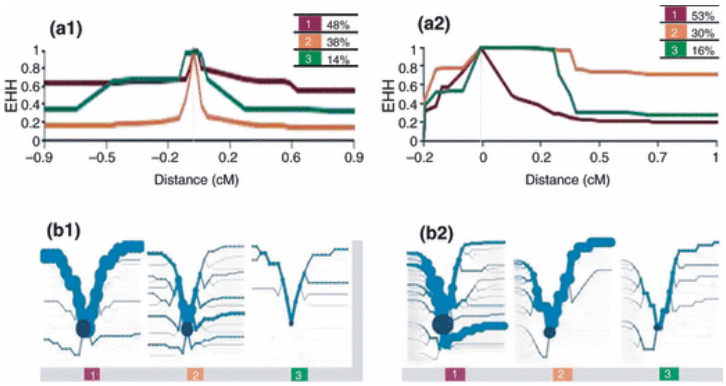


Figure 1.2 The EHH Pattern in the Casein Cluster (Left Side) and *DGAT1* (Right Side) Gene

Note: a1 and a2: EHH against distance plots showed decay of haplotype homozygosity with increasing in distance for the three most frequent core haplotypes. The haplotype frequencies for the three core haplotypes were presented on the legends in the plots; b1 and b2: Haplotype bifurcation plots of the three core haplotypes for the casein cluster and *DGAT1* regions respectively. Reproduced with permission from [38]. Copyright John Wiley and Sons. All Rights Reserved.

In the genome-wide analyses, they first determined 3741 core regions in bovine genome based on the long-range extensive LD in these regions. The relative EHH was calculated at 1 cM distances on both upstream and downstream for all core haplotypes within these core regions. Surprisingly, in the dairy cattle's genome, 702 core haplotypes were detected to have significant selection signatures with *P*-value < 0.05, and 161 ones with *P*-value < 0.01. And for the 12 most significant core

haplotype (Figure 1.3), they explored QTL databases available online to identify genes located around its core regions with published QTL in cattle. Many major QTL and functional genes related to economically important traits were found around the candidate selection regions, which met the hypothesis that selection in the economically important traits could cause the selection signature around the related QTL. Similarly, Qanbari et al. (2014) applied iHS and CLR tests to detected selection signatures in Fleckvieh cattle, a dual purpose cattle suitable for production of dairy and beef, and found that genes in 106 detected candidate selection regions were relevant to coat coloring pattern, neurobehavioral functioning and sensory perception [54]. Then, by applying GWAS, they validated these selection regions harboring genes related to coat coloring pattern in cattle.

Another example about detecting recent positive selection in cattle is a study about using XP-EHH test to detect artificial selection signatures in ten cattle breeds [55], including two dairy cattle breeds, two beef cattle breeds, four dairy-beef cattle breeds and two artificially unselected breeds. Generally, these ten breeds in this study had differences in genetic background, environment and selection purpose. In the results of XP-EHH test, the two artificially unselected breeds got obviously less selection signatures (confirmed by neighboring markers and/or multiple significant breed-comparisons) in comparison with other eight artificially selected breeds, which suggested that only natural selection influence the haplotype homozygosity in their genomes. For these artificially selected breeds, the number of confirmed selection signatures ranged from 100 to 229. Similarly to the findings in other cattle breeds mentioned above, these selection signatures included all of the six target genes that were well known with major effect on milk and beef traits. Thus, the results in

previous studies in cattle suggested that detecting recent selection signatures could be another useful method to identify QTL or candidate genes related to selected traits even without phenotype information.

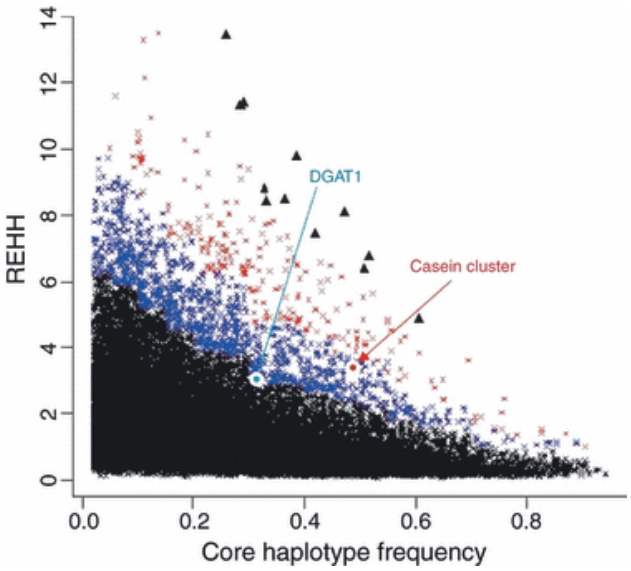


Figure 1.3 Distribution of Relative EHH against Haplotype Frequencies in the Holstein Genome

Note: Relative EHH was estimated at ±1 cM distances around all possible core haplotypes in this dairy cattle genome. Core haplotypes with significant relative EHH statistic are presented in blue (P-value < 0.05) and red (P-value < 0.01) respectively except 12 core haplotypes presented by triangles that meant their P-value < 0.001. The points displaying DGAT1 and the casein cluster are marked in the plot. Reproduced with permission from [38], Copyright John Wiley and Sons. All Rights Reserved.

In chickens, using whole-genome re-sequencing and high-density SNP chips, respectively, Rubin *et al.* (2010) and Elferink *et al.* (2012) have initially investigated selection signatures in large numbers of chicken breeds using Z-transformed pooled heterozygosity (ZHp) scores. This statistic estimates local heterozygosity depression in chromosomal regions [56, 57] and has been applied for detecting alleles that have

swept to fixation or near-fixation by long-term directional selection or during domestication [39]. However, as mention before, modern breeding practices have a more recent selection history, and have been employed to select for a more sophisticated traits, such as feed efficiency and meat yield, by modern breeding organizations. Therefore, most signatures of this more recent selection are likely yet to be uncovered in the genome of modern broiler chickens. Thus, besides investigating LD, another objective in the study is to apply both cross-population methods, XP-EHH and XP-CLR, to identify recent selection signatures in commercial broilers.

Chapter 2

LINKAGE DISEQUILIBRIUM IN CROSSBRED AND PURE LINE CHICKENS

2.1 Introduction

Progress in next-generation sequencing and high-density single nucleotide polymorphism (SNP) genotyping technologies offer unprecedented opportunities for detecting causal polymorphisms or achieving high accuracies of prediction in genomic selection [8, 11–14]. As mentioned in Chapter 1, however, taking full advantage of these new technologies may be limited in livestock populations due to the high level and extent of LD. Although LD extends over long distances in most livestock populations, comparisons of LD patterns between populations show that shared haplotype segments are much shorter when the population consists of multiple purebred populations [25–27], which indicates that LD decays more quickly in multi-breed or crossbred populations than in purebred populations. Therefore, in cases where LD does not extend over long distances, multi-breed and crossbred populations can be potentially useful for fine mapping and identification of causal polymorphisms. Furthermore, as explained in Chapter 1, using a crossbred population as reference population in genomic selection can also be advantageous, particularly in livestock species with crossbreeding programs, such as poultry, pigs and beef cattle. In this study, our aim was to characterize the consistency of LD and differences in LD between crossbred and their component purebred populations. As previous studies on LD in layer [10] and village chickens [35], we used the Illumina 60K chicken SNP

panel [49] which contains over 10 times more SNPs than that of most other studies on LD in chickens [15, 31, 32, 34, 58].

2.2 Methods

2.2.1 Animals and Data Preparation

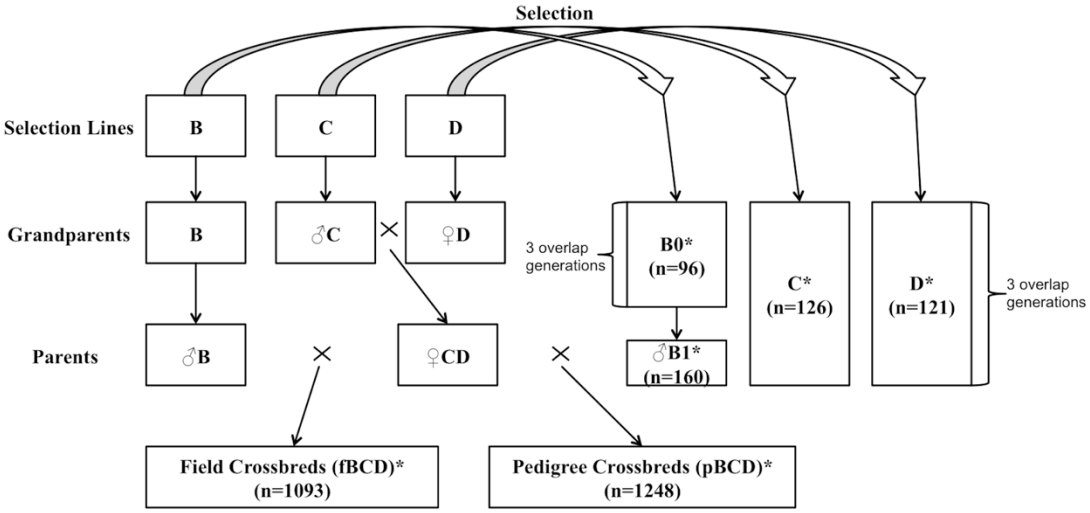


Figure 2.1 Schematic Representation of the Broiler Population Structure

Note: genotyped pure lines and crossbred individuals were sampled from pedigree pure lines (right) and two broiler-crossbreeding programs (bottom). Among the pedigree pure lines, line B is a male line and both lines C and D are female lines. All genotyped birds (indicated by "*") were sampled from male flocks. The genotyped birds from male and female lines were elite sires randomly sampled from two and three overlapping generations, respectively. A random sample of the B0 elite sires' progeny (B1) was also genotyped. The field crossbred chickens (fBCD) were end-product meat birds, whereas the pedigree crossbred chickens (pBCD) were produced for the genetic evaluation of B1 sires. The CD mothers of the field and pedigree crossbred chickens were different but from the same generation of CD parents. The line B fathers of field and pedigree crossbred chickens were diverged for two to three generations.

A total of 2844 individuals were genotyped using the Illumina 60K chicken SNP array [49]. All genotyped birds were sampled from male flocks and included 2341 crossbred and 503 purebred chickens. Among the 503 genotyped purebred chickens, 256 were sampled from a male line, i.e. line B, and 126 and 121 chickens were sampled from two female lines, i.e. lines C and D, respectively (there was no line A in this study). Individuals that were genotyped from the female lines were elite sires that were randomly sampled from three overlapping generations. Only a portion of the chickens that were genotyped from the male line (B0; n = 96) were elite sires that were also sampled from three overlapping generations; another set of genotyped chickens (B1; n = 160) was a random sample of the progeny of the B0 elite sires (Figure 2.1).

All 2341 genotyped crossbred chickens were produced by a three-way cross of B × [C×D], in which males of line B were mated with CD crossbred females, which is a crossbred female product produced by a two-way C × D cross, in which males of line C were mated with females of line D. Crossbred individuals were sampled from two broiler chicken populations: (i) broiler chickens from the field (end product meat birds), which will be referred to as field crossbred chickens (or fBCD, n = 1093) and (ii) broiler chickens from a pedigree house (produced for genetic evaluation of the pedigree B1 sires), which will be referred to as pedigree crossbred chickens (or pBCD, n = 1248) (Figure 2.1).

To assess the extent to which the LD pattern in crossbred populations can be predicted based on the genotypes of their component pure lines, we created a combined population by simply combining the genotype data of a random sample of 200 chickens of line B, 100 of line C and 100 of line D. The 2:1:1 proportion of lines

B:C:D was used to mimic the expected genetic contribution of these lines to the autosomes of crossbred individuals. This combined population will be referred to as the combined BCD (or cBCD) population.

The Illumina 60K SNP chip contains 57636 SNPs [49]. In this study, we used only SNPs with assigned positions on autosomes (based on the latest reference genome, *Gallus gallus* 4.0 UCSC, May 2012). Within each pure line and crossbred population, we discarded SNPs with a call rate less than 90%, Mendelian inconsistency greater than 0.001 and minor allele frequency (MAF) less than 0.05. Also, SNPs that strongly deviated from Hardy–Weinberg equilibrium (p value < 0.001) in the pure lines were discarded, as well as SNPs on chromosome 16 and two linkage groups because there were too few SNPs in the 60K SNP panel for these chromosomes.

2.2.2 Haplotype and Haploblock Analyses

Determining haplotype phase and frequency is necessary to estimate LD and can provide useful information about breed-specific haplotypes and the history of artificial selection. We used BEAGLE (Version 3.3.2) [59] to infer haplotype phase for each genotyped individual in each population. As in Badke et al. [13], we set BEAGLE to run 100 iterations of the phasing algorithm and to sample 100 haplotype pairs for each individual per iteration.

In theory, haplotype homozygosity is defined as the likelihood of randomly sampling two identical haplotypes from a population, which is calculated as the sum of squares of haplotype frequencies [60]. Based on the results of the haplotype phases obtained with BEAGLE, haplotype homozygosity was estimated using haplotypes frequencies for 250-kb sliding windows, with a step size of 25 kb. For each

population, Haploview (Version 4.2) [61] was used to define haplotype blocks (haploblocks) with the built-in algorithm suggested by Gabriel et al. [62]. In this model, the confidence interval of observed values of LD measured by D' was estimated to determine the upper and lower confidence bounds of D' (5% tails of the overall probability distribution of D'), and the blocking structure was determined by defining SNP pairs to be in "strong LD" if the upper confidence bound was above 0.98 and the lower bound was above 0.7 [62].

2.2.3 Estimation of Linkage Disequilibrium

We calculated *r* using the equation below and used its square, r^2, to measure LD between marker pairs that are separated by less than 5 Mb on each chromosome:

$$r_{ij} = \frac{f(MN) - f(M)f(N)}{\sqrt{f(M)f(m)f(N)f(n)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1)$$

where r_{ij} is the correlation between alleles at SNP *i* (alleles *M* and *m*) and alleles at SNP *j* (alleles *N* and *n*); *f(MN)* is the observed frequency of haplotype *MN*, which can be simply obtained from the phasing results; and *f(M), f(m), f(N)* and *f(n)* are observed frequencies of alleles *M, m, N* and *n*, respectively [63].

