2015

Whole genome analyses in cattle and chickens using Bayesian methods

Ziqing Weng
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agriculture Commons, Animal Sciences Commons, Biostatistics Commons, and the Genetics Commons

Recommended Citation
Weng, Ziqing, "Whole genome analyses in cattle and chickens using Bayesian methods" (2015). Graduate Theses and Dissertations. 14874.
https://lib.dr.iastate.edu/etd/14874

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Whole genome analyses in cattle and chickens using Bayesian methods

by

Ziqing Weng

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Animal Breeding and Genetics (Quantitative Genetics)

Program of Study Committee:
Dorian J. Garrick, Major Professor
   Alicia L. Carriquiry
   Jack C. M. Dekkers
   Rohan L. Fernando
   Daniel S. Nettleton

Iowa State University
   Ames, Iowa
2015

Copyright © Ziqing Weng, 2015. All rights reserved.
DEDICATION

Dedicated to my family and friends, for their love, support, and guidance all these years.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii
LIST OF FIGURES ............................................................... xi
ACKNOWLEDGEMENTS .......................................................... xvi
ABSTRACT ................................................................. xviii

CHAPTER 1. GENERAL INTRODUCTION ................................. 1
  1.1 Introduction .............................................................. 1
  1.2 Research Objectives ..................................................... 6
  1.3 Organization of thesis ................................................... 7
  1.4 Literature review ......................................................... 8
    1.4.1 Genetic analyses of variation in meiotic recombination ....... 8
    1.4.2 Bayesian regression models ...................................... 11
  1.5 Bibliography ............................................................. 13

CHAPTER 2. RECOMBINATION LOCATIONS AND RATES IN BEEF CAT-
TLE ASSESSED FROM PARENT-OFFSPRING PAIRS .................. 24
  2.1 Abstract ................................................................. 24
  2.2 Introduction ............................................................. 25
  2.3 Methods ................................................................. 27
    2.3.1 Genotype and phenotype ......................................... 27
    2.3.2 Halotype phasing .................................................. 28
    2.3.3 Estimating recombination events ................................. 28
    2.3.4 Estimating heritabilities ........................................ 30
2.3.5 Genome-wide association study ................................. 32
2.3.6 Imputation from low- to high-density and cross-validation of imputation ac-
curacy ................................................................. 33
2.4 Results ........................................................................ 34
  2.4.1 Phasing accuracy and crossover probability .................. 34
  2.4.2 Number of haplotypes and recombination rates ............... 35
  2.4.3 Analysis of genome-wide recombination number (GRN) ...... 36
  2.4.4 Estimated heritability and QTL for genome-wide recombination number ... 37
  2.4.5 Imputation accuracy and recombination rate ................... 38
2.5 Discussion ................................................................. 39
  2.5.1 Impact of phasing methods ........................................ 39
  2.5.2 Assumptions for the identification of recombination events .... 40
  2.5.3 Impact of homozygous segments on the identification of recombination events 41
  2.5.4 Estimation of chromosome genetic distance per Mb and heritability of GRN . 41
  2.5.5 Genomic regions associated with GRN ......................... 42
  2.5.6 Influence of recombination on imputation accuracy and haplotype phasing . 43
2.6 Conclusions ................................................................. 43
2.7 Acknowledgements ....................................................... 44
2.8 Bibliography ............................................................... 44
2.9 Tables .......................................................................... 51
2.10 Figures .................................................................... 54

CHAPTER 3. IDENTIFICATION OF RECOMBINATION HOTSPOTS AND
GENETIC VARIANTS THAT INFLUENCE RECOMBINATION RATE IN
LAYER CHICKENS ............................................................ 68
3.1 Abstract ........................................................................ 68
3.2 Introduction ................................................................. 69
3.3 Methods ...................................................................... 70
  3.3.1 Genotypes .......................................................... 70
3.3.2 Identification of recombination events ........................................ 70
3.3.3 Examining factors that could affect the observed recombination events .... 71
3.3.4 Estimating repeatability and heritability ...................................... 73
3.3.5 Identification of QTL affecting GRN .......................................... 74

3.4 Results and Discussion ................................................................. 76
3.4.1 Recombination between lines ..................................................... 76
3.4.2 Recombination between sexes ..................................................... 79
3.4.3 Chromosome size and GC content .............................................. 80
3.4.4 GHU .................................................................................. 81
3.4.5 Genome regions associated with GRN ......................................... 82

3.5 Conclusions .................................................................................. 86
3.6 Acknowledgements ........................................................................ 86
3.7 Bibliography .................................................................................. 86
3.8 Tables ........................................................................................... 93
3.9 Figures ........................................................................................... 99

CHAPTER 4. GENOME-WIDE ASSOCIATION STUDY OF GROWTH AND
BODY COMPOSITION TRAITS IN BRANGUS BEEF CATTLE ............. 109
4.1 Abstract ....................................................................................... 109
4.2 Introduction ................................................................................... 110
4.3 Methods ........................................................................................ 111
4.3.1 Phenotypes and Genotypes ....................................................... 111
4.3.2 Imputation ................................................................................ 111
4.3.3 Degressed EBVs ....................................................................... 112
4.3.4 Genome-wide association study ............................................... 113
4.4 Results and Discussion .................................................................. 115
4.4.1 Imputed HD genotypes .............................................................. 115
4.4.2 Significant windows and lead SNP ........................................... 115
4.4.3 Pleiotropic or closely linked QTL .............................................. 117
4.4.4 GO analysis of pleiotropic QTL ............................................. 119
4.4.5 Trait-specific QTL .............................................................. 120
4.4.6 QTL segregating in other breeds ........................................ 122
4.5 Conclusions ................................................................. 123
4.6 Acknowledgements ............................................................ 123
4.7 Bibliography .................................................................. 123
4.8 Tables .......................................................................... 131
4.9 Figures .......................................................................... 140

CHAPTER 5. EFFECTS OF NUMBER OF TRAINING GENERATIONS FOR
GENOMIC PREDICTION IN VARIOUS TRAITS IN A LAYER CHICKEN
POPULATION ................................................................. 145
5.1 Abstract ..................................................................... 145
5.2 Introduction ................................................................. 146
5.3 Methods ..................................................................... 148
  5.3.1 Phenotypes and genotypes ......................................... 148
  5.3.2 Statistical model ....................................................... 149
  5.3.3 Optimal number of training generations .................. 152
  5.3.4 Marker-based heritability ......................................... 152
5.4 Results and discussion .................................................. 153
  5.4.1 Prediction accuracy in progeny ............................... 153
  5.4.2 The relationship among optimal number of training generations, heritability,
       and genetic architecture ............................................ 156
5.5 Conclusions ................................................................. 157
5.6 Acknowledgements ........................................................ 157
5.7 Bibliography ................................................................ 158
5.8 Tables .......................................................................... 162
5.9 Figures .......................................................................... 167
CHAPTER 6. GENERAL DISCUSSION AND CONCLUSIONS ............... 171

6.1 General Discussion .................................................. 171
   6.1.1 Research objectives ........................................ 171
   6.1.2 Factors influencing identification of recombination ........ 172
   6.1.3 Selection on recombination .................................. 174
   6.1.4 Identification of causal QTL ................................. 175
   6.1.5 Training generations ........................................ 176

6.2 Conclusions ............................................................ 178

6.3 Bibliography ........................................................... 179
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Physical length, estimated genetic length and recombination distance per Mb of bovine autosomes.</td>
<td>51</td>
</tr>
<tr>
<td>2.2</td>
<td>Candidate windows and SNPs for genome-wide recombination number in Angus and Limousin.</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>1 Mb windows that explain a significant proportion of genetic variation (&gt;0.2%), and results for significant SNPs and positional candidate genes within these windows detected for genome-wide recombination number in Angus and Limousin.</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Imputation accuracy, number of markers, SNP density and average LD for bovine autosomes in Angus based on UMD3.1 assembly locus coordinates.</td>
<td>53</td>
</tr>
<tr>
<td>3.1</td>
<td>Average recombination rates (± SD) for 0.5-Mb windows per chromosome.</td>
<td>93</td>
</tr>
<tr>
<td>3.2</td>
<td>Description of 1-Mb regions in white layer chickens that explained more than 0.8% of genetic variance and the significant SNPs within the QTL regions for genome-wide recombination number.</td>
<td>94</td>
</tr>
<tr>
<td>3.3</td>
<td>Description of 1-Mb regions in brown layer chickens that explained more than 0.8% of genetic variance and the significant SNPs within the QTL regions for genome-wide recombination number.</td>
<td>95</td>
</tr>
<tr>
<td>3.4</td>
<td>Gene Ontology term analyses of positional candidate genes in QTL regions.</td>
<td>96</td>
</tr>
</tbody>
</table>
4.1 Heritability ($h^2$), number of genotyped animals with DEBV and mean reliabilities ($\pm$ SE) for the evaluated traits in Brangus. 131

4.2 Distribution of birth year of genotyped Brangus cattle on each sSNP panel. 132

4.3 The number of animals, number of SNPs before imputation, the mean of average additive genetic correlation (SD), and the mean of maximum additive genetic correlation (SD) between and within different groups. 133

4.4 Description of 1-Mb windows that explained more than 0.5% of genetic variance, and the significant SNPs within the window for all studied traits in Brangus. 134

4.5 The GO analysis of positional candidate genes in the identified significant 1-Mb QTL window. 136

4.6 Phenotypic (above diagonal) and genetic (below diagonal) correlation coefficients among 6 traits which associated with pleiotropic QTL. 138

4.7 The GO analysis of positional candidate genes in the identified significant 1-Mb QTL window. 139

5.1 Description of the phenotypes in 16 traits in each generation. 162

5.2 Examples of experimental design for training (T) and validation (V) sets. 163

5.3 Description of the average number of individuals with phenotypes and genotypes in each generation for early and late traits. 164

5.4 Prediction accuracies (SD) of different composition of training sets in eEW (early average weight of 3-5 eggs) using BayesB. In this analysis, G10 was used as the validation generation, and training individuals were randomly sampled from G4 to G9. 165
5.5 Estimates of pedigree-based heritabilities and marker-based heritabilities (±SE$^2$) for the 16 traits$^1$ from univariate animal models.
LIST OF FIGURES

2.1 Probability of zero (black), one (red), two (green), three (blue), and more than three (purple) crossover events per chromosome for each of the 29 bovine autosomes in Angus. Solid line represents expected probability calculated by Karlin’s map function, dashed line represents observed probability using DAGPHASE, and dotted line represents observed probability using BEAGLE. 54

2.2 Expected and observed crossover probabilities in Angus and Limousin. Description: (A) Representative example of expected and observed crossover probabilities in Angus (green) and Limousin (blue) for autosome 15. (B) Probability of zero (black), one (red), two (green), three (blue), and more than three (grey) crossover events for the 29 bovine autosomes in both Angus and Limousin. Plain line represents expected probability, dashed line represents observed probability in Angus, and dotted line represents observed probability in Limousin. 55

2.3 Representative example of expected and observed crossover probabilities on autosome 15 in Angus using UMD3.1 (green) and USDA-AIPL coordinates (blue). 56

2.4 Correlation between number of haplotypes and recombination rate within 1 Mb windows across the 29 autosomes in Angus. 57
2.5 Relationship between number of haplotypes and crossover rate within 1 Mb windows across the 29 autosomes in Angus. Number of haplotypes in hot (red) and cold windows (blue), which were defined as windows with a recombination rate $\geq 0.02$ ($\geq 1.5$ standard deviations from the mean) and windows with recombination rate $\leq 0.004$, respectively; * indicates a significant difference with $p < 0.05$, and ** indicates a significant difference with $p < 0.01$ for a paired t-test. 58

2.6 Expected and estimated genetic distance per Mb (cM/Mb) for the 29 bovine autosomes in Angus (green) and Limousin (blue). 59

2.7 Representative example of the variation in recombination rate within 1 Mb windows across bovine autosome 15. The plain line (upper) corresponds to the recombination rate estimated in Angus, while the dashed line (lower) corresponds to the recombination rate estimated in Limousin. 60

2.8 Variation in recombination rate within 1 Mb windows across the 29 bovine autosomes. The plain line (upper) corresponds to the recombination rate estimated in Angus, while the dashed line (lower) corresponds to recombination rate estimated in Limousin. 61

2.9 Frequency distribution of genome-wide recombination number (GRN) in both Angus (left) and Limousin (right). 62
2.10 Distribution of genome-wide recombination number in Angus and Limousin families. (A) Representative example of distribution of genome-wide recombination number (GRN) in Angus half-sib families. Only sires with no more than 20 offspring are presented. Sires were sorted according to the number of their offspring. (B) GRN in Angus half-sib families. Black dots correspond to GRN in sons sorted by sires and red dots correspond to the average GRN for each sire. (C) GRN in Limousin half-sib families. Black dots correspond to GRN in sons sorted by sires and red dots correspond to the average GRN for each sire. .......................... 63

2.11 Proportion of genetic variance explained by 1 Mb windows across the genome for genome-wide recombination number (GRN) in Angus and Limousin. .......................... 64

2.12 Relationship between imputation accuracy and recombination rate across the 29 autosomes in Angus. (A) Average imputation accuracy in individuals with zero (red), one (green), two (blue), and more than two (purple) crossover events across the 29 autosomes. (B) Correlation between imputation accuracy and crossover rate (cM/Mb) within 1 Mb windows. .......................... 65

2.13 Imputation accuracy in recombination hot (red) and cold windows (blue). Hot (red) and cold windows (blue), which were defined as windows with a recombination rate $\geq 0.02$ ($\geq 1.5$ standard deviations from the mean) and windows with recombination rate $\leq 0.004$, respectively; * indicates a significant difference with $p < 0.05$, and ** indicates a significant difference with $p < 0.01$ for a paired t-test. .......................... 66

2.14 Chromosome-wide recombination probabilities calculated using Karlin’s (red), Haldane’s (green) and Kosambi’s (blue) map functions. 67
3.1 Recombination rate within 0.5-Mb windows estimated on GGA1. The black solid line corresponds to the recombination rate estimated from 580K SNPs in white layers. The grey dashed line corresponds to the recombination rate estimated from 42K SNPs in brown layers.  

3.2 Variation in recombination rate within 0.5-Mb windows across the 28 autosomes. The black solid line corresponds to the recombination rate estimated from 580K SNPs in white layers. The grey dashed line corresponds to the recombination rate estimated from 42K SNPs in brown layers.  

3.3 Recombination rate within 0.5-Mb estimated using segregating 580K SNPs (black solid line) and 42K SNPs (green dashed line) in white layers on GGA1.  

3.4 The GRN in white and brown half-sib families. Black dots correspond to GRN in offspring sorted by parents and red dots correspond to the average GRN for each parent. Only parents with at least 2 offspring are presented.  

3.5 Distributions of the NHA (number of common haplotype alleles) and PHHA (proportion of homozygosity of common haplotype alleles) within 0.5-Mb window in white and brown layers.  

3.6 Recombination rate within 0.5-Mb windows estimated on GAA1 for males and females in white and brown chickens.  

3.7 Distributions of the average GRN of males (red) and females (blue) in white and brown layers.  

3.8 Relationship of GC content and CpG island density with recombination rate within 0.5-Mb windows across chromosomes in white and brown layers.  

3.9 Proportion of genetic variance explained by 1-Mb regions across the genome for GRN in white layers.  

3.10 Proportion of genetic variance explained by 1-Mb regions across the genome for GRN in brown layers.
4.1 Proportion of genetic variance explained by 1-Mb windows across the genome for the studied traits in Brangus. Birth weight direct; Birth weight maternal; Back fat thickness; Intramuscular fat thickness; Rib eye muscle area; Rump fat thickness; Mature weight; Weaning weight; Yearling weight. ................................................................. 140

5.1 Prediction accuracies of EBV across all traits using BayesB, and PBLUP with truncated (PBLUP_T) and full pedigree (PBLUP_F). 167
5.2 Prediction accuracies of EBV across different validation sets using PBLUP_T and BayesB over different number of training generations in each trait. ................................................................. 168
5.3 Optimal number of training generations using BayesB in each trait. Traits were sorted by pedigree-based heritability. .................. 169
5.4 Scatter plot of prediction accuracy using BayesB over training generations. The blue line is the regression of GEBV accuracy on training generations. The red line indicates the optimal number of training generations. 170
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my wholehearted gratitude to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, my major supervisor, Dr. Dorian Garrick, for his guidance, patience and support throughout this research and the writing of this thesis. His words of encouragement have often inspired me and bring me insights into the exciting field of animal breeding and quantitative genetics. Under his instruction in my coursework and research, I acquired different professional skills, and became an independent researcher.

I am greatly thankful to my Ph.D committee members, Drs. Rohan Fernando, Jack Dekkers, Daniel Nettleton, and Alicia Carriquiry, for their efforts and contributions to this work. Rohan has motivated me with his critical and diligent attitude toward science. Rohan has always been encouraging, and sharing his knowledge and experiences with me like a friend. I would also like to thank Jack, who provides incisive suggestions and criticism to my research. My gratitude also goes to Daniel, my statistics minor supervisor. I have benefited from his inspirational teaching style, and sharp vision of my research topics. I am especially thankful to Alicia, who has always been supportive, friendly, and helpful. I feel blessed to work with distinguished and remarkable people at Iowa State University. They are not only great professors, but also excellent mentors.

Further, I'm grateful to Dr. Mahdi Saatchi, for his guidance throughout the initial stages of my graduate career. I would like to thank Dr. Anna Wolc for sharing her knowledge of chicken genetics, and in various aspects of my research and life. I would like to thank Dr. David Habier, for sharing his knowledge in statistics and genetics, and providing me valuable suggestions on life experience and career development.

Thanks to Mengya Zhang, Dr. Huiyu Wang, Weixuan Fu, Mengguo Yan, Dr. Napapan Habier, Dr. Tian Lin, Ruolin Liu, Dr. Zaoli Xu, Dr. Xuan Ge, and my other friends in U.S., for sharing
life with me. It is grateful to be part of their life. I hope our paths will cross again in the future.

Thanks to all my friends and relatives in China and in other counties over the world. Although the distance is long, their supports are persistent and strong.

I would additionally like to thank my officemates, fellows, and colleagues, Jenelle Dunkelberger, Melanie Hess, Emily Mauch, Lydia Hartie, Brittany Shonka, Laura Siocnarf, Dr. Hailin Su, Dr. Nick Serao, Dr. Jian Zeng, Dr. Xiaochen Sun, Dr. Dawn Koltes, Dr. James Koltes, Hao Cheng, Luke Kramer, Haibo Liu, and Andrew Hess, etc. I am lucky to be a member of this awesome Genomic Selection group, and be involved in the big family of Animal Breeding & Genetics.

Finally, with my whole heart and soul, I thank my family, for their unconditioned love. I thank my husband, Zhenqi Li, for years of patience, love, and understanding. I thank my parents, especially, my mom, for raising me with curiosity of science, enthusiasm for life, and courage to face difficulties.
ABSTRACT

The major task of animal breeding is to achieve genetic improvement for traits of economic importance in livestock species. Genetic improvement of a specific trait could be measured as genetic gain per year within a population, which is determined by the selection intensity, the standard deviation of true breeding value, the accuracy of estimated breeding value (EBV), and the generation interval. One main strategy to improve genetic gain is to improve the selection accuracy in EBV. The Research topics concerning improving selection accuracy of EBV have been of great interest in livestock breeding for decades.

The advent of genomic prediction has enhanced prediction accuracy of EBV and revolutionized selection strategies in livestock. Genomic prediction utilizes statistical models to predict genomic estimated breeding values (GEBV) based on genome-wide single nucleotide polymorphisms (SNPs). The accuracy of genomic prediction is affected by various of factors, such as the size of reference population, the linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), and the genetic properties of traits. Bayesian hierarchical models, which consider all unknown parameters and estimating all SNP effects simultaneously, are widely used in whole genome analyses.

Due to LD between SNPs and QTL, genomic prediction can capture the effects of QTL for the traits of interest. Alleles of different loci cluster together in a haplotype blocks, which are passed from the parents to the progeny. Compared to SNP effect model, haplotype effect model reduces dimensions of parameters and utilizes multi-loci LD. Fitting haplotype alleles is likely to capture QTL effects better than fitting SNPs genotypes. However, the occurrence of recombination during meiosis breaks down the haplotype blocks which would influence the accuracy of both population-wide and family-wide haplotype reconstructions, and erodes the LD between SNPs and QTL. Meiotic recombination events distribute non-randomly along the genome in many species. It occurs more frequently in recombination hotspots, which are defined as short intervals with sig-
nificantly higher recombination rates than the average recombination rate on each chromosome. Characterizing recombination events will enable the identification of haplotype diversity, elucidate genetic variation along the genome, and eventually improve genomic prediction. The first objective of this thesis was to evaluate the relationship between recombination and haplotype reconstruction, to investigate factors affecting recombination, to recognize recombination hotspots, and to locate QTL influencing genome-wide recombination rate in beef cattle and layer chickens.

Genome-wide association studies were performed to identify the QTL controlling the traits of interest in livestock species. The availability of dense SNP genotypes along the genome aids the detection of causal genetic variants. Accuracy of genomic prediction can be enhanced by integrating dense SNP genotypes with the genotypes of causal QTL. The second objective of this thesis was to identify causal QTL associated with growth and body composition traits in Brangus beef cattle.

Simulation studies have shown that using distant ancestors for prediction of EBV or GEBV of young animals is less informative than using close relatives. The effect of including distant ancestral generations in the training set on prediction accuracy has not been well studied in a real population under selection. The third objective of this thesis was to assess the optimal numbers of training generations needed to yield the maximum genetic prediction accuracy for various of traits in layer chickens.

Chapter 2 studied 2775 Angus and 1485 Limousin cattle genotyped with the Illumina Bovi- neSNP50 chip. Haplotype phasing was performed using DAGPHASE and BEAGLE based on UMD3.1 assembly. Recombination events were identified by comparing the reconstructed chromosomal haplotypes between sire-offspring pairs. The average genome-wide recombination number for sires were recorded as phenotypes. The BayesB approach was used to identify QTL influencing genome-wide recombination events. Genotype imputation from a 7K subset to the 50K chip in Angus population was conducted by BEAGLE. Due to the linkage information from relatives, DAGPHASE was superior to BEAGLE in haplotype phasing. In Angus, 427 1-Mb windows containing recombination hotspots were detected and 348 recombination hotspots were identified in Limousin. The regions with high recombination rates had low accuracy of haplotype phasing and genotype imputation. Limited population sizes and half-sib family sizes, along with the occurrences of gene
conversion, genotyping errors, and map errors, hinder the identification of recombination events. Different QTL regions influencing genome-wide recombination were identified in the two breeds.

A total of 1200 white layers genotyped with 580K SNP panel and 5108 brown layers genotyped with 42K SNP panel were studied in Chapter 3. Recombination events were identified using LINKPHASE within half-sib families. The BayesB approach was used to identify QTL influencing genome-wide recombination in each line. The number of recombination hotspots detected in white and brown layers were 190 and 199, respectively. Only 28 of them were common to both lines. Recombination rates differed in lines and sexes. Family structure, marker density, inbreeding level, and haplotype structure could influence the identification of recombination events. Chromosome size, GC content, and CpG island density showed negative correlations with recombination rate. Several significant QTL windows, which harbor candidate genes were identified in the 2 lines. In general, recombination rate is a complex, breed-specific, polygenic trait. Identification of recombination provides us opportunities to improve map assembly, and enhance haplotype reconstruction. Implementing recombination information will help to improve genomic prediction in livestock breeding.

In Chapter 4, a total of 1537 Brangus beef cattle were genotyped with Bovine50K, GGPHD77K, or BovineHD770K SNP chip. FImpute was used to impute Bovine50K or GGPHD77K to BovineHD770K. The BayesB approach with weighting factors was used to map QTL in each trait. A total of 18 different QTL regions were detected across 9 studied traits. Among these QTL, 11 were trait-specific, while others were pleiotropic. Five large-effect QTL were found segregating in other breeds, including Angus and Nellore. These identified QTL will aid our understanding of the biological processes of growth and body composition traits in beef cattle and ultimately help to enhance genomic prediction across multi-breed.

Phenotypic records from 16 traits on 17 793 birds over 9 non-overlapping generations were analyzed in Chapter 5. Among these birds, 5108 of them had genotypes for 23 098 segregating SNPs. Two prediction methods, best linear unbiased prediction model with pedigree relationships, and BayesB were applied to predict EBV or GEBV in each validation generation based on varying numbers of ancestral training generations. The optimal number of training generations that resulted in the highest prediction accuracy of GEBV was obtained for each trait. The relationship between
optimal number of training generations and heritability of traits was evaluated. Prediction accuracy of EBV and GEBV increased by including close ancestral generations, but either reached an asymptote or decreased slightly when distant ancestral generations were included in training. The optimal number of training generations increased with heritability. Based on the studied dataset, 4 or 5 training generations is optimal to most of polygenic traits.

In summary, this thesis investigated genome-wide recombination mechanisms in beef cattle and layer chickens, identified positional candidate genes for growth and meat production traits in Brangus beef cattle, and assessed the effect of distant ancestral generations on genomic prediction accuracy in layer chickens. The locations of recombination hotspots and QTL which control genome-wide recombination rates were identified in both beef cattle and layer chickens. The identification of recombination patterns along the genome will aid in better defining haplotype blocks and improving genomic prediction accuracy. Selection on recombination rate is challenging, but it has potential benefits to increase genetic gain in livestock breeding system. In Brangus beef cattle, 7 pleiotropic QTL and 11 trait-specific QTL were identified which will help us better understanding the biological processes accounting for variation in growth and body composition traits in Brangus cattle. Utilizing genotypes of identified causal QTL will enhance genomic prediction. Since the effect of adding distant ancestral generations in training on prediction accuracy differed between traits, different prediction strategies should be applied based on the importance of selected traits in a specific breeding population. Implementing these findings in livestock breeding program will help to improve the genomic prediction accuracy, and ultimately to increase short-term and long-term genetic gain for selected traits.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

The main task of animal breeding is to maximize genetic improvement for economically important traits. These traits are usually quantitative or complex traits such as meat quality or egg production. The genetic level of a population for traits of interest will be improved by selecting elite animals with superior genetic merit to be parents (Lush, 1937). The improvement of genetic level due to selection is measured as genetic gain or response to selection, which can be calculated as:

\[ \Delta g = \frac{ir\sigma_A}{L}, \]  \[1\]

where \( i \) is the selection intensity, \( r \) is the selection accuracy base on the estimated breeding values (EBV), \( \sigma_A \) is the standard deviation of the true breeding values (TBV) for the population under selection, and \( L \) is the generation interval. To improve the genetic merit of a population for traits of interest, elite animals with superior TBV are selected to be parents (Lush, 1937). Because TBV are not observed in reality, selection decisions are made based on EBV. The prediction accuracy of EBV is pivotal to genetic improvement as the genetic gain is proportional to the value of the prediction accuracy.

The traditional method for estimating TBV for an individual relies on the availability of phenotypic records from itself, relatives, or both that are combined through and the numerator relationship matrix derived from pedigree. The Best Linear Unbiased Prediction (BLUP) (Henderson, 1984) is the most commonly used method for EBV, where TBV are fitted as random effects in a linear mixed
model. Over the past several decades, selection based on BLUP EBV has resulted in significant genetic improvement for economically-important traits in the livestock industry (Hill, 2014).

As the advent of a large number of single nucleotide polymorphisms (SNP), whole genome analyses including genomic prediction and genome-wide association studies (GWAS) became feasible and popular. The extensive use of genome-wide dense SNP markers has revolutionized genetic improvement programs in livestock (Goddard et al., 2010) and plants (Zhong et al., 2009), and has helped us to identify QTL for traits of interest. Compared to conventional prediction methods, such as BLUP, genomic prediction that exploits all SNPs to predict individual’s breeding value can increase the prediction accuracy of EBV, decrease the generation interval, and eventually improve the genetic gain. GWAS has enabled the discovery of a large number of QTL for important traits (Dekkers and Hospital, 2002), which substantially increased our understanding of the biological functions of genes affecting the traits. In addition, including identified QTL in the prediction equations will further enhance the accuracy of genomic prediction.

The concept of genomic selection using SNPs was introduced by Meuwissen et al. (2001). The approach of genomic selection is often performed in two steps (Meuwissen et al., 2001). First, SNP effects are estimated from individuals in a training set which have both SNP genotypes and trait phenotypes. Then, genomic estimated breeding values (GEBV) of genotyped individuals in a separate (validation) population are obtained by summing the estimated SNP effects multiplied by SNP genotypes. Selection decisions based on GEBV can be made at a much earlier stage than in traditional selections, since GEBV of young animals can be obtained once their DNA samples are collected. Based on previous studies (Daetwyler et al., 2010; Goddard et al., 2010; Habier et al., 2007; Hayes et al., 2009; Muir, 2007; Sun et al., 2011; Wolc et al., 2011), the accuracy of genomic prediction is mainly influenced by linkage and linkage disequilibrium (LD) between SNP and QTL (Goddard et al., 2010; Sun et al., 2011), reference population size (Hayes et al., 2009), pedigree relationship (Habier et al., 2007), genetic properties of traits (Daetwyler et al., 2010; Wolc et al., 2011), marker density (Daetwyler et al., 2010; Goddard et al., 2010), population structure (Muir, 2007) and other factors.
LD between SNP and QTL is a major source of genetic information that contributes to the accuracy of genomic prediction (Habier et al., 2013). In practice, QTL genotypes cannot be observed directly. Instead, only genotypes of SNP markers are observed. The genetic variance at QTL for a desired trait can be captured by SNPs due to LD between SNPs and QTL. Two-locus LD exists when genotypes at a SNP locus are predictive of genotypes at a QTL. Many algorithms have been developed based on single locus LD to explore desired traits and to investigate individual genetic merits. The Bayesian SNP effect model is the most commonly used model for genomic prediction and GWAS (Zou and Zeng, 2008). The basic form of the SNP effect model is:

\[ y = X\beta + Z\alpha + e, \]  

where \( y \) is an vector or trait phenotypes, \( X \) is an incidence matrix relating to the records of fixed effects \( \beta \), \( Z \) is a genotype covariates matrix coded as 0, 1, or 2, \( \alpha \) is a vector of random partial regression coefficients of SNPs (marker effects), and \( e \) is a vector of residuals. According to Bayesian analysis, prior distributions of \( \beta, \alpha, \) and \( e \) must be assumed. According to the Bayesian analyses, prior distributions of \( \beta, \alpha, \) and \( e \) must be specified. Inferences of these unknown parameters are drawn based on the posterior sample obtained from Markov chain Monte Carlo (MCMC). The single-site Gibbs sampler (Geman and Geman, 1984), block Gibbs sampler (e.g. Sorensen and Gianola (2002)), and Metropolis-Hastings sampling (Hasting, 1970) are often used to construct Markov chains and draw isamples from the joint posterior distributions. Other than the MCMC algorithms, the expectation-maximization algorithm (Dempster et al., 1977) and variational Bayes algorithms (e.g. Li and Sillanpää (2012)) have been proposed as alternative approaches for estimating SNP effects. In most circumstances, the number of SNPs is much larger than the number of phenotypes. Bayesian methods allow a simultaneous estimation of all SNP effects in the model with a heterogeneous shrinkage on the effects according to their prior specification (de los Campos et al., 2013). The performance of the SNP effect model is belived to be primarily driven by LD between SNPs and QTL. For some traits, the accuracy of prediction using SNP effect model is low, because the causal QTL are in weak LD with SNP markers.
In the breeding system, progeny inherit genetic merit from their elite, selected parents. Transmission of alleles from parents to progeny is not independent. Linked genes, or genes with a similar biological function tend to cluster into groups (Hurst et al., 2004) and are passed from parents to progeny in the form of a haplotype. In order to capture QTL effects effectively though multi-locus LD (LD between QTL and haplotype alleles), the haplotype model has been proposed by (Sun et al., 2014):

\[ y = X\beta + \sum_{j=1}^{n_q} H_j W_j \gamma_j + e. \]  

where \( y \) is an vector or trait phenotypes, \( X \) is an incidence matrix relating to the records of fixed effects \( \beta \), \( n_q \) is the total length of genome (cM), \( \gamma_j \) is a vector of the \( n_j \) allelic values (coded as 0, 1, and 2) at the \( j \)th centi-Morgan (cM), \( H_j \) is a matrix of covariates for \( \gamma_j \) at the \( j \)th cM, \( W_j \) is a diagonal matrix of the indicator variables, and \( e \) is a vector of residuals. It has been reported that fitting haplotype alleles is an effective strategy for capturing QTL effects, compared to fitting SNP genotypes in the prediction model in the simulated study (Sun et al., 2014). There are two advantages of the haplotype model over the SNP effect model. First, the haplotype model can utilize multi-locus LD efficiently. Secondly, the haplotype model can reduce parameter dimension in the model. In practice, the number of haplotype alleles are often found to be less than the number of SNP markers within a small genomic window if rare haplotype alleles are ignored, which are likely due to phasing errors.