Previous studies on LD in chickens showed that the extent of LD over physical distances varies greatly among the different categories of avian chromosomes: macro-chromosomes (GGA1 to 5), intermediate chromosomes (GGA6 to 10) and micro-chromosomes (GGA11 to 38) [43]. Thus, we estimated LD separately for each category of chromosomes within each population. To visualize the LD pattern for each category of chromosomes in different populations, r^2 values were ordered in ascending order based on the physical distance between the corresponding SNP pairs, and then a

rolling average of LD was calculated as the arithmetic mean of all r^2 values for SNP pairs in 25-kb intervals and plotted against physical distance between SNPs.

2.2.4 Estimation of Consistency of LD

Consistency of LD between two populations was calculated as the correlation of r between SNP pairs. We used the SNPs that were common to the populations to estimate the consistency of LD as the correlation r_{ij} between the same pairs of SNPs within a given interval in two populations. For simplicity, this will be referred to as the correlation of r. To visualize and compare the correlation of r among different pairs of populations, the pairwise correlation of r was estimated separately for each category of chromosomes in 50-kb non-overlapping intervals and plotted against physical distance.

2.3 Results

Table 2.1 Quality Control Criteria and Number of SNPs Discarded in Each Population

Quality control	Population[1]					
	All	B	C	D	fBCD	pBCD
Chromosomes not included	4522					
Mendelian inconsistency	1456					
SNPs not called[2]		706	978	978	823	839
Monomorphic SNPs		7907	9888	7768	3467	3582
SNPs with a call rate < 0.9		550	154	230	510	448
SNPs with a MAF < 0.05		4914	4096	4300	3205	3149
HWE (p value < 0.001)		121	163	194	NT[3]	NT[3]
SNPs in use		37460	36379	38188	43653	43640
Common SNPs	26350					

Note: [1]B: line B; C: line C; D: line D; fBCD: field crossbred chickens; pBCD: pedigree crossbred chickens; [2]SNPs that were genotyped but not called; [3]NT: the Hardy-Weinberg equilibrium test was not applied to crossbred chickens.

2.3.1 Marker Statistics

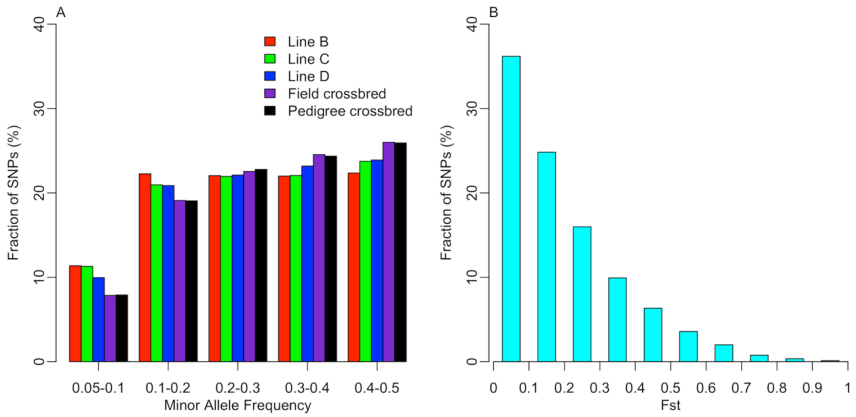


Figure 2.2 Distribution of Minor Allele Frequency and F_{st} of SNPs

Note: 2A: Distribution of MAF of SNPs after quality control in each population. Each population is represented by a different color; 2B: Distribution of F_{st} of common SNPs to the three purebred populations.

The numbers of SNPs that remained after quality control and were used in subsequent analyses for pure lines B, C and D, and field crossbred (fBCD) and pedigree crossbred populations (pBCD) ranged from 36379 to 43653 and are in Table 2.1. There were 26350 common SNPs in these five populations. SNPs that were evaluated in the combined BCD (cBCD) population were the same as those included in the field crossbred population.

Distributions of MAF for SNPs after quality control are in Figure 2.2A for each population. More than 65% of SNPs in the three purebred populations and more than 70% of SNPs in the crossbred populations had a MAF greater than 0.2. MAF distributions were mostly uniform for MAF greater than 0.05. As expected, the number of SNPs with a high MAF was larger for the crossbred populations than for

each of the pure lines. In addition, F_{st} [64] were estimated among the three purebred populations for all common SNPs after quality control, and their distributions are in Figure 2.2B. The average F_{st} was greater than 0.20, which suggests that there was substantial genetic differentiation among these purebred populations [65].

2.3.2 Linkage Disequilibrium

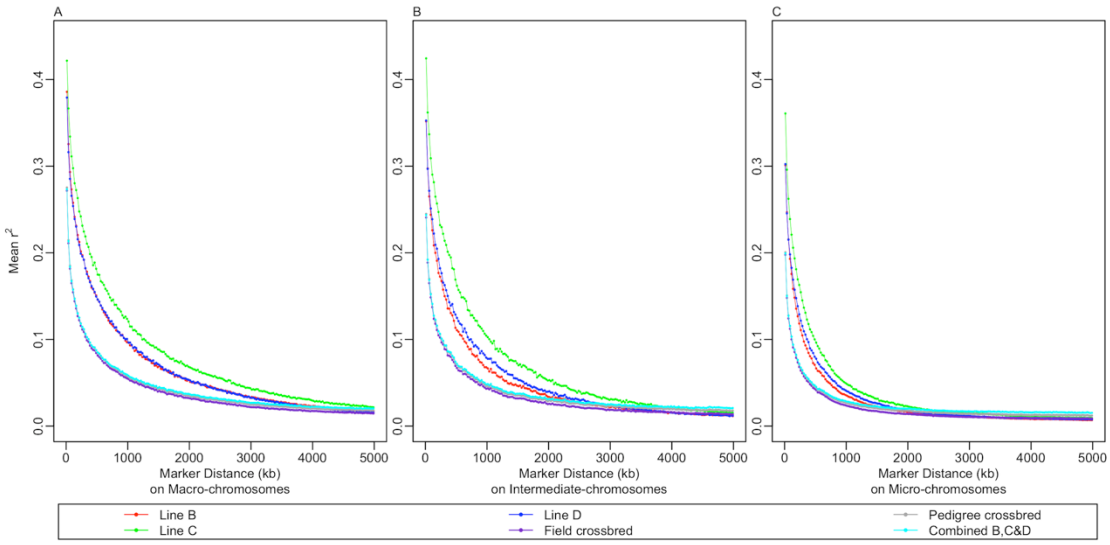


Figure 2.3 Decay of Linkage Disequilibrium with Distance on Different Categories of Chromosomes in Different Populations

Note: Each population is represented by a different color. Each point in the plots represents the mean r^2 of marker pairs in a 25-kb interval. Points representing field and pedigree crossbred and combined BCD populations (purple, grey and cyan, respectively) are almost overlapping and hard to distinguish in most areas.

As shown in Figure 2.3, LD declined as the distance between markers in all populations increased, and r^2 converged to 0.02 approximately at around 2 Mb, 4 Mb and 5 Mb on micro-, intermediate, and macro-chromosomes, respectively. At marker interval distances smaller than 1 Mb, LD differed considerably between crossbred and

purebred populations and also between chromosome categories. Micro-, intermediate, and macro-chromosomes showed the lowest, second lowest and highest mean r^2, respectively, across all populations (Figure 2.3). The crossbred populations and line C displayed the lowest and the highest mean r^2, respectively. The mean r^2 of lines B and D were similar but lower than that of line C. Compared with the pure lines, the distance at which r^2 decayed below 0.2 ($D_{0.2}$) was considerably smaller in the crossbred populations; in the crossbred populations, $D_{0.2}$ was equal to ~50, ~25, and ~15 kb for the macro-, intermediate, and micro-chromosomes, respectively while in the pure lines, $D_{0.2}$ was greater and equal to 225, 150, and 80 kb. Line C showed the largest $D_{0.2}$ for all three categories of chromosomes.

The mean r^2 between adjacent SNPs across all autosomes studied here ranged from 0.34 to 0.40 in the pure lines but was on average only equal to 0.24 in the crossbred populations. Due to the different densities of SNPs on each chromosome in the 60K SNP chip, mean r^2 values were similar in the three categories of chromosomes. Furthermore, in the pure lines, at least 53.2% of adjacent SNP pairs had an r^2 greater than 0.2 and 42.2% had an r^2 greater than 0.3, but in the crossbred populations, only 39.5% and 28.4% showed an r^2 greater than 0.2 and 0.3, respectively.

The two crossbred populations and the combined BCD population showed almost the same level of LD (Figure 2.3) and very high correlations of r (Figure 2.4) at all distances between SNPs studied here. Thus, only results of the field crossbred population *vs.* pure lines are presented here.

2.3.3 Haploblock and Haplotype Homozygosity

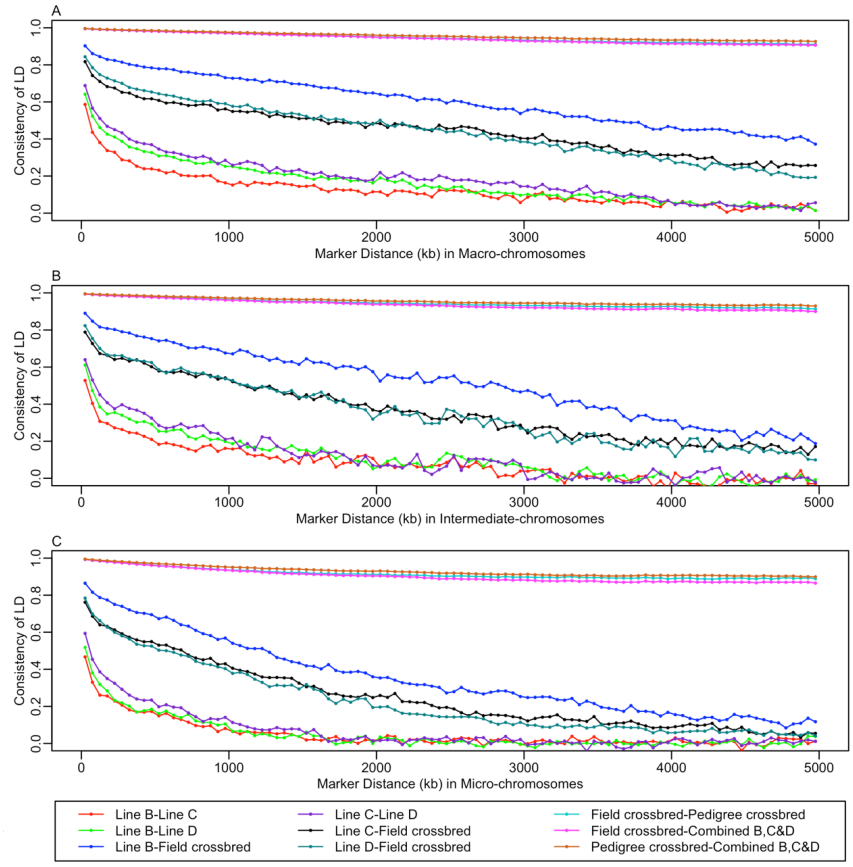


Figure 2.4 Correlation of LD on Different Chromosome Categories among the Pure Lines and Crossbred Populations

Note: Each population is represented by a different color. Each point in the plots represents the mean correlation of *r* in a 50-kb interval. Points representing the correlations of *r* between field and pedigree crossbred populations and combined BCD populations are almost overlapping and hard to distinguish in some areas.

The statistics of haploblock distributions in the different populations are in Table 2.2 and Figure 2.5. After quality control, more than 62% of SNPs formed haploblocks in the pure lines but only 30.6% in the field crossbred population. Also,

the field crossbred population had the lowest genome coverage in haploblocks and the smallest overall median haploblock size. Moreover, nearly half (49.1%) of the haploblocks in the field crossbred population were slightly smaller than 25 kb, compared with 42.6%, 36.6% and 43.5% in the pure lines B, C and D, respectively. Line C had the largest genome coverage in haploblocks and showed the largest median haploblock size. Among the three chromosome categories and in each population, macro-chromosomes had the largest average length of haploblocks, followed by the intermediate and micro-chromosomes (See Appendix A). All these results were consistent with the LD patterns observed in these populations.

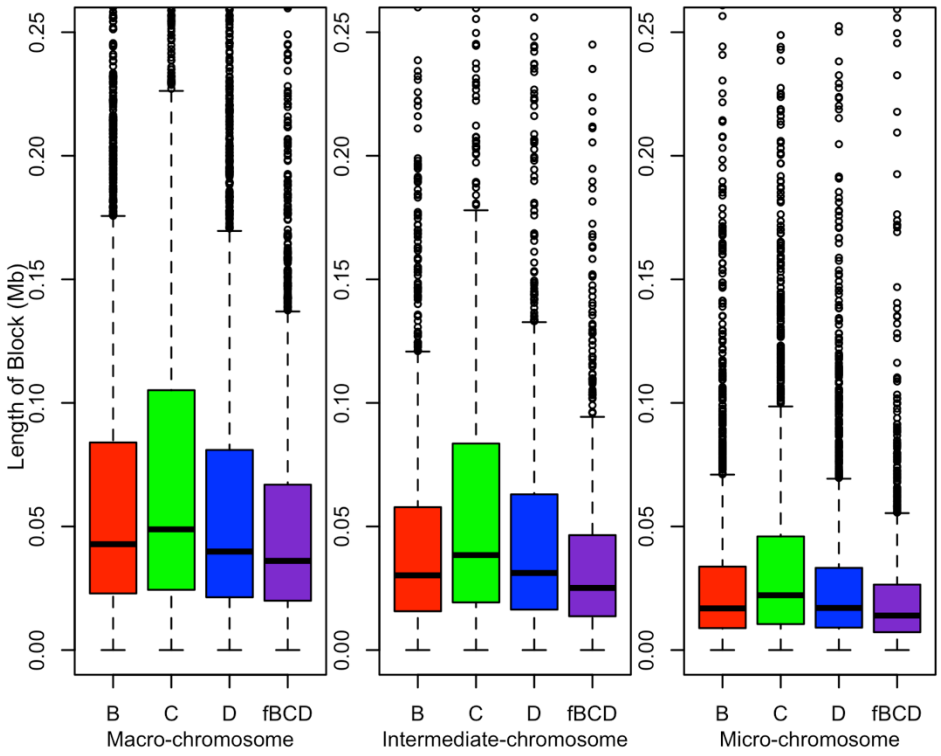


Figure 2.5 Length of Haploblocks in Different Categories of Chromosome in Different Populations

Note: B: line B; C: line C; D: line D; fBCD: field crossbred population. Different populations are represented by different colors. The ends of the whiskers represent the lowest datum within 1.5 IQR (interquartile range) of the lower quartile, and the highest datum within 1.5 IQR of the upper quartile.

Table 2.2 Summary Statistics of Haploblock Structure across Different Populations

Statistics	**Population**[1]			
	B	C	D	fBCD
Median block size (kb)	30.8	37.2	29.6	25.7
Maximum block size (kb)	3521.9	3527.6	4226.0	2737.2
Genome coverage (Mb)[2]	446.7	485.8	401.4	229.0
TSNPs[3]	26,293	25,720	23763	13,375
BSNPs (%)[4]	70.2	70.7	62.2	30.6
Mean ± SD nBSNPs[5]	3.5 ± 3.4	4.2 ± 4.6	3.5 ± 3.9	2.6 ± 2.1
Max nBSNPs[6]	86	89	118	86

Note: [1]B: line B; C: line C; D: Line D; fBCD: field crossbred chickens; [2]genome coverage with all haploblocks; [3]total number of SNPs that form haploblocks; [4]percentage of SNPs that form haploblocks; [5]mean and standard deviation of number of SNPs that form haploblocks; [6]maximum number of SNPs that form haploblocks.

Haplotype homozygosity (HH) was measured over sliding windows of 250 kb. Results for chromosomes 3, 8 and 19 are in Figure 2.6 and represent macro-, intermediate, and micro-chromosomes, respectively. Among all the populations, crossbred populations showed a lower average HH than purebred populations for all chromosome categories, with most HH values being less than 0.1. Differences between populations were not very obvious for the micro-chromosomes, because the extent of HH with a window size of 250 kb was very small for micro-chromosomes compared to macro-chromosomes. Although the overall HH pattern was consistent with the results of LD analyses in these populations, the local HH pattern on each chromosome varied among the four populations.

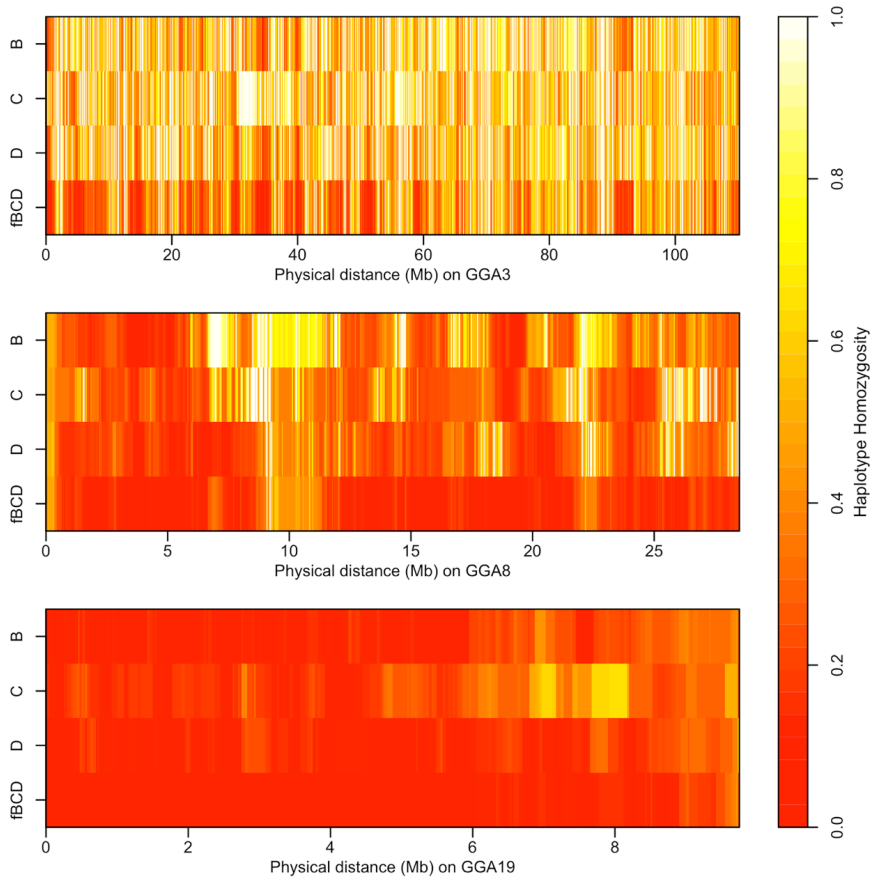


Figure 2.6 Haplotype Homozygosity on Chicken Chromosomes GGA3, GGA8 and GG19

Note: B: line B; C: line C; D: line D; fBCD: field crossbred population. The end position on each chromosome represents the physical position of its last SNP on the Illumina 60K chicken SNP panel. Each bin in the plots has a size of 25kb and represents haplotype homozygosity of a 250kb sliding window with a step size of 25 kb on each chromosome.

2.3.4 Consistency of LD

The correlation of r measures the degree of agreement of the direction and level of LD for SNP pairs between two populations [11, 13]. In principle, these correlations can range from –1 to 1: a high positive value indicates high LD and the

same haplotype phase in the two populations, and a high negative value results from a high LD between two markers but with the opposite phase in the two populations [12].

It should be noted that in this study correlations of *r* were high and positive (> 0.99) between field and pedigree crossbred populations and also between either of these real crossbred populations and the combined BCD population at distances between markers less than 50 kb (see Figure 2.4). The correlation gradually decayed as the distance between markers increased but still remained high (0.87 to 0.93), even if SNPs were about 5 Mb apart.

Our results show that correlations of *r* between the pure line C and field crossbred population (0.86 to 0.92; < 10 kb) and between the pure line D and field crossbred population (0.83 to 0.88; < 10 kb) were generally similar but lower than those between the pure line B and field crossbred population (0.91 to 0.94; < 10 kb). Among the pure lines, the correlation of r between the two female lines was slightly higher than that between the male line and either of the female lines.

As with the decay of LD, correlations of *r* decreased with physical distance and this decrease was greater for the smaller chromosomes. For example, correlations of *r* between the pure line B and field crossbred population were 0.94, 0.94 and 0.91 in the interval of 0 to 10 kb on macro-, intermediate- and micro-chromosomes, respectively, but decreased to 0.39, 0.21 and 0.13, respectively, at an interval of approximately 5 Mb.

2.4 Discussion

Studies of LD in farm animals have been mostly limited to purebred populations and there is limited information about the extent of LD in crossbred populations and the consistency of LD phase between crossbred populations and their

parental pure lines [14]. In this study, we characterized the consistency of phase and level of LD between crossbred broiler chickens and their parental pure lines. The crossbred chickens in our study were a three-way cross of B × [C × D] that was produced using three pure lines B, C and D, which contributed 50%, 25%, and 25% of genetic material, respectively, to the autosomes of the crossbred animals. Our analyses used SNP genotypes on 27 chicken autosomes (GGA1 to 28, excluding GGA16 and other micro-chromosomes because the 60K SNP chip did not include enough markers on these chromosomes). To our knowledge, there is no published report that characterizes LD in crossbred chicken populations and that compares the consistency of phase and level of LD between crossbred and their parental pure lines.

2.4.1 Rapid Decay of LD in Crossbred Populations and Micro-Chromosomes

The level and pattern of LD differed between the populations used in our study. The two field and pedigree crossbred populations had very similar levels of LD for all physical distances. However, compared with the pure lines, crossbred populations showed a small extent and rapid decay of LD by distance for all three chromosome categories. For example, for the macro-chromosomes, the mean r^2 for SNPs that were 0 to 10 kb apart was 0.32 in the field crossbred population but greater than 0.44 in the pure lines, and the mean r^2 dropped to less than 0.2 at distances between SNPs of ~50 kb in the field crossbred population, whereas in the pure lines, this drop was observed for much greater distances (~225 kb). Similarly, the extent of LD was smaller in crossbred beef cattle than in purebred Angus and Charolais cattle [14]. The rapid decay of LD by distance in crossbred populations can be useful for high resolution mapping of causal polymorphisms. Indeed, if the extent of LD is small, it is less likely

that SNPs far away from a causal polymorphism will be in LD with the polymorphism, which confines associations to SNPs that are in close vicinity of the causal polymorphism, thereby increasing map resolution. Moreover, when using a higher SNP density, the small extent of LD in crossbred populations may be an advantage for genomic selection because the tight linkage between causal polymorphisms and adjacent SNPs is less likely to be broken down by recombination, and therefore the accuracy of genomic predictions will persist over more generations [24].

An average LD (r^2) greater than 0.2 [23, 66, 67] or 0.3 [8, 68] between adjacent SNPs has been recommended to detect SNPs associated with causal polymorphisms or to achieve a reasonable accuracy of prediction in genomic selection. Although in the pure lines at least 53.2% of adjacent SNP pairs of the 60K SNP panel had an r^2 greater than 0.2 and 42.2% had an r^2 greater than 0.3, in the field crossbred population, only 28.4% of adjacent SNP pairs showed an r^2 greater than 0.3, which suggests that a higher density SNP panel would be an advantage for GWA studies or to implement genomic selection in commercial crossbred populations.

In our study, the extent of LD varied greatly between chromosome categories in different populations and decreased as the distance between SNPs increased (Figure 2.3). Consistent with previous studies [10, 15, 34], our results showed that small-size autosomes had less LD than large-size autosomes. The differences in LD between small- and large-size autosomes have been attributed mainly to differences in recombination rates per unit of physical distance, with micro-chromosomes showing the largest recombination rate per Mb (6.4 cM/Mb), followed by the intermediate chromosomes (3.9 cM/Mb), and then macro-chromosomes (2.8 cM/Mb) [43].

It was noted that SNP ascertainment bias (Figure 2.2A) could be an important factor affecting our results of LD analysis based on SNPs in the Illumina 60K chicken SNP array. SNP ascertainment bias of genotyping arrays is mainly related to the protocol used to identify SNPs and to the sampling of a limited number of non-random individuals for their detection. In general, this leads to overestimation of LD [69][48]. For example, for the Illumina 60K chicken SNP array, only four commercial breeding lines (two broilers and two layer lines) were used for SNP detection, and SNPs were identified by sequencing DNA pooled from 25 individuals from each of these commercial breeding lines [49]. Furthermore, only SNPs with medium to high MAF were selected. Each of these limitations can be a potential factor contributing to SNP ascertainment bias in our data obtained using the Illumina 60K chicken SNP array. Although SNP ascertainment bias cannot be avoided when using genotyping arrays, for the purpose of comparison, we estimated LD on intermediate chromosomes in 72 crossbred chickens that were genotyped using the recently available Affymetrix 600k chicken SNP array. This array was designed by sequencing more individuals, i.e. 243 chickens from 24 chicken lines, including 15 commercial lines (broilers or layers), inbred layers and one unselected layer line [70]. Therefore, compared to the 60K SNP array, results from the 600K SNP array should be less affected by SNP ascertainment bias. As expected, our results showed that the LD (r^2) measured at distances up to 5 Mb was slightly lower with the 600K SNP array than with the 60K SNP array. The average difference in LD at distances up to 1 Mb was, however, less than 0.016 (i.e. 19.1% reduction in average LD when using the 600K SNP array), and the differences became smaller and more stable at larger distances (Figure 2.7). Thus, although SNP ascertainment bias cannot be avoided when using genotyping arrays, we believe our

results and conclusions on differences in LD patterns in crossbred and pure line chickens are reliable.

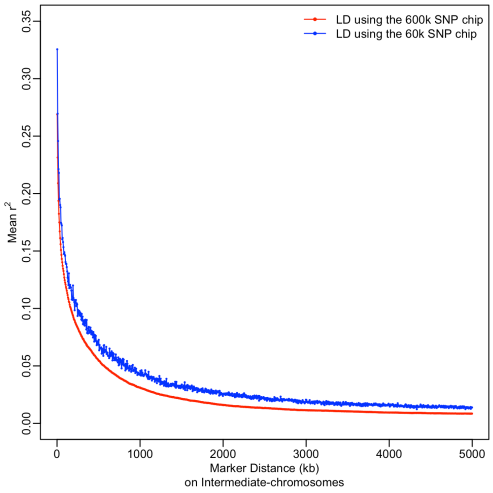


Figure 2.7 Decay of Linkage Disequilibrium with Distance on Intermediate Chromosomes in Field Crossbred Chickens using Different SNP Arrays

Note: The results of LD using the 600K and 60K SNP arrays are represented by red and blue color, respectively. Each point in the plot represents the mean r^2 of marker pairs in a 5-kb interval.

2.4.2 Small Haploblock Size in Crossbred Populations

Consistent with the results of the LD analysis, results of the haploblock analyses showed that the crossbred populations had the lowest genomic coverage (229.0Mb) in haploblocks and the smallest average number of SNPs forming haploblocks. Given the size of the autosomes (~918.9 Mb; GGA1 to 28 without GGA16; *Gallus gallus* 4.0, November 2011), this means that haploblocks covered only 24.9% of the genome of the crossbred populations, which is nearly half that of

the pure lines. In a study using commercial chickens, the genome coverage in haploblocks ranged from 337.1 to 599.4 Mb, with an F1 cross between two layer populations showing a lower genome coverage (337.1 Mb) than the layer pure line [10]. Although the pure lines in this study were not the parental lines of this F1 cross, the low genome coverage in haploblocks in this two-way layer cross is consistent with what we observed in the three-way broiler cross.