When applied to real data, the haplotype model did not always outperform the SNP effect model (Sun, 2014). Possible reasons include the effect of recombination, improper haplotype size, and problems with map assembly or haplotype phasing in real analysis. Meiotic recombination may occur between a pair of homologous chromosomes which generates crossovers by breaking the original parental haplotypes and producing new haplotypes. If recombination occurs, different haplotypes can be passed down from the same parents to offspring. As a result, the occurrence of recombination may reduce the LD between haplotype alleles and QTL alleles, given that the haplotype window harbors QTL. This will decrease the accuracy of reconstructing offspring haplotypes, affect imputation of missing genotypes within the population, and eventually influence genomic pre-
diction. In simulation studies, map position is known and haplotype phase is also known without error. Since QTL are simulated in the center of a 1-cM window, it is appropriate to define a 1-cM haplotype window. However, for analysis of real data, haplotype size is arbitrarily defined as 1-Mb. Several studies have investigated the effect of haplotype size on genomic prediction accuracy. For example, Cuyabano et al. (2014) applied the genomic prediction model based on 2-SNP haplotypes in a Nordic Holstein population. Calus et al. (2008) observed that prediction accuracy using 10-SNP haplotypes was higher than that using 2-SNP haplotypes in a simulation study. Hickey et al. (2013) found that increasing the size of haplotypes leads to a drop in prediction accuracy. Since recombination occurs in recombination hotspots, the distribution of recombination hotspots across the chromosomes can provide insights for redefining haplotype windows. Recombination hotspots are the chromosomal segments with much higher recombination rates compared to the neighboring areas. If the haplotype window is located in a recombination hotspots region, the number of haplotype alleles may be even larger than the number of SNP markers. Therefore, it is best to divide a recombination hotspots region, into many shorter haplotype windows. On the other hand, adjacent regions with low recombination rates can be combined together into a single haplotype window to decrease the number of effects to be estimated. Besides defining proper haplotype length, use of recombination information may help to detect potential map errors and improve the accuracy of haplotype phasing and genotype imputation. Therefore, the first objective of this thesis was to investigate the nature of recombination, which will aid in understanding genetic variation across the genome, optimize haplotype detection algorithms, and eventually enhance genomic prediction strategies.

As the availability of SNP genotypes has enabled identification of genes controlling traits of economic importance (Dekkers and Hospital, 2002), QTL with large effects have been successfully identified by GWAS or QTL mapping in many livestock species. For example, DGAT1 affecting milk production in dairy cattle (Grisart et al., 2002; Littlejohn et al., 2014), PLAG1 influencing body weight in cattle (Littlejohn et al., 2012; Saatchi et al., 2014), and IGF2 controlling muscle growth in pigs (Jungerius et al., 2004; Van Laere et al., 2003). Including QTL genotypes or causal mutations in selection criterion, enable us to capture Mendelian sampling variance of QTL and to
improve the prediction accuracies of corresponding traits (Dekkers, 2004). As a result, it is possible to breed better animals and to obtain greater genetic gain per year by implementing information of causal genes in the selection scheme. Another objective of this thesis was to identify causal variants influencing routinely evaluated traits and to explore the feasibility of integrating causal QTL in the breeding programme.

It is known that increasing the training data size is expected to increase genomic prediction accuracy (Daetwyler et al., 2008; Hayes et al., 2009). In reality, the number of genotyped animals in a commercial population is limited per generation. One way to increase the size of the training set is to include data from previous generations. Utilizing data from past generations may also help to avoid selection bias in populations undergoing selection (Henderson, 1975; Im et al., 1989). In Pedigree-based BLUP (PBLUP), distant generations contribute little to prediction accuracy of recent generations, since the additive genetic relationship is halved each additional generation. Compared to PBLUP, distant generations are expected to contribute more to prediction accuracy in Genomic-BLUP, provided the existence of LD between SNP and QTL (Habier et al., 2007). However, the contribution of distant ancestors is not equivalent to that of close relatives for predicting GEBV of young animals (Habier et al., 2007; Wolc et al., 2013). The impact of adding distant ancestral generations to the training data on the accuracy of genomic prediction needs to be characterized in a population undergoing selection. The last objective of this thesis was to evaluate genomic prediction accuracy when including different numbers of distant ancestral generations in the training data set, and to identify the optimal number of training generations needed for different traits.

1.2 Research Objectives

The objectives of the research presented in this thesis were 1) to identify recombination hotspots, to evaluate factors influencing recombination events, and to locate QTL associated with genome-wide recombination rates in beef cattle and layer chickens (Chapters 2 and 3); 2) to identify QTL influencing growth and body composition traits in Brangus beef cattle using 770K SNP genotypes (Chapter 4); 3) to evaluate the prediction accuracy when including distant ancestral generations...
in the training data set, and to determine the optimal number of training generations required for different traits in a layer breeding line (Chapter 5).

### 1.3 Organization of thesis

The rest of this chapter provides a brief summary of the literature on the current research status of meiotic recombination, and development of Bayesian methodologies for genomic prediction.

Chapter 2 "Recombination locations and rates in beef cattle assessed from parent-offspring pairs" was published in *Genetics Selection Evolution* (Weng et al., 2014). The relationships between recombination, haplotype phasing, and imputation were evaluated, in addition to the identification of recombination hotspots and candidate genes that influence genome-wide recombination in Angus and Limousin beef cattle. This work provides insight regarding the properties of recombination in the beef cattle genome.

Chapter 3 "Identification of recombination hotspots and genetic variants that influence recombination rate in layer chickens" will be submitted to a peer reviewed journal. We identified recombination hotspots across the chicken genome, evaluated the impact of various factors such as marker density, CG content, CpG island density, and sequence motif density on recombination rate, and identified genetic variants influencing recombination events in two purebred layer chicken lines.

Chapter 4 "Genome-wide association study of growth and body composition traits in Brangus beef cattle" was accepted by *Journal of Livestock Science*. In this paper, several polygenic and trait-specific QTLs associated with growth and body composition traits in Brangus beef cattle were identified using 770K SNP genotypes. The biological functions and related pathways of these QTLs were studied.

Chapter 5 "Effects of number of training generations for genomic prediction in various traits in a layer chicken population" was submitted to *Genetics Selection Evolution*. The effect of adding distant ancestral generations in the training data set on prediction accuracy varied among traits. Heritability and genetic architecture of the trait influenced the optimal number of training generations.
Chapter 6 consists of a general discussion and conclusions for the research presented in Chapters 2, 3, 4, and 5.

1.4 Literature review

1.4.1 Genetic analyses of variation in meiotic recombination

A haplotype is a group of alleles that tend to be inherited together from parents to progeny. Meiotic recombination, which takes place between a pair of homologous chromosomes, generates crossovers and rearranges haplotypes. The universal role of meiotic recombination in evolutionary biology is to increase the genetic diversity within a population by generating novel allelic combinations. Investigating the properties of recombination can help us understand genome-wide variation in haplotype diversity, explain genomic evolution, and improve genomic prediction accuracy. Recombination hotspots (coldspots) are regions with much higher (lower) recombination rates compared to average rate across the chromosomes. Approaches for identifying genome-wide recombination events and explanations of their manners have been reviewed in mice (Kauppi et al., 2004; Paigen and Petkov, 2010), humans (Jeffreys and Neumann, 2009; Lynn et al., 2004), and plants (Mezard, 2006).

The process of homologous recombination includes several steps. At first, two homologous chromosomes (four chromatids) are aligned with each other. Recombination starts at a hotspot when the *trans*-acting gene products, like PR domain-containing 9 (*PRDM9*), permit SPO11 to catalyze a DNA double-strand break (DSB) on one chromatid (Paigen and Petkov, 2010). Following this process, Holiday Junctions are created, where two DNA strands switch partners between two double helices. The Holiday Junctions have two crossing strands and two non-crossing strands. Heteroduplexes are produced by the exchanging of DNA strands. A heteroduplex contains DNA mismatches with an approximate length of 250 bp. There are two ways to resolve the Holiday structure: one resolution cuts the original pair of crossing strands in the same way, creating heteroduplexes.
without crossover (gene conversion), or on the crossing strands are cut in the opposite way, producing two heteroduplexes with crossover. Most crossovers occur in intergenic or non-coding regions (Mezard, 2006).

There are several approaches to infer recombination rates and map recombination hotspots. Constructing a linkage map is one way of obtaining recombination rates. With the availability of a physical map, map functions enable the transformation of map distance into recombination rates. One approach assumes a number of crossovers in an interval in a probability model, e.g. Haldane’s map function, and Karlin’s map function. The other approach models recombination events in two adjacent intervals with interference, e.g. Kosambi’s map function. Haldane assumes the probability of crossover following a Poisson distribution (Haldane, 1919). This model doesn’t take interference between crossovers into account. According to the Poisson distribution, the probability of n crossovers in an interval of \( x \) (Morgan) is:

\[
P_n = (\lambda x)^n e^{-\lambda x}, [4]
\]

where \( \lambda = 1 \), is the expected number of crossover with in \( x \). Given Haldane’s assumption, the probability of recombination in an interval of \( x \) (Morgan) is \( r_x = \frac{1}{2}(1 - e^{-2x}) \). Karlin (1984) used the binomial distribution to model crossovers, which calculates the probability of \( n \) crossovers in an interval of length \( x \) as:

\[
P_n = \binom{N}{n} (x/N)^n (1 - x/N)^{N-n}, [5]
\]

where \( N \) is the maximum number of crossovers in the interval, and \( p \) for \( P_n \) is derived as \( x/N \). Kosambi’s method considers interferences between crossovers and models recombination in two adjacent intervals (Kosambi, 1944). According to Kosambi’s map function, the probability of recombination in an interval of \( x \) (Morgan) is \( r_x = \frac{1}{2}(e^{4x} - 1)/(e^{4x} + 1) \).

Statistical models utilizing on LD, haplotype blocks, and pedigree information have been developed to investigate recombination. The coalescent model, which approximates the likelihood of the dataset (McVean et al., 2002) and adopts the Monte Carlo integration algorithm (Fearnhead
and Donnelly, 2001), is a popular approaches for estimating recombination rate. An alternative approach, which introduces the conditional likelihood of the haplotypes. The model assumes the frequencies of the new haplotypes depends on the frequencies of the previously observed haplotypes (Li and Stephens, 2003). This model out-performed the previous coalescent model in many circumstances (Jeffreys et al., 2005). The software LINKPHASE based on an improved Hidden Markov Model combining linkage, LD, and pedigree information (Druet and Georges, 2010, 2015), is widely used in livestock breeding for haplotype reconstruction and crossover identification. Other than estimating recombination rate, many methods have been developed to identify recombination hotspots. Fearnhead (2006) introduced sequenceLDhot, which is a computationally intensive method and only considers a small number of SNPs around the hotspots. LDhot uses the pairwise composite likelihood to detect recombination hotspots in a window of 200kb (McVean et al., 2004). However, this method includes a very complex simulation step. The Hotspotter developed by Li and Stephens (2003) is a powerful method to detect hotspots, but it has a relatively high false positive rate. To minimize the false positive rate, a truncated and weighted pairwise likelihood method was introduced (Li et al., 2006). The detection powers of these methods is limited by the quality of map assembly, density of markers, genotyping errors, and haplotyping errors.

The total number of recombination events on a single chromosome or across the whole genome and the rates of recombination are mediated by a number of factors. Lercher and Hurst (2002) pointed out that the correlation between recombination rate and genetic diversity exists across the entire genome. Spencer et al. (2006) observed a positive correlation between recombination rate and nucleotide diversity in humans. Recombination was found to associate with other genomic features, such as GC and CpG content (Galtier et al., 2001), and gene expression in mammals (Kauppi et al., 2004; Paigen and Petkov, 2010) and plants (Paape et al., 2012). Lynn et al. (2004) reviewed gender, age, and chromosomal effects on recombination. Recombination rate has been found differed in species (Chowdhury et al., 2009; Kong et al., 2008), breeds (Weng et al., 2014), and sexes (Kong et al., 2008, 2014). The 17q21.31 inversion (Stefansson et al., 2005), and sequence variants in \textit{RNF212} were previously reported to affect genome-wide recombination phenotypes in males and females (Chowdhury et al., 2009; Kong et al., 2008; Sandor et al., 2012). Recombination hotspots are
located in different genomic areas, some of which are concentrated in repeat regions, such as LINEs (Long Interspersed Nuclear Elements) (Kauppi et al., 2004). In mice and humans, the detected recombination hotspots are the sites of both crossovers and gene conversions (Guillon and de Massy, 2002; Jeffreys and May, 2004). The cis-acting sequence (Jeffreys and Neumann, 2009), and trans-acting gene, PRDM9, have been identified as regulators of recombination hotspots in humans (Berg et al., 2010; Myers et al., 2010), mice (Baudat et al., 2010; Parvanov et al., 2010), and dairy cattle (Sandor et al., 2012). Kong et al. (2014) identified 13 genetic variants associated with male and female recombination rate using whole-genome sequencing of 2261 Icelanders. Among these genetic variants, one could explain over 10% of genetic variance in female genome-wide recombination rate (Kong et al., 2014).

Many questions regarding meiotic recombination remain unanswered. The DNA sequences that define recombination hotspots, and the relationship between recombination, selection, and evolution are still obscure. Although many features of the regulatory systems of recombination are unknown, our understanding of meiotic recombination in mice and humans is remarkably advanced compared to those in livestock species. Whether the recombination mechanisms identified in mice and humans remains valid in a variety of livestock species is open to debate.

1.4.2 Bayesian regression models

When Meuwissen et al. (2001) proposed genomic prediction of breeding values using genome-wide molecular markers, a new era began for animal and plant breeders, and various of Bayesian methods based on multi-locus association models have been developed. There are two ways of methods used to fit SNP genotypes in statistical models. The first is the breeding value model (BVM), which utilizes the marker-derived genomic relationship matrix instead of pedigree-based relationship matrix in a mixed model (e.g. GBLUP) (de los Campos et al., 2013; Habier et al., 2007; Nejati-Javaremi et al., 1997; VanRaden, 2008). The second is the SNP effect model, which is based on Markov chain Monte Carlo (MCMC) sampling and Bayesian hierarchical models. It has been shown that BVM and the SNP effect models result in equivalent predictions of breeding
values (Fernando, 1998; Strandén and Garrick, 2009). The equation of SNP effect model is presented in equation [2].

Most Bayesian models differ in the prior distributions of SNP effects ($\alpha$). BayesA is a model with $t$ priors for SNP effects (Meuwissen et al., 2001), which assume that SNP effects follow identical and independent univariate-$t$ distributions with mean zero, scale parameter $S_\alpha^2$, and degree of freedom $\nu$. The degree of freedom ($\nu$) determines the amount of shrinkage on the estimated SNP effect. The assumption of $t$ priors for SNP effects is equivalent to an independent univariate normal distribution with a null mean and locus specific variance for each SNP effect (Fernando and Garrick, 2013). BayesA has been criticized for using an improper degree of freedom for the prior (Gianola et al., 2009).

Meuwissen et al. (2001) proposed a variable selection model based on BayesA, which is referred to as BayesB. In BayesB, the prior distributions of SNP effects are identical independent mixture distributions with a point mass at zero with probability $\pi$ and a univariate-$t$ distribution elsewhere. The parameter $\pi$ is assumed known. As in BayesA, the univariate-$t$ distribution has a null mean, scale parameter $S_\alpha^2$, and degree of freedom $\nu$. BayesB more heavily shrinks small SNPs effects towards zero relative to those with large SNP effects.

According to the BayesC method (Kizilkaya et al., 2011), the prior assumption of SNP effects is a mixture distribution with a point mass at zero with probability $\pi$ and a univariate-normal distribution with a null mean and common variance $\sigma_\alpha^2$ elsewhere. The effect variance, $\sigma_\alpha^2$, has a scaled inverse chi-square distribution with scale parameter $S_\alpha^2$ and degree of freedom $\nu_\alpha$. If $\pi$ is treated as an unknown with a uniform prior, BayesC is called BayesC$\pi$ (Habier et al., 2011).

Erbe et al. (2012) proposed BayesR method. The priors for SNP effects follow a mixture of four normal distributions. Later this model was extended to BayesRC. In BayesRC, SNPs are grouped according to their biological functions and within each group, SNP effects follow the same mixture of four normal distributions (Hayes et al., 2014; MacLeod et al., 2014). MacLeod et al. (2014) demonstrated that the modified Bayesian models are powerful for identifying causal variants using large sequencing data.
BayesD which uses the gamma distribution as prior assumption on the scale parameter of SNP effect variances was proposed by Habier et al. (2011). Bayesian LASSO (Tibshirani, 1996) assumes a double exponential prior distributions for SNP effects (Usai et al., 2009). SSVS (stochastic search variable selection) is similar to BayesB and assumes that the conditional prior distribution of SNP effects follows a mixture of two normal distributions (Verbyla et al., 2009). Ante-BayesA and ante-BayesB which consider the correlation between SNP effects (Yang and Tempelman, 2012). BayesA and BayesB haplotype models fit haplotype alleles rather than SNP genotypes (Sun et al., 2014), and the Bayesian QTL model fits effects for unobserved QTL genotypes and utilizes information from LD, cosegregation, and pedigree (Zeng, 2015), were also applied in genomic selection. In addition to the parametric Bayesian models mentioned above, semi- and non-parametric machine learning methods have also been used for genomic prediction (de los Campos et al., 2009; Gianola et al., 2006).

There is debate over the necessity of implementing additional components in models for improving prediction accuracy, such as epistatic effects, polygenic effects, and interactions between genotype and environment. The prediction accuracy of GEBV is influenced by many factors, including choice of model, size of the training population, extent of LD, marker density, trait heritability, trait architecture, and the relationship between individuals in the training and validation populations. Some of these factors cannot be controlled, while others are unexpected and unknown in the real data analyses (de los Campos et al., 2013). It is always interesting to evaluate the effects of different factors and to investigate the performance of novel prediction models of both simulation and empirical studies.

1.5 Bibliography


CHAPTER 2. RECOMBINATION LOCATIONS AND RATES IN BEEF CATTLE ASSESSED FROM PARENT-OFFSPRING PAIRS

Zi-Qing Weng\textsuperscript{1}, Mahdi Saatchi\textsuperscript{1}, Robert Schnabel\textsuperscript{2}, Jeremy Taylor\textsuperscript{2}, Dorian Garrick\textsuperscript{1}

\textsuperscript{1}Department of Animal Science, Iowa State University, Ames, IA 50010, USA
\textsuperscript{2}Division of Animal Science, University of Missouri, Columbia, MO 65211, USA

A paper published in \textit{Genetics Selection Evolution 2014, 46:34}

2.1 Abstract

Recombination events tend to occur in hotspots and vary in number among individuals. The presence of recombination influences the accuracy of haplotype phasing and the imputation of missing genotypes. Genes that influence genome-wide recombination rate have been discovered in mammals, yeast, and plants. Our aim was to investigate the influence of recombination on haplotype phasing, locate recombination hotspots, scan the genome for Quantitative Trait Loci (QTL) and identify candidate genes that influence recombination, and quantify the impact of recombination on the accuracy of genotype imputation in beef cattle. 2775 Angus and 1485 Limousin parent-verified sire/offspring pairs were genotyped with the Illumina BovineSNP50 chip. Haplotype phasing was performed with DAGPHASE and BEAGLE using UMD3.1 assembly SNP (single nucleotide polymorphism) coordinates. Recombination events were detected by comparing the two reconstructed chromosomal haplotypes inherited by each offspring with those of their sires. Expected crossover probabilities were estimated assuming no interference and a binomial distribution for the frequency
of crossovers. The BayesB approach for genome-wide association analysis implemented in the GenSel software was used to identify genomic regions harboring QTL with large effects on recombination. BEAGLE was used to impute Angus genotypes from a 7K subset to the 50K chip. DAGPHASE was superior to BEAGLE in haplotype phasing, which indicates that linkage information from relatives can improve its accuracy. The estimated genetic length of the 29 bovine autosomes was 3097 cM, with a expected genome-wide genetic distance averaging 1.23 cM/Mb. 427 and 348 windows containing recombination hotspots were detected in Angus and Limousin, respectively, of which 166 were in common. Several significant SNPs and candidate genes, which influence genome-wide recombination were localized in QTL regions detected in the two breeds. High-recombination rates hinder the accuracy of haplotype phasing and genotype imputation. In conclusion, small population sizes, inadequate half-sib family sizes, recombination, gene conversion, genotyping errors, and map errors reduce the accuracy of haplotype phasing and genotype imputation. Candidate regions associated with recombination were identified in both breeds. Recombination analysis may improve the accuracy of haplotype phasing and genotype imputation from low- to high-density SNP panels.

### 2.2 Introduction

The meiotic exchange of DNA between homologous chromosomes is known as recombination. Recombination events do not take place randomly throughout the genome, but tend to occur in recombination hotspots (Jeffreys et al., 2005), which are usually small regions in which recombination rate is significantly higher than in surrounding regions. Rates of recombination on different chromosomes are sex-specific, with rates in females being higher near centromeres and rates in males being higher near telomeres (Kong et al., 2002). Methodologies for discovering recombination hotspots and descriptions of their properties have been reviewed from the perspectives of mammals (Kauppi et al., 2004; Paigen and Petkov, 2010) and plants (Mezard, 2006). Elucidating the characteristics of recombination might help understand the creation and loss of haplotypes and explain genome-wide variation in linkage disequilibrium (LD).
Recombination rates are related to distance from the centromere (Mezard, 2006; Jensen-Seaman et al., 2004), and regional GC content (Galtier et al., 2001; Paape et al., 2012). The location and activity of recombination hotspots is regulated both by cis and trans acting genes (Paigen and Petkov, 2010). Trans-acting genes, such as PRDM9 control hotspot activation in mice (Parvanov et al., 2010) and humans (Jeffreys et al., 2005; Paigen and Petkov, 2010), and allow a meiosis-specific protein, SPO11 to initiate recombination (Paigen and Petkov, 2010). Moreover, REC8 (Parvanov et al., 2010), RNF212 (Jeffreys et al., 2005; Paigen and Petkov, 2010; Sandor et al., 2012) and other loci (Paigen and Petkov, 2010) have been found to influence genome-wide recombination activity in cattle (Sandor et al., 2012) and humans (Chowdhury et al., 2009; Kong et al., 2008b).

Genome-wide association studies (GWAS) associate genomic variants with a trait of interest to identify positional candidate loci (Weller, 2009; Lango and Weedon, 2008). Haplotype-based association tests and imputation from low- to high-density genotyping panels can both improve the power of GWAS to detect QTL (Browning, 2008) and most methods for haplotype phasing can also be used for genotype imputation. Furthermore, the estimation of haplotype phase can use LD information (Browning and Browning, 2007a) and/or pedigree structure (Druet and Georges, 2010). Statistical models used to infer haplotype phase and impute missing genotypes include Hidden Markov models (Scheet and Stephens, 2006; Browning and Browning, 2007b), rule-based approaches (Habier et al., 2010), long-range phasing algorithms (Kong et al., 2008a), and other methods. The importance of haplotype phase estimation and genotype imputation is increasing as large-scale sequencing projects generate genome-wide genotype information.

Effectiveness of genotype phasing and imputation are influenced by marker density, extent of LD, effective population size, marker minor allele frequency (MAF), size of the training population, position on the chromosome, and the extent of pedigree relationships between training and imputed populations (Hickey et al., 2012; Zhang and Druet, 2010; Druet et al., 2010). Kirk and Cardon (2002) pointed out that a small number of genotyping errors can significantly decrease the apparent haplotype frequency and the accuracy of haplotype reconstruction. Haplotype frequencies and counts are also affected by recombination.
Although Sandor et al. (2012) have reported estimated heritabilities of recombination rate and the identification of recombination hotspots and quantitative trait loci (QTL) in dairy cattle, recombination rates have been less investigated in cattle than in mice, humans and other mammals. In our study, we quantified recombination rates and their impact on phasing accuracy in two purebred beef cattle populations i.e. Angus and Limousin. Our goals were to: (i) examine the impact of pedigree information, phasing method, and single nucleotide polymorphism (SNP) location errors on the inference of haplotypes, (ii) quantify the impact of recombination on haplotype phasing, (iii) locate recombination hot windows and QTL which influence genome-wide recombination numbers (GRN), and (iv) evaluate the relationship between recombination rate and accuracy of genotype imputation in beef cattle.

2.3 Methods

2.3.1 Genotype and phenotype

A total of 3570 Angus bulls born between 1955 and 2008, and 2275 Limousin cattle (1319 bulls and 955 daughters) born between 1974 and 2007 that were genotyped with the BovineSNP50 BeadChip (Illumina, San Diego, CA) were used in this study. Genotypes were obtained using DNA samples extracted from semen or hair samples and did not require an approved animal use and care protocol. Genome-wide Mendelian consistency was tested on sire-offspring pairs, and those that failed or had genotype call rates (CR) below 95% were removed. After selection, 2778 Angus and 1485 Limousin parent-verified sire-offspring pairs remained. The average sizes of the 604 Angus and 235 Limousin half-sib families were 4.6 (between 1 and 103) and 6.3 (between 1 and 135), respectively. Individual SNPs with a CR less than 0.95, a MAF less than 0.01, a p value for a Hardy Weinberg equilibrium test less than 0.001, or a Mendelian inconsistency rate greater than 0.0024 (95% quantile) were removed. After quality control, 40 990 SNPs across 29 Bos taurus (BTA) autosomes in Angus and 38 815 SNPs in Limousin remained, of which 34 788 were in common. Missing 50K genotypes (0.45% and 0.02% of all Angus and Limousin genotypes, respectively) were imputed using BEAGLE3.3 (Browning and Browning, 2007b). In order to evaluate the relationship
between recombination rate and genotype imputation, a cross-validation study was conducted to quantify the imputation accuracy. Imputation from low- to high-density SNP panels was performed in Angus based on the use of only that subset of 7345 SNPs from the 50K panel that were on the GGP7K panel (GeneSeek, Lincoln, NE) for those animals used in the validation.

2.3.2 Halotype phasing

Phasing of haplotypes was performed one chromosome at a time using either the unrelated option in BEAGLE 3.3 (Browning and Browning, 2007b) or DAGPHASE2.4 (Druet and Georges, 2010). Phasing was first performed using SNP coordinates from the UMD3.1 assembly, which is known to contain some errors. The USDA-AIPL linkage map (http://www.ars.usda.gov/) that was constructed from linkage analysis based on the UMD3.1 assembly coordinates was used as an alternative. The comparison of phasing accuracies obtained from these two alternative marker orders was determined for BTA15 as a representative example. The hidden Markov model based on the Viterbi algorithm (Rabiner, 1989) implemented in BEAGLE was used to reconstruct haplotypes and to impute from low- to high-density SNP panels. In order to increase haplotype phasing accuracy, BEAGLE was set to run 20 iterations of the phasing algorithm and to sample 20 haplotype pairs per individual per iteration. Before using DAGPHASE, the assembly coordinates for the markers were used to generate a genetic map assuming that 1 Mb equals 1 cM. Then, DAGPHASE was used to reconstruct haplotypes based on the output of a directed acyclic graph (DAG) from BEAGLE with scale and shift parameters set at 2.0 and 0.1, respectively.

The number of different haplotypes in every 1 Mb window was counted for each chromosome in the Angus population. Phasing errors can result in erroneous haplotypes that might occur at low frequency, so only haplotypes with a frequency greater than 5% (>5% quantile) in each window were included.

2.3.3 Estimating recombination events

Recombination events were identified as phase changes in the transmitted gametes by comparing the two reconstructed haplotypes inherited by each offspring with the two reconstructed haplotypes
of their sire. Each recombination event was localized to a recombination interval defined by a pair of informative markers for which the phase was known. Haplotype mismatches were not common but were identified when the putative paternally-inherited haplotype of the offspring identified by BEAGLE or DAGPHASE was not identical to either of the haplotypes of the sire. Double crossover events that occurred in intervals less than 2 Mb, animals with more than three crossover events per chromosome, parent-offspring pairs with a haplotype mismatch rate greater than 0.05, crossover events occurring in 1 Mb windows for which the estimated recombination rate was significantly greater than 0.025 or which contained SNPs with a haplotype mismatch rate greater than 0.05 were ignored. Such unlikely crossover events were attributed to either genotyping or phasing errors.

The GRN for each parent-offspring pair was calculated as the summation of observed crossover events across the 29 autosomes. On average, one crossover event occurs on a chromosome of size 1 Morgan (M) (Ott, 1991). Accordingly, the average genome-wide recombination distance per Mb was calculated as the GRN divided by the total length of the 29 bovine autosomes. We found that GRN decreased with increasing family size and that haplotype phasing error rates were inflated in smaller families. As a result, only half-sib families with at least three offspring were retained in the following analysis.

The observed probabilities of 0, 1, 2, 3, >3 crossover events were separately calculated for every autosome. For a given number of crossovers, after removing unlikely crossover events described above, the number of parent-offspring pairs with that number of crossovers was divided by the total number of parent-offspring pairs in the analyzed population, to obtain the observed crossover probability for that chromosome. This produced five observed crossover probabilities for every autosome.

The five expected crossover probabilities were calculated for each autosome based on its length in base pairs assuming that crossover events follow a binomial distribution (Karlin, 1984). Ott (1991) pointed out that N = 4 would be a reasonable maximum crossover number for chromosomes for which the maximum recombination rate is less than 0.5. The equation to calculate the expected crossover probabilities was:
\[ p = \binom{N}{t} (x/N)^t (1-x/N)^{N-t}, \]  

where \( p \) is the expected crossover probability, \( N \) is the maximum number of crossover events per chromosome, \( t \) is the observed number of crossover events (0/1/2/3/4) per chromosome, and \( x \) is the length (M) of the corresponding chromosome, assuming 100 Mb is 1 M.

The expected genetic length of each chromosome (M) was computed as \( \sum_{i=1}^{4} i p_i \), where \( i \) is the number of crossover events (1/2/3/4) on the corresponding chromosome, and \( p_i \) is the expected probability of crossover \( i \). The expected chromosomal genetic distance per Mb was calculated as the expected genetic length (cM) divided by the physical length (Mb) of the corresponding chromosome.

Recombination rate was estimated for every non-overlapping 1 Mb window to identify recombination hot windows. Some recombination intervals for a particular recombination event could not be localized to positions strictly within a single 1 Mb window. In those cases, a part of the recombination event was considered to have occurred in each window that spanned the recombination interval. The recombination rate in a defined 1 Mb window was computed as:

\[ c_w = \left( \sum_{k=1}^{n} x_k / r_k \right) / T, \]

where \( c_w \) is the observed window recombination rate, \( n \) is the total number of recombination events observed on the corresponding chromosome, \( x_k \) is the overlap (in Mb) between the 1 Mb window and recombination interval \( k \), \( r_k \) is the length (in Mb) of the recombination interval, and \( T \) is the total number of sire-offspring pairs.

### 2.3.4 Estimating heritabilities

Genome-wide recombination numbers of sires were treated as phenotypes, thus sires with multiple offspring had repeated records. Narrow sense heritabilities (\( h^2 \)) of GRN were estimated separately for each breed using a repeatability model in ASReml3.0 (Gilmour et al., 2009). The model equation was:
\[ y = 1\mu + Zu + Zp + e, [3] \]

where \( y \) is the vector of repeated genome-wide recombination phenotypes for sires, \( \mu \) represents the unknown mean treated as a fixed effect, \( u \) is the vector of random animal effects with \( Var(u) = A\sigma^2_u \), where \( A \) is the pedigree relationship matrix among sires, \( p \) is the vector of permanent environmental effects, \( 1 \) and \( Z \) are design matrices, and \( e \) is the vector of residual effects.

A marker-based heritability was estimated using a BayesC model (Kizilkaya et al., 2011; Habier et al., 2011) as implemented in GENSEL4.0 software (Fernando and Garrick, 2013). BayesC assumes that all SNP effects have a common variance, and the prior for that variance has a scaled inverse Chi-square distribution. The model equation was:

\[ y_i = \mu + \sum_{j=1}^{k} z_{ij} s_j + e_i, [4] \]

where \( y_i \) is the average GRN for sire \( i \), \( \mu \) is the population mean, \( k \) is the number of SNP, \( z_{ij} \) is genotype code (0/1/2) for SNP \( j \) in sire \( i \), \( s_j \) is the random effect for SNP \( j \) with \( s_j \overset{i.i.d.}{\sim} N(0, \sigma^2_s) \), and \( e_i \) is a weighted residual effect. Parameter \( \pi \) was set to 0 in this study. BayesC with \( \pi \) equal to 0 is equivalent to GBLUP (Genomic Best Linear Unbiased Predictor), except that the variance components are treated as unknown with scaled inverse chi-squared priors. Markov chain Monte Carlo (MCMC) sampling with 41 000 iterations in which the first 1000 samples were discarded for burn-in, was used to make inferences about variance components and heritability. The weighting factor \( (w_n) \) (Garrick et al., 2009) for residual variance was calculated as:

\[ w_n = \frac{1 - h^2}{ch^2 + \frac{1+(n-1)c}{n} - h^2}, [5] \]

where \( h^2 \) is the narrow sense heritability estimated from pedigree, \( c \) is the proportion of genetic variation that could not be explained by markers, \( t \) is the repeatability, and \( n \) is the number of
observations for the sire. In this study, $c$ was assumed to be equal to 0.40 for both Angus and Limousin, according to Saatchi et al. (2012).

### 2.3.5 Genome-wide association study

Mapping QTL that influence sire mean GRN was undertaken using the BayesB method (Meuwissen et al., 2001) with weighting factors defined as for the above model implemented in GENSEL4.0 software (Fernando and Garrick, 2013). BayesB assumes that each SNP effect is drawn from a distribution with a locus-specific variance with scaled inverse Chi-square prior distributions, and that a fraction $(1 - \pi)$ of the markers have non-zero effects. Parameter $\pi$ was assumed to be equal to 0.995, which results in about 0.5% of the SNPs fit in the model at each iteration. Based on simulations, Sun et al. (2011) showed that the BayesB method could precisely map QTL. The genome was divided into non-overlapping 1 Mb windows and the posterior distribution of the percentage of genetic variance attributed to each window was constructed from the MCMC samples (e.g. Wolc et al. (2012)). The expected percentage of genetic variance explained by each of the $\sim$2600 1 Mb windows is about 0.04% under a polygenic model. Windows that explained at least 0.2% (5 fold the expected percentage) of the genetic variance (Wolc et al., 2012), and extended regions on either side of these windows (±2 Mb) were considered to represent QTL. Unpublished simulations using beef cattle genotypes showed that the location of a QTL can be up to 2 Mb up- or downstream of a 1 Mb window that explains a high proportion of genetic variance. The window posterior probabilities of association (WPPA) of candidate windows (i.e. with at least 0.2% genetic variance) was at least 1.5-fold greater than the average WPPA of 1 Mb windows across the genome. WPPA is the posterior probability that a window harbors a QTL, which is the proportion of samples for which at least one SNP in the window was included with a non-zero effect. The SNP that had the highest posterior probability of inclusion (PPI) and explained the largest proportion of genetic variance in each central candidate window was identified as a candidate SNP. The PPI is estimated as the percentage of MCMC samples in which a given SNP had a non-zero effect. The proportion of genetic variance explained by each candidate SNP was assessed as the difference in genetic variance explained by the window when it included or excluded the candidate SNP. Significance of the effect
of the candidate SNP was evaluated in an animal model with ASReml3.0 (Gilmour et al., 2009), by fitting the SNP genotype as a fixed class effect. Bonferroni adjustment was applied to \( p \) values from that single SNP analysis by accounting for the number of effective chromosome segments across the genome \( (M_e) \), which was calculated as Goddard et al. (2011):

\[
M_e = 2N_e L k / \log(N_e L), \quad [6]
\]

where \( N_e \) is the effective population size, \( L \) is the average length of a chromosome in Morgan (~1 M), and \( k \) is the number of chromosomes \( (k = 30) \). In this study, \( N_e \) was assumed to be equal to 545 for Angus and 91 for Limousin, based on Stachowicz et al. (2013).

Using the human-bovine comparative map implemented in VCMap3.0 (http://www.animalgenome.org/VCmap/manual/v3), orthologous human genome regions corresponding to candidate bovine windows were located. Positional candidate genes within these orthologous human regions were identified using the NCBI Human Genome Overview Build 36.3. A list of previously published human candidate genes related to meiosis, recombination, or the cell cycle were extracted from OMIM (http://www.omim.org). Using VCMap3.0 or Ensembl, locations of the bovine orthologs of these genes were mapped to the bovine genome. These locations were used to test for concordance between locations of candidate genes and identified QTL.

### 2.3.6 Imputation from low- to high-density and cross-validation of imputation accuracy

Cross-validation was used to determine the accuracy of BEAGLE imputation from 7K to 50K SNPs in the Angus dataset. The genotyped bulls were clustered into five groups using a K-means clustering method based on additive genetic relationships between animals (Saatchi et al., 2011). The aim of this method was to increase within-group and decrease between-group relationships. Four testing groups were used for phasing haplotypes from the 50K SNP genotypes, while imputation from the 7K SNP panel was performed in the fifth validation group. This was repeated with each of the five groups being treated once as the validation group. The 7K SNP genotypes were extracted from the 50K SNP genotypes for the validation group.
Accuracies of imputation were quantified per marker and summarized per chromosome and per animal. The imputation accuracy was evaluated as the fraction of the imputed genotypes that were identical to the original genotypes on the 50K SNP panel in the validation group. Imputation accuracy was also quantified separately in every 1 Mb window along each chromosome.

Levels of LD between every two adjacent SNPs were evaluated as $r^2$, the squared simple correlation between genotypes of two adjacent SNPs using R software. These measures of LD between adjacent markers were averaged to provide a single measure of LD for each 1 Mb window.

2.4 Results

2.4.1 Phasing accuracy and crossover probability

Figure 2.1 shows that the probability of zero crossover events per chromosome increased with decreasing chromosome length, while the probability of two or more crossover events decreased. The observed proportion of individuals with more than three crossover events was higher than the expected value for all autosomes. In general, DAGPHASE, which uses linkage information from parent-offspring relationships, produced a distribution of crossovers for which the observed frequencies were closer to the expected values, indicating that it was superior for phasing compared to BEAGLE with pedigree ignored.

The distributions of observed numbers of crossovers on BTA15 were close to the expected values for both the Angus and Limousin breeds, except for the proportion of more than two crossover events per meiosis (Figure 2.2A), which exceeded the expected values. Possible reasons for the phasing errors that likely caused the discrepancies between expected and observed crossover probabilities are small half-sib family sizes (median size was 2 in both populations), limited numbers of parent-offspring pairs, and errors in some mapped SNP locations. Both the Angus (41.4%) and Limousin (36.6%) breeds had a large proportion of half-sib families represented by only one son. Compared to Angus, the Limousin breed had a higher probability of more than two crossover events per meiosis, probably because of its smaller sample size which reduces the accuracy of haplotype phasing. The higher accuracy in Angus compared to Limousin was also observed for autosomes other than BTA15.
(Figure 2.2B). It has been found that the larger the phasing sample size, the greater the haplotype phasing accuracy (Browning, 2008).

Figure 2.3 compares observed probabilities of crossover events on Angus BTA15 using UMD3.1 versus USDA-AIPL locus coordinates. A total of 1304 SNPs were assigned based on the UMD3.1 and 1262 SNPs based on the USDA-AIPL, with 1234 common SNPs. The estimated probability of more than two crossover events using USDA-AIPL coordinates was smaller than that using the UMD3.1 coordinates, which suggests that a better genome assembly can improve the accuracy of phasing.

2.4.2 Number of haplotypes and recombination rates

The average number of unique haplotypes per 1 Mb windows (±SD) was 37.1±13.8 in Angus. Despite the presence of outliers, there was a linear relationship between number of haplotypes and recombination rate (Figure 2.4). The number of haplotypes declined with decreasing window-wide recombination rates, because new haplotypes are formed by recombination. Recombination hot or cold windows were defined as windows with recombination rates greater than 0.02 (≥ 1.5 standard deviations from the mean) and lower than 0.004, respectively. These definitions differ from those previously used for dairy cattle (60 Kb window) (Sandor et al., 2012) and humans (<2 Kb window) (Paigen and Petkov, 2010), because of the different lengths of the defined window. The average number of SNPs (±SD) was 17.8 (±4.5) in hot windows and 13.7 (±5.8) in cold windows. The average numbers of haplotypes (±SD) in hot and cold windows were adjusted for the corresponding average number of SNPs. The number of haplotypes was equal to 50.9±13.3 (ranging from 23.4 to 99.5) in hot windows, and 24.5±6.8 (ranging from 7.4 to 47.2) in cold windows, respectively (Figure 2.5). The correlation coefficient between the average number of unique haplotypes within each window and the recombination rate in that 1 Mb window was 0.64. All autosomes showed significant disparities in numbers of haplotypes per 1 Mb windows between hot and cold windows.
2.4.3 Analysis of genome-wide recombination number (GRN)

A total of 76,186 and 32,052 informative crossover events were identified in Angus and Limousin gametes, respectively. The physical length of the 29 bovine autosomes is 2511.4 Mb (UMD3.1), which corresponds to an average genetic length of 3097 cM (Table 2.1). On average, the expected genome-wide genetic distance per Mb across the 29 autosomes was 1.23 cM/Mb. BTA20 had the lowest, and BTA23 the highest cM per Mb ratio. Figure 2.6 shows the expected and estimated genetic distances per Mb for the 29 bovine autosomes. Genetic distances per Mb differed between chromosomes; short chromosomes had greater genetic distances per Mb than long chromosomes (Kong et al., 2002). Chromosomal genetic distances per Mb estimated in the two breeds were similar, with a correlation coefficient of 0.84 between Angus and Limousin. However, the genetic distances per Mb were lower than the expected values for most autosomes, which suggests that conservative filtering of unlikely crossover events leads to an underestimation of the chromosome-specific recombination distance per Mb.

Taking BTA15 as an example, the correlation between window recombination rates in the Angus and Limousin breeds was equal to 0.56, and recombination rates in a 1 Mb window varied from 0 to over 0.02 (Figure 2.3). A large number of recombination hot and cold windows were detected across the chromosome. Since bovine chromosomes are acrocentric, with the centromere at the proximal chromosome end, recombination rates were relatively low in that region. Reduced information at the proximal end of the chromosome could also lead to a low accuracy of detected recombination events. As shown in Figure 2.8, the location of hot and cold windows for recombination was consistent for the two breeds across the genome, although, in some instances, window shifts existed, such that a higher recombination rate for Angus corresponds to a lower recombination rate for Limousin and vice versa. Across the genome, the correlation of 1 Mb window recombination rate between the Angus and Limousin breeds was high, with a correlation coefficient of 0.49. The average window recombination rates per 1 Mb (±SD) were equal to 0.0099±0.0052 and 0.0088±0.0053 in Angus and Limousin breeds, respectively. A total of 427 and 348 hot windows were identified in Angus and Limousin, respectively, of which 166 were in common. Hot windows were found in both the
proximal and distal chromosome ends, while cold windows clustered around the middle of each chromosome and the proximal chromosome end.

The average number of recombination events per chromosome differed between autosomes. Longer autosomes tended to have more recombination events. The average GRN (±SD) was equal to 27.4±5.0 in Angus and 26.9±4.8 in Limousin. These values were close to the paternal recombination numbers of 27.6 reported by Chowdhury et al. (2009) and 27.0 reported by Kong et al. (2008a) in humans. GRN did not differ significantly between the breeds (Figure 2.9). Estimates of GRN slightly decreased with increasing family size (with a correlation coefficient near -0.1), as did the observed variation of GRN across families, which is probably due to an increase of phasing errors in small families (Figure 2.10).

2.4.4 Estimated heritability and QTL for genome-wide recombination number

The pedigree-based estimates of heritability of GRN (±SE) by ASReml3.0 (Gilmour et al., 2009) were equal to 0.26±0.030 and 0.23±0.042 and estimates of repeatability were equal to 0.33±0.027 and 0.30±0.038 in Angus and Limousin, respectively. However, estimates of marker-based heritability of GRN (±SE) by BayesC in GENSEL4.0 software (Fernando and Garrick, 2013) were slightly lower, i.e. 0.17±0.039 in Angus and 0.14±0.031 in Limousin. Results reported in Saatchi et al. (2011) demonstrate that the marker-based heritability of routinely recorded traits (e.g. calving ease) of American Angus beef cattle was sometimes lower than the value of the pedigree-based heritability. This suggests that markers only captured a proportion of the genetic variance estimated from pedigree.

Manhattan plots of the proportion of genetic variance explained by each 1 Mb window across the genome for GRN in Angus and Limousin are in Figure 2.11. The number of windows explaining at least 0.2% of the additive genetic variance was 35 in Angus and 22 in Limousin. The cumulative variance explained by those windows was equal to 17.8% in Angus and 8.2% in Limousin. Windows that exceeded 0.2% additive genetic variance and had 1.5-fold average WPPA were considered to be significant for further study (Table 2.2). Different candidate SNPs were identified within each window in Angus and Limousin. The highest proportion of genetic variance (3.48%) was
explained by a 1 Mb window located at 67 Mb on BTA21 for Angus, which had a high WPPA (0.45) and a significant SNP accounting for 3.42% of the genetic variance. The most significant region in the Limousin breed was a 1 Mb window located at 89 Mb on BTA4 and explained 2.55% of genetic variance. Positional candidate genes (Table 2.2), that have been reported to be involved in meiotic recombination, DNA replication, DNA repair or the cell cycle (http://www.omim.org) were detected within or near (±2 Mb) significant windows but only in Angus; *RAD51C*, *RAD52C*, and *XRCC3* are involved in both meiotic recombination and repair of damaged DNA, while *PRMT8* is only involved in DNA repair, whereas *PTPRM* and *RAD17* regulate cellular processes, such as differentiation and cell cycle checkpoint control.

### 2.4.5 Imputation accuracy and recombination rate

In Angus, the average imputation accuracy across animals was equal to 0.981, ranging from 0.81 to 1.00, and the average imputation accuracy per chromosome was also equal to 0.981, ranging from 0.97 to 0.99 (Table 2.3). BTA21 had the lowest imputation accuracy (0.973), while BTA4 had the highest accuracy (0.985). The average marker density (i.e. average distance in kb between two adjacent markers) was equal to 61.0 kb, ranging from 54.6 to 70.9 kb, and the average r² between adjacent markers within each 1 Mb window was 0.237, ranging from 0.192 to 0.269. Imputation accuracy increased slightly as marker density and r² increased.

In Figure 2.12A, Angus bulls were grouped according to the number of observed crossover events per chromosome. The average imputation accuracy (±SD) in groups with no, one, two and more than two crossover events was equal to 0.986±0.00835, 0.983±0.0191, 0.981±0.0203, and 0.980±0.0215, respectively. Imputation accuracy decreased only slightly as the number of crossover events increased. Taking BTA1 as an example, imputation accuracy was highest in individuals with no crossover events, and lowest in individuals with more than two observed crossover events due to a higher risk of phasing errors.

Window-wide imputation accuracy decreased with increasing recombination rate (See Figure 2.12B). The correlation coefficient between window-wide imputation accuracy and recombination rate was equal to -0.49 and the regression coefficient of imputation accuracy on recombination rate
was equal to -1.0. Average imputation accuracies of 0.975 (ranging from 0.913 to 0.995), and 0.990 (ranging from 0.927 to 1.00) were found in hot and cold windows, respectively. Figure 2.13 shows that imputation accuracy was lower in recombination hot windows than in cold windows.

2.5 Discussion

2.5.1 Impact of phasing methods

Accuracy of phasing haplotypes was quantified in relation to recombination events. DAGPHASE (Druet and Georges, 2010), which uses linkage information from parent-offspring relationships was superior to using BEAGLE (Browning and Browning, 2007b) with relationships for phasing haplotypes ignored. DAGPHASE extracts both population LD and linkage information for phasing, rather than relying on LD alone (Druet and Georges, 2010). To infer haplotypes of offspring with both genotyped parents, parental haplotypes and linkage information were used. For offspring with one genotyped parent, linkage and LD information were jointly used by DAGPHASE, while only LD information extracted from the BEAGLE-produced DAG file was used to determine haplotypes of individuals with both parents non-genotyped. DAGPHASE uses a diploid Hidden Markov model (HMM) and Viterbi algorithm with linkage and LD information to improve phase reconstruction (Druet and Georges, 2010). BEAGLE also assumes a HMM but uses EM-style updating that locally clusters haplotypes (Browning and Browning, 2007b). BEAGLE with the options applied in this study phased related individuals as if they were unrelated. Ignoring additive genetic relationships can lead to inconsistent haplotypes between related individuals. Thus, the accuracy of estimating haplotypes can be improved by using linkage information from relatives (Browning, 2008). However, DAGPHASE does not reconstruct haplotypes of parents, while LINKPHASE, another algorithm from the Phasebook software package (Druet and Georges, 2010), could use offspring information to modify phase reconstruction in parents. Further investigation is needed to verify whether the combination of both DAGPHASE and LINKPHASE may lead to more robust results in large families.
2.5.2 Assumptions for the identification of recombination events

Haldane’s (e.g., Kauppi et al. (2004); Sandor et al. (2012)) and Kosambi’s map functions (e.g., Kong et al. (2002); Arias et al. (2009)) are often used to calculate recombination probabilities and to estimate the genetic length of each chromosome. In contrast, in this study crossover events were assumed to follow a binomial distribution following Karlin (1984). Haldane (1919) assumed a Poisson distribution and that crossovers in adjacent intervals occur independently. Kosambi’s function (Kosambi, 1944) makes strong assumptions about interference between nearby crossovers. Kosambi’s function appears to produce more realistic map distances than Haldane’s function (Ott, 1991). Both these map functions postulate that theoretically an unlimited number of crossovers can occur per chromosome. However, Karlin assumed that, at most, \( N \) crossovers could be independently distributed in an interval, with the number of events following a binomial distribution.

In Figure 2.14, the autosomes were sorted by their genetic length (M), and chromosomal recombination rates produced using Karlin’s formula (with \( N = 4 \)) were intermediate to those produced by the Kosambi and Haldane functions. Lian et al. (2008) reported that crossover interference increases with decreasing chromosome length. Since strong positive crossover interference exists, quantifying the level of crossover interference on each chromosome could improve the estimation of recombination rates. For example, Broman and Weber (2000) found that in human family data a gamma distribution better characterized chromosome-specific crossover interference than did four other distributions.

Many instances of double crossover events over a short distance (i.e. within 1 Mb) and individuals with excessive numbers of recombination events were observed. Genotyping, phasing, and map errors can cause overestimation of recombination rates and lead to biases in determining haplotypes from genotypic data. The data were carefully filtered for the presence of apparent double crossover events. Two crossover events separated by a small distance (<2 Mb) were attributed to phasing errors and were ignored from calculation of crossover probabilities and GRN. Other phasing or genotyping errors, such as when a sire had a recombination rate significantly higher than 0.025 in a certain window, or when all sons in a family showed two or more crossover events at the same location, were also ignored in subsequent analyses. More than two crossover events located nearby
suggests a marker order error. Rather than removing the unlikely recombination event as we did here, a reordering of the markers should be considered for further study (Broman et al., 1998). The existence of gene conversion across the chromosome is another possible cause of apparent double crossover events. During meiosis, heteroduplexes are generated in the form of either crossover or non-crossover events. Gene conversion is the non-crossover form of a heteroduplex, which is the consequence of mismatched base pairs in a heteroduplex region corrected by DNA repair mechanisms (Alberts et al., 2008). Heteroduplex regions can extend for several kb and can contain some mismatched base pairs (Alberts et al., 2008). The resolution of the SNP panel used in this study did not allow heteroduplex regions to be confirmed. Some of the double crossovers over a short distance (i.e. within 1 Mb), which we ignored may represent gene conversion. There is evidence that crossover hot spots are hot spots for gene conversion in mice (Guillon and de Massy, 2002) and humans (Jeffreys and May, 2004).

2.5.3 Impact of homozygous segments on the identification of recombination events

Each recombination event was identified within a recombination interval, which is the segment of homozygous non-informative loci that could have belonged to the haplotype of either parent. Recombination that occurs within a long homozygous segment cannot be localized. The average length of recombination intervals was 1.38 Mb (\(\sim\)23 SNPs) for Angus and for Limousin. The average number of SNPs in homozygous segments in sires was equal to 3.82\(\pm\)12.34 (\(\sim\)0.23\(\pm\)0.74 Mb) in Angus and 3.42\(\pm\)6.73 (\(\sim\)0.20\(\pm\)0.40 Mb) in Limousin. Long homozygous segments were defined as those containing more than 20 SNPs. On average, an autosome contained 3.4\(\pm\)1.3 such long segments in Angus and 1.5\(\pm\)0.7 in Limousin. Therefore, the impact of long homozygous segments on the identification of recombination events is not expected to be a factor in this study.

2.5.4 Estimation of chromosome genetic distance per Mb and heritability of GRN

The average chromosome-specific cM per Mb ratio increased with chromosome size, as in previous studies (Kong et al., 2002; Arias et al., 2009). Kong et al. (2002) reported an average genomic genetic distance per Mb of 1.19 cM/Mb in humans. Our estimate of 1.23 cM/Mb was
similar to the 1.25 cM/Mb value reported by Arias et al. (2009) based on the Btau4.0 assembly with a total physical length of 2468.3 Mb for the 29 autosomes, rather than 2511.4 Mb for the UMD3.1 assembly. Inconsistencies in chromosome lengths and marker order led to different chromosome genetic lengths.

The pedigree-based heritabilities in Angus (0.26) and Limousin (0.23) were slightly higher than that (0.22) reported by Sandor et al. (2012) in a sample of 13,975 Dutch Holstein-Friesian bulls within three-generation paternal half-sib families. A repeatability model with GRN records for sires was considered in our study. Sandor et al. (2012) fitted genome-wide recombination rates corrected for family size in an animal model. Differences in generation structures, sample sizes, and estimation models could lead to disparities in heritability estimates. Kong et al. (2004) estimated a heritability of recombination rate of 0.30 in humans, which indicates that a large genetic component underlies variation in recombination rate and that the heritability of GRN differs between breeds and species.

2.5.5 Genomic regions associated with GRN

GWAS have been widely applied in humans (Donnelly, 2008) and livestock (Goddard and Hayes, 2009). Because inferences from Bayesian methods are based on the joint posterior distribution, they are useful for GWAS (Zou and Zeng, 2008). Regardless of the method used, detection of large-effect QTL is easier than detection of small-effect QTL. Unpublished simulations using beef cattle genotypes shows that causal mutations may lie in regions upstream or downstream of the window that has the strongest association. Although flanking regions near the most strongly associated windows (±2 Mb) were investigated, further analyses are needed to confirm our results. Significant windows associated with genome-wide recombination were located on different chromosomes in the Angus and Limousin breeds. However, Saatchi and Garrick (2013) identified QTL of growth and production traits with consistent effects across multiple breeds. Further investigation is needed to verify whether the location and impact of recombination QTL differs between breeds. Two regions, one on BTA6 for the Angus breed and one on BTA4 for the Limousin breed, were found to explain a significant proportion of the genetic variance (>0.2%), but SNPs with the highest PPI within these two windows were not significant. However, the previous validated genes RNF212 (Sandor
et al., 2012; Chowdhury et al., 2009; Kong et al., 2008b) and SPO11 (Paigen and Petkov, 2010), were 4 Mb downstream from the window detected on BTA6 in Angus and 4 Mb upstream from the window detected on BTA4 in Limousin, respectively. Differences in mapping results for genome-wide recombination number (or rate) in plants (Paape et al., 2012), humans (Chowdhury et al., 2009), dairy cattle (Sandor et al., 2012), and beef cattle suggest that genome-wide recombination could be regulated in a species-specific manner, that the effects of QTL differ between species, and that the genetic determinism of regulation of recombination is probably polygenic.

2.5.6 Influence of recombination on imputation accuracy and haplotype phasing

Imputation accuracy was higher in the regions with denser markers and higher LD levels ($r^2$). With denser markers, better imputation accuracy is expected (Druet et al., 2010; Weng et al., 2013) and stronger LD between SNPs improves the reconstruction of haplotypes (Yu and Schaid, 2007). Higher recombination rates reduced the accuracy of haplotype phasing and genotype imputation and conversely, imputation accuracy was lowest in recombination hot windows.

The use of haplotypes is advantageous for genomic prediction and GWAS (Browning and Browning, 2007a; Calus et al., 2008) provide accurately phased haplotypes. Marker location errors within a genome assembly can be detected by recombination analysis. An improved genome assembly leading to a more accurate reflection of true meiotic recombination could be produced by reordering the markers. Similarly, the accuracy of haplotype phasing and imputation from low- to high-density SNP panels could be improved by using recombination results. How to implement recombination information in haplotype phasing and imputation remains a challenging question.

2.6 Conclusions

This study investigated the relationships between recombination, haplotype phasing, and imputation in two breeds of cattle. The accuracy of phasing using DAGPHASE was superior to BEAGLE, which did not use linkage information from parent-offspring. The major reasons for the detection of unlikely recombination events are gene conversion and phasing errors. Gene conver-
sion is caused by mismatch correction in heteroduplex regions. Phasing errors can be influenced by limited sample size, small half-sib families, low marker density, and marker location errors in the genome assembly. The QTL mapping results for genome-wide recombination number in Angus differed from those in Limousin, which suggests that recombination is under polygenic control. High levels of recombination decrease the accuracy of phasing and genotype imputation. These results suggest that recombination analysis can detect location errors within the genome assembly, and could be used to improve the inference of haplotype phase and the accuracy of genotype imputation from low- to high-density panels.

2.7 Acknowledgements

This project was supported by National Research Initiative grants number 2008-35205-04687 and 2008-35205-18864 from the USDA Cooperative State Research, Education and Extension Service and National Research Initiative grant number 2009-65205-05635 from the USDA National Institute of Food and Agriculture. We acknowledge the constructive comments of two reviewers.

2.8 Bibliography


## 2.9 Tables

Table 2.1  **Physical length, estimated genetic length and recombination distance per Mb of bovine autosomes.**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Expectation of genetic length (cM)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Physical length (Mb)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>cM/Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>166.0</td>
<td>158.3</td>
<td>1.05</td>
</tr>
<tr>
<td>2</td>
<td>148.0</td>
<td>137.1</td>
<td>1.08</td>
</tr>
<tr>
<td>3</td>
<td>141.8</td>
<td>121.4</td>
<td>1.17</td>
</tr>
<tr>
<td>4</td>
<td>132.5</td>
<td>120.2</td>
<td>1.10</td>
</tr>
<tr>
<td>5</td>
<td>130.0</td>
<td>121.2</td>
<td>1.07</td>
</tr>
<tr>
<td>6</td>
<td>134.2</td>
<td>119.5</td>
<td>1.12</td>
</tr>
<tr>
<td>7</td>
<td>125.5</td>
<td>112.6</td>
<td>1.11</td>
</tr>
<tr>
<td>8</td>
<td>124.4</td>
<td>113.4</td>
<td>1.10</td>
</tr>
<tr>
<td>9</td>
<td>110.3</td>
<td>105.7</td>
<td>1.04</td>
</tr>
<tr>
<td>10</td>
<td>118.9</td>
<td>104.3</td>
<td>1.14</td>
</tr>
<tr>
<td>11</td>
<td>129.9</td>
<td>107.3</td>
<td>1.21</td>
</tr>
<tr>
<td>12</td>
<td>117.3</td>
<td>91.2</td>
<td>1.29</td>
</tr>
<tr>
<td>13</td>
<td>118.3</td>
<td>84.2</td>
<td>1.40</td>
</tr>
<tr>
<td>14</td>
<td>127.4</td>
<td>84.7</td>
<td>1.51</td>
</tr>
<tr>
<td>15</td>
<td>110.3</td>
<td>85.3</td>
<td>1.29</td>
</tr>
<tr>
<td>16</td>
<td>112.4</td>
<td>81.7</td>
<td>1.38</td>
</tr>
<tr>
<td>17</td>
<td>97.0</td>
<td>75.2</td>
<td>1.29</td>
</tr>
<tr>
<td>18</td>
<td>103.2</td>
<td>66.0</td>
<td>1.56</td>
</tr>
<tr>
<td>19</td>
<td>100.8</td>
<td>64.1</td>
<td>1.57</td>
</tr>
<tr>
<td>20</td>
<td>73.7</td>
<td>72.0</td>
<td>1.02</td>
</tr>
<tr>
<td>21</td>
<td>90.2</td>
<td>71.6</td>
<td>1.26</td>
</tr>
<tr>
<td>22</td>
<td>91.4</td>
<td>61.4</td>
<td>1.49</td>
</tr>
<tr>
<td>23</td>
<td>90.0</td>
<td>52.5</td>
<td>1.71</td>
</tr>
<tr>
<td>24</td>
<td>85.8</td>
<td>62.7</td>
<td>1.37</td>
</tr>
<tr>
<td>25</td>
<td>62.0</td>
<td>42.9</td>
<td>1.45</td>
</tr>
<tr>
<td>26</td>
<td>69.8</td>
<td>51.7</td>
<td>1.35</td>
</tr>
<tr>
<td>27</td>
<td>60.9</td>
<td>45.4</td>
<td>1.34</td>
</tr>
<tr>
<td>28</td>
<td>57.3</td>
<td>46.3</td>
<td>1.24</td>
</tr>
<tr>
<td>29</td>
<td>68.0</td>
<td>51.5</td>
<td>1.32</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3097.3</strong></td>
<td><strong>2511.4</strong></td>
<td><strong>1.23</strong></td>
</tr>
</tbody>
</table>

<sup>a</sup>Estimated using Karlin’s map function; <sup>b</sup>bovine UMD3.1 genome assembly.
Table 2.2  Candidate windows and SNPs for genome-wide recombination number in Angus and Limousin.  1 Mb windows that explain a significant proportion of genetic variation (>0.2%), and results for significant SNPs and positional candidate genes within these windows detected for genome-wide recombination number in Angus and Limousin.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Significant 1 Mb window</th>
<th>Significant SNP</th>
<th>Candidate Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BTA Window (Mb)</td>
<td>% genetic variance</td>
<td>Number of SNPs</td>
</tr>
<tr>
<td>Angus</td>
<td>5</td>
<td>106</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>115</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>10</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>12-13</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>67</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>3</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>39</td>
<td>1.71</td>
</tr>
<tr>
<td>Limousin</td>
<td>4</td>
<td>89</td>
<td>2.55</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>55</td>
<td>0.20</td>
</tr>
</tbody>
</table>

WPPA, window posterior probability of association; PPI, posterior probability of inclusion of the selected SNP; p-value adjusted by Bonferroni test, significance of the selected SNPs in single SNP analysis; distance, approximate distance (Mb) between the positions of the candidate gene and significant SNP.
Table 2.3  **Imputation accuracy, number of markers, SNP density and average LD for bovine autosomes in Angus based on UMD3.1 assembly locus coordinates.**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Imputation Accuracy</th>
<th># of markers</th>
<th>Marker density (kb)</th>
<th>$r^2$ in 1 Mb window</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.983± 0.007</td>
<td>2603</td>
<td>60.83</td>
<td>0.260± 0.303</td>
</tr>
<tr>
<td>2</td>
<td>0.984± 0.007</td>
<td>2221</td>
<td>61.71</td>
<td>0.234± 0.290</td>
</tr>
<tr>
<td>3</td>
<td>0.984± 0.008</td>
<td>2026</td>
<td>59.94</td>
<td>0.259± 0.305</td>
</tr>
<tr>
<td>4</td>
<td>0.985± 0.006</td>
<td>1984</td>
<td>60.58</td>
<td>0.248± 0.295</td>
</tr>
<tr>
<td>5</td>
<td>0.985± 0.007</td>
<td>1710</td>
<td>70.87</td>
<td>0.257± 0.307</td>
</tr>
<tr>
<td>6</td>
<td>0.984± 0.008</td>
<td>2027</td>
<td>58.93</td>
<td>0.258± 0.306</td>
</tr>
<tr>
<td>7</td>
<td>0.983± 0.007</td>
<td>1750</td>
<td>64.36</td>
<td>0.269± 0.311</td>
</tr>
<tr>
<td>8</td>
<td>0.983± 0.009</td>
<td>1803</td>
<td>62.89</td>
<td>0.254± 0.300</td>
</tr>
<tr>
<td>9</td>
<td>0.983± 0.007</td>
<td>1650</td>
<td>64.07</td>
<td>0.232± 0.293</td>
</tr>
<tr>
<td>10</td>
<td>0.980± 0.008</td>
<td>1731</td>
<td>60.26</td>
<td>0.217± 0.284</td>
</tr>
<tr>
<td>11</td>
<td>0.982± 0.009</td>
<td>1830</td>
<td>58.64</td>
<td>0.269± 0.305</td>
</tr>
<tr>
<td>12</td>
<td>0.983± 0.008</td>
<td>1344</td>
<td>67.83</td>
<td>0.234± 0.284</td>
</tr>
<tr>
<td>13</td>
<td>0.978± 0.011</td>
<td>1440</td>
<td>58.40</td>
<td>0.261± 0.303</td>
</tr>
<tr>
<td>14</td>
<td>0.983± 0.008</td>
<td>1455</td>
<td>58.18</td>
<td>0.268± 0.310</td>
</tr>
<tr>
<td>15</td>
<td>0.984± 0.008</td>
<td>1304</td>
<td>65.41</td>
<td>0.226± 0.287</td>
</tr>
<tr>
<td>16</td>
<td>0.981± 0.009</td>
<td>1352</td>
<td>60.45</td>
<td>0.236± 0.296</td>
</tr>
<tr>
<td>17</td>
<td>0.983± 0.011</td>
<td>1273</td>
<td>59.04</td>
<td>0.216± 0.272</td>
</tr>
<tr>
<td>18</td>
<td>0.980± 0.009</td>
<td>1061</td>
<td>62.21</td>
<td>0.251± 0.294</td>
</tr>
<tr>
<td>19</td>
<td>0.975± 0.011</td>
<td>1081</td>
<td>59.26</td>
<td>0.231± 0.277</td>
</tr>
<tr>
<td>20</td>
<td>0.984± 0.008</td>
<td>1223</td>
<td>58.91</td>
<td>0.232± 0.285</td>
</tr>
<tr>
<td>21</td>
<td>0.973± 0.010</td>
<td>1115</td>
<td>64.21</td>
<td>0.233± 0.286</td>
</tr>
<tr>
<td>22</td>
<td>0.980± 0.008</td>
<td>1041</td>
<td>59.02</td>
<td>0.236± 0.290</td>
</tr>
<tr>
<td>23</td>
<td>0.980± 0.008</td>
<td>857</td>
<td>61.30</td>
<td>0.197± 0.254</td>
</tr>
<tr>
<td>24</td>
<td>0.984± 0.007</td>
<td>1036</td>
<td>60.54</td>
<td>0.251± 0.300</td>
</tr>
<tr>
<td>25</td>
<td>0.980± 0.009</td>
<td>786</td>
<td>54.59</td>
<td>0.214± 0.267</td>
</tr>
<tr>
<td>26</td>
<td>0.975± 0.010</td>
<td>879</td>
<td>58.80</td>
<td>0.211± 0.279</td>
</tr>
<tr>
<td>27</td>
<td>0.981± 0.009</td>
<td>769</td>
<td>59.05</td>
<td>0.232± 0.285</td>
</tr>
<tr>
<td>28</td>
<td>0.976± 0.011</td>
<td>773</td>
<td>59.91</td>
<td>0.199± 0.246</td>
</tr>
<tr>
<td>29</td>
<td>0.979± 0.009</td>
<td>866</td>
<td>59.47</td>
<td>0.192± 0.243</td>
</tr>
</tbody>
</table>
Figure 2.1  Probability of zero (black), one (red), two (green), three (blue), and more than three (purple) crossover events per chromosome for each of the 29 bovine autosomes in Angus. Solid line represents expected probability calculated by Karlin’s map function, dashed line represents observed probability using DAGPHASE, and dotted line represents observed probability using BEAGLE.
Figure 2.2  **Expected and observed crossover probabilities in Angus and Limousin.**