The percentage of SNPs forming haploblocks differed between populations (Table 2.2). Only 31.4% of SNPs formed haploblocks in the field crossbred population, compared to more than 60% in the three pure lines. This finding shows that, because of the small extent of LD, most markers did not form haploblocks in the crossbred populations. The small percentage of SNPs forming haploblocks and the small haploblock size in the crossbred populations also indicate that the 60K SNP chip used in our study does not have an adequate SNP density for high-resolution characterization of the haplotype structure in crossbred chickens. Megens et al. [34] investigated the LD and haplotype diversity on four ~1-cM regions on macro- and micro-chromosomes and suggested that whole-genome marker assays would need to contain at least 100K informative SNPs to exploit haplotype information in commercial chicken populations. Consistent with this, a study on the haplotype structure of traditional and village chickens suggested using at least 90 to 110K SNPs to construct a whole-genome haplotype map for these populations [35]. Therefore, the recently available 600K SNP chicken panel [70] is likely to provide a higher resolution haplotype map in crossbred chickens.

Haplotype homozygosity is a measure of haplotype diversity in a population. In a previous study, a relatively small number (1 to 7) of haplotypes accounted for

most of the haplotype diversity (> 90%) found on the macro-chromosomes, but not on the micro-chromosomes [34]. These results are consistent with our findings across all populations. Within 250-kb windows, the pure-line chickens showed limited haplotype sharing, which is consistent with a moderate correlation of r (< 0.5) at the same distance between these populations (Figure 2.6). In some chromosomal regions, high levels of HH extended over longer distances in one pure line but not in the others. Because one of the key characteristics of positive selection is the presence of unusually long-range HH [37], these differences in HH patterns could be considered as evidence of recent positive selections in these pure lines. Therefore, it would be interesting to conduct an analysis of selection sweeps in these pure lines.

2.4.3 Consistency of LD from Pure Lines to Crossbreds

Because of the pyramidal structure of conventional chicken breeding programs, there is approximately four years of lag time from pedigree pure line birds to end-product crossbred birds [71]. To assess if LD persists between pedigree pure lines and commercial crossbred birds, we estimated the correlation r between pure lines in the pedigree program (top of the pyramid) and crossbred chickens that were sampled from the field (bottom of the pyramid). Correlations of r were high (0.83 to 0.94) between these populations for closely spaced SNPs (0 to 10 kb) but these correlations decreased as the distance between SNPs increased; correlations dropped by 4 to 9% from 0 to 10 kb to 10 to 50 kb distances between markers across the three chromosome categories. In our study, the 60K SNP panel provided an average marker spacing of one informative SNP per ~25 kb; therefore, it is expected that using a higher density SNP panel (such as the recently available 600K SNP array) will increase the accuracy of genomic selection of pure lines for crossbred performance.

This conclusion is consistent with results from two simulation studies [24, 28] that showed that training in crossbred populations led to slightly less accurate predictions of true breeding values of selection candidates in purebred populations compared with training only in the selected purebred population. However, by increasing the density of the SNP panel, differences in accuracies became much smaller.

The correlation of *r* between crossbred populations and pure lines differed in the three chromosome categories, with macro-chromosomes showing the highest levels of correlation and micro-chromosomes showing the lowest. These results indicated that, in GWA studies or in genomic selection programs, micro- and intermediate chromosomes would require a higher SNP density per kb than macro-chromosomes. In our study, each pure line showed a different level of correlation of *r* with the crossbred populations and, of all comparisons, this value was the highest between line B and the crossbred populations. As mentioned before, line B was the terminal male line for crossbred chickens B × [C × D], which means this line contributed 50% of genetic material to the autosomes of crossbred individuals; whereas the female lines C and D were expected to contribute each 25%. As for female lines, the correlation of *r* was greater between line D and the crossbred populations than between line C and the crossbred populations, which could be due to a greater correlation of *r* between lines D and B than between lines C and B.

2.4.4 Consistency of LD Phase among Pure Lines

The correlation of *r* can reflect the relative degree of similarities and divergences between purebred animals. In a large-scale genome-wide survey of SNP variation in cattle breeds, the correlation of *r* declined as the divergence between breeds increased [72]. In our study, the correlation of *r* between female lines was

slightly but consistently higher than that between female and male lines, which suggests that the two female lines may share a more similar genetic background than each of these lines does with the male line. This is consistent with the fact that male and female lines originated from different breeds, i.e. the male line from Cornish, a meat type breed, and the female lines from dual-purpose breeds. Furthermore, since selection goals of male and female lines are different, this may have contributed to the similarity between female lines and to the divergence between the female lines and the male line. Unlike the male line, which was selected primarily for growth-related traits, the female lines were selected for both reproductive and growth traits.

Using markers on chromosomes 1 and 4, Andreescu et al. (2007) estimated correlations of LD within 500 kb among nine purebred chicken lines from a commercial broiler breeding company; the correlations over all pairs of lines ranged from 0.21 to 0.94, with an average correlation of 0.52 [32]. Badke et al. (2012) reported a correlation of LD for distances between markers less than 10 kb that was equal to 0.92 between Landrace and Yorkshire breeds and 0.87 between these breeds and the Duroc breed; these values decreased to 0.41 to 0.57 for distances between markers around 1 Mb [13]. Moreover, a study in cattle found that the correlation of LD for distances between markers less than 10 kb was 0.97 between Dutch black-and-white Holstein-Friesian vs. Dutch red-and-white Holstein-Friesian and New Zealand Friesian vs. Zealand Jersey [11]. In our study, none of the chromosome categories reached this high level of consistency of LD between pure lines for distances between markers less than 10 kb, and correlations of r ranged from 0.73 to 0.82 for distances between markers less than 10 kb, even on macro-chromosomes. It is likely that the genetic diversity of the chicken lines in our study was greater (average $F_{st} > 0.20$) than

that of both the cattle and pig breeds used in the aforementioned studies. Another possible explanation for this difference is the overall higher recombination rates per unit of physical distance on chicken chromosomes compared with the average ratio of 1cM/1Mb in mammalian livestock animals.

2.4.5 Consistency of LD between Crossbred Populations and the Combined BCD Population

To assess the extent to which the LD pattern in crossbred populations can be predicted using the genotypic information of their component pure lines, we created a combined BCD population (see Methods section) and studied the differences in LD between this hypothetical population and the actual crossbred populations. Across all three chromosome categories, levels of LD were almost the same in these crossbred and the combined BCD populations (Figure 2.3). This was also reflected by the consistency of LD between these populations, since the correlations of r were very high (> 0.99) for distances between markers less than 50 kb between the crossbred and the combined BCD populations, and decayed gradually as the distance between markers increased, but still remained high (0.87 to 0.93) for markers that were about 5 Mb apart (see Figure 2.4). These results indicate that by using only genotype information on the pure lines, one can predict the LD in crossbred populations with very high accuracy, as well as the correlation r between crossbred populations and their component pure lines.

2.5 Conclusions

In conclusion, our study characterized the extent and consistency of LD in commercial broiler populations from different angles and showed that, between

crossbred populations and their component pure lines, the consistency of the level and phase of LD for short distances between markers (0 to 10 kb) is remarkably high. Compared with the pure lines, the crossbred populations showed a considerably lower level of LD and a smaller haploblock size, which suggests that using crossbred animals as a reference population can be an advantage for high-resolution mapping of causal polymorphisms in GWA studies and to achieve better persistence of the accuracy of genomic estimated breeding values over generations in genomic selection programs. However, our results also suggest that a higher SNP density, particularly on micro-chromosomes, is necessary to take full advantage of crossbred populations in GWA studies or in genomic selection programs. Finally, our results prove that LD for short and long distances between markers and haplotype phase for short distances between markers in a crossbred population can be predicted with very high accuracy using genotype information of its parental pure lines.

Chapter 3

DETECTION OF RECENT SELECTION SIGNATURES IN COMMERCIAL BROILER CHICKENS

3.1 Introduction

Artificial selection is the primary factor in the domestication and breeding history of livestock species. As mentioned in Chapter 1, modern broiler chickens have been under strong artificial selection, mostly for traits of economic importance for farmers, such as growth rate, feed efficiency and body composition [73]. And these traits are complex traits that are controlled by many genes. Consequently, it is highly likely that strong selection for these traits has worked simultaneously on multiple causal genes across the genome. Therefore, high throughput methods are required to screen the whole genome for signatures of positive selection. With the availability of high throughput genotyping tools, such as high-density SNP arrays and next-generation sequencing, it has become possible to conduct genome-wide studies for detecting such genomic footprints of artificial selection.

In this study, with the same SNP array mentioned in Chapter 2, we applied both XP-EHH and XP-CLR methods in five commercial broiler purebred populations, including three male lines and two female lines, to detect the signatures of recent selection in these commercial broiler stocks. The findings here help to improve our understanding of the biological mechanisms controlling economically important traits in modern commercial broiler chickens.

3.2 Methods

3.2.1 Animals and Data Preparation

A total of 565 chickens from five commercial pure lines were genotyped using the Illumina 60k chicken SNP arrays [49]. All genotyped birds sampled were males and included 318 and 247 birds from male and female lines, respectively. Our sample included 24, 256 and 38 chickens, respectively, from three male lines with somewhat distinct breeding goals, labelled as lines ML1, ML2 (Line B in Chapter 2) and ML3, respectively, as well as 126 and 121 chickens from two female lines, labelled as lines FL1 (Line C in Chapter 2) and FL2 (Line D in Chapter 2), respectively. ML1, FL1 and FL2 genotyped chickens were elite sires randomly sampled from each purebred population. In ML2, only a portion of the genotyped male lines chickens (ML2_0; n = 96) were elite sires, and these were randomly sampled from three overlapping generations; another portion (M2_1; n = 160) was a random sample of the progeny of the ML2_0 elite sires. ML3 genotyped chickens were a random sample of male chickens from this purebred population. Male and female lines originated from different breeds, i.e. the male line from Cornish, a meat type breed, and the female lines from White Rock, a dual-purpose breed. Each of these five lines came from a different source to Heritage Breeders, and all lines, except FL1, have been reproductively isolated for more than 40 generations. A one-time crossbreeding with ML2 and then backcrossing with FL1 happened in the history of FL1, and the resulting new FL1 population has been reproductively isolated for more than 25 generations. In each generation within each pure line, approximately 50 to 80 male and 500 to 800 female birds have been selected for reproducing the next generation.

The Illumina 60k SNP chip contains a total of 57,636 SNPs [49]. For the purpose of this study, we used only SNPs with assigned positions on the current

chicken genome (based on the latest reference genome, Gallus gallus 4.0 UCSC, May 2012). Within each pure line population, we discarded SNPs with a call rate < 90% and Mendelian inconsistency > 0.001. We also did not use SNPs that were monomorphic among all the pure lines or SNPs on chromosomes 16 and W and two linkage groups, as there were too few SNPs in the 60k SNP panel for these chromosomes. After quality control, 48,950 SNPs were used in subsequent analyses for five populations (Table 3.1).

Because a linkage map was required for the XP-CLR method, we calculated the genetic position of all the markers in the 60k SNP chip using a subset of markers with known genetic position (previously provided by Groenen *et al*. (2009) [74]), assuming that the recombination rates between two markers with known genetic positions were uniformly distributed. We used BEAGLE (Version 3.3.2) [59] to impute missing genotypes and infer haplotype phase by chromosome in each purebred line respectively, for further detecting selection signatures and identifying breed-specific haplotypes under artificial selection.

Table 3.1 Quality Control of Genotype among Five Populations

Total SNPs	**Quality control**					**SNPs used**
	NI[1]	**MI**[2]	**UG**[3]	**MO**[4]	**LCR**[5]	
57,636	1,507	1,478	873	4,292	536	48,950

Note: [1] SNPs on on GGA16, W and two linkage groups (LGE22C19W28_E50C23 and LEG64) or SNPs with unknown positions on Galgal4; [2] Mendelian inconsistency; [3] Ungenotyped; [4] Monomorph; .[5] Low call rate.

3.2.2 XP-EHH Test

As mentioned in Chapter 1, the XP-EHH test uses the integrated EHH (*iHH*) of a core SNP in two populations. We used the software developed by Pickrell *et al*. [75] to

estimate unstandardized XP-EHH statistics for all SNPs (after quality control) in all five purebred lines with cross-population comparison of each purebred line with the four remaining lines: for example, ML1 *vs*. ML2, ML3, FL1 or FL2 (four cross-population tests for each line). The unstandardized XP-EHH statistics were adjusted using their means and variances in each purebred comparison, and then the standardized XP-EHH statistics were used to estimate *P*-values using standard normal distribution [45]. For each purebred comparison, we determined the candidate regions under positive selection by clustering the significant core SNPs (*P*-value < 0.05) with a distance of less than 200kb.

3.2.3 XP-CLR Test

To confirm selection signatures detected by the XP-EHH analysis, we applied the XP-CLR test based on the change in the allele frequency spectrum, since it has the advantage of enlarging signals to allow the resolution of more precise regions [50]. In our study, grid points at the putative selected allele positions were set along each chicken chromosome with a spacing of 2 kb, and sliding window size was set as 0.5 cM around the grid points. To reduce the contribution of SNPs in high LD to the likelihood function, the cut-off level of absolute pairwise correlation coefficient of two SNPs was set to 0.9 for estimation of the weight factor (w in equation 1.9) [50].

For each cross-population comparison, the cutoff threshold of 0.5% XP-CLR scores was applied to determine windows with strong signals across the whole genome. We then determined the candidate selection regions by clustering these windows, such that windows with genetic distances less than 1 cM constitute a candidate selection

region. The selection regions detected by both statistical methods for each purebred line were determined as candidate regions under positive selection. A Karyogram layout of candidate selection regions detected by both tests was created using the ggbio R package [76]. We used the genomic database search engine BioMart (http://www.biomart.org/) to identify genes in the candidate selection regions.

3.3 Results

In total, 1,079 putative selection regions were detected with *P*-values < 0.05 using XP-EHH test, and 1,018 putative selection regions were detected using the criterion of a 0.5% cutoff of XP-CLR scores. Regions detected using XP-EHH overlapped 31.53% of the regions that were identified using XP-CLR. Even though 328 overlapped regions were detected by both methods among five purelines, some regions detected by either the XP-EHH or XP-CLR tests were wide enough to overlap with more than one region detected by the other test. Therefore, in total, 224 and 321 unique regions were detected using XP-EHH and XP-CLR tests, respectively. In each line, approximately 11.09% of chicken genome was covered by regions detected by XP-EHH methods, while approximately 2.58% of chicken genome was covered by regions detected by XP-CLR methods. The overlapped selection regions in both methods only represented approximately 1.45% of chicken genome in each line. Selection regions detected by XP-EHH were much wider, mainly because the EHH test is an LD-based method, and LD is expected to extend over longer distances in regions under recent selection [77]. In contrast, regions detected by XP-CLR tests showed overall narrower and perhaps more accurate candidate regions than those detected by the XP-EHH tests. Thus, to narrow down regions that overlapped between

the two methods, we considered the 321 regions based on the XP-CLR test as the candidate selection regions. Their ranges of 321 regions are visualized in Figure 3.1.

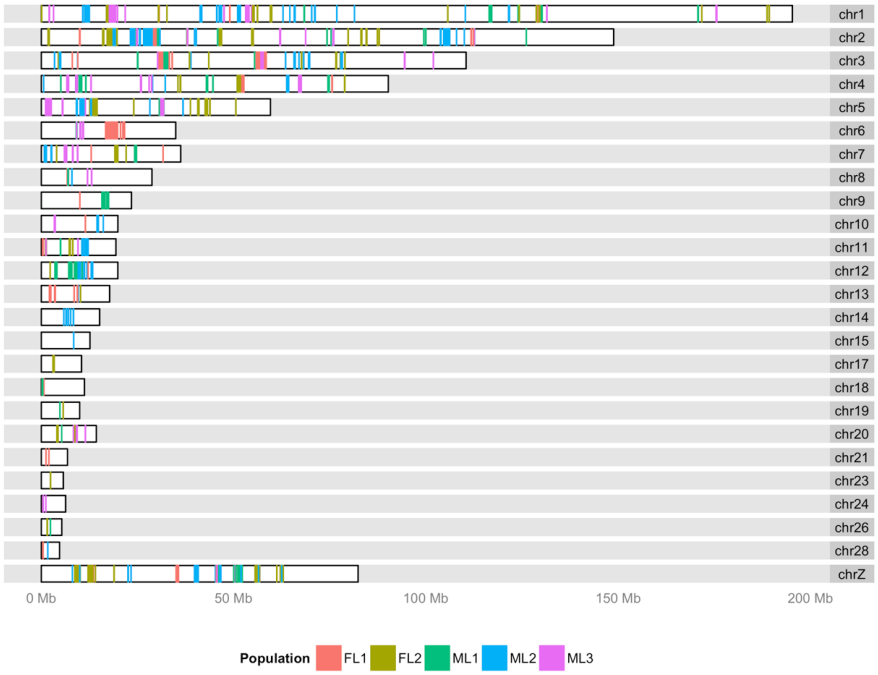


Figure 3.1 Candidate Selection Regions Detected by XP-EHH and XP-CLR Tests

Note: For each pureline, the overlapped regions detected by the XP-EHH and XP-CLR tests were presented based on the ranges from XP-CLR test. Each population is denoted by a different color.

By further examining these 321 regions, we identified 42 regions that were shared by two or more purebred lines (Appendix B). Because only common regions between two or more lines were counted in the overlapped regions detected in multiple lines, most of them (20 out of 42 regions) were smaller than 50kb while 4 regions were larger than 0.5Mb. Using BioMart, 91 genes could be found in the 42 regions (Appendix B) among which 9 regions were located at gene deserts and 19 regions only

harbored 1 or 2 genes. For the 9 regions located at gene deserts, the genes close to them (± 100kb) were listed on table in Appendix B.

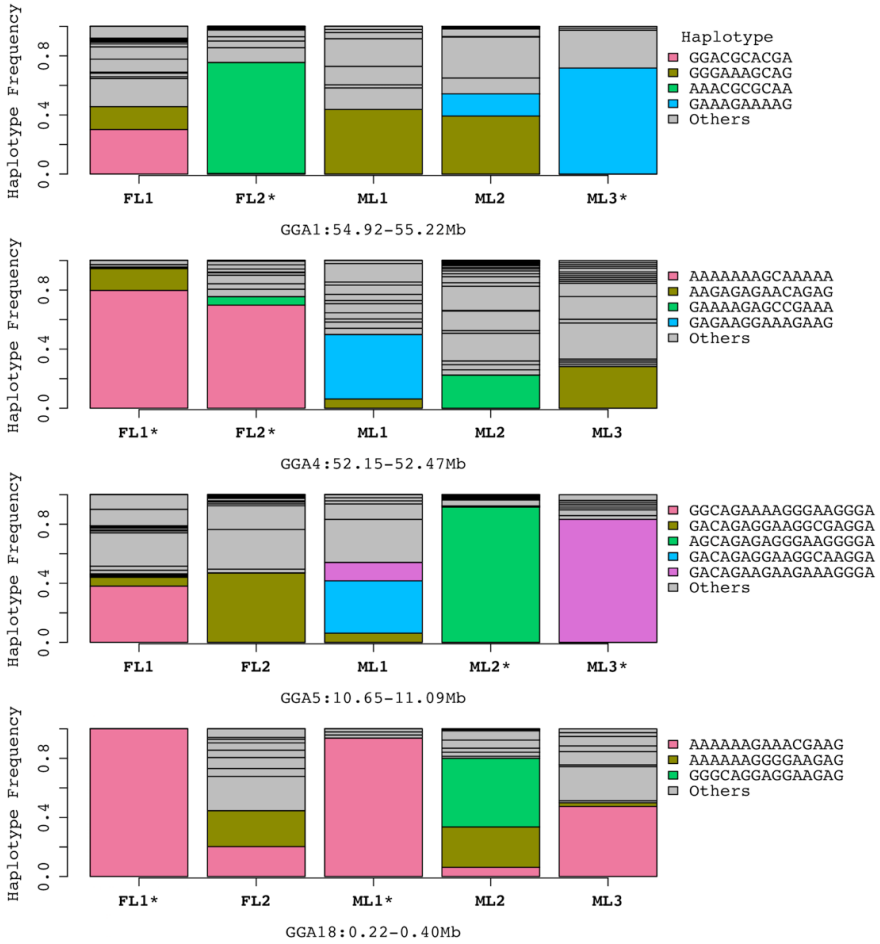


Figure 3.2 Haplotype Frequencies of Four Selection Regions Detected in Multiple Lines

Note: For a given region, major haplotype in each line was assigned a different color from grey. Selection signatures were detected in lines marked with "*".

To gain insight into population differences in the overlapped candidate regions, we constructed haplotypes and estimated haplotype frequencies in these regions in

each population (Appendix C). This analysis was performed only for 15 genomic regions with at least 5 SNPs in our genotype data. Figure 3.2 represents the results of haplotype analysis in four selection regions with 10 to 20 SNPs in our genotype data. As demonstrated in Figure 3.2, in these genomic regions, haplotypes with high frequency are present in the lines with a selection signature. For example, in a selection region on GGA4 (52.15-52.47Mb), the same haplotype showed high frequency in FL1 and FL2, although the range of this region was more than 300kb. Another interesting example is a ~240kb region on GGAZ (45.49-45.73Mb); male lines had the same major haplotype, which was different from the major haplotypes in the female lines.

3.4 Discussion

In modern broiler breeding, the practice of selective mating is utilized to influence the expression of economically important traits in subsequent generations. Through such selection, the "beneficial" alleles tend to become more frequent in populations over time. In our study, we applied XP-EHH and XP-CLR tests to detect the genomic regions under selection by measuring the characteristics of extended haplotype homozygosity and change in allele frequency spectrum, respectively. By cross-population comparisons of five commercial broiler purebred lines, we identified the genomic regions that are most likely to harbor genes related to traits of economic importance in broiler chickens.

It should be mentioned that genetic drift and genetic bottleneck have potential to influence the results of selection signature studies such as this one. However, with modeling genetic drift in reference population by Brownian motion, the XP-CLR method applied in this study was designed to identify the selection signatures where

differences in derived allele frequency between objective and reference populations were too large to be due to genetics drift [50]. Thus, to minimize influence of genetic drift, we only kept candidate selection regions detected by both XP-EHH and XP-CLR methods. Furthermore, bottleneck events did not occur in lines used in the present study for more than 40 generations. A potential limitation of our study is inherent to cross-population methods, which may fail to detect a selection signature where the desirable allele has been under similar level of positive or negative selection pressure in all the purebred lines that were studied in our study. However, this limitation should not be a major concern in our study since the lines that were used differed in the primary traits of selection. For example, the primary traits are growth rate for ML1, breast muscle yield for ML2 and feed efficiency for ML3. Therefore, alleles are expected to be under unequal selection pressure, or as explained below some alleles may have been selected in opposite direction between male and female lines.

3.4.1 Candidate Selection Regions

We compared the genes in the candidate selection regions in our study with those from two previous studies on detecting selection sweeps in chickens. Among 91 genes in the 42 regions in our study, only two genes (*SOX6* and *GJD2*) are in genomic regions detected by Rubin *et al* (2010) in commercial broilers. Also, only two genes (*GAS7* and *STXBP6*) in our list are among 366 genes (based on Ensembl gene ID) detected by Elferink et al. (2012). This low extent of overlap with previous studies is likely be related to the different methods that we used for detecting selection signatures in the present study. As mentioned before, the cross-population methods in our study are aimed at detecting recent selection signatures, whereas the ZHp method used in two previous studies is primarily focused on detecting older selection

signatures such as those accumulated during domestication. For better comparison, we estimated ZH scores (Z transformed average heterozygosity) over sliding 5-marker windows on autosomes using data from our study (Appendix D) and found that 31 out of 81 genes harbored in 41 selection regions detected by ZH scores were also detected in broilers from two pervious studies (Table D1 and Table D2 in Appendix D) although our resource populations were much different to those in the two previous studies.

In our results from two cross-population methods (XP-EHH and XP-CLR), 42 regions were detected by both methods in multiple populations, which might indicate that gene(s) in these regions have been independently selected in multiple populations, i.e., parallel selection. To examine haplotype diversity among the 5 purebred lines in these overlapped regions, we identified major haplotypes and estimated their frequency within each population. As shown in Figure 3.2, in a given selection region, the same major haplotypes may be shared by multiple lines, which may reflect selection for the same allele of a gene in these lines. However, in 9 of 15 candidate selection regions in the table of Appendix C, such as GGA1: 54.92-55.22Mb and GGA13: 9.38-9.44Mb, the major haplotype varied greatly between lines. The difference in major haplotype may represent high diversity of genetic background among these pure lines [77]. Alternatively, it is possible that selection has acted on different alleles of a gene in these lines, i.e., divergent selection. For example, previous studies found that fertility was reduced in chickens under strong selection for body weight due to the negative genetic correlation between reproduction and growth traits [78–80]. Overall, selection for reproduction traits has been more emphasized in female lines, whereas selection for high feed-efficiency and increased skeletal muscle

growth has been the major focus in male lines. Thus, the frequency of alleles benefiting reproduction traits but adversely affecting growth traits are expected to be relatively higher in female lines as compared with the male lines.

3.4.2 Candidate Genes in Regions Detected in Mulitple Populations

Chr	Start (bp)	End (bp)	Gene	Function or association
2	37844468	37965174	*cTR*	Growth, development and homeostasis
5	10549851	10784968	*SOX6*	Development of chondrocytes and skeletal muscle
5	31546121	31552902	*ACTC1*	Muscle development
5	31693846	31748417	*STXBP6*	Bone allocation and fecundity traits
18	297153	334809	*MYH13*	Skeletal muscle development
Z	56844264	56903872	*CAST**	Growth and muscle quality

Table 3.2 A Partial List of Candidate Genes Detected in/near the 42 Overlapped Regions

Note: * It means that the gene is located outside but close to a candidate selection region

In the 42 overlapped regions, we identified several genes related to growth, development, feed efficiency and reproduction in chickens (Table 3.2). A few of these genes are discussed below as examples of candidate genes under recent selection in multiple broiler chicken populations. For example, *MYH13 (myosin heavy chain (MyHC) 13)* gene is located in an overleaped region on GGA18 (0.22-0.40Mb), and other four genes (*MYH1A, MYH1B, MYH1C, MYH1E*) belong to MyHC family are located very close to this regions. Previous studies found that MyHC genes play important roles in skeletal muscle development [81–83], and the polymorphisms in *MYH3* were significantly associated with growth and carcass traits in Qinchuan cattle [84, 85]. Another muscle related gene, *SOX6 (Sex Determining Region Y-Box 6),* was

located in a candidate selection region (GGA5: 10.65-11.09Mb) detected in two male lines, ML2 and ML3, in our study. This gene encodes a Sry-related transcription factor that promotes early chondroblast differentiation and plays a critical role in differentiation and proliferation of chondrocytes as well as normal fiber type differentiation of fetal skeletal muscle in mice [86–88]. Moreover, although no gene was found inside an overlapped region around 56.76Mb on GGAZ due to its small size (8kb), *CAST* (*calpastatin*) is located around the region. Calpastatin is a specific inhibitor of endogenous calpain, and calpain family plays an important role in embryonic development and muscle growth [89–91]. Many studies have found the polymorphisms in *CAST* had significantly associated with growth traits and meat quality traits in livestock animals [92–97].

Moreover, another muscle-related gene, *ACTC1 (actin, alpha, cardiac muscle 1)*, was found in one of the selection regions on GGA5 (31.06-31.82Mb). This gene encodes cardiac muscle alpha actin in chickens and plays an important role in fetal development as well as cell survival, differentiation and development of muscle [98–101]. *Syntaxin binding protein 6* (*STXBP6)* is another gene in this candidate selection region on GGA5. A previous study has indicated *STXBP6* had potential pleiotropic effect on bone tissue and fecundity traits in chickens [102]. Interestingly, this selection sweep which was detected in two male lines (ML1 and ML3) in our study, was also found in a previous study in layer chickens [103]. One possible reason why this selection sweep is shared by broiler (meat-type) and layer (egg-type) chickens is that some economically important traits, such as body weight, are shared by layers' and broilers' breeding plan although their breeding strategies can be much different: breeder improved meat production in broilers by selection on high body weight and

growth rate while they improved feed efficiency in layers by selection on increased egg mass and low body weight [104–106]. Alternatively, this shared selection sweep may be explained by the pleiotropic effect of *STXBP6* on both bone tissue and fecundity traits.