**Description:**  (A) Representative example of expected and observed crossover probabilities in Angus (green) and Limousin (blue) for autosome 15. (B) Probability of zero (black), one (red), two (green), three (blue), and more than three (grey) crossover events for the 29 bovine autosomes in both Angus and Limousin. Plain line represents expected probability, dashed line represents observed probability in Angus, and dotted line represents observed probability in Limousin.
Figure 2.3  Representative example of expected and observed crossover probabilities on autosome 15 in Angus using UMD3.1 (green) and USDA-AIPL coordinates (blue).
Figure 2.4 Correlation between number of haplotypes and recombination rate within 1 Mb windows across the 29 autosomes in Angus.
Figure 2.5  **Relationship between number of haplotypes and crossover rate within 1 Mb windows across the 29 autosomes in Angus.** Number of haplotypes in hot (red) and cold windows (blue), which were defined as windows with a recombination rate ≥ 0.02 (≥ 1.5 standard deviations from the mean) and windows with recombination rate ≤ 0.004, respectively; * indicates a significant difference with $p < 0.05$, and ** indicates a significant difference with $p < 0.01$ for a paired t-test.
Figure 2.6  Expected and estimated genetic distance per Mb (cM/Mb) for the 29 bovine autosomes in Angus (green) and Limousin (blue).
Figure 2.7 Representative example of the variation in recombination rate within 1 Mb windows across bovine autosome 15. The plain line (upper) corresponds to the recombination rate estimated in Angus, while the dashed line (lower) corresponds to the recombination rate estimated in Limousin.
Figure 2.8  Variation in recombination rate within 1 Mb windows across the 29 bovine autosomes. The plain line (upper) corresponds to the recombination rate estimated in Angus, while the dashed line (lower) corresponds to recombination rate estimated in Limousin.
Figure 2.9  Frequency distribution of genome-wide recombination number (GRN) in both Angus (left) and Limousin (right).
Figure 2.10  Distribution of genome-wide recombination number in Angus and Limousin families. (A) Representative example of distribution of genome-wide recombination number (GRN) in Angus half-sib families. Only sires with no more than 20 offspring are presented. Sires were sorted according to the number of their offspring. (B) GRN in Angus half-sib families. Black dots correspond to GRN in sons sorted by sires and red dots correspond to the average GRN for each sire. (C) GRN in Limousin half-sib families. Black dots correspond to GRN in sons sorted by sires and red dots correspond to the average GRN for each sire.
Figure 2.11  Proportion of genetic variance explained by 1 Mb windows across the genome for genome-wide recombination number (GRN) in Angus and Limousin.
Figure 2.12  Relationship between imputation accuracy and recombination rate across the 29 autosomes in Angus. (A) Average imputation accuracy in individuals with zero (red), one (green), two (blue), and more than two (purple) crossover events across the 29 autosomes. (B) Correlation between imputation accuracy and crossover rate (cM/Mb) within 1 Mb windows.
Figure 2.13  **Imputation accuracy in recombination hot (red) and cold windows (blue).** Hot (red) and cold windows (blue), which were defined as windows with a recombination rate $\geq 0.02$ ($\geq 1.5$ standard deviations from the mean) and windows with recombination rate $\leq 0.004$, respectively; * indicates a significant difference with $p < 0.05$, and ** indicates a significant difference with $p < 0.01$ for a paired t-test.
Figure 2.14  Chromosome-wide recombination probabilities calculated using Karlin’s (red), Haldane’s (green) and Kosambi’s (blue) map functions.
CHAPTER 3. IDENTIFICATION OF RECOMBINATION HOTSPOTS AND GENETIC VARIANTS THAT INFLUENCE RECOMBINATION RATE IN LAYER CHICKENS

Ziqing Weng$^1$, Anna Wolc$^{1,2}$, Rohan L. Fernando$^1$, Jack C.M. Dekkers$^1$, Jesus Arango$^2$, Petek Settar$^2$, Janet E. Fulton$^2$, Neil P. O’Sullivan$^2$ and Dorian J. Garrick$^1$

$^1$Department of Animal Science, Iowa State University, Ames, IA, USA

$^2$Hy-Line International, Dallas Center, IA, USA

3.1 Abstract

The frequency of recombination events varies across the genome and among individuals, which may be related to genomic features in many species. The objective of this study was to assess the frequency of recombination events and identify quantitative trait loci (QTL) for recombination rate in two purebred layer chicken lines. A total of 1200 white layers were genotyped with 580K and 5108 brown layers were genotyped with 42K single nucleotide polymorphism panels. Recombination events were detected within half-sib families by LINKPHASE. The number of recombination events within each 0.5-Mb window was calculated. The top 10% of 0.5-Mb windows for recombination rate per chromosome were considered recombination hotspots. The BayesB model was used for genome-wide association studies on genome-wide recombination numbers in each line. Windows that explained >0.8% genetic variance were considered to harbor a quantitative trait locus. There were 14,746, and 215,808 recombination events detected across the genomes in 1301 and 9292 meioses in white and brown layers, respectively. The identified recombination rates differed between lines,
which might due to distinct population inbreeding coefficients and different haplotype structure across genome. There were 190 and 199 recombination hotspots detected in white and brown layers, 28 of which were common to both lines. Females had higher window recombination rates and more genome-wide recombination number per meiosis than males in the 2 lines. Window recombination rate had a strong negative correlation with chromosome size and strong positive correlation with GC content, and CpG island density across the genomes in both lines. Several significant QTL regions, which harbor candidate genes influencing genome-wide recombination, were identified in the 2 lines.

3.2 Introduction

Meiotic recombination occurs between homologous chromosomes and produces crossovers and gene conversions. Characterizing patterns and rates of recombination is essential for understanding genome-wide genetic diversity. recombination events are not evenly distributed among the genome, but rather their locations are strongly controlled by cis and trans acting genes (Paigen and Petkov, 2010). Recombination events occur more frequent in hotspots, which are defined as short intervals with significantly greater recombination rates, compared to surrounding regions.

Mechanisms regulating genome-wide recombination patterns are of interest. Recombination rates have been reported to differ by sex (Kong et al., 2002; Petkov et al., 2007), species (Jensen-Seaman et al., 2004), and breed (Groenen et al., 2009; Weng et al., 2014). Relationships have been found between estimated recombination rate and chromosome size (Kong et al., 2002), distance from the centromere (Jensen-Seaman et al., 2004; Mezard, 2006), GC content (Galtier et al., 2001; Groenen et al., 2009), family structure (Druet and Georges, 2014), and map assembly (Weng et al., 2014). The strength of these relations varies by species.

Genome-wide association studies (GWAS) have identified several quantitative trait loci (QTL) that regulate genome-wide recombination in humans (Chowdhury et al., 2009; Coop et al., 2008), mice (Brunschwig et al., 2012), cattle (Sandor et al., 2012; Weng et al., 2014), and plants (Paape et al., 2012). These QTL include genetic variants in RNF212 (Kong et al., 2008; Sandor et al., 2012).
2012), and PRDM9 (Parvanov et al., 2010; Sandor et al., 2012). Groenen et al. (2009) identified genome sequence features correlated with recombination rate by genotyping 10K single nucleotide polymorphisms (SNPs) in 3 chicken populations. Recombination rates assessed from high density SNP panels have been less thoroughly investigated in chickens than in other organisms. The objectives of this study were to identify genome-wise recombination hotspots and genetic variants that influence recombination events in 2 purebred layer chicken lines.

3.3 Methods

3.3.1 Genotypes

The data analyzed in this study consists of genotypic records from 1200 birds from a white egg laying population (hatched between 2006-2012) genotyped with a 580K single nucleotide polymorphism (SNP) panel, and 5108 brown egg laying birds (hatched between 2003-2011) genotyped with a 42K SNP panel. Missing genotypes (~0.006% in white layers and ~0.01% in brown layers) were imputed using FImpute (Sargolzaei et al., 2014). A total of 173,224 and 23,098 segregating SNPs (5510 overlapped between the panels) across 28 Gallus gallus autosomes (GGA) and sex chromosome were remained on corresponding 580K and 42K SNP panels after removing SNPs with call rate <0.95, minor allele frequency (MAF) <0.025, or Mendelian inconsistency rate between parent-offspring >0.05.

3.3.2 Identification of recombination events

Recombination events were determined only on autosomes within half-sib families in white and brown layers separately using LINKPHASE3.0 (Druet and Georges, 2015). Only half-sib families with at least 2 offspring were used in the analysis. LINKPHASE utilizes linkage, LD, and pedigree information, and applies diploid Hidden Markov model and Baum-Welch algorithm to improve haplotype reconstruction (Druet and Georges, 2010, 2015). Because LINKPHASE could detect putative map errors, markers with >0.9 map confidence score which combines information of recombination rates, parental genotyping errors, and genotype discrepancies in offspring (Druet and Georges, 2015),
were considered map errors and removed from the map assembly. Recombination intervals defined as a pair of heterozygous markers were reported for each parent-offspring pair. Total recombination rate was estimated for every non-overlapping 0.5-Mb window across macrochromosomes (GGA1-GGA5), intermediate chromosomes (GGA6-GGA10) and microchromosomes (Consortium, 2004; Axelsson et al., 2005; Megens et al., 2009). The 0.5-Mb window recombination rate was computed as:

\[ c_w = \frac{\sum_{i=1}^{n} x_i/r_i}{T}, [1] \]

where \( c_w \) is the window recombination rate, \( n \) is the total number of recombination events observed on the corresponding chromosome, \( x_i \) is the overlap (in Mb) between the 0.5 Mb window and recombination interval \( i \), \( r_i \) is the length (in Mb) of the recombination interval, and \( T \) is the total number of sire-offspring pairs. Window recombination rate were also estimated overall and for males and females separately.

For each chromosome, the top 10% of 0.5-Mb windows with the highest recombination rates were considered recombination hotspots, and those with a recombination rate of 0 were considered to be cold spots. For each parent, the proportion of recombination events falling within hotspots across the genome, which defined as the genome-wide hotspot usage (GHU), was calculated for both white layer parents (segregating 580K genotypes) and brown layer parents (segregating 42K genotypes).

Recombination events for each parent were summed over 28 autosomes. Genome-wide recombination number (GRN) was calculated across 28 autosomes in white layer parents using segregating 580K genotypes and in brown layer parents using segregating 42K genotypes, respectively.

### 3.3.3 Examining factors that could affect the observed recombination events

#### 3.3.3.1 Marker density

In order to test the impact of marker density on identification of recombination events, genotypes of white layers were randomly sampled and trimmed down within chromosome (the overlapped SNP genotypes on two panels were retained) to a subset of 23K SNPs with similar marker densities to
the segregating 42K genotypes in brown layers. Whole-genome recombination number and 0.5-Mb window recombination rate were calculated in white layers using segregating trimmed 23K genotypes. Results were compared with those obtained from white layers using segregating 580K genotypes and from brown layers using segregating 42K genotypes, separately.

### 3.3.3.2 Family structure

There were 448 half-sib families (282 male parents and 166 female parents) averaging 3.5±2.7 birds per family in white layers, which contained 969 male-offspring pairs and 332 female-offspring pairs. In brown layers there were a total of 1717 half-sib families (621 male parents and 1096 female parents) averaging 6.0±4.9 birds per family including 4719 male-offspring pairs and 4573 female-offspring pairs, respectively. The effect of family size on observed GRN was assessed in a simple regression model in both white and brown layers. A sample of 20 parental half-sib families (10 male and 10 female families) with 10 offspring was randomly selected from white and brown layers, respectively. In order to avoid sample bias, sampling was repeated 5 times. Window recombination rate within 0.5-Mb and GRN were obtained and compared between the two samples in white and brown layers.

### 3.3.3.3 Genomic inbreeding coefficient

A genomic measure of individual inbreeding coefficient was calculated using runs of homozygosity (ROH), which was proposed by McQuillan et al. (2008). ROH, which represents chromosome segments of continuous homozygous SNPs and provides information about inbreeding levels, were detected using PLINK v1.07 (Purcell et al., 2007). To minimize the probability of identifying spurious ROH, a LD-pruned SNP data was used for white and brown layers separately. After removing SNPs with $r^2>0.5$ within each 50-SNP window, a total of 20 707 and 3413 SNPs were remained on white (580K) and brown (48K) layers respectively. A sliding window with a length of 0.5 Mb was used to scan ROH along the genome. Genomic inbreeding coefficient was calculated for each individual.
3.3.3.4 Haplotype structure

Two measurements were used to determine haplotype structure. First, number of common haplotype alleles (NHA), which is the number of haplotype alleles with no less than 1% frequency in a single window. Furthermore, proportion of homozygosity of common haplotype alleles (PHHA), which is the sum of the frequencies of the homozygous common haplotype alleles in a single window. Both NHA and PHHA were determined within each 0.5-Mb window in the 2 lines using C++ script.

3.3.3.5 Chromosome size and GC content

The relationships between recombination rate, size of chromosome, GC content, and CpG island density within 0.5-Mb windows were assessed in the 2 lines. Information of the GC content and CpG island density were obtained using hgTables tool from the UCSC genome browser (Rao et al., 2013). The average GC content and CpG island density were calculated for each 0.5-Mb window on each chromosome.

3.3.4 Estimating repeatability and heritability

Because parents with multiple offspring had repeated records, heritability and repeatability of GHU and GRN were estimated separately for each line using a repeatability model in ASReml3.0 (Gilmour et al., 2009). The model equation was:

\[ y = Xb + Zu + Zp + e, [2] \]

where \( y \) is the vector of repeated GHU or GRN for every parent, \( b \) represents the unknown mean and gender of parent treated as a fixed effect, \( u \) is the vector of random animal effects with \( Var(u) = A\sigma_a^2 \), where \( A \) is the pedigree relationship matrix among parents, and \( \sigma_a^2 \) is the additive genetic variance, \( p \) is the vector of permanent environmental effects with \( Var(p) = I\sigma_p^2 \), where \( \sigma_p^2 \) is the permanent environment variance, \( X \) and \( Z \) are design matrices, and \( e \) is the vector of residual effects with \( Var(e) = I\sigma_e^2 \), where \( \sigma_e^2 \) is the residual variance.

A marker-based heritability was estimated using average GRN in a weighted BayesC model with \( \pi \) equal to 0 (Habier et al., 2011; Kizilkaya et al., 2010) implemented in GENSEL4.0 software.
(Fernando and Garrick, 2013; Garrick and Fernando, 2013). The prior assumption of BayesC is that marker effects have identical and independent mixture distributions. The model equation was:

\[ y = Xb + Z\alpha + e, \]  

where \( y \) is the average GRN for parents, \( X \) is an incidence matrix for the fixed effect, \( b \) is a vector of fixed effects (mean and gender), \( Z \) is a \( n \times m \) matrix for SNP genotype covariates (coded as 0, 1, or 2), \( n \) is the number of parents, \( m \) is the number of SNP loci, \( \alpha \) is a vector of random allele substitution effects of SNP with mixture priors, and \( e \) is a weighted residual effect. It is assumed that the SNP effect in equation \([3]\) follows a mixture distribution as below:

\[ \alpha_j \overset{i.i.d.}{\sim} \begin{cases} 0 & \text{with probability } \pi \\ N(0, \sigma^2_\alpha) & \text{with probability } 1 - \pi \end{cases} \]  

where \( \alpha_j \) is the marker effect for SNP \( j \), and \( \sigma^2_\alpha \) is the common variance. The priors of the genetic and residual variance components required for BayesC were obtained from the ASReml analysis. Markov chain Monte Carlo (MCMC) sampling with 55,000 iterations in which the first 5,000 samples were discarded, was used to estimate variance components and heritability. The weighting factor \( w_n \) (Garrick et al., 2009; Wolc et al., 2011) to account for heterogeneity of residual variance due to differences in the number of GRN observations contributing to the average GRN was calculated as:

\[ w_n = \frac{1 - h^2}{1 + (n-1)t - h^2}, \]  

where \( h^2 \) is the narrow sense heritability estimated from pedigree, \( t \) is the repeatability, and \( n \) is the number of observations contributing to the average GRN for that parent.

3.3.5 Identification of QTL affecting GRN

Genome-wide association study of average GRN was performed for each line by a weighted BayesB method (Meuwissen et al., 2001) implemented in GenSel4.4 software (Fernando and Garrick, 2013; Garrick and Fernando, 2013). The model equations were equivalent to equation \([3]\), \([4]\), and
[5], except that BayesB assumes a fraction $\pi$ of SNPs have zero effects, while the rest of SNP effects follow identical and independent univariate-$t$ distributions with null means, an inverse chi-squared prior for the SNP variances with degrees of freedom of $\nu_a$, and scale parameter $S_a^2$. The priors of the SNP and residual scale factors were obtained from genetic and residual variance estimated from the ASReml analysis. The MCMC sampling had 55 000 iterations with the first 5000 samples as burn-in. GRN were identified across 28 autosomes. GWAS was conducted across the genome (28 autosomes and sex chromosomes) in white layers using segregating 580K genotypes, and in brown layers using segregating 42K genotypes, respectively. The values of $\pi$ were 0.999 in white layers and 0.99 in brown layers.

The expected proportion of genetic variance (GV%) explained by each non-overlapping 1-Mb region was approximately 0.04% under a polygenic model. 1-Mb regions that explained at least 0.8% (> 20 folds of the expected percentage) of the genetic variance or their flanking regions on both sides of these regions (± 1 Mb), were considered to harbor QTL (Fernando and Garrick, 2013; Serão et al., 2014; Weng et al., 2014). The proportion of genetic variance explained by each 1-Mb region was sampled in each iteration of the MCMC chains (Fernando and Garrick, 2013).

The posterior probabilities of association (PPA) of each QTL region, which is the percentage of samples for that region that contained at least one non-zero effect SNP, were also calculated. The SNP with the highest SNP posterior probability of inclusion (SPPI) were selected as the lead SNP within each QTL region. The SPPI was defined as the proportion of MCMC samples in which a SNP had with non-zero effects were involved. Each lead SNP was fitted separately from the other SNP in the QTL region to estimate the GV% explained by that SNP (Serão et al., 2014). Single SNP GWAS analyses were assessed using an animal model with ASReml v3.0 (Gilmour et al., 2009), in order to evaluate the significance of the lead SNP (Wolc et al., 2014). The NCBI gene database was used to obtain positional candidate genes near lead SNP (±1Mb). Gene ontology (GO) term enrichment analysis on positional candidate genes in both white and brown layers was performed using the GO database and web tool g:Profiler (Reimand et al., 2011). Only GO terms associated with biological processes with significant P value of 0.01 (Benjamini-Hochberg FDR corrected) were retained.
3.4 Results and Discussion

3.4.1 Recombination between lines

Table 3.1 shows the average window-wide recombination rate across 28 chromosomes in the 2 lines. The 0.5-Mb window recombination rate varied from 0 to over 0.025 in white layers and ranged from 0 to 0.047 in brown layers. The average window recombination rate across 28 chromosomes differed between the 2 lines. In white layers, the average window recombination rate across the genome was 0.0070 ± 0.010, about half that estimated in brown layers 0.014 ± 0.012. In general, recombination rates were higher in brown layers than these in white layers (p<0.0001). Figure 3.1 uses GGA1 as an example to show window recombination rates in white and brown layers. The average window recombination rate on GGA was 0.013 ± 0.0070 in white layers and in 0.014 ± 0.0096 brown layers. Window recombination rate changed along the chromosome. In some cases, windows with higher recombination rates for white layers corresponded to a lower recombination rate in brown layers and vice versa. Different window recombination rate and recombination landscapes across the genome between the 2 lines (Figure 3.2).

A large number of recombination hotspots were detected in white (190) and brown layers (199), 28 of which were common to both lines. There were 551 and 45 recombination coldspots detected in white and brown layers, 22 of which were common to both lines. Overall, 14 746 and 215 808 recombination events were detected across the genomes in 1301 and 9292 meioses in white and brown layers, respectively. The average GRN per parent per meiosis were 11.89 ± 4.13 in white and 23.78 ± 4.37 in brown layers. In order to investigate the inconsistent results between the 2 lines, the following factors were examined.

3.4.1.1 Marker density

GGA1 was used as an example (Figure 3.3). The average window recombination rate on GGA1 in white layers using trimmed 42K SNPs was 0.011 ± 0.0056, which was significantly lower (p=0.0046) than using 580K SNPs (0.013 ± 0.0070). The correlation of recombination rates on GGA1 based these two sets of segregating SNPs in white layers was 0.68. The average number of SNPs
within a 0.5-Mb window was 103.3, 12.5, and 12.2 on GGA1 in white layers with 580K genotypes, 
white layers with 42K genotypes, and brown layers with 42K genotypes, respectively. The average 
GRN in white layers using 42K genotypes was 8.58±0.045, which was significantly lower \((p<0.0001)\) 
than in white layers using 580K SNPs \((11.89±4.13)\), and in brown layers \((23.78±4.37)\).

The observed window recombination rates and GRN obtained from segregating 580K genotypes 
were higher than 42K genotypes in white layers, which confirmed the influence of marker density 
on the identification of recombination events in this study. However, marker density cannot explain 
the distinct recombination results between 2 lines.

The accuracy of identification of recombination events could be improved by using a denser 
marker panel. However, it is important to be aware of the impact of gene conversion, when using 
sequence data (or equivalently dense marker panels) to detect recombination. Gene conversion 
is a non-crossover event during meiosis (Alberts et al., 2008). Without a careful filtering process, 
gene conversions might be detected as double-crossovers and inflate recombination rates. Due to the 
limited density of the SNP genotypes used in this study, gene conversion regions can not be detected 
Alternative ways to increase the precision of uncovering recombination events, such as correcting 
marker locations in the map assembly, or applying map function to account for interaction between 
adjunct crossovers, should be investigated further.

3.4.1.2 Family structure

In order to minimize the impact of sample size and family size on identification of recombination 
events, window recombination rate on GGA1 and GRN were calculated using the sample of 20 half-
sib families (10 offspring per family) in both white and brown layers. The average recombination 
rate within 0.5-Mb window was higher \((p=0.011)\) in brown layers. White layers \((12.08±0.34)\) had 
significant less \((p<0.0001)\) GRN than brown layers \((22.73±0.15)\). The similar results were observed 
using all half-sib families with at least 2 offspring per family in the 2 lines.

Druet and Georges (2015) pointed out the family size could influence the identification of re-
combination events. In this study the effect of family size on identification of recombination events 
was not significant in white \((p=0.28)\) and brown layers \((p=0.082)\), although variations of GRN were
observed between half-sib families in the two lines using the whole datasets (Figure 3.4). Based on the results, family structure is not the causal factor introducing the differences of observed recombination events between the 2 lines.

### 3.4.1.3 Genomic inbreeding coefficient

The estimated average genomic inbreeding coefficient was $0.082 \pm 0.054$ in white layers, and was $0.031 \pm 0.028$ in brown layers. The correlation between individual GRN and genomic inbreeding coefficient was -0.19 in white layers and -0.47 in brown layers. Recombination that occurs within a long chromosome segment of homozygous SNPs cannot be identified. Individual with higher genomic inbreeding coefficient has more numbers of ROH and longer lengths of ROH, which might hinder the identification of recombination events. The 2 lines had distinct genomic inbreeding coefficient, which could attribute to the different recombination results between lines.

### 3.4.1.4 Haplotype structure

Figure 3.5 shows the distribution of NHA and PHHA within 0.5-Mb window across 28 autosomes in both white and brown layers. The NHA and PHHA varied among chromosomes. The average NHA in white layers was $2.98 \pm 0.068$, which was significantly lower ($p<0.0001$) than that in brown layers ($10.61 \pm 0.067$). The average PHHA was significantly different ($p<0.0001$) in white (0.69±0.0046) and brown layers (0.18±0.0045). The correlation between NHA was 0.44 in white layers and was 0.72 in brown layers. The correlation between PHHA was -0.32 in white layers and was -0.47 in brown layers. Recombination hotspots had significantly higher NHA and lower PHHA, compared to coldspots, in both white and brown layers. In white layers, NHA were 4.52±0.14 and 2.11±0.080, and PHHA were 0.56±0.017 and 0.79±0.010, in hotspots and coldspots, respectively. In brown layers, NHA were 14.46±0.22 and 4.20±0.46, and PHHA were 0.12±0.0079 and 0.48±0.017, in hotspots and coldspots, respectively.

Recombination could break down haplotype alleles inherited from parental generations and create new haplotypes. Less haplotype alleles were observed and more homozygous haplotype alleles were preserved in the windows with lower recombination rate. Brown layers had much
higher recombination rate and corresponding larger number of common haplotype alleles and lower proportion of homozygous common haplotype alleles within 0.5-Mb window than white layers. The results of PHHA were consistent with genomic inbreeding coefficient in the 2 lines. The existence of homozygous haplotype alleles would influence the identification of recombination events.

The observed window recombination rates and GRN were differed significantly between the 2 lines. Groenen et al. (2009) observe significant heterogeneity in recombination rate across 3 different chicken populations. Weng et al. (2014) reported distinct recombination results between Angus and Limousin beef cattle. It is likely that the breed-specific characteristics of recombination might due to population structure (e.g. inbreeding levels), and genomic structure (e.g. property of haplotype alleles).

### 3.4.2 Recombination between sexes

Table 3.1 also presents 0.5-Mb window recombination rates in male and female parents separately. The difference in recombination rates between genders was significant for white layers (male=0.0099±0.0077; female=0.013±0.010; \( p<0.0001 \)) and brown layers (male=0.017±0.013; female=0.016±0.013; \( p=0.0076 \)). For GGA1 for example, females (0.016±0.010) had significantly higher \( (p=0.0001) \) recombination rate than males (0.012±0.0060) in white layers. The recombination rate of males (0.012±0.0057) and females (0.014±0.0081) were different \( (p=0.051) \) for GGA1 in brown layers. In Figure 3.6, the recombination patterns in males and females were slightly different on GGA1 in both lines. Recombination rate varied in males and females, and the locations of recombination hot and cold spots were inconsistent for both sexes across the genome in white and brown layers, respectively. Figure 3.7 shows the average number of recombination events per meiosis of males and females in white and brown layers. The average GRN per meiosis were different between sexes in both white \( (p<0.0001) \) and brown layers \( (p=0.042) \).

In white layers, the average size of half-sib families was 4.0±3.0 in males, and 2.5±1.5 in females. For brown layers, the average number of birds in male families was 8.0±5.7 and 4.7±4.0 in female families. The random samples of 10 male and 10 female half-sib families (10 offspring per family) were used to recalculate window recombination rate and GRN in males and females separately,
in order to avoid the impacts of numbers of observations and family sizes on the identification of recombination events between sexes. The same as observed above, average recombination rate within 0.5-Mb window was higher in females than males in both white ($p=0.0001$) and brown layers ($p=0.0052$). Females (16.36±1.10) had higher GRN than males (11.11±0.52) in white layers. In brown layers, GRN was 23.27±0.20 in females, and 22.38±0.16 in males.

Genomic inbreeding coefficients between sexes were similar in white ($p=0.67$) and brown layers ($p=0.45$). The inbreeding coefficients were 0.0083±0.052 for females and 0.0081±0.056 for males in white layers, and were 0.030±0.026 and 0.033±0.028 in brown layers. Also, similar results of NHA and PHHA within 0.5-Mb window were obtained between sexes in the 2 lines.

In this study, the observed window recombination rates and average GRN were significantly greater in females than in males in 2 lines, which confirmed the sex-specific property of recombination in layer chickens. Different recombination rates across sexes, i.e. females have higher recombination rates than males, have been identified in *Drosophila* (Tsai et al., 2011), mice (Petkov et al., 2007), and humans (Coop et al., 2008). The genetic variants which control recombination in females and males might be diverse. Kong et al. (2014) identified 3 variants associated with male genome-wide recombination rate only, and 7 variants associated with female recombination only using the Icelandic genealogy database.

### 3.4.3 Chromosome size and GC content

The average window recombination rates were of the same order of magnitude on macrochromosomes 0.0067 ± 0.0092 (0.011 ± 0.0086), intermediate chromosomes 0.0069 ± 0.0091 (0.016 ± 0.013), and microchromosomes 0.0085 ± 0.013 (0.022 ± 0.018) in white (brown) layers. The differences between recombination rate on macrochromosomes and microchromosomes were significant in both white ($p=0.011$) and brown layers ($p<0.0001$). Recombination rate is negatively correlated with chromosome size in white ($r^2=-0.16$) and brown layers ($r^2=-0.50$). Rodionov (1996), and Groenen et al. (2009) also observed higher recombination rates in microchromosomes compared to macrochromosomes. In this study, GGA1-GGA5 were considered as macrochromosomes, while GGA11-GGA28 were defined as microchromosomes (Consortium, 2004; Axelsson et al., 2005;
Megens et al., 2009). Although Rodionov (1996), and Groenen et al. (2009) considered GGA1-8 as macrochromosomes, results still indicate that recombination rate is negatively correlated with chromosome size (Consortium, 2004). Kong et al. (2002) and Weng et al. (2014) found similar results in humans and beef cattle respectively.

Correlations between chromosome size and GC content (-0.57), CpG island density (-0.60) were negative. Figure 3.8 shows the relationship between recombination rates and GC density. In white layers, the correlation between recombination rate and GC content was 0.13, and between recombination rates and CpG island density was 0.13. In brown layers, correlations between recombination rate with GC content and CpG island density were both 0.29 ($p<0.0001$). Recombination hotspots differed significantly from cold spots in GC content ($p=0.046$) and CpG island density ($p=0.042$) in both lines. The strength of association between recombination and GC content and CpG island density was stronger in brown layers than in white layers. Although it is known that recombination rate is related to genome structure, the strength of this correlation varies among organisms. For example, GC content is positively correlated with recombination rate in some organisms (Galtier et al., 2001; Jensen-Seaman et al., 2004), but shows weak or no correlations in others (Mezard, 2006; Paape et al., 2012). The 0.5-Mb windows with higher recombination rate have higher GC content. According to the “biased gene conversion” hypothesis (Marais, 2003), recombination hotspots become GC rich regions (Consortium, 2004).

### 3.4.4 GHU

The GHU differed significantly between individuals in both white ($p=0.0034$; range: 0-50.0%) and brown layers ($p=0.0038$; range: 0%-83.8%). The average GHU in white layers was 9.6% ± 5.9%, which was significantly different ($p < 2.2 \times 10^{-16}$) from brown layers (20.3%±5.4%). The differences of GHU between genders was not significantly different in both lines (white: $p=0.021$; brown: $p=0.4$). The average GHU of males was 9.1% ± 6.2% (20.2% ± 4.6%), and 10.4% ± 5.2% (20.4% ± 5.8%) for females in white (brown layers). Heritability and repeatability estimates were $0.10 \pm 0.041$ ($0.10 \pm 0.015$) and $0.14 \pm 0.037$ ($0.14 \pm 0.014$) in white (brown) layers.

The estimated heritabilities of GHU in the 2 lines were lower than the estimates computed in
cattle. Sandor et al. (2012) reported GHU to have a heritability of 0.21 in Dutch Holstein Friesians. Several studies have shown that the PRDM9 gene controls activation of mammalian recombination hotspots (Parvanov et al., 2010; Sandor et al., 2012). Particularly, Sandor et al. (2012) observed large peaks on the chromosomes X cattle at the position of two PRDM9 paralogues. It is likely that GHU is regulated by a sex chromosome since causal genes were identified on chromosome X in cattle. In this study, GWAS on GHU identified 6 and 11 significant QTL regions (GV% > 0.5%) in white and brown layers (results were not reported). In white layers, significant QTL regions were distributed on GGA3, 9, 10, 20, 21, and 28. In brown layers, all 11 significant QTL regions were located on Z chromosome. No significant QTL regions overlapped between the 2 lines and no significant candidate genes were identified due to limited annotation information. Genetic variations controlling GHU may differ between breeds and species.