A growth related gene, *Thyroid hormone receptor beta* (*cTR),* was found in or near four overlapped regions on GGA2 (37.90-38.07Mb). Thyroid hormone can regulate animal growth, development and homeostasis [107], and its receptor mediates thyroid hormone actions [108]. Mice with homozygous mutant *cTR* gene manifest low weight gain and decreased bone development compared to normal mice [109]. In a ~40kb-length candidate selection region (GGA2: 38.03-38.07Mb) which were detected in 4 pure lines (FL1, FL2, ML2 and ML3), there are 3 SNPs in our dataset, which construct the same major haplotype (AAA) in two female lines and ML2, but the major haplotype (GGG) in ML3 is completely different. It is possible that in this candidate region, *cTR* gene has been under divergent selection among these lines. For example, one allele has been selected in two female lines and ML2 while an alternative allele was selected in ML3. This assumption can be further supported considering diverse functions of thyroid hormone: it has been reported that thyroid hormone also plays a critical role in fertility, but excessive amounts of this hormone in hyperthyroidism has a negative effect on reproduction in humans [110, 111]. Therefore, the divergent effects of thyroid hormone on reproduction and growth traits may explain why the receptor gene, *cTR*, may have been under divergent selection among these lines.

3.5 Conclusions

Using 60k SNP genotypes of 565 chickens from 5 commercial pure lines and with cross-population comparison using two methods, our study uncovered novel candidate regions for recent selection in broiler chickens. Based on the biological function of genes in the candidate regions, several genes, such as *SOX6* and *cTR*, have possibly made large contributions to economically important traits in chickens. Our findings suggest that recent selection in broiler breeding has had large impact on frequency of genes controlling important traits, such as weight gain, muscle mass, feed efficiency and reproduction. Finally, since most of the candidate genes identified in the present study are novel and have been probably under recent selection, they should be of great interest for future research into the genetic architecture of traits relevant to modern broiler breeding.

REFERENCES

1. Havenstein GB, Ferket PR, Qureshi MA: **Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets.** *Poult Sci* 2003, **82**:1500–1508.

2. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES: **Linkage disequilibrium in the human genome.** *Nature* 2001, **411**:199–204.

3. Roach JC, Glusman G, Smit AF a, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328**:636–639.

4. Kimura M: *The Neutral Theory of Molecular Evolution*. Cambridge University Press; 1984.

5. Gillespie JH: *Population Genetics: A Concise Guide*. JHU Press; 2010.

6. Hayes B, Visscher P: **Novel multilocus measure of linkage disequilibrium to estimate past effective population size**. *Genome ...* 2003(March):635–643.

7. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM, Unit HG, Hospital WG, Eh E, Kingdom U, Biology E: **Recent human effective population size estimated from linkage disequilibrium**. 2007, **2**:520–526.

8. 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.

9. Stachowicz K, Sargolzaei M, Miglior F, Schenkel FS: **Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle.** *J Dairy Sci* 2011, **94**:5160–5175.

10. Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H: **Linkage disequilibrium reveals different demographic history in egg laying chickens.** *BMC Genet* 2010, **11**:103.

11. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: **Linkage disequilibrium**

and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 2008, **179**:1503–1512.

12. Uimari P, Tapio M: **Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds.** *J Anim Sci* 2011, **89**:609–614.

13. Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP: **Estimation of linkage disequilibrium in four US pig breeds.** *BMC Genomics* 2012, **13**:24.

14. Lu D, Sargolzaei M, Kelly M, Li C, Vander Voort G, Wang Z, Plastow G, Moore S, Miller SP: **Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle.** *Front Genet* 2012, **3**:152.

15. Aerts J, Megens HJ, Veenendaal T, Ovcharenko I, Crooijmans R, Gordon L, Stubbs L, Groenen M: **Extent of linkage disequilibrium in chicken.** *Cytogenet Genome Res* 2007, **117**:338–345.

16. Dekkers JCM: **Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons.** *J Anim Sci* 2004, **82 E-Suppl**.

17. Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP, VanRaden PM, Viaud-Martinez K a, Wiggans GR: **Design of a bovine low-density SNP array optimized for imputation.** *PLoS One* 2012, **7**:e34130.

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

19. Wolc A, Stricker C, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Habier D, Fernando R, Garrick DJ, Lamont SJ, Dekkers JCM: **Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model.** *Genet Sel Evol* 2011, **43**:5.

20. Fulton JE: **Genomic selection for poultry breeding**. *Anim Front* 2012, **2**:30–36.

21. Schaeffer LR: **Strategy for applying genome-wide selection in dairy cattle.** *J Anim Breed Genet* 2006, **123**:218–223.

22. Goddard ME, Hayes BJ: **Genomic selection.** *J Anim Breed Genet* 2007, **124**:323–330.

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

24. Toosi A, Fernando RL, Dekkers JCM: **Genomic selection in admixed and crossbred populations.** *J Anim Sci* 2010, **88**:32–46.

25. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin C-W, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, et al.: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438**:803–819.

26. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ, Comstock KE, Keller ET, Mesirov JP, von Euler H, Kämpe O, Hedhammar A, Lander ES, Andersson G, Andersson L, Lindblad-Toh K: **Efficient mapping of mendelian traits in dogs through genome-wide association.** *Nat Genet* 2007, **39**:1321–1328.

27. Setoguchi K, Furuta M, Hirano T, Nagao T, Watanabe T, Sugimoto Y, Takasuga A: **Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (CW-2) on chromosome 6 and the Ile-442-Met substitution in NCAPG as a positional candidate.** *BMC Genet* 2009, **10**:43.

28. Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM: **Genomic selection of purebreds for crossbred performance.** *Genet Sel Evol* 2009, **41**:12.

29. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177**:2389–2397.

30. 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, **41**:51.

31. Heifetz EM, Fulton JE, O'Sullivan N, Zhao H, Dekkers JCM, Soller M: **Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations.** *Genetics* 2005, **171**:1173–1181.

32. Andreescu C, Avendano S, Brown SR, Hassen A, Lamont SJ, Dekkers JCM: **Linkage disequilibrium in related breeding lines of chickens.** *Genetics* 2007, **177**:2161–2169.

33. Abasht B, Sandford E, Arango J, Settar P, Fulton JE, O'Sullivan NP, Hassen A, Habier D, Fernando RL, Dekkers JCM, Lamont SJ: **Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations**. *BMC Genomics* 2009, **10**(Suppl 2):S2.

34. Megens H-J, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH, Hanotte O, Groenen MAM: **Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken.** *BMC Genet* 2009, **10**:86.

35. Wragg D, Mwacharo JM, Alcalde JA, Hocking PM, Hanotte O: **Analysis of genome-wide structure, diversity and fine mapping of Mendelian traits in traditional and village chickens.** *Heredity (Edinb)* 2012, **109**:6–18.

36. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES: **Positive natural selection in the human lineage.** *Science* 2006, **312**:1614–1620.

37. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J V, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 2002, **419**:832–837.

38. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi a R, Simianer H: **A genome-wide scan for signatures of recent selection in Holstein cattle.** *Anim Genet* 2010, **41**:377–389.

39. Utsunomiya YT, Pérez O'Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, Mészáros G, Sölkner J, Garcia JF: **Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods.** *PLoS One* 2013, **8**:e64280.

40. Lichten M, Goldman a S: **Meiotic recombination hotspots.** *Annu Rev Genet* 1995, **29**:423–444.

41. Petes TD: **Meiotic recombination hot spots and cold spots.** *Nat Rev Genet* 2001, **2**:360–369.

42. Arnheim N, Calabrese P, Nordborg M: **Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved.** *Am J Hum Genet* 2003, **73**:5–16.

43. Chicken I, Sequencing G: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**:695–716.

44. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**:e72.

45. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, et al.: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913–918.

46. Kim Y, Stephan W: **Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome**. 2002, **777**(February):765–777.

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

48. Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, Megens H-J, Zhang H, Okimoto R, Vereijken A, Jungerius A, Albers GAA, Lawley CT, Delany ME, MacEachern S, Cheng HH: **Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds.** *Proc Natl Acad Sci U S A* 2008, **105**:17312–17317.

49. Groenen MAM, Megens H-J, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, Vereijken A, Okimoto R, Muir WM, Cheng HH: **The development and characterization of a 60K SNP chip for chicken.** *BMC Genomics* 2011, **12**:274.

50. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps**. *Genome Res* 2010, **20**:393–402.

51. Box GEP, Draper NR: *Empirical Model-Building and Response Surfaces*. *Volume 424*. Wiley New York; 1987.

52. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R: **Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition.** *Genome Res* 2002, **12**:222–231.

53. Boettcher PJ, Caroli A, Stella A, Chessa S, Budelli E, Canavesi F, Ghiroldi S, Pagnacco G: **Effects of casein haplotypes on milk production traits in Italian Holstein and Brown Swiss cattle.** *J Dairy Sci* 2004, **87**:4311–4317.

54. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, Nielsen R, Simianer H: **Classic selective sweeps revealed by massive sequencing in cattle.** *PLoS Genet* 2014, **10**:e1004148.

55. Rothammer S, Seichter D, Förster M, Medugorac I: **A genome-wide scan for**

signatures of differential artificial selection in ten cattle breeds. *BMC Genomics* 2013, **14**:908.

56. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L: **Whole-genome resequencing reveals loci under selection during chicken domestication.** *Nature* 2010, **464**:587–591.

57. Elferink MG, Megens H-J, Vereijken A, Hu X, Crooijmans RPM a, Groenen M a M: **Signatures of selection in the genomes of commercial and non-commercial chicken breeds.** *PLoS One* 2012, **7**:e32720.

58. Abasht B, Sandford E, Arango J, Settar P, Fulton JE, O'Sullivan NP, Hassen A, Habier D, Fernando RL, Dekkers JCM, Lamont SJ: **Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations.** *BMC Genomics* 2009, **10**:S2.

59. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**:210–223.

60. Sabatti C, Risch N: **Homozygosity and linkage disequilibrium.** *Genetics* 2002, **160**:1707–1719.

61. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**:263–265.

62. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225–2229.

63. Devlin B, Risch N: **A comparison of linkage disequilibrium measures for fine-scale mapping.** *Genomics* 1995, **29**:311–322.

64. Wright S: **The genetical structure of populations.** *Ann Eugen* 1951, **15**:323–354.

65. Hartl DL, Clark AG: *Principles of Population Genetics*. *Volume 116*. Sunderland: Sinauer associates; 1997.

66. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, Ostrander EA, Wayne RK: **Linkage disequilibrium and demographic history of wild and**

domestic canids. *Genetics* 2009, **181**:1493–1505.

67. Shea J, Agarwala V, Philippakis AA, Maguire J, Banks E, Depristo M, Thomson B, Guiducci C, Onofrio RC, Kathiresan S, Gabriel S, Burtt NP, Daly MJ, Groop L, Altshuler D: **Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction.** *Nat Genet* 2011, **43**:801–805.

68. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3**:299–309.

69. Nielsen R, Signorovitch J: **Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium**. *Theor Popul Biol* 2003, **63**:245–255.

70. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P, Hocking PM, Fife M, Salmon N, Fulton J, Strom TM, Haberer G, Weigend S, Preisinger R, Gholami M, Qanbari S, Simianer H, Watson KA, Woolliams JA, Burt DW: **Development of a high density 600K SNP genotyping array for chicken.** *BMC Genomics* 2013, **14**:59.

71. Pollock DL: **A geneticist's perspective from within a broiler primary breeder company.** *Poult Sci* 1999, **78**:414–418.

72. Consortium TBH: **Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds**. *Science* 2009, **324**:528–532.

73. Crawford RD: *Poultry Breeding and Genetics*. Elsevier; 1990.

74. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H, Crooijmans RPMA, Besnier F, Lathrop M, Muir WM, Wong GK-S, Gut I, Andersson L: **A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate.** *Genome Res* 2009, **19**:510–519.

75. Pickrell JJK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW: **Signals of recent positive selection in a worldwide sample of human populations**. *Genome Res* 2009, **19**:826–837.

76. Yin T, Cook D, Lawrence M: **ggbio: an R package for extending the grammar of graphics for genomic data**. *Genome Biol* 2012, **13**:R77.

77. Fu W, Dekkers JCM, Lee WR, Abasht B: **Linkage disequilibrium in crossbred and pure line chickens**. *Genet Sel Evol* 2015, **47.1**:11.

78. Liu G, Dunnington EA, Siegel PB: **Correlated responses to long-term divergent selection for eight-week body weight in chickens: growth, sexual maturity, and egg production.** *Poult Sci* 1995, **74**:1259–1268.

79. Anthony NB, Dunnington E a., Siegel PB: **Egg Production and Egg Composition of Parental Lines and F1 and F2 Crosses of White Rock Chickens Selected for 56-Day Body Weight**. *Poult Sci* 1989, **68**:27–36.

80. Jaap RG, Muir F V.: **Erratic Oviposition and Egg Defects in Broiler-type Pullets**. *Poult Sci* 1968, **47**:417–423.

81. Lagrutta AA, McCarthy JG, Scherczinger CA, Heywood SM: **Identification and developmental expression of a novel embryonic myosin heavy-chain gene in chicken**. *DNA* 1989, **8**:39–50.

82. Lyons G, Ontell M, Cox R: **The expression of myosin genes in developing skeletal muscle in the mouse embryo.** *J cell ...* 1990, **111**(October):1465–1476.

83. Contribution O: **Embryonic myosin heavy-chain mutations cause distal arthrogryposis and developmental myosin myopathy that persists postnatally**. 2014, **65**:1083–1091.

84. Wang L, Liu X, Niu F: **Single nucleotide polymorphisms , haplotypes and combined genotypes in MYH3 gene and their associations with growth and carcass traits in Qinchuan cattle**. 2013, **3**:417–426.

85. Niu F, Wang L, Liu X, Wang H, Yang J: **Genetic diversity of MYH3 gene associated with growth and carcass traits in Chinese Qinchuan cattle**. *Mol Biol Rep* 2013, **40**:5635–5643.

86. Smits P, Li P, Mandel J, Zhang Z, Deng JM, Behringer RR, de Crombrugghe B, Lefebvre V: **The transcription factors L-Sox5 and Sox6 are essential for cartilage formation.** *Dev Cell* 2001, **1**:277–290.

87. Smits P, Dy P, Mitra S, Lefebvre V: **Sox5 and Sox6 are needed to develop and maintain source, columnar, and hypertrophic chondrocytes in the cartilage growth plate.** *J Cell Biol* 2004, **164**:747–758.

88. Hagiwara N, Yeh M, Liu A: **Sox6 is required for normal fiber type differentiation of fetal skeletal muscle in mice.** *Dev Dyn* 2007, **236**:2062–2076.

89. Goll DE, Thompson VF, Taylor RG, Christiansen JA, Biology M: **Role of the caipain system in muscle growth**. 1992:225–237.

90. Huang J, Forsberg NE: **Role of calpain in skeletal-muscle protein degradation**. *Proc Natl Acad Sci* 1998, **95**:12100–12105.

91. Arthur JSC, Elce JS, Hegadorn C, Williams K, Greer PA: **Disruption of the Murine Calpain Small Subunit Gene , Capn4 : Calpain Is Essential for Embryonic Development but Not for Cell Growth and Division**. 2000, **20**:4474–4481.

92. Lindholm-Perry AK, Rohrer GA, Holl JW, Shackelford SD, Wheeler TL, Koohmaraie M, Nonneman D: **Relationships among calpastatin single nucleotide polymorphisms, calpastatin expression and tenderness in pork longissimus1**. *Anim Genet* 2009, **40**:713–721.

93. Hu Y-D, Zhang Z-R, Zhu Q: **Identification and Association of the Single Nucleotide Polymorphisms in Calpastatin (CAST) Gene with Carcass Traits in Chicken**. *J Anim Vet Adv* 2011, **10**:2968–2974.

94. Allais S, Levéziel H, Hocquette JF, Lepetit J, Rousset S: **Effects Of Polymorphisms In The Calpastatin And μ- Calpain Genes On Meat Tenderness In Three French Beef Breeds**. 2000.

95. Byun SO, Zhou H, Forrest RHJ, Frampton CM, Hickford JGH: **Association of the ovine calpastatin gene with birth weight and growth rate to weaning**. *Anim Genet* 2008, **39**:572–573.

96. Greguła-Kania M: **Effect of calpastatin gene polymorphism on lamb growth and muscling**. *Ann Anim Sci* 2012, **12**:63–72.

97. Tait RG, Shackelford SD, Wheeler TL, King DA, Casas E, Thallman RM, Smith TPL, Bennett GL: **μ-Calpain , calpastatin , and growth hormone receptor genetic effects on preweaning performance , carcass quality traits , and residual variance of tenderness in Angus cattle selected to increase minor haplotype and allele frequencies 1 , 2 , 3**. 2014:456–466.

98. Wu T, Zhang Z, Yuan Z, Lo LJ, Chen J, Wang Y, Peng J: **Distinctive Genes Determine Different Intramuscular Fat and Muscle Fiber Ratios of the longissimus dorsiMuscles in Jinhua and Landrace Pigs.** *PLoS One* 2013, **8**.

99. Gunning P, O'neill G, Hardeman E: **Tropomyosin-based regulation of the actin cytoskeleton in time and space**. *Physiol Rev* 2008, **88**:1–35.

100. Kuninger D, Kuzmickas R, Peng B, Pintar JE, Rotwein P: **Gene discovery by microarray: identification of novel genes induced during growth factor-mediated muscle cell survival and differentiation**. *Genomics* 2004, **84**:876–889.

101. McNally E, Dellefave L: **Sarcomere mutations in cardiogenesis and ventricular noncompaction**. *Trends Cardiovasc Med* 2009, **19**:17–21.

102. Johnsson M, Rubin C, Höglund A, Sahlqvist A, Jonsson KB, Kerje S, Ekwall O, Kämpe O, Andersson L, Jensen P: **The role of pleiotropy and linkage in genes affecting a sexual ornament and bone allocation in the chicken**. *Mol Ecol* 2014, **23**:2275–2286.

103. Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, Turner F, Burt DW, Preisinger R, Gianola D, Simianer H: **A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens**. *PLoS One* 2012, **7**:49525.

104. Muir WM, Aggrey SE: *Poultry Genetics, Breeding, and Biotechnology*. CABI; 2003.

105. Thiruvenkadan AK, Panneerselvam S, Prabakaran R: **Layer breeding strategies : an overview**. 2010, **66**(September):477–502.

106. Thiruvenkadan a. K, Prabakaran R, Panneerselvam S: **Broiler breeding strategies over the decades: an overview**. *Worlds Poult Sci J* 2011, **67**:309–336.

107. Zhou H, Mitchell AD, McMurtry JP, Ashwell CM, Lamont SJ: **Insulin-like growth factor-I gene polymorphism associations with growth, body composition, skeleton integrity, and metabolic traits in chickens.** *Poult Sci* 2005, **84**:212–219.

108. Lazar MA: **Thyroid hormone receptors: multiple forms, multiple possibilities**. *Endocr Rev* 1993, **14**:184–193.

109. Kaneshige M, Kaneshige K, Zhu X, Dace A, Garrett L, Carter TA, Kazlauskaite R, Pankratz DG, Wynshaw-boris A, Refetoff S, Weintraub B, Willingham MC, Barlow C, Cheng S: **Mice with a targeted mutation in the thyroid hormone β receptor gene exhibit impaired growth and resistance to thyroid hormone**. 2000, **97**.

110. Dittrich R, Beckmann MW, Oppelt PG, Hoffmann I, Lotz L, Kuwert T, Mueller A: **Thyroid hormone receptors and reproduction**. *J Reprod Immunol* 2011, **90**:58–66.

111. Krassas GE, Poppe K, Glinoer D: **Thyroid function and human reproductive health.** *Endocr Rev* 2010, **31**:702–755.

112. Abasht B, Dekkers JCM, Lamont SJ: **Review of quantitative trait loci identified in the chicken.** *Poult Sci* 2006, **85**:2079–2096.

113. Florini JR, Ewton DZ, Coolican SA: **Growth hormone and the insulin-like growth factor system in myogenesis**. *Endocr Rev* 1996, **17**:481–517.

114. Froesch ER, Schmid C, Schwander JT, Zapf J: **Actions of insulin-like growth factors**. *Annu Rev Physiol* 1985, **47**:443–467.

115. Hammerman M: **The growth hormone-insulin-like growth factor axis in kidney**. *Am J Physiol* 1989, **257**(4 Pt 2):F503–F514.

116. Mohan S, Baylink DJ: **Impaired skeletal growth in mice with haploinsufficiency of IGF-I: genetic evidence that differences in IGF-I expression could contribute to peak bone mineral density differences.** *J Endocrinol* 2005, **185**:415–420.

117. Liu JL, Grinberg A, Westphal H, Sauer B, Accili D, Karas M, LeRoith D: **Insulin-like growth factor-I affects perinatal lethality and postnatal development in a gene dosage-dependent manner: manipulation using the Cre/loxP system in transgenic mice.** *Mol Endocrinol* 1998, **12**:1452–1462.

118. Liu JP, Baker J, Perkins AS, Robertson EJ, Efstratiadis A: **Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r).** *Cell* 1993, **75**:59–72.

119. Derek R Le, Flier JS, Underhill LH, Le Roith D: **Insulin-like growth factors**. *N Engl J Med* 1997, **336**:633–640.

120. Amills M, Jiménez N, Villalba D, Tor M, Molina E, Cubiló D, Marcos C, Francesch A, Sànchez A, Estany J: **Identification of three single nucleotide polymorphisms in the chicken insulin-like growth factor 1 and 2 genes and their associations with growth and feeding traits.** *Poult Sci* 2003, **82**:1485–1493.

121. Bennett a K, Hester PY, Spurlock DEM: **Polymorphisms in vitamin D receptor, osteopontin, insulin-like growth factor 1 and insulin, and their associations with bone, egg and growth traits in a layer--broiler cross in chickens.** *Anim Genet* 2006, **37**:283–286.

122. Mul JD, Yi C, van den Berg SAA, Ruiter M, Toonen PW, van der Elst MCJ, Voshol PJ, Ellenbroek BA, Kalsbeek A, la Fleur SE, Cuppen E: **Pmch expression during early development is critical for normal energy homeostasis.** *Am J Physiol Endocrinol Metab* 2010, **298**:E477–E488.

123. Sun G, Li M, Li H, Tian Y, Chen Q, Bai Y, Kang X: **Molecular cloning and SNP association analysis of chicken PMCH gene.** *Mol Biol Rep* 2013, **40**:5049–55.

Appendix A

SUMMARY OF HAPLOBLOCK STRUCTURE IN DIFFERENT CHROMOSOME CATEGORIES

Table A.1 Summary Statistics of Haploblock Structure in Different Chromosome Categories

Chromosome Type	Statistics	Populations[1]			
		B	C	D	fBCD
Macro	MedBS[2] (kb)	42.8	48.8	40.0	36.1
	MaxBS[3] (kb)	3,521.9	3,527.6	4,226.0	2,737.2
	TSNPs[4]	13,535	13,353	12,324	6,956
Intermediate	MedBS[2] (kb)	30.2	38.4	31.2	25.1
	MaxBS[3] (kb)	1,810.0	1,868.0	1,640.0	1,421.0
	TSNPs[4]	4,402	4,255	4,007	2,108
Micro	MedBS[2] (kb)	16.9	22.2	17.0	14.0
	MaxBS[3] (kb)	1,330.0	975.0	774.2	794.3
	TSNPs[4]	8,356	8,112	7,432	4,311

Note: Statistics of haplotype structure are presented for different chromosome categories in chicken genome.

[1] B: line B; C: line C; D: Line D; fBCD: field crossbred chickens;

[2] Median haploblock size;

[3] Maximum haplolock size;

[4] Total number of SNPs forming haploblocks.

Appendix B

CANDIDATE SELECTION REGIONS IN MULTIPLE POPULATIONS

Table B.1 Candidate Selection Regions Detected by XP-EHH and XP-CLR Tests in Multiple Populations

Chr.	Region Start (bp)	Region End (bp)	Population	No. Gene	Gene Symbol
1	30507595	30545595	FL1,FL2	1	*ANO6*
1	54923595	55215595	ML3,FL2	1	*STAB2*
1	124221595	124223595	ML1,FL2	3*	*MID1;CLCN4;WWC3*
1	124241595	124291595	ML1,FL2	2	*CLCN4;WWC3*
1	128735595	128769595	FL1,FL2	1*	*CRLF2*
1	129549595	129629595	ML1,FL2	2	*RP2;SLC9A7*
2	18696746	19226746	ML2,FL2	6	*ENSGALG00000007962;Diet1;ENSGALG00000026391;ARL5B;NSUN6;CACNB2*
2	24796746	24964746	ML1,ML3	6	*ASNS;C1GALT1;COL28A1;MIOS;RPA3;gga-mir-1685*
2	29210746	29986746	ML2,FL1	6	*HDAC9;TWISTNB;ENSGALG00000019648;TMEM196;ENSGALG00000026059;FERD3L*
2	37896746	37952746	ML2,FL2	1	*cTR*
2	37952746	38002746	ML2,FL1,FL2	2*	*cTR;uc_338*
2	38002746	38028746	ML2,FL2	3*	*cTR;RARB;uc_338*
2	38028746	38066746	ML2,FL1, ML3,FL2	3*	*cTR;RARB;uc_338*
2	112576746	112600746	ML1,ML3	3*	*RAB2A;CHD7;gga-mir-1557*
3	9477236	9489236	ML2,FL1	1	*VPS54*
3	31915236	32869236	ML1,FL1	2	*SULT6B1;CEBPZ*
3	38561236	38579236	ML2,FL2	1	*SIPA1L2*
3	38795236	38903236	ML2,FL1	1	*DISC1*
3	68203236	68245236	ML2,FL2	1	*PREP*
4	8966567	9078567	ML1,ML3	2	*RPS6KA6;ENSGALG00000020298*
4	9268567	9366567	ML1,ML3	5	*ENSGALG00000007110;ENSGALG00000007118;ENSGALG00000007123;SH3BGRL;ENSGALG00000007131*

Table B Candidate Selection Regions Detected by XP-EHH and XP-CLR in Multiple Populations (Continued)

4	52154567	52474567	FL1,FL2	1	*ANKRD50*
5	10653054	11089054	ML2,ML3	3	*SOX6;C5H11ORF58;PLEKHA7*
5	12813054	13015054	ML2,FL2	4	*KCNQ1;TRPM5;TSSC4;TAPA1*
5	31057054	31817054	ML1,ML3	7	*DPH6;ZNF770;AQR;ACTC1;GJD2;STXBP6;ENSGALG00000026078*
6	9181500	9189500	ML1,ML3	1	*Band*
8	6908377	7040377	ML1,ML2, FL1	3	*RFWD2;TNR;SCARNA3*
11	1095919	1147919	FL1,FL2	3	*FAM65A;CTCF;RLTPR*
11	1233919	1237919	ML2,ML3	1	*KCTD19*
12	10631819	10643819	ML1,ML2	2	*CFAP100;gga-mir-1731*
13	9380016	9440016	ML2,FL1	8	*SFXN1;KIF20A;BRD8;NHP2;N4BP3;B4GAL-T7;LMAN2; ENSGALG00000003134*
18	221750	403750	ML1,FL1	4	*GAS7;MYH13;ENSGALG00000027323;ENSGALG00000028102*
Z	9535533	9727533	FL1,FL2	5	*GOLPH3;MTMR12;ENSGALG00000003235;SUB1;SNORA66*
Z	10001502	10025533	ML2,FL2	1	*ADAMTS12*
Z	45489533	45725533	ML2,FL1	4	*SLC27A6;YTHDC2;MCC;U2*
Z	45725533	45781533	ML2,FL1, ML3	1	*DCP2*
Z	46107533	46277533	ML1,ML3	3	*EPB41L4A;NREP;SNORA13*
Z	46387533	46389533	ML2,ML3	1	*CAMK4*
Z	50415533	50461502	ML1,ML3	1*	*ST8SIA4*
Z	51937502	51939502	ML1,ML2	3*	*NANS;CLTA;GNE*
Z	56761533	56769533	ML2,ML3, FL2	3*	*fbn2;ERAP1;CAST*
Z	62481533	62525533	ML2,FL2	1	*VCAN*

Note: Some genes are overlapped among different selection regions because some regions are close to each others; * the genes close to the region (± 100kb) and none gene is annotated inside the region.