3.4.5 Genome regions associated with GRN

GRN was repeatable, with repeatability estimates of 0.24 ± 0.020 and 0.21 ± 0.0033 in white and brown layers, respectively. Heritability was 0.17 ± 0.022 for white layers, and 0.16 ± 0.0037 for brown layers. Marker-based heritability estimates of GRN were similar to the pedigree-based estimates being 0.15 ± 0.0014 and 0.13 ± 0.0013 in white and brown layers, respectively. The estimated heritabilities in layer chickens were lower than those reported in beef cattle, dairy cattle, or human. Weng et al. (2014) reported the pedigree-based heritability of GRN was 0.26 in Angus and 0.23 in Limousin. Sandor et al. (2012) reported a heritability of 0.22 in Dutch Holstein-Friesian bulls. Heritability of recombination rate was 0.30 in humans (Kong et al., 2004).

The number of recombination events for every parent was treated as a quantitative trait to map the genetic variants that influence recombination rate in the 2 lines separately. The proportions of genetic variance explained by each 1-Mb region across the genome in white and brown layers are presented in Figures 3.9 and 3.10. Tables 3.2 and 3.3 show the genetic variance explained, and posterior probability of association of significant QTL regions, MAF, physical position, significance of lead SNPs, and list of nearby candidate genes. Since it has been shown that the location of the causal mutation could be extended to 1 Mb on either side of the informative 1 Mb QTL region,
neighboring regions were combined for analysis (Fernando and Garrick, 2013; Serão et al., 2014; Weng et al., 2014). In general, 14 QTL on 8 different chromosomes influencing recombination were identified in white layers. Only 6 QTL on 4 chromosomes were identified in brown layers. The GV% explained by significant QTL regions ranged from 1.00% to 8.39% in white layers, and from 0.83% to 19.74% in brown layers. No common QTL regions were identified across both lines.

A total of 20 candidate genes were identified within and/or near the QTL regions in white layers, while 10 candidate genes were identified in brown layers. No candidate genes were identified on chromosome1 at 158Mb, chromosome 2 at 81-82Mb, or GAA13 at 3Mb in white layers.

The GO term analyses of the positional candidate genes identified in the 2 lines are shown in Table 3.4. Significant enrichments for biological process, including cellular metabolic process, nuclear division, mitotic nuclear division, cell proliferation, and other functions related to cell process and cell division. Candidate genes were clustered according to their biological functions or pathways involved. Mostly enriched GO terms included positional candidate genes identified in both lines, except mitotic cell cycle, mitotic cell cycle process, and negative regulation of endothelial cell proliferation.

3.4.5.1 Candidate genes identified in white layers

One of the strongest candidate genes, SPO11 (SPO11 meiotic protein covalently bound to DSB homolog), was identified on chromosome 20 near 10Mb. SPO11 produces a meiosis-specific protein, which could initiate recombination (Paigen and Petkov, 2010). Guillon et al. (2005) reported that both crossover and non-crossovers pathways were SPO11 dependent in mice. Weng et al. (2014) identified SPO11 as a positional candidate gene on genome-wide recombination number in Limousin cattle.

Positional candidate genes on GAA1 at 54Mb include NFYB (nuclear transcription factor Y, beta), located at 54.80-54.82Mb, and TDG (thymine-DNA glycosylase) located at 54.85-54.86Mb. NFYB is a subunit of a conserved transcription factor that regulates DNA binding (Chen et al., 2011). Tini et al. (2002) reported that TDG initiates repair of G/T and G/U mismatches, and is associated with CpG islands. Two candidate genes were identified on GAA1 at 84 Mb: RPL24
(ribosomal protein L24), which is involved in the cellular process of translation, and integration of the membrane and ribosomes; *IMPG2* (interphotoreceptor matrix proteoglycan 2) which is an integral component of the membrane. The region of GAA1 at 162-165Mb contains three candidate genes: *RGCC* (regulator of cell cycle), *WBP4* (WW domain binding protein 4), and *LECT1* (leukocyte cell derived chemotaxin 1). They are involved in the process of cell cycling, RNA splicing and cell differentiation. The last candidate gene identified on GAA1, *CCNA1* (cyclin A1) plays a role in the cell cycle process, it is located at 171.87-171.88Mb. Only two candidate genes were identified on chromosome 2 at 76-77Mb. *MRPL36* (mitochondrial ribosomal protein L36) involved in the cellular process of translation, and *CLPTM1L* (CLPTM1-like), an integral component of the membrane. One candidate gene was identified on GGA3, *PRIM2* (primase, DNA, polypeptide 2), which is located at 86.05-86.13Mb. It was reported that *PRIM2* is active in both the initiation of DNA replication and synthesis (Shiratori et al., 1995). Three candidate genes were detected on chromosome 4 at 44Mb: *EPGN* (epithelial mitogen), *AREGB* (amphiregulin B), and *EREG* (epiregulin). These three genes belong to the ErbB signaling pathway, which regulates diverse biologic responses, including cell proliferation, differentiation, motility, and survival (OMIM). The positional candidate genes on chromosome 5 at 13Mb were *IGF2* (insulin-like growth factor 2), *MRPL23* (mitochondrial ribosomal protein L23), *RAG1* (recombination activating gene 1), and *RAG2* (recombination activating gene 2). *IGF2* is a growth factor that controls hormone activity, regulation of mitosis, and cell differentiation. Van Laere et al. (2003) identified a mutation in a regulatory site of *IGF2* underlying a major QTL effect for muscle growth in pigs. Jungerius et al. (2004) showed that a paternally expressed QTL for backfat thickness in a Meishan-European white pig cross is caused by that same mutation in *IGF2*. *MRPL23* is involved in the cellular process of translation and integration of the membrane and ribosomes. The proteins *RAG1* and *RAG2* carry out the formation of double-stranded breaks at recombination signal sequences (McBlane et al., 1995). The candidate gene *RPS17L* (ribosomal protein S17-like) located on GAA10 at 0.83-0.84Mb, has a biological function related to membrane and ribosome function. *YWHAB* (tryptophan 5-monooxygenase activation protein, beta polypeptide) identified on chromosome 20 at 5 Mb, plays a role in the cell cycle process.
3.4.5.2 Candidate genes identified in brown layers

The signal near \textit{RNF212} was identified on chromosome 4 influencing GRN in brown layers. The sequence variants in \textit{RNF212} (ring finger protein 212) affect genome-wide recombination rate in males and females in humans (Chowdhury et al., 2009; Kong et al., 2008). Its significant impact on recombination rate also has been detected in cattle (Sandor et al., 2012; Weng et al., 2014). Three other candidate genes were identified in the same region: \textit{MAEA} (macrophage erythroblast attacher) located at 84.06-84.11Mb, is associated with the cell cycle and cell division; \textit{FGFRL1} (fibroblast growth factor receptor-like 1) located at 84.63-84.79Mb, is an integral component of the membrane; \textit{MRPL35} (mitochondrial ribosomal protein L35) located at 85.06-85.07Mb, is a component of the mitochondrial ribosome (Koc et al., 2001).

Two candidate genes were identified on GAA1, \textit{CBLL1} (Cbl proto-oncogene, E3 ubiquitin protein ligase-like 1) located at 14.62-14.63Mb and \textit{RECQL} (RecQ protein-like) located at 65.47-65.49Mb. \textit{CBLL1} is involved in positive regulation of cell migration and negative regulation of cell adhesion. \textit{RECQL} is involved in various types of DNA repair, including mismatch repair, nucleotide excision repair, and direct repair (Puranam and Blackshear, 1994). The QTL region containing \textit{RECQL} was associated with GRN in Angus cattle (Weng et al., 2014). Chromosome 2 also had two candidate genes: \textit{CDH10} (cadherin 10, type 2) located at 72.18-72.27Mb, which is an integral component of the membrane, and \textit{MYC} (v-myc avian myelocytomatosis viral oncogene homolog) located at 139.31-139.32, which plays a critical role in DNA replication, cell growth and cell cycle progression. The candidate genes \textit{FANCD2} (Fanconi anemia, complementation group D2) and \textit{GNAI2} (guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2) were identified on GAA12 at 2Mb. Through a mechanism related to homologous recombination, \textit{FANCD2} promotes gene conversion and DNA repair, while \textit{GNAI2} is involved in the cell cycle, cell division, and cell proliferation processes.

Only two previously identified genes, \textit{SPO11} and \textit{RNF212}, were detected in white or brown layers separately. Other promising candidate genes include \textit{RAG1}, \textit{RAG2}, and \textit{RECQL}, although further investigation is required. The different mapping results between white and brown layers,
and between chickens and other organisms (e.g. cattle, humans, mice and plants) suggest that recombination is a species-specific polygenic trait.

3.5 Conclusions

Genome-wide recombination patterns and rates were characterized in both white and brown layers. A large number of recombination events, and recombination hotspots were identified. In-breeding coefficient, and haplotype structure could influence the identification of recombination events, which lead to the differences in the observed recombination rates (numbers) and hotspot locations between lines. Recombination events were found to be sex-specific. Females have higher window recombination rate and more genome-wide recombination number per meiosis than males. Window recombination rate showed a strong negative correlation with chromosome size and positive correlation with GC content, and CpG island density in the 2 lines. Several different QTL regions associated with genome-wide recombination were identified in the two breeds, which indicates that recombination event is a complex polygenic trait.

3.6 Acknowledgements

This study was supported by Hy-Line Int., the EW group, and Agriculture and Food Research Initiative competitive grants 2009-35205-05100 and 2010-65205-20341 from the USDA National Institute of Food and Agriculture Animal Genome Program.

3.7 Bibliography


OMIM. Online mendelian inheritance in man (http://www.omim.org).


### 3.8 Tables

Table 3.1  Average recombination rates (± SD) for 0.5-Mb windows per chromosome.

<table>
<thead>
<tr>
<th>GGA Length (Mb)(^a)</th>
<th>White layers</th>
<th>Brown layers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>1 195.28</td>
<td>0.013 ± 0.0070</td>
<td>0.012 ± 0.0057</td>
</tr>
<tr>
<td>2 148.81</td>
<td>0.0047 ± 0.0079</td>
<td>0.0075 ± 0.0052</td>
</tr>
<tr>
<td>3 110.45</td>
<td>0.0075 ± 0.010</td>
<td>0.011 ± 0.0035</td>
</tr>
<tr>
<td>4 90.22</td>
<td>0.0062 ± 0.0082</td>
<td>0.0088 ± 0.0045</td>
</tr>
<tr>
<td>5 59.58</td>
<td>0.0074 ± 0.0083</td>
<td>0.010 ± 0.0056</td>
</tr>
<tr>
<td>6 34.95</td>
<td>0.0095 ± 0.0095</td>
<td>0.014 ± 0.0061</td>
</tr>
<tr>
<td>7 36.25</td>
<td>0.0013 ± 0.0038</td>
<td>0.00035 ± 0.00056</td>
</tr>
<tr>
<td>8 28.76</td>
<td>0.0054 ± 0.0067</td>
<td>0.0073 ± 0.0053</td>
</tr>
<tr>
<td>9 23.44</td>
<td>0.0092 ± 0.0093</td>
<td>0.012 ± 0.0058</td>
</tr>
<tr>
<td>10 19.91</td>
<td>0.012 ± 0.012</td>
<td>0.016 ± 0.0067</td>
</tr>
<tr>
<td>11 19.4</td>
<td>0.0049 ± 0.0086</td>
<td>0.0062 ± 0.0048</td>
</tr>
<tr>
<td>12 19.9</td>
<td>0.0053 ± 0.0078</td>
<td>0.0057 ± 0.0069</td>
</tr>
<tr>
<td>13 17.76</td>
<td>0.0070 ± 0.0089</td>
<td>0.0091 ± 0.0067</td>
</tr>
<tr>
<td>14 15.16</td>
<td>0.010 ± 0.011</td>
<td>0.013 ± 0.0091</td>
</tr>
<tr>
<td>15 12.66</td>
<td>0.0068 ± 0.011</td>
<td>0.0098 ± 0.0072</td>
</tr>
<tr>
<td>16 0.54</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17 10.45</td>
<td>0.0082 ± 0.010</td>
<td>0.094 ± 0.0090</td>
</tr>
<tr>
<td>18 11.22</td>
<td>0.012 ± 0.014</td>
<td>0.016 ± 0.012</td>
</tr>
<tr>
<td>19 9.98</td>
<td>0.0018 ± 0.0030</td>
<td>0.00068 ± 0.00078</td>
</tr>
<tr>
<td>20 14.3</td>
<td>0.0069 ± 0.0074</td>
<td>0.0075 ± 0.0069</td>
</tr>
<tr>
<td>21 6.8</td>
<td>0.016 ± 0.020</td>
<td>0.023 ± 0.018</td>
</tr>
<tr>
<td>22 4.08</td>
<td>0.00060 ± 0.0011</td>
<td>0.00011 ± 0.00025</td>
</tr>
<tr>
<td>23 5.72</td>
<td>0.019 ± 0.020</td>
<td>0.023 ± 0.022</td>
</tr>
<tr>
<td>24 6.32</td>
<td>0.015 ± 0.014</td>
<td>0.018 ± 0.014</td>
</tr>
<tr>
<td>25 2.19</td>
<td>0.0052 ± 0.011</td>
<td>0.0031 ± 0.0063</td>
</tr>
<tr>
<td>26 5.33</td>
<td>0.015 ± 0.016</td>
<td>0.016 ± 0.012</td>
</tr>
<tr>
<td>27 5.21</td>
<td>0.025 ± 0.031</td>
<td>0.032 ± 0.034</td>
</tr>
<tr>
<td>28 4.74</td>
<td>0.0035 ± 0.0073</td>
<td>0.0020 ± 0.0037</td>
</tr>
</tbody>
</table>

\(^a\)Chromosomes sizes based on galGal4 from the UCSC website
Table 3.2  **Description of 1-Mb regions in white layer chickens** that explained more than 0.8% of genetic variance and the significant SNPs within the QTL regions for genome-wide recombination number.

<table>
<thead>
<tr>
<th>GGA_Mb</th>
<th>No. SNP</th>
<th>GV%</th>
<th>PPA</th>
<th>Lead SNP</th>
<th>SPPI</th>
<th>GV% SNP</th>
<th>MAF</th>
<th>Position (Mb)</th>
<th>p valuef</th>
<th>Candidate Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.54</td>
<td>126</td>
<td>1</td>
<td>0.27</td>
<td>AX-80768024</td>
<td>0.02</td>
<td>0.05</td>
<td>0.132</td>
<td>54.56</td>
<td>0.003</td>
<td>TDG, NFYB</td>
</tr>
<tr>
<td>1,84</td>
<td>112</td>
<td>8.39</td>
<td>0.54</td>
<td>AX-80984549</td>
<td>0.52</td>
<td>8.38</td>
<td>0.151</td>
<td>84.06</td>
<td>0.04</td>
<td>RPL24, IMPG2</td>
</tr>
<tr>
<td>1,158</td>
<td>165</td>
<td>1</td>
<td>0.23</td>
<td>AX-75313537</td>
<td>0.01</td>
<td>0.42</td>
<td>0.349</td>
<td>158.73</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>1,162-165</td>
<td>801</td>
<td>4.39</td>
<td>0.89</td>
<td>AX-75325622</td>
<td>0.02</td>
<td>1.3</td>
<td>0.189</td>
<td>165.07</td>
<td>0.006</td>
<td>RGCC, WBP4, LECT1</td>
</tr>
<tr>
<td>1,171</td>
<td>198</td>
<td>1</td>
<td>0.34</td>
<td>AX-75341236</td>
<td>0.02</td>
<td>0.1</td>
<td>0.316</td>
<td>171.68</td>
<td>0.03</td>
<td>CCNA1</td>
</tr>
<tr>
<td>2,76-77</td>
<td>408</td>
<td>4.13</td>
<td>0.76</td>
<td>AX-76154814</td>
<td>0.02</td>
<td>0.15</td>
<td>0.346</td>
<td>77.43</td>
<td>0.06</td>
<td>ANKH</td>
</tr>
<tr>
<td>2,81-82</td>
<td>196</td>
<td>3.91</td>
<td>0.64</td>
<td>AX-76163810</td>
<td>0.01</td>
<td>0.38</td>
<td>0.312</td>
<td>82.25</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>3,85</td>
<td>468</td>
<td>3.92</td>
<td>0.79</td>
<td>AX-76572482</td>
<td>0.01</td>
<td>0.16</td>
<td>0.357</td>
<td>86.15</td>
<td>0.04</td>
<td>PRIM2</td>
</tr>
<tr>
<td>4,44</td>
<td>183</td>
<td>7.93</td>
<td>0.56</td>
<td>AX-7669056</td>
<td>0.48</td>
<td>7.88</td>
<td>0.126</td>
<td>44.57</td>
<td>0.08</td>
<td>EPGN, AREGB, EREG</td>
</tr>
<tr>
<td>5,13</td>
<td>398</td>
<td>2.23</td>
<td>0.64</td>
<td>AX-76779213</td>
<td>0.01</td>
<td>0.26</td>
<td>0.355</td>
<td>13.59</td>
<td>0.02</td>
<td>IGF2, MRPL23, RAG1, RAG2</td>
</tr>
<tr>
<td>10,1</td>
<td>382</td>
<td>1.08</td>
<td>0.39</td>
<td>AX-75574022</td>
<td>0.01</td>
<td>0.1</td>
<td>0.472</td>
<td>1.06</td>
<td>0.03</td>
<td>RPS17L</td>
</tr>
<tr>
<td>13,3</td>
<td>166</td>
<td>2.37</td>
<td>0.5</td>
<td>AX-75754286</td>
<td>0.01</td>
<td>0.77</td>
<td>0.465</td>
<td>3.62</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>20,5</td>
<td>283</td>
<td>1.02</td>
<td>0.35</td>
<td>AX-76217814</td>
<td>0.01</td>
<td>0.2</td>
<td>0.316</td>
<td>5.09</td>
<td>0.02</td>
<td>YWHAB</td>
</tr>
<tr>
<td>20,10</td>
<td>189</td>
<td>1.15</td>
<td>0.29</td>
<td>AX-76199322</td>
<td>0.03</td>
<td>0.12</td>
<td>0.385</td>
<td>10.96</td>
<td>0.005</td>
<td>SPO11</td>
</tr>
</tbody>
</table>

*GV%, proportion of the genetic variance explained by window or SNP.  
PPA, posterior probability of association of 1-Mb region.  
Lead SNP, SNP with the highest SPPI in the 1-Mb region.  
SPPI, SNP posterior probability of inclusion.  
MAF, minor allele frequency.  
p value, the significance of single SNP analysis in ASREML.*
Table 3.3 **Description of 1-Mb regions in brown layer chickens** that explained more than 0.8% of genetic variance and the significant SNPs within the QTL regions for genome-wide recombination number.

<table>
<thead>
<tr>
<th>GGA_Mb</th>
<th>No. SNP</th>
<th>GV%(^a)</th>
<th>PPA(^b)</th>
<th>Lead SNP(^c)</th>
<th>SPPI(^d)</th>
<th>GV% SNP</th>
<th>MAF(^e)</th>
<th>Position (Mb)</th>
<th>p value(^f)</th>
<th>Candidate Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,15</td>
<td>22</td>
<td>0.83</td>
<td>0.49</td>
<td>Gga_rs13713650</td>
<td>0.34</td>
<td>0.1</td>
<td>0.38</td>
<td>15.28</td>
<td>0.022</td>
<td>CBLL1</td>
</tr>
<tr>
<td>1,66</td>
<td>19</td>
<td>5</td>
<td>0.57</td>
<td>Gga_rs13879050</td>
<td>0.45</td>
<td>0.1</td>
<td>0.39</td>
<td>66.67</td>
<td>0.06</td>
<td>RECQL</td>
</tr>
<tr>
<td>2,73</td>
<td>16</td>
<td>19.74</td>
<td>0.8</td>
<td>Gga_rs14203992</td>
<td>0.74</td>
<td>0.2</td>
<td>0.35</td>
<td>73.74</td>
<td>0.07</td>
<td>CDH10</td>
</tr>
<tr>
<td>2,318,140-141</td>
<td>72</td>
<td>12.16</td>
<td>1.29</td>
<td>Gga_rs13794969</td>
<td>0.17</td>
<td>2.24</td>
<td>0.39</td>
<td>138.2</td>
<td>0.005</td>
<td>MYC</td>
</tr>
<tr>
<td>4,85</td>
<td>33</td>
<td>2.21</td>
<td>0.8</td>
<td>Gga_rs14498387</td>
<td>0.67</td>
<td>0.54</td>
<td>0.29</td>
<td>85.48</td>
<td>0.07</td>
<td>RNF212, MAEA, FGFRL1, MRPL35</td>
</tr>
<tr>
<td>12,2</td>
<td>23</td>
<td>0.83</td>
<td>0.61</td>
<td>Gga_rs14032471</td>
<td>0.5</td>
<td>0.46</td>
<td>0.071</td>
<td>2.72</td>
<td>0.1</td>
<td>FANC D2, GNAI2</td>
</tr>
</tbody>
</table>

\(^a\)GV%, proportion of the genetic variance explained by window or SNP.
\(^b\)WPPA, posterior probability of association of 1-Mb region.
\(^c\)Lead SNP, SNP with the highest SPPI in the 1 Mb region.
\(^d\)SPPI, SNP posterior probability of inclusion.
\(^e\)MAF, minor allele frequency.
\(^f\)p value, the significance of single SNP analysis in ASREML.
<table>
<thead>
<tr>
<th>Term ID</th>
<th>Term name</th>
<th>p value</th>
<th>Lists of genes</th>
<th>GGA_Mb (White)</th>
<th>GGA_Mb (Brown)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:004260</td>
<td>Cellular</td>
<td>5.15E-05</td>
<td>*RPS17L,FANCD2,YWHAB,IGF2,MRPL23,RAG1,RAG2,EPGN,EREGR,AREGB,</td>
<td>1.54, 1.71, 1.84,</td>
<td>1.15, 1.66,</td>
</tr>
<tr>
<td></td>
<td>macromolecule</td>
<td></td>
<td>*NFYB,TDG,RECQL,RPL24,PRIM2,MYC,WBP4,RGCC,CCNA1,MRPL35,CBLL1</td>
<td>1.162-165, 3, 8.5, 4.4,</td>
<td>2, 138, 140-141,</td>
</tr>
<tr>
<td></td>
<td>metabolic process</td>
<td></td>
<td></td>
<td>5.13, 10.1, 20.5</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0051301</td>
<td>Cell division</td>
<td>8.03E-05</td>
<td>*GNAI2,IGF2,EPGN,EREGR,AEAE,RGCC</td>
<td>1.162-165, 4.4, 5.13,</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0007088</td>
<td>Mitotic nuclear division</td>
<td>1.09E-04</td>
<td>*IGF2,EPGN,EREGR,RGCC</td>
<td>1.162-165, 4.44</td>
<td>4.85</td>
</tr>
<tr>
<td>GO:0007169</td>
<td>Transmembrane receptor protein tyrosine</td>
<td>1.80E-04</td>
<td>*IGF2,EPGN,EREGR,AEG,FGFRL1,LECT1</td>
<td>1.162-165, 4.44</td>
<td>4.85</td>
</tr>
<tr>
<td></td>
<td>kinase signaling pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0034641</td>
<td>Cellular nitrogen compound metabolic process</td>
<td>2.22E-04</td>
<td>*RPS17L,FANCD2,IGF2,MRPL23,RAG1,RAG2,EREGR,AREGB,NFYB,TDG,RECQL,RPL24,PRIM2,</td>
<td>1.54, 1.84, 1.62-165,</td>
<td>1.66, 2.138, 140-141,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*MYC,WBP4,RGCC,MRPL35</td>
<td>3.85, 4.4, 5.13, 10.1,</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>Macromolecule</td>
<td>2.27E-04</td>
<td>*RPS17L,FANCD2,YWHAB,IGF2,MRPL23,RAG1,RAG2,EPGN,EREGR,AREGB,NFYB,TDG,RECQL,</td>
<td>1.54, 1.71, 1.84,</td>
<td>1.15, 1.66,</td>
</tr>
<tr>
<td></td>
<td>metabolic process</td>
<td></td>
<td>*RPL24,PRIM2,MYC,WBP4,RGCC,CCNA1,MRPL35,CBLL1</td>
<td>1.162-165, 3, 8.5</td>
<td>2, 138, 140-141,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4, 5.13, 10.1, 20.5</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0008283</td>
<td>Cell proliferation</td>
<td>2.60E-04</td>
<td>*GNAI2,IGF2,RAG1,RAG2,AREGB,EPGN,EREGR,AEG,FGFRL1,LECT1,RGCC</td>
<td>1.162-165, 4.4, 5.13,</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0090068</td>
<td>Cell cycle process</td>
<td>3.50E-04</td>
<td>*FANCD2,IGF2,EPGN,EREGR,RGCC,CCNA1,MYC,YWHAB,GNAI2,MAEA</td>
<td>1.71, 1.62-165, 4.44,</td>
<td>2, 138, 140-141,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.13, 20.5</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0006807</td>
<td>Nitrogen compound metabolic process</td>
<td>4.44E-04</td>
<td>*RPS17L,FANCD2,IGF2,MRPL23,RAG1,RAG2,EREGR,AREGB,NFYB,TDG,RECQL,RPL24,PRIM2,</td>
<td>1.54, 1.84, 1.62-165,</td>
<td>1.66, 2.138, 140-141,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*MYC,WBP4,RGCC,MRPL35</td>
<td>3.85, 4.4, 5.13, 10.1,</td>
<td>4.85, 12.2</td>
</tr>
<tr>
<td>GO:0000280</td>
<td>Nuclear division</td>
<td>6.67E-04</td>
<td>$FANCD2, IGF2, EPGN, EREG, RGCC$</td>
<td>1,162-165, 4,44, 5,13</td>
<td>12.2</td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>----------</td>
<td>----------------------------------</td>
<td>------------------------</td>
<td>------</td>
</tr>
<tr>
<td>GO:0007346</td>
<td>Mitotic cell cycle</td>
<td>8.45E-04</td>
<td>$IGF2, EPGN, EREG, RGCC, CCNA1$</td>
<td>1,71, 1,162-165, 4,44, 5,13</td>
<td></td>
</tr>
<tr>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
<td>1.30E-03</td>
<td>$FANCD2, RAG1, RAG2, AREGB, TDG, RECQL, PRIM2$</td>
<td>1,54, 1,162-165, 1,66, 12.2</td>
<td></td>
</tr>
<tr>
<td>GO:0090304</td>
<td>Nucleic acid metabolic process</td>
<td>1.30E-03</td>
<td>$RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC$</td>
<td>3,85, 4,44, 5,13, 10,1</td>
<td>12.2</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>Cellular metabolic process</td>
<td>1.60E-03</td>
<td>$RPS17L, FANCD2, YWHAB, IGF2, MRPL23, RAG1, RAG2, EPGN, EREG, AREGB, NFYB, TDG, ANKH, PRIM2, MYC, WBP4, RGCC, CCNA1, MRPL35, CBLL1$</td>
<td>1,54, 1,71, 1,84, 1,162-165, 3,85, 1,128,140-141, 4,44, 5,13, 10,1, 20.5</td>
<td></td>
</tr>
<tr>
<td>GO:0009987</td>
<td>Cellular process</td>
<td>1.77E-03</td>
<td>$RPS17L, FANCD2, YWHAB, GNAI2, IGF2, MRPL23, RAG1, RAG2, EPGN, EREG, AREGB, NFYB, TDG, ANKH, RECQL, MAEA, RPL24, FGFRL1, PRIM2, MYC, LECT1, WBP4, RGCC, CCNA1, MRPL35, CBLL1$</td>
<td>1,54, 1,71, 1,84, 1,162-165, 2,76-77, 2,138,140-141, 3,85, 4,44, 5,13, 4,85, 12.2</td>
<td></td>
</tr>
<tr>
<td>GO:0044238</td>
<td>Primary metabolic process</td>
<td>1.84E-03</td>
<td>$RPS17L, FANCD2, YWHAB, IGF2, MRPL23, RAG1, RAG2, EPGN, EREG, AREGB, NFYB, TDG, RECQL, RPL24, PRIM2, MYC, WBP4, RGCC, CCNA1, MRPL35, CBLL1$</td>
<td>1,54, 1,71, 1,84, 1,162-165, 3,85, 1,128,140-141, 4,44, 5,13, 10,1, 20.5</td>
<td>4,85, 12.2</td>
</tr>
<tr>
<td>GO:0007167</td>
<td>Enzyme linked receptor protein signaling pathway</td>
<td>1.89E-03</td>
<td>$IGF2, EPGN, EREG, AREGB, FGFRL1, LECT1$</td>
<td>1,162-165, 4,44, 5,13</td>
<td>4,85</td>
</tr>
<tr>
<td>GO:0071704</td>
<td>Organic substance metabolic process</td>
<td>3.44E-03</td>
<td>$RPS17L, FANCD2, YWHAB, IGF2, MRPL23, RAG1, RAG2, EPGN, EREG, AREGB, NFYB, TDG, RECQL, RPL24, PRIM2, MYC, WBP4, RGCC, CCNA1, MRPL35, CBLL1$</td>
<td>1,54, 1,71, 1,84, 1,162-165, 3,85, 1,128,140-141, 4,44, 5,13, 10,1, 20.5</td>
<td>4,85, 12.2</td>
</tr>
<tr>
<td>GO:1903047</td>
<td>Mitotic cell cycle process</td>
<td>3.75E-03</td>
<td>$IGF2, EPGN, EREG, RGCC, CCNA1$</td>
<td>1,71, 1,162-165, 4,44, 5,13</td>
<td></td>
</tr>
<tr>
<td>GO:0034645</td>
<td>Cellular macromolecule biosynthetic process</td>
<td>3.99E-03</td>
<td>RPS17L, IGF2, MRPL23, EREG, AREGB, NFYB, TDG, RPL24, PRIM2, MYC, RGCC, MRPL35</td>
<td>1.54, 1.84, 1.162-165, 1.162-165, 4.44, 5.13, 4.85</td>
<td></td>
</tr>
<tr>
<td>GO:0006139</td>
<td>Nucleobase-containing compound metabolic process</td>
<td>4.66E-03</td>
<td>RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC</td>
<td>1.54, 1.162-165, 4.44, 5.13, 10.1, 12.2, 2.138, 140-141, 2.138, 140-141, 4.85</td>
<td></td>
</tr>
<tr>
<td>GO:0009059</td>
<td>Macromolecule biosynthetic process</td>
<td>5.14E-03</td>
<td>RPS17L, IGF2, MRPL23, EREG, AREGB, NFYB, TDG, RPL24, PRIM2, MYC, RGCC, MRPL35</td>
<td>1.54, 1.84, 1.162-165, 1.162-165, 4.44, 5.13</td>
<td></td>
</tr>
<tr>
<td>GO:0046483</td>
<td>Heterocycle metabolic process</td>
<td>6.04E-03</td>
<td>RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC</td>
<td>1.54, 1.162-165, 4.44, 5.13, 10.1, 12.2, 2.138, 140-141, 2.138, 140-141, 4.85</td>
<td></td>
</tr>
<tr>
<td>GO:0006725</td>
<td>Cellular aromatic compound metabolic process</td>
<td>6.29E-03</td>
<td>RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC</td>
<td>1.54, 1.162-165, 4.44, 5.13, 10.1, 12.2, 2.138, 140-141, 2.138, 140-141, 4.85</td>
<td></td>
</tr>
<tr>
<td>GO:0008150</td>
<td>Biological process</td>
<td>6.42E-03</td>
<td>RPS17L, FANCD2, YWHA5B, GNAI2, IGF2, MRPL23, RAG1, RAG2, RNF212, SPO11, IMPG2, EPGN, EREG, AREGB, NFYB, TDG, CDH10, ANKH, RECQL, MAEA, RPL24, FGFR1L, PRIM2, MYC, LECT1, WBP4, RGCC, CCNA1, MRPL35, CBL1</td>
<td>1.54, 1.71, 1.84, 1.162-165, 2.138, 140-141, 2.138, 140-141, 3.85, 4.44, 5.13, 10.1, 20.5, 20.10</td>
<td></td>
</tr>
<tr>
<td>GO:0001937</td>
<td>Negative regulation of endothelial cell proliferation</td>
<td>6.52E-03</td>
<td>RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC, CCNA1, MRPL35, CBL1, LECT1, RGCC</td>
<td>1.162-165</td>
<td></td>
</tr>
<tr>
<td>GO:1901360</td>
<td>Organic cyclic compound metabolic process</td>
<td>8.64E-03</td>
<td>RPS17L, FANCD2, IGF2, RAG1, RAG2, EREG, AREGB, NFYB, TDG, RECQL, PRIM2, MYC, WBP4, RGCC</td>
<td>1.54, 1.162-165, 4.44, 5.13, 10.1, 12.2, 2.138, 140-141, 2.138, 140-141</td>
<td></td>
</tr>
<tr>
<td>GO:0098609</td>
<td>Cell-cell adhesion</td>
<td>9.43E-03</td>
<td>IGF2, RAG1, RAG2, CDH10, FGFR1L, RGCC</td>
<td>1.162-165, 5.13, 2.73, 4.85</td>
<td></td>
</tr>
</tbody>
</table>
3.9 Figures

Figure 3.1 **Recombination rate within 0.5-Mb windows estimated on GGA1.** The black solid line corresponds to the recombination rate estimated from 580K SNPs in white layers. The grey dashed line corresponds to the recombination rate estimated from 42K SNPs in brown layers.
Figure 3.2  Variation in recombination rate within 0.5-Mb windows across the 28 autosomes. The black solid line corresponds to the recombination rate estimated from 580K SNPs in white layers. The grey dashed line corresponds to the recombination rate estimated from 42K SNPs in brown layers.
Figure 3.3  Recombination rate within 0.5-Mb estimated using segregating 580K SNPs (black solid line) and 42K SNPs (green dashed line) in white layers on GGA1.
Figure 3.4 **The GRN in white and brown half-sib families.** Black dots correspond to GRN in offspring sorted by parents and red dots correspond to the average GRN for each parent. Only parents with at least 2 offspring are presented.
Figure 3.5  Distributions of the NHA (number of common haplotype alleles) and PHHA (proportion of homozygosity of common haplotype alleles) within 0.5-Mb window in white and brown layers.
Figure 3.6  Recombination rate within 0.5-Mb windows estimated on GAA1 for males and females in white and brown chickens.
Figure 3.7 Distributions of the average GRN of males (red) and females (blue) in white and brown layers.
Figure 3.8  Relationship of GC content and CpG island density with recombination rate within 0.5-Mb windows across chromosomes in white and brown layers.
Figure 3.9  Proportion of genetic variance explained by 1-Mb regions across the genome for GRN in white layers.
Figure 3.10  Proportion of genetic variance explained by 1-Mb regions across the genome for GRN in brown layers.
CHAPTER 4. GENOME-WIDE ASSOCIATION STUDY OF GROWTH AND BODY COMPOSITION TRAITS IN BRANGUS BEEF CATTLE

Ziqing Weng¹, Hailin Su¹, Mahdi Saatchi², Jungjae Lee¹, Milton G. Thomas³, Jenelle R. Dunkelberger¹, Dorian Garrick¹

¹Department of Animal Science, Iowa State University, Ames, IA 50010, USA
²Genus plc, Hendersonville, TN 37075, USA
³Department of Animal Science, Colorado State University, Fort Collins, CO 80523, USA

A paper accepted by Journal of Livestock Science

4.1 Abstract

The availability of high-density single nucleotide polymorphism (SNP) genotypes, such as Boveine HD770K, provides opportunities to identify genomic regions associated with traits in cattle. The objective of this study was to identify quantitative trait loci (QTL) associated with growth and body composition traits in Brangus beef cattle using actual or imputed 770K SNP genotypes. A total of 1537 Brangus beef cattle were genotyped with the Bovine50K, GGPHD77K, or BovineHD770K SNP chip and deregressed estimated breeding values were derived and fitted as observations in analyses. BayesB approach was used to map QTL for each trait, and significant windows and SNPs were identified. A total of 18 QTL were identified, in which 7 were associated with more than one trait, while the remained 11 QTL were trait-specific. One pleiotropic QTL of particularly large-effect was identified on chromosome 6 at 38 Mb, which influences birth weight, weaning weight, and
yearling weights, and harbors growth-related genes NCAPG and LCORL. Biological pathways of each identified QTL were also performed using gene ontology term enrichment analysis. The QTL mapping results obtained from this study will aid in better understanding the biological processes accounting for variation in growth and body composition traits in Brangus cattle.

4.2 Introduction

Genome-wide association studies (GWAS) are widely used to detect chromosomal regions that are responsible for genetic variation for a trait of interest (Weller, 2009). Increasing marker density, enlarging the sample size, and use of markers in strong linkage disequilibrium (LD) with quantitative trait loci (QTL), are ways to improve the accuracy of GWAS. Bayesian regression methods are useful for GWAS (Zou and Zeng, 2008) because they account for uncertainty in parameters required to construct posterior distributions for QTL inference. Recently, several regions of the bovine genome associated with growth traits have been reported in beef cattle. One region is on Bovine autosome (BTA) 6, which harbors candidate genes including NCAPG and LCORL (Saatchi et al., 2014; Santana et al., 2014; Snelling et al., 2010; Nishimura et al., 2012). Another region, located on BTA 14, contains PLAG1, CHCHD7, and other related genes, all of which affect growth in cattle (Nishimura et al., 2012; Pausch et al., 2011; Saatchi et al., 2014) and humans (Pryce et al., 2011; Utsunomiya et al., 2013).

Brangus is a composite breed comprising Brahman and Angus (3/8 Bos indicus × 5/8 Bos taurus). QTL mapping using Brangus might identify QTL that are segregating in both Angus and Brahman. Peters et al. (2012) and Saatchi et al. (2014) mapped QTL for growth traits using the lower density BovineSNP50 in two subsets (835 heifers and 1328 cattle) of this same dataset of U.S. Brangus cattle, respectively. However, GWAS of Brangus using high density of SNP genotypes have not been fully addressed. The objective of this study was to identify QTL associated with growth and body composition traits in a Brangus population using imputed BovineHD770K SNP genotypes, and comparing the results with previous studies.
4.3 Methods

4.3.1 Phenotypes and Genotypes

Estimated breeding values (EBVs) and heritability estimates (Table 4.1) of growth and body composition traits in 1537 Brangus beef cattle (983 heifers and 554 bulls from 168 farms) were obtained from the International Brangus Breeders Association. Traits studied included direct birth weight, maternal birth weight, back fat thickness, intramuscular fat thickness, rib eye muscle area, rump fat thickness, mature weight, weaning weight, and yearling weight.

A subset of these Brangus cattle (1121) were genotyped with Bovine50K (Illumina, San Diego, CA), while another 243 cattle were genotyped with BovineHD770K (Illumina, San Diego, CA), and the rest 173 cattle were genotyped with GGPHD77K (Gene Seek, Lincoln, NE). These Brangus cattle were born between 1976 to 2012. Distribution of their birth year was shown in Table 4.2. Most animals genotyped with BovineHD770K and Bovine50K were born between 2001 to 2010, while majority of animals genotyped with GGPHD77K were born after 2010. Genotypes were obtained using DNA extracted from blood, semen, or hair samples and did not require an approved animal care and use statement.

4.3.2 Imputation

All of the animals genotyped with the Bovine50K or GGPHD77K were imputed to BovineHD770K using FImpute (Sargolzaei et al., 2014). Before imputation, 77 1089 SNPs on BovineHD770K with call rate >0.90 (6173 SNPs were removed) and could be mapped to SNPchiMP (Nicolazzi et al., 2014) were retained as HD reference panel. SNPs aligned to HD reference panel with call rate >0.90 were remained on Bovine50K (48 740 SNPs) and GGPHD77K (73 989 SNPs), respectively. All of 1537 animals had call rate >0.90.

After imputation, 736 053 segregating markers remained after quality control and limiting markers to those that were uniquely assigned to 29 bovine autosomes and X chromosome on UMC assembly (University of Missouri). In total, 31 506 SNP with minor allele frequency <0.005, 1565 SNPs with Mendelian inconsistency rate between parent and progeny >0.05, and 1965 unmapped SNP
were removed, respectively. Genotype quality control was conducted using PLINK v1.0 (Purcell et al., 2007).

Genotyped animals were clustered into three groups according to SNP genotype panels (Bovine50K or GGPHD77K or BovineHD770K). The average additive genetic relationship coefficient (\(a_{mean}\)) and maximum additive genetic relationship coefficient (\(a_{max}\)) were calculated between each animal and other animals from the same group or other groups using pedigree information. To qualify the relationship between different groups, the mean values of \(a_{mean}\) and \(a_{max}\) within and between groups were calculated across all animals in each group.

A simulation study was conducted to evaluate the imputation accuracy of FImpute. A sample of 10 animals was randomly selected from HD770K group. The remained 233 animals were used as reference population. Imputation from 77K to 770K SNP panels was performed in these selected animals based on the use of only that subset of 73,989 SNPs from the 770K panel that were on the 77K. The same strategy was conducted to impute genotypes from 50K to 770K SNP panels only using the overlapped 48,740 SNPs. Each scenario was repeated 5 times, in order to avoid sample bias. Allelic r² was used to qualify imputation accuracy, which depends less on SNP allele frequency compared with concordance rate (Browning and Browning, 2009). It was calculated as the squared correlation between the imputed genotypes and original genotypes on the 770K SNP panel in the selected animals.

4.3.3 Degressed EBVs

Individual EBVs combine information from the performance of the individual, any progeny, and its parents, based on pedigree relationships. In this study, deregressed estimated breeding values (DEBVs) were derived according to Garrick et al. (2009) and used as response variables to estimate SNP effects in a weighted Bayesian analysis. Weighting factors were used to account for heterogeneous variance caused by differences in reliabilities of individual DEBVs. Only the DEBVs with reliability > 0.8\(h^2\) were retained in the analysis which eliminated animals without individual or offspring information contributing to their EBVs, causing the number of genotyped animals with DEBVs to vary between traits (Table 4.1).
4.3.4 Genome-wide association study

Mapping of QTL associated with the performance traits of interest was conducted using the BayesB method (Meuwissen et al., 2001) with weighting factors implemented in GENSEL4.4 software (Fernando and Garrick, 2013; Garrick and Fernando, 2013). According to the BayesB method, each SNP effect follows an independent, univariate \( t \)-distribution with null mean, degree of freedom \( \nu_u \), and scale parameter \( S_u^2 \), which is equivalent to a univariate normal distribution with null mean and locus specific variance (Fernando and Garrick, 2013). For each trait, DEBVs were used as response variables in the following model:

\[
y_i = \mu + \sum_{j=1}^{k} z_{ij} u_j + e_i, \quad [1]
\]

where \( y_i \) is the DEBV for animal \( i \), \( \mu \) is the population mean, \( k \) is the number of SNP loci, \( z_{ij} \) is marker genotype code (0/1/2) for SNP \( j \) in animal \( i \), \( u_j \) is the substitution effect for SNP \( j \) with \( u_j \overset{i.i.d.}{\sim} \begin{cases} 0 & \text{with prob. } \pi \\ N(0, \sigma_j^2) & \text{with prob. } 1 - \pi \end{cases} \) and \( e_i \) is a residual effect with heterogeneous variance.

Weighting factors (Garrick et al., 2009) were calculated as:

\[
w_i = \frac{1 - h^2}{c + (1 - r_i^2)/r_i^2 h^2}, \quad [2]
\]

where \( c \) is the proportion of the genetic variance (GV\%) not explained by markers (assumed to be 0.40) (Saatchi et al., 2013), \( h^2 \) is the heritability of the trait, and \( r_i^2 \) is the reliability of the DEBV for animal \( i \). Parameter \( \pi \) was assumed to be 0.999, corresponding to 0.1\% of SNPs to be fitted in the model each iteration. The estimates of genetic and residual variances were obtained from a BayesC\( \pi \) analysis (Habier et al., 2011), in which \( \pi \) was treated as an unknown parameter. Markov chain Monte Carlo (MCMC) sampling with 44 000 iterations were used to estimate posterior means of SNP substitution effects, with the first 4000 samples discarded as burn-in.

The genome was divided into 2600 non-overlapping 1-Mb windows, with the expected GV\% explained by each 1-Mb window being approximately 0.04\% under an infinitesimal polygenic model. Windows that explained at least 0.5\% (> 10 folds of the expected percentage) of genetic variance, or their adjacent windows 1-Mb up or downstream, were considered to harbor QTL (Garrick and
Fernando, 2013; Weng et al., 2014). The proportion of genetic variance in any particular iteration explained by each 1-Mb window was defined as:

\[ q_w = \frac{\hat{\sigma}_{gw}^2}{\hat{\sigma}_g^2} \]

where \( \hat{\sigma}_{gw}^2 \) is the sampled genetic variance explained by markers in genomic window \( w \), and \( \hat{\sigma}_g^2 \) is the sampled total genetic variance. See Fernando and Garrick (2013) for details in sampling \( \hat{\sigma}_{gw}^2 \) and \( \hat{\sigma}_g^2 \). The window posterior probabilities of association (WPPA) for each candidate window were also calculated, reflecting the proportion of samples for which there was at least one non-zero effect SNP. The squared correlations (\( r^2 \)) between genotypes of every two adjacent SNPs within 1-Mb window were calculated as a measure of LD using R software (R Core Team, 2014).

The lead SNP was defined as the SNP with the highest SNP posterior probability of inclusion (SPPI) within the candidate window (window that explained at least 0.5% of GV). The SPPI was estimated as the percentage of MCMC samples in which a given SNP had a non-zero effect. The GV% explained by each lead SNP was assessed by fitting it separately from the candidate window in the GWAS models (Serão et al., 2014). Significance of every lead SNP was evaluated in an animal model with ASReml v3.0 (Gilmour et al., 2009), by fitting the single SNP genotype (0/1/2 or AA/AB/BB) as a fixed class effect (e.g. Wolc et al. (2014)). A full pedigree contains 17 865 animals over 10 generations was used in the analysis. Least square means (LSM) for the genotypes of the lead SNP were estimated from the same model. The additive effect of each lead SNP was calculated as half of the difference between two homozygotes (\( BB - AA \)), whereas the dominance effect was calculated as LSM of heterozygote minus the average LSM of homozygotes (\( AB - \frac{BB + AA}{2} \)). Significance of additive and dominance effect of every lead SNP was also assessed using ASReml. Positional candidate genes harbor or near (±1Mb) lead SNP that might control expression of the traits were identified using Ensembl, and NCBI gene database.

Significant QTL windows, which associated with more than one trait, were defined as pleiotropic QTL. Genetic correlations between traits associated with pleiotropic QTL were estimated in a multi-traits animal model using ASReml. Gene ontology (GO) term enrichment analysis on genes within/near pleiotropic QTL was performed using the GO database and web tool g:Profiler (Reimand
et al., 2011). Only GO terms with significant P value of 0.05 (Benjamini-Hochberg FDR corrected), and associated with biological processes and Kyoto encyclopedia of genes and genomes (KEGG) pathway were retained.

4.4 Results and Discussion

4.4.1 Imputed HD genotypes

The additive relationship coefficients within and between the three groups genotyped by BovineHD770K (group 1), GGPHD77K (group 2), or Bovine50K (group 3) are shown in Table 4.3. Their averaged \(a_{mean}\) within group were similar, which ranged from 0.10 to 0.14. Group 3 had the highest average within group \(a_{max}\) (0.53±0.16), where as group 2 had the lowest coefficient (0.20±0.10). The averaged \(a_{mean}\) and \(a_{max}\) between group 1 and group 2 were 0.10±0.046 and 0.17±0.11, respectively. The averaged \(a_{mean}\) and \(a_{max}\) between group 1 and group 3 were 0.12±0.045 and 0.38±0.16, respectively.

The average allelic \(r^2\) from 77K to 770K SNP genotypes among 50 randomly selected animals (10 animals per repeat) was 0.97±0.022. The average allelic \(r^2\) from 50K to 770K SNP genotypes was 0.96±0.034. Sargolzaei et al. (2014) reported the allelic \(r^2\) was 0.85 for imputation from 3k to 50k, was 0.94 for 6k to 50k, and was 0.96 for 50k to 300k. In their simulation study, 100 animals were included in the reference population, and 2000 or 500 target individuals were used for imputation from 3k/6k to 50k and from 50k to 300k, respectively.

Though imputation accuracies using FImpute in the simulation studies were high, and genotyped individuals between different groups were close to each other, the size of HD770K group (reference population) was limited in the real data analysis. Since imputation could affect the GWAS results, further study is needed to validate the QTL identified in this study.

4.4.2 Significant windows and lead SNP

Table 4.4 shows descriptive statistics of significant windows associated with different studied traits, such as number of SNP included in the window, explained GV%, and WPPA, as well as
descriptions of lead SNPs (e.g. SPPI, MAF, physical position, and explained SNP GV%). Positional
candidate genes for growth, reproduction, and production in Brangus cattle were also detected
within the significant windows. 8 QTL were obtained for weaning weight, but only 1 QTL for fat
thickness. No significant QTL were identified for maternal birth weight. The largest QTL window,
which explained 6.95% of genetic variance, was captured on BTA6 at 38 Mb in direct birth weight.
The QTL window with smallest effect size (0.51%) was identified in rib eye muscle area. Manhattan
plots showing the proportion of genetic variance explained by each 1-Mb window across 9 traits are
in Figure 4.1.

A total of 18 significant QTL were associated with growth and body composition traits in
Brangus beef cattle using imputed Bovine770K SNP genotypes. Of these 18 QTL, 7 were common
to all traits studied, and the other 11 were trait-specific. Compared to previous studies, Peters et al.
(2012) and Saatchi et al. (2014), four novel pleiotropic QTL and two novel trait-specific QTL were
identified in Brangus beef cattle in this study. These four novel pleiotropic QTL, which located on
BTA5 at 47-48 Mb, BTA10 at 33 Mb, BTA12 at 88 Mb, and BTA 20 at 7-8 Mb, were found to
influence direct birth weight, weaning weight, yearling weight, mature weight, intramuscular fat,
and/or rib eye muscle area in Brangus beef cattle. Two novel trait-specific QTL were, on BTA5
at 15 Mb associated with rump fat thickness, and on BTA9 at 101 Mb associated with back fat
thickness. However, further study is required to explore the potential biological functions of these
two novel QTL. Other trait-specific QTL, such as QTL on BTA3 at 27Mb, on BTA12 at 82Mb,
on BTA18 at 54Mb, and on BTA19 at 26 and 29Mb, were previously reported only in Bos taurus
breeds, but their signal were captured in Brangus beef cattle in this study.

Four large-effect lead SNP, which explained more than 1% of GV%, were identified. The lead
SNP on BTA6 at 38M b, rs133837477, explained 6.9% of genetic variance on direct birth weight,
1.16% on weaning weight, and 1.82% on yearling weight. The other large-effect lead SNP identified
in multiple traits (rib eye muscle area, weaning weight and yearling weight) was rs1354677277
located on BTA12 at 88Mb. On BTA5 at 48Mb, rs135392994 was responsible for 1.15% of genetic
variance on intramuscular fat thickness, whereas rs42356467 on BTA20 at 7Mb explained 1.13% of
genetic variance on yearling weight.
Table 4.5 presents the LSM of traits for the genotypes, and estimates of additive and dominance effects of all lead SNP from single SNP analysis. All lead SNP had significant effect with P value less than 0.001 on target traits. LSM of target traits for homozygotes and heterozygotes were significantly different in most lead SNP. Majority of lead SNP had significant additive effects. Only two lead SNP, rs13383747 and rs133487364 on BTA 6 at 38Mb, showed significant dominance effects on weaning weight, and yearling weight, or direct birth weight, respectively. Favorable and unfavorable genotypes of lead SNP for target traits could be identified from estimates of LSM. For example, rs135392994, AA animals showing favorable performance (-0.13) for intramuscular fat thickness, but BB animals showing unfavorable performance (0.41).

4.4.3 Pleiotropic or closely linked QTL

Seven large-effect pleiotropic or closely linked QTL, meaning that the QTL had an effect on multiple traits, were identified, located on BTA5 at 47-48 Mb, BTA6 at 38 Mb, BTA7 at 93 Mb, BTA10 at 33 Mb, BTA12 at 88 Mb, BTA 20 at 4 Mb, and BTA 20 at 7-8 Mb. Two adjacent candidate windows were combined for analysis (i.e., BTA5 at 47 Mb combined with BTA5 at 48 Mb, and BTA20 at 7 Mb combined with BTA20 at 8 Mb), because it was shown that the location of the QTL could be in the flanking regions (±2 Mb) on either side of the 1-Mb QTL window (Garrick and Fernando, 2013; Weng et al., 2014).

A QTL on BTA5 at 47-48 Mb was associated with birth weight, weaning weight, ultrasound intramuscular fat, and ultrasound rib eye muscle area. Three candidate genes were identified within this region: *HMGA2* (high mobility group protein A2) located at 48.05-48.20 Mb, which is associated with human height (Weedon et al., 2007); *MSRB3* (methionine-S-sulphoxide reductase) located at 48.56-48.74 Mb and associated with gestational age at birth in humans (Lee et al., 2012); *LEMD3* (LEM domain containing 3) located at 48.77-48.84 Mb and associated with bone density disorder in humans (Ben-Asher et al., 2005). Other QTL associated with bovine growth and body weight have also been reported in this region (Maltecca et al., 2009; McClure et al., 2010). Saatchi et al. (2014) also identified QTL in this region associated with birth weight, calving ease, marbling, and rib eye muscle area in Brangus (a subset of this data set) using 50K SNP genotypes.
The QTL on BTA6 at approximately 38 Mb associated with birth weight direct, weaning weight and yearling weight, had a greater average GV% (3.64%) than other pleiotropic QTL. This region explained 6.94% of additive genetic variance in birth weight direct, 1.58% in weaning weight, and 2.4% in yearling weight. Snelling et al. (2010) found in crossbred beef cattle that most SNPs associated with direct growth are located on BTA6. Nishimura et al. (2012) identified major QTL for carcass weight in Japanese black cattle. Many other studies have also reported QTL associated with average daily gain (Santana et al., 2014), reproductive traits (Daetwyler et al., 2008), and milk production (Olsen et al., 2005) on BTA6 in beef cattle or dairy cattle. This region harbors 6 promising candidate genes: ABCG2 (ATP-binding cassette, sub-family G, member 2), IBSP (integrin-binding sialoprotein), LAP3 (leucine aminopeptidase 3), LCORL (ligand dependent nuclear receptor corepressor-like), MEPE (Matrix extracellular phosphoglycoprotein), NCAPG (non-SMC condensing I complex, subunit G), PKD2 (polycystic kidney disease 2), and SPP1 (secreted phosphoprotein 1). ABCG2 is evolved in iron transportation and metabolism, IBSP and SPP1 function in bone mineralization, and PKD2 plays an role in calcium channel activity, calcium binding, and ion binding (OMIM). LAP3 has an impact on calving ease (Bongiorni et al., 2012), LCORL affects stature in humans and cattle (Pryce et al., 2011), MEPE is differentially expressed in RNA-Seq analysis of heifer pregnancy from pre- and post-puberty Brangus heifers (Cánovas et al., 2014), and NCAPG is associated with growth traits in cattle (Weikard et al., 2010; Widmann et al., 2015).

The positional candidate gene on BTA7 at 93Mb was ARRD3 (arresting domain containing 3), which is located at 93.24-93.25 Mb. Oka et al. (2006) reported that ARRD3 interacts with plasma membrane, endosomes, and lysosomes during endocytosis. This QTL window was also detected in an association study with composition of fatty acids in beef cattle (Alexander et al., 2007). Saatchi et al. (2014) indicated that this QTL has large-effects on birth weight, calving ease, carcass weight, rib eye muscle area, and weaning weight across several breeds.

A pleiotropic QTL on BTA10 at approximately 33 Mb was associated with mature weight, weaning weight, yearling weight, and ultrasound rib eye area. SPRD1 (sprouty-related, EVH1 domain containing 1) regulates several growth factors (Nonami et al., 2004) and is the positional
candidate gene. (McClure et al., 2010) found QTL within the same region associated with mature weight in Angus.

The QTL on BTA12 at approximately 88 Mb explained significant GV% in mature weight, weaning weight, yearling weight, and ultrasound rib eye area. The lead SNP (rs135467227) was the same across these 4 traits, and is located near COL4A1 (collagen, type IV, alpha 1). COL4A1 is related to developmental biology, protein digestion, and protein absorption (OMIM). A QTL associated with mature weight in Angus was also reported in this region (McClure et al., 2010).

A pleiotropic QTL on BTA20 at 4 Mb was associated with mature weight and weaning weight. The lead SNPs detected in mature weight and weaning weight were rs42661324 and rs42661321, respectively. Two positional candidate genes exist in this region, ERGIC1 (endoplasmic reticulum-Golgi intermediate compartment protein 1) located at 4.51-4.62 Mb and SH3PXD2B (SH3 and PX domain 2B) located at 4.01-4.14 Mb. Saatchi et al. (2014) found associations of this QTL with body weight and calving ease in Angus, Shorthorn, Hereford, Red Angus, and Simmental cattle.

Two pleiotropic genes, UTP15 (U3 small nucleolar ribonucleoprotein, homolog), and TNPO1 (transportin 1), were detected on BTA20 at 7-8 Mb associated with birth weight (GV%=0.53%), mature weight (GV%=0.56%), and yearling weight (GV%=1.44%). UTP15 is located at 7.97-7.99 Mb and TNPO1 is located at 8.65-8.74 Mb. Both genes are evolved in biological process including cellular macromolecule localization and cellular protein localization (Table 4.7). Peters et al. (2012) detected QTL associated with longissimus muscle (GV%=1.3%), 365-d weight (GV%=0.14%), and average daily gain (GV%=0.62%), located on BTA20 at 6-8 Mb in Brangus using 50K SNP genotypes with Bayesian C approach.

4.4.4 GO analysis of pleiotropic QTL

Correlation coefficients between 6 traits which associated with pleiotropic QTL were estimated (Table 4.6). Except correlations between rib eye muscle area and intramuscular fat thickness, other phenotypic and genetic correlations were relatively strong. Genetic correlation among direct birth weight, rib eye muscle area, mature weight, weaning weight, and yearling weight ranged from 0.31 to 0.95. Intramuscular fat thickness had negative genetic correlation with other traits, ranged from
-0.13 to -0.21, except with rib eye muscle area (0.083). The effects of pleiotropic QTL, which control the performance of multiple traits, result genetic correlations among those traits.

The GO term analyses of the positional candidate genes within pleiotropic QTL window are shown in Table 4.7. Significant enrichments for biological process and KEGG pathway analyses, including adipose tissue development, fat pad development, cellular localization, cellular response to chemical stimulus, the extracellular matrix (ECM)-receptor interaction, and focal adhesion were obtained. Candidate genes were clustered according to their biological functions or pathways involved. For example, a pleiotropic QTL located on BTA5 at 47-48 Mb, is associated with birth weight, intramuscular fat thickness, rib eye muscle area, and weaning weight. Its positional candidate gene, \textit{HMGA2}, showed GO enrichments in the process of adipose tissue development, fat pad development, cellular localization, and cellular response to chemical stimulus. These biological processes have potential impacts on the performance of associated traits, especially directly related to intramuscular fat thickness.

4.4.5 Trait-specific QTL

In total, 11 trait-specific QTL were identified for 5 traits. A QTL for birth weight was identified on BTA5 at 5 Mb, which harbors \textit{OSBPL8} (oxysterol binding protein-like 8), a member of the \textit{OSBP} family of intracellular lipid receptors (OMIM). Peters et al. (2012) reported a QTL near BTA5 at 6 Mb which has an effect on birth weight in Brangus using 50K SNP genotypes.

The QTL on BTA9 at 101-Mb was found only for ultrasound fat thickness. \textit{PDE10A} (phosphodiesterase 10A) located in this region, mediates signal transduction by regulating the intracellular concentration of cyclic nucleotides (Fujishige et al., 2000). Further study is needed to investigate its potential biological function on fat thickness.

Ultrasound intramuscular fat had 2 trait-specific QTL. One was located on BTA18 at 54 Mb, and the other on BTA6 at 81-Mb. Positional candidate genes on BTA18 at 54 Mb include \textit{AP2S1} (adaptor-related protein complex 2, sigma 1 subunit), \textit{DCT3} (dishevelled-binding antagonist of beta-catenin 3), and \textit{NAPA} (N-ethylmaleimide-sensitive factor attachment protein, alpha). \textit{AP2S1} is a calcium-sensing receptor regulator and associates with the plasma membrane. Recurrent mis-
sense mutations in *AP2S1* is a cause of isolated hypoparathyroidism (Lambert et al., 2014). *DACT3* is expressed in embryonic development (Fisher et al., 2006), and *NAPA* encodes membrane protein. Saatchi et al. (2014) identified QTL in the same region associated with calving ease direct in Simmental cattle. McClure et al. (2010) reported QTL for birth weight on BTA18 at 56 Mb. A positional candidate gene on BTA6 at 81-Mb is *TECLR* (trans-2, 3-enoyl-CoA reductase-like), which regulates lipid metabolic process (Talmud et al., 2009), and expressed differentially on age at puberty in Brangus heifers (Cánovas et al., 2014).

The QTL associated with ultrasound rump fat thickness were all trait-specific. QTL on BTA3 at 27 Mb harbors *NHLH2* (nescient helix loop helix 2) and *CASQ2* (calsequestrin 2). Brennan et al. (2006) identified that *NHLH2* is a central nervous system regulator of adult body weight and has a putative role in body weight management in cattle. The *CASQ2* protein serves as the major calcium ion reservoir and forms a protein complex with ryanodine receptor (OMIM). A Region near BTA3 at 27 Mb was reported to be associated with weaning weight (McClure et al., 2010). A QTL associated with weaning weight was found on BTA3 at 26 Mb in Angus (McClure et al., 2010). *RASSF9* (Ras association domain family member 9), which is located on BTA5 at 15 Mb and is involved in mitosis, is another positional candidate gene. Though no positional genes were found on BTA9 at 5 Mb, this region is related to average daily gain in Brangus (Peters et al., 2012) and mature weight in Angus (McClure et al., 2012). Regions on BTA19 at 26 and 29 Mb harbor the positional candidate genes *MIS12* (MIND kinetochore complex component) and *STX8* (syntaxin 8), respectively. *MIS12* is essential for chromosome segregation and alignment (Kiyomitsu et al., 2007), and *STX8* cycles in plasma membrane. McClure et al. (2010) reported QTL between 26-29 Mb on BTA19 associated with mature weight and yearling weight.

Two trait-specific QTL were identified for weaning weight. One QTL was located on BTA12 at 82 Mb and harbors the positional candidate gene *METTL21C* (methyltransferase like 21C). *METTL21C*, located at 82.91-82.92 Mb, is a member of the protein lysine methyltransferase family and methylate nonhistone proteins (OMIM). Saatchi et al. (2014) found the same QTL related to calving ease direct in Maine-Anjou. McClure et al. (2010) reported QTL associated with mature weight near 81-Mb. Two QTL were identified located on BTA14 approximately at 26 Mb. The
positional candidate genes \textit{CYP7A1} (cytochrome P450, family 7, subfamily A, polypeptide 1) and \textit{SDCBP} (syndecan binding protein) are located at 25.35-25.36 Mb at 26.41-26.45 Mb, respectively. \textit{CYP7A1} is involved in the synthesis of cholesterol, steroids, and other lipids and is associated with neuromyelitis optica (Zhao et al., 2013). \textit{SDCBP} encodes protein that binds to a variety of trans-membrane proteins. Pleiotropic QTL located at 23-26 Mb were associated with body weight and calving ease and were found to segregate in Brangus, Gelbvieh, and Simmental cattle (Saatchi et al., 2014). The region on BTA14 at 25 Mb, which is 1-Mb down-stream of the identified QTL in this study, contains large-effect genes such as \textit{PLAG1} (pleiomorphic adenoma gene 1) and \textit{CHCHD7} (coiled-coil-helix-coiled-coil-helix domain containing 7). Pausch et al. (2011) and Pryce et al. (2011) identified major loci affecting growth-related traits on BTA14 at 25 Mb in cattle and its orthologous region on human chromosome 8, which have been shown to be associated with adult height (Gudbjartsson et al., 2008). Littlejohn et al. (2012) validated genetic effects of \textit{PLAG1} on animal size and growth in Holstein-Friesian dairy calves. Cánovas et al. (2014) reported two genes harboring SNP associated with Brangus heifer fertility traits.

\subsection*{4.4.6 QTL segregating in other breeds}

In this study, we identified 5 large-effect QTL in Brangus beef cattle located on BTA5 at 47-48 Mb, on BTA6 at 38 Mb, on BTA7 at 93 Mb, on BTA14 at 26 Mb, and on BTA20 at 4 Mb, which were found segregating across different cattle breeds (Saatchi et al., 2014). Saatchi et al. (2014) reported that the QTL harbored in BTA5 at 48 Mb were segregating in Angus and Brangus, the QTL within BTA6 at 37-39Mb, BTA7 at 93 Mb, and BTA20 at 4-5 Mb were segregating across Red Angus, Simmental, Gelbvieh, Hereford, Limousin, and Shorthorn, while the QTL in BTA 14 at 24-26Mb were segregating across Simmental, Gelbvieh, and Brangus. Identified QTL previously found in other breeds may provide important information for future across-breed genomic prediction analyses.
4.5 Conclusions

In summary, 18 significant QTL were associated with growth and body composition traits in Brangus beef cattle using actual and imputed Bovine770K SNP genotypes. Among the identified QTL, 7 QTL had pleiotropic effects and 11 were trait-specific. Compared to previous studies, 4 novel pleiotropic QTL, 2 novel trait-specific QTL, and 5 QTL reported only in *Bos taurus* studies, were identified in Brangus beef cattle. Further analysis is needed to validate the identified QTL in this study. In conclusion, the identified QTL and their corresponding functional annotation aids in understanding the biological processes of regions that effect growth and body composition in Brangus beef cattle.

4.6 Acknowledgements

Financial support provided by USDA-AFRI (Grant no.2008-35205-18751 and 2009-35205-05100). We gratefully acknowledge the data provided by the International Brangus Breeders Association.

4.7 Bibliography


OMIM. Online mendelian inheritance in man (http://www.omim.org).