Appendix C

MAJOR HAPLOTYPE AND ITS FREQUENCY IN SELECTION SIGNATURES

Table C.1 Major Haplotype and Its Frequency in Selection Signatures

Chr.	Region Start (bp)	Region End (bp)	Population	No. SNP	FL1		FL2		ML1		ML2		ML3	
					MH	fMH	MH	fMH	MH	fMH	MH	fMH	MH	fMH
1	54923595	55215595	FL2,ML3	10	H1	0.30	H2	0.75	H3	0.44	H3	0.39	H4	0.72
2	18696746	19226746	FL2,ML2	22	H1	0.34	H2	0.89	H3	0.38	H2	0.83	H2	0.24
2	24796746	24964746	ML1,ML3	7	H1	0.34	H2	0.44	H1	0.83	H3	0.55	H1	0.91
2	29210746	29986746	FL1,ML2	32	H1	0.78	H2	0.27	H3	0.44	H4	0.74	H4	0.26
3	31915236	32869236	FL1,ML1	38	H1	0.77	H2	0.59	H2	0.46	H3	0.21	H4	0.28
4	52154567	52474567	FL1,FL2	14	H1	0.80	H1	0.70	H2	0.44	H3	0.22	H4	0.28
5	10653054	11089054	ML2,ML3	18	H1	0.38	H2	0.47	H3	0.35	H4	0.92	H5	0.83
5	12813054	13015054	FL2,ML2	5	H1	0.27	H2	0.97	H3	0.44	H4	0.61	H5	0.35
5	31057054	31817054	ML1,ML3	29	H1	0.46	H1	0.39	H1	0.96	H1	0.57	H1	0.82
8	6908377	7040377	FL1,ML1, ML2	8	H1	0.73	H2	0.34	H2	0.77	H2	0.99	H2	0.95
13	9380016	9440016	FL1,ML2	7	H1	0.49	H2	0.33	H3	0.69	H4	0.63	H4	0.19
18	221750	403750	FL1,ML1	15	H1	1.00	H2	0.24	H1	0.94	H3	0.46	H1	0.47
Z	9535533	9727533	FL1,FL2	6	H1	0.98	H1	0.95	H2	0.35	H2	0.66	H2	0.59
Z	45489533	45725533	FL1,ML2	7	H1	0.96	H2	0.41	H3	0.65	H3	1.00	H3	0.96

Note: MH: major haplotype symbol; fMH: Major haplotype frequency.

Appendix D

DETECTING SELECTION SIGNATURES USING ZH SCORES

To improve our comparison with previous studies by Rubin et al. (2010) and Elferink et al. (2012) [56, 57], we estimated ZH scores (Z transformed average heterozygosity) over sliding 5-marker windows on autosomes using data from our study. For estimation of ZH scores, we used the similar equation that was adopted by Elferink et al. (2012):

$$H_i = \frac{2 \sum n_{MAJ} \sum n_{MIN}}{\left(\sum n_{MAJ} + \sum n_{MIN}\right)^2}$$

$$ZH_i = \frac{H_i - \mu H}{\sigma H}$$

where H_i is the heterozygosity of pureline i; $\sum n_{MAJ}$ and $\sum n_{MIN}$ are the sum of major and minor allele frequencies, respectively, within a window; μH is the overall average heterozygosity and σH is the standard deviation for all windows. Unlike the previous study in which allele frequencies were estimated based on the 60k SNP genotyping of DNA pools from 13 broiler lines, we estimated allele frequencies based on individual genotypes, obtained using the same SNP array, of 565 birds from 5 broiler lines. To calculate the ZH score, we first estimated allele frequencies of SNPs within each pure line, and then averaged the allele frequencies across all 5 lines.

In total, we identified 41 significant candidate selection regions with a ZH score smaller than -4, and 12 of these regions (29.3%) overlapped with regions detected by Elferink *et al*. (2012) (Table D1 below). Also, 81 genes could be identified

in the 41 selection regions detected by ZH scores. Of these 81 genes, 22 genes overlapped with findings in Robin *et al*. (2010) and 20 genes overlapped in the Elferink *et al*. study (2012). In summary, 31 genes (38%) overlapped with these previous findings (Table D2 below), and 11 of these 31 genes are detected in two prior studies as well as in our study, including *IGF1, PMCH, PARPBP, NUP37, CCDC53, DRAM1, GNPTAB*, *TBXAS1, TPK1, HNF4G* and *CTK1*. Of these 11 genes, the first 8 genes are located at 55.43–56.14Mb on GGA1. This region on GGA1 is also known for QTL affecting body weight, abdominal fat and thigh muscle weight in experimental broiler chickens [112]. The most likely candidate genes under selection in this well-known QTL could be *insulin-like growth factor 1 (IGF1)* and *pro-melanin-concentrating hormone (PMCH*). Insulin-like growth factor I (IGF-I), encoded by *IGF1,* is a polypeptide hormone that stimulates the proliferation, differentiation and metabolism of myogenic cell lines in different species [113]. The importance of IGFs in growth and maintenance of various tissues has been well-established [114–119]. Previous studies showed that polymorphisms in *IGF1* were significantly associated with many important traits in broilers, such as growth, body composition and feeding traits [107, 120, 121]. On the other hand, in rat, loss of *PMCH* affected energy expenditure and resulted in a 20% lower set point for body weight [122]. Polymorphisms in *PMCH* were found to be significantly associated with growth and meat quality traits in chickens [123].

It should be noted our resource populations (5 broiler purelines) were much different from those in the two previous studies. There were 13 broiler purelines, including 8 male (sire) lines and 5 female (dam) lines, in Elferink *et al.* (2012) and only two commercial broiler lines in Rubin *et al.* (2010). Also, based on our records,

these broiler pure lines in our study have been reproductively isolated for more than 40 generations, except a one-time crossbreeding with ML2 in FL1, which means these purelines have no recent intercross with the chicken lines used in two previous study. And our results showed that the selection regions detected by the cross-population methods (XP-EHH and XP-CLR) were not consistent with the regions by the ZHp method in two previous studies. However, selection regions detected by ZH scores were notably overlapped with findings in previous studies. The cross-population methods in our study are aimed at detecting recent selection signatures, whereas the ZHp method used in two previous studies is primarily focused on detecting older selection signatures such as those accumulated during domestication. Thus, the most possible reason of high overlap using ZH scores is that some older selection signatures during chicken domestication were shared with commercial broilers used in our study as well as two previous studies.

Table D.1 Candidate Selection on Autosomes Regions Using ZH Scores

Chr.	Region	Min ZH Scores*	Overlaped Regions on Galgal 4.0 by Elfernick et al (2012)	Overlaped Regions on Galgal 2.1 by Elfernick et al (2012)
1	17325206-17446734	-4.418184661		
1	55318088-55688710	-4.375274826	55270842-55742636	57160808-57640392
1	55890067-56127275	-5.125910409		
1	94794472-94910370	-5.014038076	94765422-94910370	98803036-98949817
1	96848298-96934530	-5.058494147		
1	100172772-100256420	-4.86413964		
1	101247403-101332268	-4.087064733	101247403-101332268	105347850-105433972
2	51162211-51283333	-5.624307335		
2	52985076-53150206	-4.309988163	53019497-53182786	54058591-54230105
2	54950540-55117267	-4.601780555		
2	118411866-118805768	-6.071187532	118363604-118805768	123456560-123902432
2	123103337-123190653	-4.154940939		
2	123245460-123367332	-4.099824187		
2	132700572-132866538	-5.206119561		

Table D.1 Candidate Selection on Autosomes Regions Using ZH Scores (Continued)

Chr.	Region	Min ZH Scores*	Overlaped Regions on Galgal 4.0 by Elfernick et al (2012)	Overlaped Regions on Galgal 2.1 by Elfernick et al (2012)
2	132700572-132866538	-5.206119561		
2	134352222-134442514	-4.40592918		
4	19743236-19835780	-4.684396983	19743236-19835780	21353385-21446397
4	25764337-25859452	-4.511174785	25764337-25859452	27390369-27486328
4	28035422-28230422	-4.187252371		
4	41084642-41198563	-4.545590581		
4	55064448-55217732	-4.776249313		
5	2313223-2476419	-6.641726542	2313223-2476419	2344535-2509232
5	17167644-17318490	-4.604209916		
5	28368731-28506092	-4.240303547		
5	29108062-29266586	-4.080978837	29135573-29295758	32166885-32328626
5	31025580-31147825	-4.220523368		
5	31507804-31832717	-5.80300688		
5	33007063-33120931	-4.497099368		
7	5643745-5905521	-4.068939155		
7	6684032-6889263	-5.738147146		
7	7509112-7650094	-5.436397921		
7	8676490-8758220	-4.212718653		
7	34593699-34664942	-4.987138098		
9	9038758-9119497	-4.095622623		
11	15147420-15179611	-4.069899318		
12	19551322-19635007	-4.267147764	19488793-19625157	20149599-20289120
19	9671672-9749988	-4.024241821	9671672-9781558	9652495-9762375
20	11419918-11452543	-4.470396511		
25	1020499-1038371	-4.037848079		
27	1800098-1822575	-4.729582369		
28	3718756-3767132	-4.048501113		
28	3880366-4012896	-5.039573006	3958560-4012896	3747817-3803959

Note: * Minimum ZH Scores among multiple 5-marker windows

Table D.2 Overlap Candidate Genes with Previous Studies

Ensembl Gene ID	Gene Name	Chr.	Gene Start (bp)	Gene End (bp)	Reference
ENSGALG00000012755	*IGF-I*	1	55432530	55480957	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012757	*PMCH*	1	55523694	55525080	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012756	*PARPBP*	1	55520474	55569389	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012759	*NUP37*	1	55569414	55593570	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012760	*CCDC53*	1	55601447	55616701	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012761	*DRAM1*	1	55643576	55658917	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012763	*GNPTAB*	1	55671166	55703413	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000012788	*PARP12*	1	55884056	55900238	Elferink et al., 2012
ENSGALG00000012791	*TBXAS1*	1	55904115	56140985	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000015737	*NCAM2*	1	100155276	100281469	Elferink et al., 2012
ENSGALG00000012369	*TPK1*	2	52757783	53010513	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000015670	*HNF4G*	2	118421432	118442821	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000016111	*CSMD3*	2	132812344	133381395	Elferink et al., 2012
ENSGALG00000009732	*PCDH18*	4	28118550	28128788	Rubin et al., 2010
ENSGALG00000003777	*NELL1*	5	2243250	2526700	Elferink et al., 2012
ENSGALG00000009845	*GJD2*	5	31568439	31572498	Rubin et al., 2010
ENSGALG00000009847	*STXBP6*	5	31693846	31748417	Elferink et al., 2012
ENSGALG00000005974	*COL6A1*	7	6739939	6761011	Rubin et al., 2010
ENSGALG00000006126	*COL6A2*	7	6781274	6810349	Rubin et al., 2010
ENSGALG00000006131	*FTCD*	7	6812146	6818978	Rubin et al., 2010
ENSGALG00000006133	-	7	6832281	6832803	Rubin et al., 2010
ENSGALG00000025432	-	7	6843583	6843666	Rubin et al., 2010
ENSGALG00000004363	*MCM3AP*	7	6849401	6877820	Rubin et al., 2010
ENSGALG00000006141	*FUT13*	7	6880141	6893211	Rubin et al., 2010
ENSGALG00000008473	*PLXND1*	12	19511694	19575586	Elferink et al., 2012
ENSGALG00000008487	*TMCC1*	12	19596795	19653917	Elferink et al., 2012
ENSGALG00000003579	*CTK-1*	28	3878507	3912897	Rubin et al., 2010; Elferink et al., 2012
ENSGALG00000003703	*USE1*	28	3920735	3924954	Rubin et al., 2010
ENSGALG00000003717	*MYO9B*	28	3926426	3956759	Rubin et al., 2010
ENSGALG00000003727	*HAUS8*	28	3966089	3971791	Elferink et al., 2012
ENSGALG00000003742	*CPAMD8*	28	3981682	4015690	Elferink et al., 2012

Note: "-" means no assoicated gene name

Appendix E

REPRINT PERMISSION LETTER

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Mar 17, 2016

This Agreement between Weixuan Fu ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	3831470177115
License date	Mar 17, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Animal Genetics
Licensed Content Title	A genome-wide scan for signatures of recent selection in Holstein cattle
Licensed Content Author	S. Qanbari,E. C. G. Pimentel,J. Tetens,G. Thaller,P. Lichtner,A. R. Sharifi,H. Simianer
Licensed Content Date	Jan 21, 2010
Pages	13
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	2
Original Wiley figure/table number(s)	Figure 2 Figure 4
Will you be translating?	Yes, including English rights
Number of languages	1
Languages	English

Title of your thesis / dissertation	Linkage disequilibrium and recent selection signatures in commercial broilers
Expected completion date	Jun 2016
Expected size (number of pages)	80
Requestor Location	Weixuan Fu 404 STAMFORD DR APT 2A NEWARK, DE 19711 United States Attn: Weixuan Fu
Billing Type	Invoice
Billing Address	Weixuan Fu 404 STAMFORD DR APT 2A NEWARK, DE 19711 United States Attn: Weixuan Fu
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the

grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,**You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE

MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in

writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles

under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Libraryhttp://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.