### 4.8 Tables

Table 4.1  **Heritability ($h^2$), number of genotyped animals with DEBV and mean reliabilities (± SE) for the evaluated traits in Brangus.**

<table>
<thead>
<tr>
<th>Trait</th>
<th>$h^2$</th>
<th>N</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight direct</td>
<td>0.42</td>
<td>1401</td>
<td>0.58±0.14</td>
</tr>
<tr>
<td>Birth weight maternal</td>
<td>0.15</td>
<td>864</td>
<td>0.40±0.16</td>
</tr>
<tr>
<td>Back fat thickness</td>
<td>0.35</td>
<td>1065</td>
<td>0.47±0.19</td>
</tr>
<tr>
<td>Intramuscular fat thickness</td>
<td>0.23</td>
<td>1152</td>
<td>0.34±0.19</td>
</tr>
<tr>
<td>Rib eye muscle area</td>
<td>0.39</td>
<td>1160</td>
<td>0.49±0.19</td>
</tr>
<tr>
<td>Rump fat thickness</td>
<td>0.38</td>
<td>860</td>
<td>0.49±0.20</td>
</tr>
<tr>
<td>Mature weight</td>
<td>0.27</td>
<td>1176</td>
<td>0.39±0.27</td>
</tr>
<tr>
<td>Weaning weight</td>
<td>0.24</td>
<td>1426</td>
<td>0.46±0.14</td>
</tr>
<tr>
<td>Yearling weight</td>
<td>0.27</td>
<td>1332</td>
<td>0.45±0.15</td>
</tr>
</tbody>
</table>

$^a$Heritabilities reported by American Brangus Breeders Association. Parent average contributions have been removed to calculate the reliabilities.
Table 4.2  Distribution of birth year of genotyped Brangus cattle on each sSNP panel.

<table>
<thead>
<tr>
<th>Birth year</th>
<th>Total animals</th>
<th>Bovine HD770K</th>
<th>GGP HD770K</th>
<th>Bovine 77K</th>
<th>Bovine 50K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976 to 1990</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>1991 to 1995</td>
<td>25</td>
<td>7</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>1996 to 2000</td>
<td>96</td>
<td>11</td>
<td>4</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>2001 to 2005</td>
<td>365</td>
<td>80</td>
<td>20</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>2006 to 2010</td>
<td>878</td>
<td>130</td>
<td>44</td>
<td>704</td>
<td></td>
</tr>
<tr>
<td>2011 to 2012</td>
<td>154</td>
<td>12</td>
<td>105</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>SNP panel</td>
<td># of animals</td>
<td># of SNPs</td>
<td>$a_{mean}$</td>
<td>$a_{max}$</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1</td>
<td>BovineHD770K</td>
<td>243</td>
<td>771089</td>
<td>0.11 (0.046)</td>
<td>0.40 (0.17)</td>
</tr>
<tr>
<td>2</td>
<td>GGPHD77K</td>
<td>173</td>
<td>73989</td>
<td>0.10 (0.040)</td>
<td>0.20 (0.10)</td>
</tr>
<tr>
<td>3</td>
<td>BOVINE50K</td>
<td>1121</td>
<td>48740</td>
<td>0.14 (0.056)</td>
<td>0.53 (0.16)</td>
</tr>
</tbody>
</table>

Table 4.3  The number of animals, number of SNPs before imputation, the mean of average additive genetic correlation (SD), and the mean of maximum additive genetic correlation (SD) between and within different groups.
Table 4.4  Description of 1-Mb windows that explained more than 0.5% of genetic variance, and the significant SNPs within the window for all studied traits in Brangus.

<table>
<thead>
<tr>
<th>Trait</th>
<th>BTA</th>
<th>No.</th>
<th>GV%</th>
<th>WP</th>
<th>Lead SNP</th>
<th>SPPI</th>
<th>GV%</th>
<th>MAF</th>
<th>Pos (Mb)</th>
<th>p value</th>
<th>Candidate Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW</td>
<td>6.38</td>
<td>276</td>
<td>6.94</td>
<td>1.00</td>
<td>rs133837477</td>
<td>1.00</td>
<td>6.90</td>
<td>0.10</td>
<td>38.67</td>
<td>&lt;.001</td>
<td>ABCG2, IBSP, LAP3, LCORL, PKD2, SPP1, NCAPG</td>
</tr>
<tr>
<td>BW</td>
<td>5.47</td>
<td>341</td>
<td>4.38</td>
<td>0.90</td>
<td>rs134538071</td>
<td>0.13</td>
<td>0.04</td>
<td>0.37</td>
<td>47.72</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>BW</td>
<td>5.93</td>
<td>232</td>
<td>1.68</td>
<td>0.91</td>
<td>rs110253449</td>
<td>0.11</td>
<td>0.33</td>
<td>0.42</td>
<td>93.28</td>
<td>&lt;.001</td>
<td>ARRDC3</td>
</tr>
<tr>
<td>BW</td>
<td>5.5</td>
<td>310</td>
<td>0.87</td>
<td>0.58</td>
<td>rs133487364</td>
<td>0.12</td>
<td>0.34</td>
<td>0.10</td>
<td>5.05</td>
<td>&lt;.001</td>
<td>OSBP1</td>
</tr>
<tr>
<td>BW</td>
<td>5.48</td>
<td>231</td>
<td>0.83</td>
<td>0.39</td>
<td>rs136104627</td>
<td>0.04</td>
<td>0.22</td>
<td>0.37</td>
<td>48.00</td>
<td>&lt;.001</td>
<td>MSRB3, HMGA2, LEMD3</td>
</tr>
<tr>
<td>BW</td>
<td>20.8</td>
<td>294</td>
<td>0.53</td>
<td>0.48</td>
<td>rs137268783</td>
<td>0.09</td>
<td>0.16</td>
<td>0.30</td>
<td>8.97</td>
<td>&lt;.001</td>
<td>TNPO1</td>
</tr>
<tr>
<td>BF</td>
<td>9.101</td>
<td>403</td>
<td>0.70</td>
<td>0.57</td>
<td>rs133152185</td>
<td>0.02</td>
<td>0.01</td>
<td>0.48</td>
<td>101.93</td>
<td>&lt;.001</td>
<td>PDE10A</td>
</tr>
<tr>
<td>IF</td>
<td>5.48</td>
<td>231</td>
<td>4.22</td>
<td>0.87</td>
<td>rs135392994</td>
<td>0.11</td>
<td>1.15</td>
<td>0.41</td>
<td>48.21</td>
<td>&lt;.001</td>
<td>MSRB3, HMGA2, LEMD3</td>
</tr>
<tr>
<td>IF</td>
<td>6.81</td>
<td>359</td>
<td>0.98</td>
<td>0.80</td>
<td>rs132939555</td>
<td>0.0*</td>
<td>0.04</td>
<td>0.44</td>
<td>81.71</td>
<td>&lt;.001</td>
<td>TECRL</td>
</tr>
<tr>
<td>IF</td>
<td>18.54</td>
<td>296</td>
<td>0.62</td>
<td>0.55</td>
<td>rs29009652</td>
<td>0.05</td>
<td>0.07</td>
<td>0.44</td>
<td>54.47</td>
<td>&lt;.001</td>
<td>AP2S1, DACT3, NAPA</td>
</tr>
<tr>
<td>REA</td>
<td>12.88</td>
<td>350</td>
<td>2.13</td>
<td>0.74</td>
<td>rs135467227</td>
<td>0.60</td>
<td>1.85</td>
<td>0.12</td>
<td>88.31</td>
<td>&lt;.001</td>
<td>COLAA1</td>
</tr>
<tr>
<td>REA</td>
<td>7.93</td>
<td>232</td>
<td>1.87</td>
<td>0.92</td>
<td>rs135607340</td>
<td>0.11</td>
<td>0.29</td>
<td>0.45</td>
<td>93.23</td>
<td>&lt;.001</td>
<td>ARRDC3</td>
</tr>
<tr>
<td>REA</td>
<td>5.48</td>
<td>231</td>
<td>1.25</td>
<td>0.65</td>
<td>rs10530078</td>
<td>0.12</td>
<td>0.20</td>
<td>0.37</td>
<td>48.02</td>
<td>&lt;.001</td>
<td>MSRB3, HMGA2, LEMD3</td>
</tr>
<tr>
<td>REA</td>
<td>10.33</td>
<td>260</td>
<td>0.51</td>
<td>0.57</td>
<td>rs41646991</td>
<td>0.05</td>
<td>0.09</td>
<td>0.46</td>
<td>33.66</td>
<td>&lt;.001</td>
<td>SPRED1</td>
</tr>
<tr>
<td>RMP</td>
<td>19.29</td>
<td>315</td>
<td>0.73</td>
<td>0.47</td>
<td>rs137606216</td>
<td>0.04</td>
<td>0.16</td>
<td>0.39</td>
<td>29.20</td>
<td>&lt;.001</td>
<td>STX8</td>
</tr>
<tr>
<td>RMP</td>
<td>5.15</td>
<td>316</td>
<td>0.57</td>
<td>0.49</td>
<td>rs137614275</td>
<td>0.04</td>
<td>0.15</td>
<td>0.49</td>
<td>15.28</td>
<td>&lt;.001</td>
<td>RASSF9</td>
</tr>
<tr>
<td>RMP</td>
<td>9.5</td>
<td>223</td>
<td>0.57</td>
<td>0.35</td>
<td>rs42554052</td>
<td>0.10</td>
<td>0.24</td>
<td>0.37</td>
<td>5.96</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>RMP</td>
<td>19.26</td>
<td>384</td>
<td>0.56</td>
<td>0.53</td>
<td>rs109321381</td>
<td>0.02</td>
<td>0.02</td>
<td>0.31</td>
<td>26.66</td>
<td>&lt;.001</td>
<td>MIS12</td>
</tr>
<tr>
<td>RMP</td>
<td>3.27</td>
<td>391</td>
<td>0.52</td>
<td>0.54</td>
<td>rs10182820</td>
<td>0.03</td>
<td>0.06</td>
<td>0.33</td>
<td>27.69</td>
<td>&lt;.001</td>
<td>NHHL2, CASQ2</td>
</tr>
<tr>
<td>MW</td>
<td>12.88</td>
<td>350</td>
<td>3.08</td>
<td>0.80</td>
<td>rs135467227</td>
<td>0.74</td>
<td>3.03</td>
<td>0.12</td>
<td>88.31</td>
<td>&lt;.001</td>
<td>COLAA1</td>
</tr>
<tr>
<td>MW</td>
<td>10.33</td>
<td>260</td>
<td>1.96</td>
<td>0.83</td>
<td>rs137403448</td>
<td>0.08</td>
<td>0.24</td>
<td>0.26</td>
<td>33.67</td>
<td>&lt;.001</td>
<td>SPRED1</td>
</tr>
<tr>
<td>MW</td>
<td>7.93</td>
<td>232</td>
<td>1.11</td>
<td>0.77</td>
<td>rs10809597</td>
<td>0.03</td>
<td>0.04</td>
<td>0.29</td>
<td>93.39</td>
<td>&lt;.001</td>
<td>ARRDC3</td>
</tr>
<tr>
<td>MW</td>
<td>20.4</td>
<td>308</td>
<td>0.96</td>
<td>0.58</td>
<td>rs42661324</td>
<td>0.10</td>
<td>0.24</td>
<td>0.16</td>
<td>4.92</td>
<td>&lt;.001</td>
<td>ERGIC1, SH3PXD2B</td>
</tr>
<tr>
<td>MW</td>
<td>20.7</td>
<td>350</td>
<td>0.56</td>
<td>0.50</td>
<td>rs41605386</td>
<td>0.10</td>
<td>0.17</td>
<td>0.30</td>
<td>7.37</td>
<td>&lt;.001</td>
<td>UTP15</td>
</tr>
<tr>
<td>WW</td>
<td>7.93</td>
<td>232</td>
<td>2.10</td>
<td>0.94</td>
<td>0.15±0.21</td>
<td>rs136675904</td>
<td>0.10</td>
<td>0.12</td>
<td>0.42</td>
<td>93.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>6.38</td>
<td>276</td>
<td>1.58</td>
<td>0.68</td>
<td>0.16±0.21</td>
<td>rs133837477</td>
<td>0.37</td>
<td>1.16</td>
<td>0.10</td>
<td>38.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>20.4</td>
<td>308</td>
<td>1.22</td>
<td>0.70</td>
<td>0.10±0.17</td>
<td>rs42661321</td>
<td>0.14</td>
<td>0.27</td>
<td>0.16</td>
<td>4.92</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>12.88</td>
<td>350</td>
<td>0.98</td>
<td>0.54</td>
<td>0.09±0.16</td>
<td>rs135467227</td>
<td>0.40</td>
<td>0.94</td>
<td>0.10</td>
<td>88.31</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>12.82</td>
<td>381</td>
<td>0.96</td>
<td>0.67</td>
<td>0.11±0.16</td>
<td>rs133716197</td>
<td>0.22</td>
<td>0.45</td>
<td>0.11</td>
<td>82.09</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>5.48</td>
<td>231</td>
<td>0.91</td>
<td>0.57</td>
<td>0.20±0.25</td>
<td>rs133143386</td>
<td>0.21</td>
<td>0.36</td>
<td>0.32</td>
<td>48.62</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>14.26</td>
<td>303</td>
<td>0.70</td>
<td>0.64</td>
<td>0.28±0.29</td>
<td>rs42406038</td>
<td>0.02</td>
<td>0.02</td>
<td>0.23</td>
<td>26.86</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>WW</td>
<td>10.33</td>
<td>260</td>
<td>0.68</td>
<td>0.62</td>
<td>0.11±0.18</td>
<td>rs137403448</td>
<td>0.06</td>
<td>0.04</td>
<td>0.26</td>
<td>33.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>YW</td>
<td>6.38</td>
<td>276</td>
<td>2.40</td>
<td>0.79</td>
<td>0.16±0.21</td>
<td>rs133837477</td>
<td>0.61</td>
<td>1.82</td>
<td>0.10</td>
<td>38.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>YW</td>
<td>12.88</td>
<td>350</td>
<td>2.11</td>
<td>0.77</td>
<td>0.09±0.16</td>
<td>rs135467227</td>
<td>0.74</td>
<td>2.33</td>
<td>0.11</td>
<td>88.31</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>YW</td>
<td>7.93</td>
<td>232</td>
<td>1.68</td>
<td>0.91</td>
<td>0.15±0.21</td>
<td>rs110253449</td>
<td>0.06</td>
<td>0.06</td>
<td>0.42</td>
<td>93.28</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>YW</td>
<td>20.7</td>
<td>350</td>
<td>1.44</td>
<td>0.70</td>
<td>0.13±0.18</td>
<td>rs42356467</td>
<td>0.43</td>
<td>1.13</td>
<td>0.30</td>
<td>7.35</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>YW</td>
<td>10.33</td>
<td>260</td>
<td>0.82</td>
<td>0.66</td>
<td>0.11±0.18</td>
<td>rs109185879</td>
<td>0.12</td>
<td>0.20</td>
<td>0.16</td>
<td>33.81</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

\(^a\) BW, birth weight direct; BWM, birth weight maternal; BF, ultrasound fat thickness; IF, ultrasound intramuscular fat; REA, ultrasound rib eye muscle area; RMP, ultrasound rump fat thickness; MW, mature weight; WW, weaning weight; YW, yearling weight.

\(^b\) BTA\_Mb, Bovine chromosome and the \(n^{th}\) 1-Mb window on the chromosome based on the UMC assembly.

\(^c\) GV\%, proportion of the genetic variance explained by window or SNP.

\(^e\) r\(^2\), square of correlation between two adjacent SNPs within the 1-Mb window.

\(^f\) WPPA, window posterior probability of association.

\(^g\) SPPI, SNP posterior probability of inclusion.

\(^h\) MAF, minor allele frequency.

\(^i\) p value, the significance of single SNP analysis in ASREML.
Table 4.5  The GO analysis of positional candidate genes in the identified significant
1-Mb QTL window.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Lead SNP Name</th>
<th>Genotype $^b$</th>
<th>Estimate of fixed effect $^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AA (0)</td>
<td>AB (1)</td>
</tr>
<tr>
<td>BW</td>
<td>rs133837477</td>
<td>-4.74 $^A$ (2.72)</td>
<td>1.71 $^B$ (1.23)</td>
</tr>
<tr>
<td>BW</td>
<td>rs134538017</td>
<td>6.35 $^A$ (1.14)</td>
<td>4.24 $^B$ (1.09)</td>
</tr>
<tr>
<td>BW</td>
<td>rs110253449</td>
<td>6.37 $^A$ (1.14)</td>
<td>4.22 $^B$ (1.09)</td>
</tr>
<tr>
<td>BW</td>
<td>rs133487364</td>
<td>3.90 $^A$ (1.09)</td>
<td>7.91 $^B$ (1.20)</td>
</tr>
<tr>
<td>BW</td>
<td>rs136104627</td>
<td>6.06 $^A$ (1.17)</td>
<td>4.01 $^B$ (1.23)</td>
</tr>
<tr>
<td>BW</td>
<td>rs137268783</td>
<td>7.51 $^A$ (1.33)</td>
<td>5.05 $^B$ (1.12)</td>
</tr>
<tr>
<td>BF</td>
<td>rs133152185</td>
<td>0.0014 $^A$ (0.0077)</td>
<td>0.011 $^B$ (0.0073)</td>
</tr>
<tr>
<td>IF</td>
<td>rs135392994</td>
<td>-0.13 $^A$ (0.12)</td>
<td>0.11 $^B$ (0.11)</td>
</tr>
<tr>
<td>IF</td>
<td>rs132933955</td>
<td>0.23 $^A$ (0.12)</td>
<td>0.038 $^B$ (0.12)</td>
</tr>
<tr>
<td>IF</td>
<td>rs29009652</td>
<td>-0.054 $^A$ (0.12)</td>
<td>0.17 $^B$ (0.12)</td>
</tr>
<tr>
<td>REA</td>
<td>rs135467227</td>
<td>1.08 $^A$ (0.18)</td>
<td>1.31 $^B$ (0.18)</td>
</tr>
<tr>
<td>REA</td>
<td>rs135607340</td>
<td>1.44 $^A$ (0.19)</td>
<td>1.21 $^B$ (0.18)</td>
</tr>
<tr>
<td>REA</td>
<td>rs110530078</td>
<td>1.47 $^A$ (0.19)</td>
<td>1.37 $^B$ (0.19)</td>
</tr>
<tr>
<td>REA</td>
<td>rs41646991</td>
<td>1.72 $^A$ (0.20)</td>
<td>1.57 $^B$ (0.39)</td>
</tr>
<tr>
<td>RMP</td>
<td>rs137606216</td>
<td>-0.081 $^A$ (0.015)</td>
<td>-0.067 $^B$ (0.014)</td>
</tr>
<tr>
<td>RMP</td>
<td>rs137614275</td>
<td>-0.049 $^A$ (0.015)</td>
<td>-0.068 $^B$ (0.014)</td>
</tr>
<tr>
<td>RMP</td>
<td>rs42554052</td>
<td>-0.053 $^A$ (0.014)</td>
<td>-0.070 $^B$ (0.014)</td>
</tr>
<tr>
<td>RMP</td>
<td>rs109321381</td>
<td>-0.094 $^A$ (0.016)</td>
<td>-0.072 $^B$ (0.014)</td>
</tr>
</tbody>
</table>
Table 4.5 continued

| SNP     | RMP   | MW    | MW    | MW    | MW    | MW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    | WW    |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|         | rs11912820 | -0.080A (0.014) | 0.061B (0.014) | 0.017** | 0.002 |
|         | rs135467227 | 113.99A (12.52) | 105.30B (11.01) | 87.35B (11.16) | -13.32* | 4.63 |
|         | rs137403448 | 77.46A (13.06) | 94.82B (11.08) | 112.11C (11.14) | 17.32* | 0.035 |
|         | rs110809597 | 137.58A (12.48) | 98.50AB (25.84) | 91.43B (10.85) | -23.08** | -16.01 |
|         | rs42661324 | 92.02A (10.92) | 114.87AB (11.32) | 126.73B (18.47) | 17.36* | 5.5 |
|         | rs41605386 | 113.99A (12.06) | 105.30B (11.01) | 87.35B (11.16) | -13.32* | 4.63 |
|         | rs136675904 | 88.85A (6.31) | 98.22B (6.15) | 113.11C (7.15) | -6.42 | 22.96* |
|         | rs13837477 | 91.57AB (6.01) | 108.11B (6.59) | 78.74A (29.96) | -6.42 | 22.96* |
|         | rs42661321 | 84.95A (10.92) | 114.87AB (11.32) | 126.73B (18.47) | 17.36* | 5.5 |
|         | rs135467227 | 81.97A (7.10) | 90.73A (6.14) | 100.61B (6.19) | 9.32** | -0.56 |
|         | rs13716197 | 102.71A (6.42) | 108.11B (6.59) | 78.74A (29.96) | -6.42 | 22.96* |
|         | rs133143386 | 113.52A (7.31) | 92.95AB (13.76) | 90.37B (6.01) | -11.58 | -9 |
|         | rs42060383 | 99.58A (6.20) | 90.54B (6.11) | 85.15B (7.77) | -7.22 | -1.83 |
|         | rs137403448 | 119.49A (10.14) | 101.55A (6.33) | 88.75B (6.11) | -15.37** | -2.57 |
|         | rs13837477 | 158.12A (9.79) | 185.67B (10.71) | 179.04AB (49.92) | 10.46* | 17.09* |
|         | rs137467227 | 175.64A (10.47) | 158.30B (10.00) | 148.13C (10.46) | -13.76* | -3.59 |
|         | rs110253449 | 169.50A (9.92) | 150.00B (10.19) | 126.17B (14.86) | 21.67** | 2.17 |
|         | rs135467227 | 196.52A (11.44) | 174.87AB (22.62) | 155.05B (9.75) | -20.74** | -0.92 |
|         | rs109185879 | 183.52A (11.18) | 162.36B (10.07) | 146.72C (10.24) | 18.40* | 7.82 |

*a* BW, birth weight direct; BWM, birth weight maternal; BF, ultrasound fat thickness; IF, ultrasound intramuscular fat; REA, ultrasound rib eye muscle area; RMP, ultrasound rump fat thickness; MW, mature weight; WW, weaning weight; WW, yearling weight.  
*b* Different letter within SNP denote significant at P value <0.01.  
*c* *Estimates significant at P value <0.01; **estimates significant at P value <0.001.
Table 4.6  Phenotypic (above diagonal) and genetic (below diagonal) correlation coefficients among 6 traits which associated with pleiotropic QTL.

<table>
<thead>
<tr>
<th>Trait</th>
<th>BW</th>
<th>IF</th>
<th>REA</th>
<th>MW</th>
<th>WW</th>
<th>YW</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW</td>
<td>-0.21</td>
<td>0.15</td>
<td>0.39</td>
<td>0.68</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>IF</td>
<td>-0.17</td>
<td>-0.052</td>
<td>-0.19</td>
<td>-0.23</td>
<td>-0.21</td>
<td></td>
</tr>
<tr>
<td>REA</td>
<td>0.34</td>
<td>0.083</td>
<td>0.47</td>
<td>0.37</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>MW</td>
<td>0.31</td>
<td>-0.21</td>
<td>0.57</td>
<td>0.82</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>WW</td>
<td>0.49</td>
<td>-0.13</td>
<td>0.57</td>
<td>0.85</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>YW</td>
<td>0.43</td>
<td>-0.16</td>
<td>0.63</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.7  The GO analysis of positional candidate genes in the identified significant 1-Mb QTL window.

<table>
<thead>
<tr>
<th>GO</th>
<th>Term ID</th>
<th>p-value(^a)</th>
<th>Term name</th>
<th>List of genes</th>
<th>BTA Mb(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>GO:0008150</td>
<td>5.0 × 10(^{-3})</td>
<td>Biological process</td>
<td>UTP15, TNPO1, SPRED1, SPP1, SH3PD2B, PKD2,</td>
<td>5,47-48, 6,38, 7,93, 10,33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCAPG, MSRB3, LEMD3, LCOL4, LAP3, IBSP,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HMG2A, EGGIC1, COL4A1, ARRDC3, ABCG2</td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>GO:0060612</td>
<td>1.1 × 10(^{-3})</td>
<td>Adipose tissue development</td>
<td>SH3PD2B, HMG2A, ARRDC3</td>
<td>5,47-48, 7,93</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0060613</td>
<td>5.8 × 10(^{-2})</td>
<td>Fat pad development</td>
<td>HMG2A, ARRDC3</td>
<td>5,47-48, 7,93</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0051641</td>
<td>5.0 × 10(^{-2})</td>
<td>Cellular localization</td>
<td>TNPO1, SH3PD2B, PKD2, HMG2A, EGGIC1</td>
<td>5,47-48, 6,38, 20,4, 20,7-8</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0070887</td>
<td>5.0 × 10(^{-2})</td>
<td>Cellular response to</td>
<td>SPP1, PKD2, LEMD3, IBSP, HMG2A, COL4A1</td>
<td>5,47-48, 6,38, 12,88,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chemical stimulus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ke</td>
<td>KEGG:04512</td>
<td>3.8 × 10(^{-2})</td>
<td>ECM-receptor interaction</td>
<td>SPP1, IBSP, COL4A1</td>
<td>6,38, 12,88</td>
</tr>
<tr>
<td>ke(^d)</td>
<td>KEGG:04512</td>
<td>3.8 × 10(^{-3})</td>
<td>EMC-receptor interaction</td>
<td>SPP1, IBSP, COL4A1</td>
<td>6,38, 12,88</td>
</tr>
<tr>
<td>ke</td>
<td>KEGG:04510</td>
<td>5.0 × 10(^{-2})</td>
<td>Focal adhesion</td>
<td>SPP1, IBSP, COL4A1</td>
<td>6,38, 12,88</td>
</tr>
</tbody>
</table>

\(^a\) p value, Benjamini-Hochberg FDR corrected p value.

\(^b\) BTA Mb, Bovine chromosome and the nth 1-Mb window on the chromosome based on the UMC assembly.

\(^c\) BP, Biological process.

\(^d\) ke, KEGG pathway.
4.9 Figures

Figure 4.1  Proportion of genetic variance explained by 1-Mb windows across the genome for the studied traits in Brangus. Birth weight direct; Birth weight maternal; Back fat thickness; Intramuscular fat thickness; Rib eye muscle area; Rump fat thickness; Mature weight; Weaning weight; Yearling weight.
Figure 4.1 continued
Figure 4.1 continued
Figure 4.1 continued
CHAPTER 5. EFFECTS OF NUMBER OF TRAINING GENERATIONS FOR GENOMIC PREDICTION IN VARIOUS TRAITS IN A LAYER CHICKEN POPULATION

Ziqing Weng¹, Anna Wolc¹,², Xia Shen³,⁴, Rohan L. Fernando¹, Jack C.M. Dekkers¹, Jesus Arango², Petek Settar², Janet E. Fulton², Neil P. O’Sullivan², Dorian J. Garrick¹

¹Department of Animal Science, Iowa State University, Ames, IA, USA
²Hy-Line International, Dallas Center, IA, USA
³Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden
⁴MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom

A paper submitted to Genetics Selection Evolution

5.1 Abstract

Genomic estimated breeding values (GEBV) obtained using single nucleotide polymorphism (SNP) genotypes are widely implemented in animal improvement programs. It is typically assumed that more animals in the training set would result in higher prediction accuracy of GEBV. The aim of this study was to quantify the prediction accuracy from including successive ancestral generations in training set, and discover the optimal numbers of training generations for different traits in an elite layer breeding line. Data were from 16 traits on 17 793 birds with phenotypic records. All
parents and some of the selection candidates from 9 non-overlapping generations were genotyped for 23,098 segregating SNPs. An animal model with pedigree relationships (PBLUP), and BayesB were applied to predict EBV or GEBV in each validation generation (progeny of the most recent training generation) based on varying numbers of immediately preceding ancestral generations. Prediction accuracy of EBV or GEBV was assessed as the correlation between EBV and phenotypes adjusted for fixed effects, divided by square root heritability. The optimal number of training generations that produced the highest prediction accuracy of GEBV was obtained for each trait. The relationship among optimal number of training generations and heritability were investigated. On average, accuracies of BayesB models were higher than for PBLUP. Prediction accuracies of genomic prediction increased with inclusion of close ancestral generations, but reached an asymptote or slightly decreased when distant ancestral generations were used in training. The optimal number of training generations increased with heritability of traits. The optimal number of generations in training set for highly heritable traits is above 4. Whereas GEBVs of less heritable traits were best predicted by using individuals in the training/validation datasets that are closely related. The effect of adding distant ancestral generations in training on prediction accuracy differed between traits. The optimal number of training generations is associated with heritability of the trait.

5.2 Introduction

Genomic prediction is rapidly becoming the preferred way to evaluate individual genetic merit given the availability of massively parallel genotyping panels for domestic animals. Genomic selection is considered promising, as it can yield higher rates of genetic gain and lower rates of inbreeding per generation than pedigree-based best linear unbiased prediction (PBLUP) (Daetwyler et al., 2007; Sonesson and Meuwissen, 2009), which is the traditional approach for calculating estimated breeding values (EBV) from phenotype and pedigree (Henderson, 1984). Simulated and real data analyses have shown that accuracies of both genomic prediction and PBLUP can be influenced by heritability of the trait, the nature of fixed effects, and the extent of additive genetic relationships between phenotyped individuals and selection candidates (Wolc et al., 2011b). Genomic prediction
accuracies are affected by marker density (Meuwissen et al., 2001), the number of animals in the training population (Daetwyler et al., 2008; Hayes et al., 2009), size and number of quantitative trait loci (QTL) (Daetwyler et al., 2010; Kizilkaya et al., 2010), amount of linkage disequilibrium (LD)/linkage between markers and QTL (Habier et al., 2010). Collectively the latter two factors characterize the genomic architecture of the trait.

Based on simplistic theory, the more animals used in training, the greater the expected accuracy of genomic prediction (Daetwyler et al., 2008; Hayes et al., 2009). Inclusion of data from past generations will increase the size of training data set. As briefly described below, another reason for using data from all past generations is to avoid selection bias (Henderson, 1975; Im et al., 1989). Under random mating, the joint distribution between the phenotypic values and breeding values can be specified using theory of covariance between relatives. This joint distribution is used to predict breeding values from phenotypes. In population undergoing selection, this joint distribution is altered in a way that depends on the type and intensity of selection. Thus, in population undergoing selection prediction of breeding values becomes difficult. However, when inference is based on conditional distributions and conditioning is on data that includes all the information used for selection, it has been shown that the selection process can be ignored (Im et al., 1989; Fernando and Gianola, 1990; Sorensen et al., 2001). Pedigree-based additive genetic covariance between a candidate and its direct ancestor is halved by each additional generation. Thus, in PBLUP, under random mating, data from distant generations contribute little to the accuracy of prediction. In a simulated population undergoing selection, it has been shown that use of the last 2 generations compared to using the full pedigree resulted in the same response to selection (Mehrabani-Yeganeh et al., 1999). This should be examined in a real population undergoing selection. In contrast to PBLUP, in GBLUP, given high LD between markers and QTL, even distant generations are expected to contribute to prediction accuracy (Habier et al., 2007). Lourenco et al. (2014) evaluated the benefit of past generations on accuracy of GEBV using single-step GBLUP.

The impact of adding distant generations to the training data on the accuracy of genomic prediction needs to be further characterized. In Lourenco et al. (2014), results were based on single-step GBLUP, where the genomic relationship matrix was blended with the pedigree-based relationship
matrix. As mentioned previously, the pedigree-based relationship decays quickly, relative to genomic relationship matrix, which may impact performance of single-step GBLUP when distant relatives are included. Further, only one set of individuals was used for validation in Lourenco et al. (2014). Therefore, results on the relationship between prediction accuracy and pedigree depth may be confounded with environmental factors.

The objective of this study was to examine the effect of including successive generations in training on genomic prediction accuracy across different validation sets, and to assess the optimal number of training generations for routinely recoded traits. Genomic predictions were obtained from BayesB method (Meuwissen et al., 2001), and these results were compared with those from PBLUP. Data from an elite line of layer chickens were used.

5.3 Methods

5.3.1 Phenotypes and genotypes

Data contained phenotypic records from 17,793 birds from an experimental brown-egg laying population, representing 11 generations hatched between 2002 and 2011. Among those, 5108 birds (including all parents used for breeding) from the most recent 9 generations (from G3 to G11) were genotyped with a custom 40K SNP panel (Illumina, San Diego, CA). A total of 23,098 segregating SNPs across 28 chromosomes remained after removing SNPs with call rate <0.95, minor allele frequency <0.025, or Mendelian inconsistency rate between parent-offspring >0.05. The following 16 traits were analyzed: early and late albumen height (eAH, lAH, mm), shell color of the first 3 eggs (eC3, index units), weight of the first 3 eggs (eE3, g), early and late egg color (eCO, lCO, index units), early and late average egg weight (eEW, lEW, g), early and late egg production rate (ePD, lPD), early and late shell puncture score (ePS, lPS, g/s), early and late yolk weight (eYW, lYW, g), body weight (lBW, kg) and age at sexual maturity (eSM, d). Measurements of early and late traits were taken at 26-28 and 42-46 weeks, respectively, except eC3 and eE3 were measured when hens reached sexual maturity. In total, there were 136,243 phenotypic records for early traits, and 45,242 phenotypic records for late traits. The pedigree-based heritability (narrow sense $h^2$) of each trait
was estimated using all animals with single-trait animal models fitted in ASREML (Gilmour et al., 2009). Genomic information was applied in this selection program starting in 2009 (G7, generation 7), before which conventional selection based on EBV from multiple traits PBLUP was conducted. The numbers of females and males selected per generation were 360 and 120 (out of ∼2000 birds) during the period of conventional selection, but were reduced to 50 for both genders (out of ∼600 birds) during genomic selection from G7 to G11. The basic description of the phenotypic records collected is in Table 5.1.

5.3.2 Statistical model

The following three single trait models were used to predict EBV or GEBV:

1) PBLUP - single trait animal models using pedigree relationships and all available phenotype records fitted using ASREML3.0 (Gilmour et al., 2009). The model equation was:

\[ y = X\beta + Za + e, [1] \]

where \( y \) is the vector of trait phenotypes in the training set, \( \beta \) represents the vector of fixed class effects (hatch within generation), \( a \) is the vector of animal breeding values with \( Var(a) = A\sigma_a^2 \), where \( A \) is the pedigree relationship matrix and \( \sigma_a^2 \) is the additive genetic variance estimated using ASREML (Gilmour et al., 2009), \( X \) and \( Z \) are design matrices, and \( e \) is the vector of residual effects with \( Var(e) = I\sigma_e^2 \), where \( \sigma_e^2 \) is the residual variance estimated using ASREML. In the pedigree-based analyses, the relationship matrix was calculated from either the full pedigree including all animals from 11 generations, or from truncated pedigrees that used only ancestors 2 generations prior to the training set. By solving the following mixed model equation Henderson (1975), the EBV of individuals in the validation population, whose phenotypes were masked, were obtained:

\[
\begin{pmatrix}
X'X & X'Z \\
Z'X & Z'Z + A^{-1}\lambda
\end{pmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{a}
\end{bmatrix} =
\begin{bmatrix}
X'y \\
X'y
\end{bmatrix}, [2]
\]

where \( \lambda = \sigma_e^2 / \sigma_a^2 \), \( \hat{\beta} \) is the estimates of fixed class effects, and \( \hat{a} \) is the vector EBV of animals included in the pedigree.
2) BayesB (Meuwissen et al., 2001; Fernando and Garrick, 2013) analyzing only genotyped individuals that had their own phenotypic records (i.e. only females), performed using GenSel4.4 software (Garrick and Fernando, 2013; Fernando and Garrick, 2013). BayesB assumes a fraction $\pi$ of SNPs have zero effects and $1 - \pi$ of SNP effects have univariate-$t$ distribution with null mean, degrees of freedom $\nu_a$, and scale parameter $S^2_a$. This prior assumption of SNP effects is equivalent to assuming each SNP effect has a univariate normal distribution with null mean and locus specific variance (Fernando and Garrick, 2013). Each locus specific variance has a scaled inverse Chi-square prior distribution with degrees of freedom $\nu_a = 4.2$ and scale parameter $S^2_a$ derived from $\frac{\hat{\sigma}^2_a(\nu_a - 2)}{\nu_a}$, where $\hat{\sigma}^2_a$ is the variance of the additive effect for a randomly sampled locus calculated as $\frac{\hat{\sigma}^2_a}{(1-\pi)\sum_{j=1}^k 2p_j(1-p_j)}$, where $\hat{\sigma}^2_a$ is the additive-genetic variance explained by SNPs, $p_j$ is the allele frequency of SNP $j$ (Habier et al., 2011). The priors for the genetic and residual variances for each trait were obtained from the single-trait pedigree-based REML analysis. Markov chain Monte Carlo (MCMC) sampling with 55 000 iterations, of which the first 5000 were discarded for burn-in, was used to estimate the posterior means of marker effects. The convergence of MCMC samples of genetic variance, residual variance, and marker heritability were assessed using the Heidelberger and Welch’s test (Heidelberger and Welch, 1983) in R/coda package (Plummer et al., 2006). The model equation is:

$$y_{im} = \beta_m + \sum_{j=1}^k z_{ij}u_j + e_i, [3]$$

where $y_{im}$ is the phenotype for genotyped individual $i$ in the training set in hatch within generation class $m$, $\beta_m$ is the effect of hatch within generation $m$, $k$ is the number of SNP loci, $z_{ij}$ is the allelic state at SNP $j$ in genotyped individual $i$ coded as 0, 1, and 2, $u_j$ is the random effect of SNP $j$ with $u_j \sim i.i.d. 0$ with prob. $\pi \sim N(0, \sigma^2_j)$ with prob. $1 - \pi$, in which $\sigma^2_j$ is the variance of the additive effect for SNP $j$, and $e_i$ is the residual effect distributed $e_i \sim N(0, \sigma^2_e)$. The assumed value of $\pi$ was 0.95. The GEBV of individual $i$ ($GEBV_i$) in the validation population was derived as,
\[ \text{GEBV}_i = \sum_{j=1}^{k} z_{ij} \hat{u}_j, \] \[ [4] \]

where \( z_{ij} \) is the allelic state at SNP \( j \) in genotyped individual \( i \), and \( \hat{u}_j \) is the posterior mean of the substitution effect of SNP \( j \) estimated over 50000 post burn-in samples.

The effect of using different training generations, including animals with own genotypes and phenotypes (~300 per generation), was assessed from G5 to G11. The training sets consisted of animals from successive numbers of ancestral generations immediately prior to the validation generation. Table 5.2 uses an example to illustrate the assignment of validation and training sets in this study. Different validation sets (from G5 to G11) with different numbers of training generations were assessed. If only G11 was used for validation, spurious environmental effects, such as heat stress in a particular year, would be confounded with the distance between the training and validation generations, and this can lead to misleading results. Thus, different validation generations were used to avoid this confounding. The maximum number of training generations for pedigree-based and marker-based analysis was 10 and 8, respectively. The numbers of phenotypic records within each generation are in Table 5.1. Table 5.3 gives the average number of available genotyped individuals in early and late traits for each generation. Predictive performance of each model was evaluated by prediction accuracy. The prediction accuracy was determined in the validation generation based on the correlation between EBV and phenotypes adjusted for fixed effects, standardized by dividing by the square root of heritability (Legarra et al., 2008; Wolc et al., 2011a).

In order to separate the impact of size of training data set and number of training generations on prediction accuracies of GEBV, additional training scenarios were considered on studied traits (eEW was used as an example) using BayesB model 5.4. In the analysis, G10 was used as the validation set, and different numbers of genotyped animals (125 or 250) were randomly sampled from 1 to 6 training generations (G4-G9). The training scenarios differed in the total number of animals and the number of generations contributing to the training set. Some scenarios had the same size of training set, but differed in the number of generations included in training. For example, scenarios
1 and 5 had 250 genotyped animals in training, but in scenario 1 all these 250 animals were from G9, whereas in scenario 5, 125 animals were from G8 and the remaining 125 animals were from G9. Each scenario was repeated 5 times, in order to avoid sample bias.

5.3.3 Optimal number of training generations

The optimal number of training generations for each trait and method was defined as the maximum from a polynomial regression fitted to all the predictions obtained for that method for that trait. We fitted a second-order polynomial relationship between the number of training generations and the prediction accuracy in the following model:

\[ y_{i,k} = a_i k^2 + b_i k + c_i + e_{i,k}, \]  

where \( y_{i,k} \) is the prediction accuracy of GEBV obtained from BayesB for trait \( i \) with \( k \) ancestor generations included in the training set, \( a_i \) and \( b_i \) are the regression coefficients, \( c_i \) is the intercept, and \( e_{i,k} \) is the residual. The significance of regression coefficients was tested in each trait. Most of regression coefficients were significant (\( p<0.01 \)), except second-order polynomial regression coefficients in eCO, eC3, eSM, lAH, lYW, and lPS. The optimal number of training generations is \( \min\left(-\frac{b_i}{2a_i}, 8\right) \), because the dataset included at most 8 generations.

5.3.4 Marker-based heritability

Marker-based heritability \( h^2_q \) was defined as the genetic variance explained by the markers divided by the total phenotypic variance. BayesC with \( \pi = 0 \) implemented in GenSel4.4 software (Fernando and Garrick, 2013; Garrick and Fernando, 2013) was used to estimate \( h^2_q \), which assumes that all the SNPs have non-zero effects, and each SNP effect is drawn from a normal distribution with a common variance. The priors for the genetic and residual variance components were obtained from the single-trait pedigree-based REML analysis for each trait. MCMC sampling with 55 000 iterations (discarding the first 5000 as burn-in) was used to make an inference on \( h^2_q \).
5.4 Results and discussion

5.4.1 Prediction accuracy in progeny

5.4.1.1 Differences between prediction methods

Figure 5.1 shows boxplots of prediction accuracies of pedigree- and marker-based methods separately for different training generations. The bar within each box represents median of prediction accuracy. Prediction accuracies of PBLUP quickly plateaued with increases in the number of training generations. The slight fluctuation in prediction accuracies in PBLUP might be due to genetic drift. Prediction accuracies for PBLUP with a truncated pedigree (including animals in training and validating sets, and their relatives tracked back 2 generations, PBLUP,T) compared to the full pedigree (including all animals in 11 generations, PBLUP,F) across validation generations were very similar. Results indicates that using truncated or full pedigree to construct the pedigree-based relationship matrix has no significant effect on the accuracy of PBLUP in terms of ranking the current cohort of candidates, in a layer chicken population undergoing selection. Mehrabani-Yeganeh et al. (1999) reported that the use of just the last 2 generations compared to the full pedigree resulted in the same selection response in a simulated closed nucleus broiler line. Lourenco et al. (2014) also found that depth of pedigree had very small impacts on the reliability of genomic evaluations in either US dairy cattle or pig data, for both PBLUP and single-step GBLUP.

The advantage of BayesB over the pedigree-based method is obvious (Figure 5.1), because marker-based method can utilize LD, linkage, and co-segregation between markers and QTL (Habier et al., 2007). Prediction accuracy obtained from PBLUP plateaued much quickly than those from BayesB with increases in the number of training generations, because pedigree-based relationship decays faster than genomic relationship (Habier et al., 2007; Wolc et al., 2011b). In this study, all MCMC samples of genetic variance, residual variance, and marker heritability using BayesB were convergent based on Heidelberger and Welch convergence diagnostics. A fixed \( \pi \) (0.95) was used for the BayesB analyses for all traits. Although using an estimated \( \pi \) based on the Bayes C\( \pi \) method (Habier et al., 2011) may result in better prediction for some analyses, using a fixed \( \pi \) for the BayesB method will not affect comparison results. Compared to single-step GBLUP used
in (Lourenco et al., 2014), BayesB method, which utilized animals with known phenotypes and genotypes in this study, does not use pedigree-based relationship matrix, or genomic relationship matrix blended with pedigree-based relationship matrix.

### 5.4.1.2 Differences between traits and training generations

In general, for the first few training generations, prediction accuracies of PBLUP, BayesB, SNP-BLUP, and HEM increased and then plateaued, or dropped slightly when adding more distant ancestral generations (Figure 5.2). The impact on prediction accuracy of GEBV of adding ancestor generations in the training set differed among the studied traits. These differences might be caused by different trait heritabilities, genetic architecture and the number of available genotypes or phenotypes. For some traits (e.g. eAH) prediction accuracy increased with increasing numbers of training generations, while other traits slightly decreasing accuracies were estimated with increasing numbers of distant generations in training (e.g. eEW).

In this layer chicken population, data from distant generations (>4 training generations) contribution little to prediction accuracy of PBLUP. For most traits, distant ancestral generations still contribute to the accuracy of genomic prediction, but their contributions were less than those of closely related generations. Wolc et al. (2013) reduced the genomic relationships between pairs of individuals when the pedigree relationship was less than 0.45, which effectively reduced the impact of distant relatives, and showed increased prediction accuracy for egg production in laying hens using GBLUP.

To avoid confounding between environmental effect (e.g. heat stress), with distance between training and validation generations, different validation sets were used in this study. However, variation in environmental effects, distinct population structure, dissimilar genomic relationship between training and validation sets, genetic drift, or interactions between genotypes and environment would contribute the fluctuation in prediction accuracies over training generations that were observed in the study. For example, the prediction accuracy of eEW ranged from 0.39 to 0.69 when 4 generations were included in training set (Figure 5.4).
In this study, the size of the validation set, the number of generations, and the density of marker panel were limited. Further study is needed to validate the effect of adding distant ancestral generations in training on genomic prediction accuracy. A bigger population with denser SNP panel would help to avoid unexpected results and to better identify the contribution of each ancestral generation.

5.4.1.3 Size and composition of training set

Table 5.4 presents prediction accuracies on eEW for 8 scenarios that differed in the total number of training animals and the number of generations contributing to the training set. As expected, prediction accuracies increased with the size of training set (Daetwyler et al., 2008; Hayes et al., 2009). For example, when the number of training animals from the same generation increased from 125 (scenario 4) to 250 (scenario 1), prediction accuracy of GEBV in validation set (G10) increased from 0.23 to 0.46.

Though the numbers of animals in training were the same between scenarios 2 and 7, prediction accuracy in scenario 2 was higher than scenario 7 (Table 5.4). The difference was more obvious when the size of training set became larger (comparison between scenarios 3 and 9). In scenario 3, all 750 training animals were from the 3 preceding generations, whereas in scenario 9, half the animals were from more distant generations. Individuals from closely related generations can better predict GEBV of validation animals compared to those from more distant generations (Habier et al., 2007; Wolc et al., 2013).

The number of genotyped animals per generation is limited in livestock species. Though increasing the number of training generations is not equivalent to increasing the training size, inclusion of successive ancestral generations is an alternative approach to enlarge the size of training population. However, the impact of including ancestral generations in training set on genomic prediction accuracies of their descendants would vary among different traits and different population structures.
5.4.2 The relationship among optimal number of training generations, heritability, and genetic architecture

Table 5.5 presents the estimates of pedigree-based heritability and marker-based heritability of each trait. Marker-based heritabilities were smaller than pedigree-based heritabilities because markers could not explain all genetic variation.

Figure 5.3 shows the number of training generations which generated the highest predicted accuracy of GEBV in each trait using different methods. All traits were sorted by pedigree-based heritability, from lowest (IPS) to highest (eCO). The blue line in Figure 5.3 presents the linear relationship between optimal training generation and pedigree-based heritability. The correlation between the optimal training generation and pedigree-based heritability was 0.65, whereas the correlation between the optimal training generation and marker-based heritability was 0.55. Figure 5.4 shows the details of regressions of prediction accuracy on the number of training generations using BayesB in each trait. In general, the highly heritable traits had higher optimal number of training generations than the lowly heritable traits. If the LD between SNPs and QTL is low for a particular trait, co-segregation information is expected to contribute more than LD (Habier et al., 2013; Sun et al., 2014; Zeng, 2015) in predicting the QTL effect. Simulation studies have shown that when the LD between SNPs and QTL was low, prediction accuracy of a model fitting SNP genotypes, which only considers LD (LD model; i.e. BayesB and BayesC) decreased as the size of the training data set increased by adding multiple contemporary groups across families and generations (Sun et al., 2014; Zeng, 2015). However, when LD between SNPs and QTL was high, prediction accuracy of the LD model increased as size of the training data set increased, until the accuracy reached a plateau (Sun et al., 2014; Zeng, 2015). The observation in this study that the optimal number of training generations increased as pedigree-based heritability increased, could be a result of the fact that lowly heritable traits are controlled by QTL in a lower level of LD with SNPs, relative to highly heritable traits.

The estimation of optimal training generation might vary according to the assumption of statistical model, and/or the density/location of markers. For some traits, if the assumption of statistical model is inappropriate, it won’t capture the effects of QTL, even by increasing the size of training
population. A simulation study has suggested that modeling co-segregation could improve prediction accuracy using training population consisted of multiple families and generations when LD between SNPs and QTL is low (Sun et al., 2014). If the causal variant or QTL is observed, accuracy of genomic prediction for an additive model will increase by adding more distant generations in training until the prediction accuracy reaches a plateau. When QTL is not observed, with a high-density panel, LD is better preserved since markers will be closer to the QTL. Thus, it is expected that the accuracy of genomic prediction will not be reduced with high-density panels and sufficiently large datasets when distant generations are used for training.

Based on this study, the optimal number of generations in the training set for highly heritable traits is more than 4. Whereas GEBVs of lowly heritable traits are best predicted by using individuals in the training/validation datasets that are more closely related. Two strategies that might be useful in population with multi-trait selection program: varying the number of training generations for different traits or obtaining a weighted optimal number of training generations for all traits. The weight for each trait is determined on its importance for the breeding program.

### 5.5 Conclusions

The change in prediction accuracy with increases in the number of training generations depends on the trait. The optimal number of training generations in genomic prediction is influenced by heritability of a trait. Given the data used in this study, traits with lower heritability had smaller optimal number of training generations than traits with higher heritability. In practice, the decision of optimal number of training generations in a multi-trait selection population should be based on the importance of selected traits in the specific breeding scheme.

### 5.6 Acknowledgements

This study was supported by Hy-Line Int., the EW group, and Agriculture and Food Research Initiative competitive grants 2009-35205-05100 and 2010-65205-20341 from the USDA National
Institute of Food and Agriculture Animal Genome Program. XS is funded by a Swedish Research Council grant (2014-371).

5.7 Bibliography


### 5.8 Tables

<table>
<thead>
<tr>
<th>Gen</th>
<th>N</th>
<th>eCO</th>
<th>eEW</th>
<th>eC3</th>
<th>eE3</th>
<th>eSM</th>
<th>eAH</th>
<th>eYW</th>
<th>ePD</th>
<th>ePS</th>
<th>lCO</th>
<th>lEW</th>
<th>lBW</th>
<th>lAH</th>
<th>lYW</th>
<th>lPD</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>436</td>
<td>436</td>
<td>410</td>
<td>440</td>
<td>436</td>
<td>0</td>
<td>440</td>
<td>414</td>
<td>414</td>
<td>440</td>
<td>414</td>
<td>414</td>
<td>414</td>
<td>414</td>
<td>0</td>
<td>414</td>
<td>411</td>
</tr>
<tr>
<td>Mean</td>
<td>56.3</td>
<td>45.4</td>
<td>73</td>
<td>63</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>5.8</td>
<td>4.7</td>
<td>11.1</td>
<td>9.1</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>1669</td>
<td>1669</td>
<td>1667</td>
<td>1667</td>
<td>1669</td>
<td>1667</td>
<td>1667</td>
<td>1657</td>
<td>1668</td>
<td>588</td>
<td>588</td>
<td>588</td>
<td>586</td>
<td>582</td>
<td>588</td>
<td>588</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>55.2</td>
<td>45.3</td>
<td>71</td>
<td>63</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>4.7</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>2738</td>
<td>2738</td>
<td>2729</td>
<td>2729</td>
<td>2738</td>
<td>2729</td>
<td>2729</td>
<td>2738</td>
<td>2729</td>
<td>649</td>
<td>649</td>
<td>647</td>
<td>649</td>
<td>646</td>
<td>635</td>
<td>649</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>56.8</td>
<td>43.6</td>
<td>71.2</td>
<td>62.1</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>2771</td>
<td>2772</td>
<td>2752</td>
<td>2752</td>
<td>2772</td>
<td>2772</td>
<td>2772</td>
<td>2772</td>
<td>2772</td>
<td>794</td>
<td>794</td>
<td>793</td>
<td>794</td>
<td>793</td>
<td>784</td>
<td>794</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>57.5</td>
<td>46.7</td>
<td>71.1</td>
<td>62.2</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>5.1</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G5</td>
<td>290</td>
<td>290</td>
<td>278</td>
<td>278</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>782</td>
<td>782</td>
<td>781</td>
<td>782</td>
<td>781</td>
<td>778</td>
<td>782</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>58.1</td>
<td>47.3</td>
<td>74</td>
<td>63</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.5</td>
<td>4.6</td>
<td>9.1</td>
<td>4.6</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G6</td>
<td>2117</td>
<td>2117</td>
<td>2103</td>
<td>2103</td>
<td>2117</td>
<td>2117</td>
<td>2117</td>
<td>2117</td>
<td>2117</td>
<td>794</td>
<td>794</td>
<td>793</td>
<td>794</td>
<td>793</td>
<td>784</td>
<td>794</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>57.2</td>
<td>45.2</td>
<td>71</td>
<td>62.1</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G7</td>
<td>290</td>
<td>290</td>
<td>278</td>
<td>278</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>290</td>
<td>782</td>
<td>782</td>
<td>781</td>
<td>782</td>
<td>781</td>
<td>778</td>
<td>782</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>59.2</td>
<td>45.3</td>
<td>71.4</td>
<td>62.1</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.6</td>
<td>9.1</td>
<td>4.6</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G8</td>
<td>251</td>
<td>252</td>
<td>275</td>
<td>275</td>
<td>272</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>250</td>
<td>782</td>
<td>782</td>
<td>781</td>
<td>782</td>
<td>781</td>
<td>778</td>
<td>782</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>56.8</td>
<td>45.2</td>
<td>71</td>
<td>62.1</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.6</td>
<td>9.1</td>
<td>4.6</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G9</td>
<td>300</td>
<td>300</td>
<td>299</td>
<td>299</td>
<td>302</td>
<td>300</td>
<td>296</td>
<td>302</td>
<td>300</td>
<td>794</td>
<td>794</td>
<td>793</td>
<td>794</td>
<td>793</td>
<td>784</td>
<td>794</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>59.1</td>
<td>44.1</td>
<td>71.9</td>
<td>61.8</td>
<td>59.4</td>
<td>2.1</td>
<td>5.6</td>
<td>67.5</td>
<td>1383.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G10</td>
<td>724</td>
<td>724</td>
<td>828</td>
<td>828</td>
<td>850</td>
<td>708</td>
<td>850</td>
<td>850</td>
<td>850</td>
<td>835</td>
<td>835</td>
<td>835</td>
<td>835</td>
<td>835</td>
<td>832</td>
<td>832</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>56.8</td>
<td>44.7</td>
<td>139.2</td>
<td>8.6</td>
<td>14.3</td>
<td>8.1</td>
<td>1514.1</td>
<td>77.8</td>
<td>61</td>
<td>2</td>
<td>7.7</td>
<td>17.3</td>
<td>77.6</td>
<td>1403.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.7</td>
<td>4.7</td>
<td>9.1</td>
<td>4.7</td>
<td>10.1</td>
<td>0.9</td>
<td>0.9</td>
<td>9.1</td>
<td>14.3</td>
<td>30.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. early (e) and late (l) CO (egg color, index units), EW (average weight of 3-5 eggs, g), C3 (color of first 3 eggs, index units), E3 (weight of first 3 eggs, g), AH (albumen height, mm), PD (egg production rate), PS (puncture score, g/s), and YW (yolk weight, g); eSM (age at sexual maturity, d); IBW (body weight, kg).
Table 5.2  Examples of experimental design for training (T) and validation (V) sets.

<table>
<thead>
<tr>
<th>Generations</th>
<th>Generation used for validation</th>
<th># of training generations (from 1 to 8/10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T V G11</td>
<td>G11</td>
<td>1</td>
</tr>
<tr>
<td>T V G10</td>
<td>G10</td>
<td>1</td>
</tr>
<tr>
<td>T V G9</td>
<td>G9</td>
<td>1</td>
</tr>
<tr>
<td>T V G8</td>
<td>G8</td>
<td>1</td>
</tr>
<tr>
<td>T T V G11</td>
<td>G11</td>
<td>2</td>
</tr>
<tr>
<td>T T V G10</td>
<td>G10</td>
<td>2</td>
</tr>
<tr>
<td>T T V G9</td>
<td>G9</td>
<td>2</td>
</tr>
<tr>
<td>T T V G8</td>
<td>G8</td>
<td>2</td>
</tr>
<tr>
<td>T T T V G11</td>
<td>G11</td>
<td>3</td>
</tr>
<tr>
<td>T T T V G10</td>
<td>G10</td>
<td>3</td>
</tr>
<tr>
<td>T T T V G8</td>
<td>G8</td>
<td>3</td>
</tr>
<tr>
<td>T T T T V G11</td>
<td>G11</td>
<td>4</td>
</tr>
<tr>
<td>T T T T V G10</td>
<td>G10</td>
<td>4</td>
</tr>
<tr>
<td>T T T T V G9</td>
<td>G9</td>
<td>4</td>
</tr>
<tr>
<td>T T T T V G8</td>
<td>G8</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Table 5.3  Description of the average number of individuals with phenotypes and genotypes in each generation for early and late traits.

<table>
<thead>
<tr>
<th>Generations</th>
<th>Number of genotyped with own records Early/Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>0/0</td>
</tr>
<tr>
<td>G2</td>
<td>0/0</td>
</tr>
<tr>
<td>G3</td>
<td>295/295</td>
</tr>
<tr>
<td>G4</td>
<td>322/323</td>
</tr>
<tr>
<td>G5</td>
<td>295/295</td>
</tr>
<tr>
<td>G6</td>
<td>360/357</td>
</tr>
<tr>
<td>G7</td>
<td>287/278</td>
</tr>
<tr>
<td>G8</td>
<td>260/268</td>
</tr>
<tr>
<td>G9</td>
<td>300/291</td>
</tr>
<tr>
<td>G10</td>
<td>240/277</td>
</tr>
<tr>
<td>G11</td>
<td>300/290</td>
</tr>
</tbody>
</table>
Table 5.4  Prediction accuracies (SD) of different composition of training sets in eEW (early average weight of 3-5 eggs) using BayesB. In this analysis, G10 was used as the validation generation, and training individuals were randomly sampled from G4 to G9.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Distribution of training animals across generations</th>
<th># of generations in training</th>
<th># animals in training</th>
<th>Prediction accuracy (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>G9=250</td>
<td>1</td>
<td>250</td>
<td>0.46 (0.089)</td>
</tr>
<tr>
<td>2</td>
<td>G9=G8=250</td>
<td>2</td>
<td>500</td>
<td>0.60 (0.019)</td>
</tr>
<tr>
<td>3</td>
<td>G9=G8=G7=250</td>
<td>3</td>
<td>750</td>
<td>0.64 (0.017)</td>
</tr>
<tr>
<td>4</td>
<td>G9=125</td>
<td>1</td>
<td>125</td>
<td>0.23 (0.021)</td>
</tr>
<tr>
<td>5</td>
<td>G9=G8=125</td>
<td>2</td>
<td>250</td>
<td>0.45 (0.088)</td>
</tr>
<tr>
<td>6</td>
<td>G9=G8=G7=125</td>
<td>3</td>
<td>375</td>
<td>0.57 (0.038)</td>
</tr>
<tr>
<td>7</td>
<td>G9=G8=G7=G6=125</td>
<td>4</td>
<td>500</td>
<td>0.57 (0.021)</td>
</tr>
<tr>
<td>8</td>
<td>G9=G8=G7=G6=G5=125</td>
<td>5</td>
<td>625</td>
<td>0.58 (0.010)</td>
</tr>
<tr>
<td>9</td>
<td>G9=G8=G7=G6=G5=G4=125</td>
<td>6</td>
<td>750</td>
<td>0.58 (0.013)</td>
</tr>
</tbody>
</table>
Table 5.5  **Estimates of pedigree-based heritabilities and marker-based heritabilities (±SE²) for the 16 traits¹ from univariate animal models.**

<table>
<thead>
<tr>
<th>Early Traits</th>
<th>eCO</th>
<th>eEW</th>
<th>eC3</th>
<th>eE3</th>
<th>eSM</th>
<th>eAH</th>
<th>eYW</th>
<th>ePD</th>
<th>ePS</th>
</tr>
</thead>
<tbody>
<tr>
<td>pedigree-(h^2)</td>
<td>0.71±0.0017</td>
<td>0.69±0.017</td>
<td>0.65±0.018</td>
<td>0.61±0.018</td>
<td>0.54±0.018</td>
<td>0.51±0.018</td>
<td>0.46±0.019</td>
<td>0.34±0.019</td>
<td>0.21±0.015</td>
</tr>
<tr>
<td>marker-(h^2)</td>
<td>0.55±0.013</td>
<td>0.53±0.013</td>
<td>0.47±0.015</td>
<td>0.44±0.014</td>
<td>0.31±0.015</td>
<td>0.36±0.015</td>
<td>0.30±0.017</td>
<td>0.16±0.017</td>
<td>0.15±0.018</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Late Traits</th>
<th>lCO</th>
<th>lEW</th>
<th>lBW</th>
<th>lAH</th>
<th>lYW</th>
<th>lPD</th>
<th>lPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>pedigree-(h^2)</td>
<td>0.68±0.025</td>
<td>0.61±0.026</td>
<td>0.56±0.026</td>
<td>0.48±0.027</td>
<td>0.46±0.028</td>
<td>0.25±0.025</td>
<td>0.20±0.028</td>
</tr>
<tr>
<td>marker-(h^2)</td>
<td>0.58±0.019</td>
<td>0.50±0.018</td>
<td>0.48±0.020</td>
<td>0.34±0.021</td>
<td>0.37±0.020</td>
<td>0.19±0.021</td>
<td>0.10±0.024</td>
</tr>
</tbody>
</table>

¹early (e) and late (l) CO (egg color, index units), EW (average weight of 3-5 eggs, g), C3 (color of first 3 eggs, index units), E3 (weight of first 3 eggs, g), AH (albumen height, mm), PD (egg production rate), PS (puncture score, g/s), and YW (yolk weight, g); eSM (age at sexual maturity, d); lBW (body weight, kg).

²SE: standard error.
5.9 Figures

Figure 5.1 Prediction accuracies of EBV across all traits using BayesB, and PBLUP with truncated (PBLUP_T) and full pedigree (PBLUP_F).
Figure 5.2 Prediction accuracies of EBV across different validation sets using PBLUP\_T and BayesB over different number of training generations in each trait.
Figure 5.3  **Optimal number of training generations using BayesB in each trait.**

Traits were sorted by pedigree-based heritability.
Figure 5.4 Scatter plot of prediction accuracy using BayesB over training generations. The blue line is the regression of GEBV accuracy on training generations. The red line indicates the optimal number of training generations.
CHAPTER 6. GENERAL DISCUSSION AND CONCLUSIONS

6.1 General Discussion

The general idea of applying genomic selection to livestock species is to improve the genetic gain of quantitative traits of economic importance. Improving the prediction accuracy can improve the genetic gain within a population. The prediction accuracy is influenced by many factors, such as the occurrence of recombination, identification of causal mutations, size of the reference population, genetic architecture of traits, and the relationship between the reference and validation populations. This thesis investigated genome-wide recombination mechanisms, identified positional candidate genes on growth and body composition traits, and assessed the effect of distant ancestral generations on genomic prediction. Chapter 6 provides general discussion and conclusions for the research presented in Chapters 2, 3, 4, and 5.

6.1.1 Research objectives

The objectives of the research presented in this thesis were 1) to investigate the characteristics of recombination, evaluate factors influencing recombination events, and identify QTL associated with genome-wide recombination number in beef cattle and layer chickens, 2) to identify genetic variants influencing growth and body composition traits in Brangus beef cattle, 3) and to evaluate prediction accuracy when including distant ancestral generations in the training dataset for different traits in a layer chicken breeding line. In Chapter 2, recombination hotspots and quantitative trait loci (QTL) that influence genome-wide recombination were identified in Angus and Limousin cattle,
respectively. The relationships among recombination, haplotype phasing, and genotype imputation were also presented in Chapter 2. The effects of sex, marker density, family structure, haplotype structure, chromosome size, GC content, and CpG island density on genome-wide recombination were further investigated in Chapter 3 using white and brown layer chickens. In Chapter 4, Bayesian genome-wide association study (GWAS) was applied to Brangus cattle data using real and imputed 770K single nucleotide polymorphism (SNP) genotypes. Gene Ontology of pleiotropic QTL was also investigated. The potential benefit of using these QTL genotypes in a breeding scheme requires further investigation. In Chapter 5, various numbers of training generations were assessed for 16 traits in layer chickens. The effect of trait heritability was also considered in determining the optimal number of training generations.

6.1.2 Factors influencing identification of recombination

Druet and Georges (2015) found that family structure, the number of genotyped parents, the number of genotyped offspring, and marker density could influence the identification of recombination events. The effects of family structure, marker density, inbreeding level, and haplotype structure on identification of recombination events were assessed in two layer chicken lines. In this study, only the parental families with more than two offspring were retained in the analysis. The effect of family size was not significant for either white layers or brown layers. Accuracy of identification of recombination events was improved by increasing the density of markers. Higher recombination rates were observed in white layers using segregating 580K SNP genotypes than using 42K SNP genotypes. This is because denser markers provide more information, which helps to identify recombination events within more precise intervals, as well as to uncover more recombination events that are unobservable with sparse markers.

Weng et al. (2014) observed different recombination patterns and identified distinct QTL associated with genome-wide recombination numbers between Angus and Limousin beef cattle. The breed-specific manner of recombination was also observed in the two layer lines. However, the observed differences between the two lines may be a result of different levels of genomic inbreeding coefficients and distinct proportions of homozygous haplotypes across chromosomes. White layers
had significantly higher inbreeding coefficients and higher proportions of homozygous haplotype alleles than brown layers. Individuals with higher genomic inbreeding coefficients are likely to have more homozygous haplotype alleles and longer runs of homozygosity. The definition of a recombination region is the chromosomal segment of homozygous loci (between two informative heterozygous loci) that can belong to either a paternal or maternal haplotype. The existence of homozygous regions may hinder the identification of recombination events. Although different QTL for recombination number were identified between the two lines, further study is needed to determine whether the breed-specific manner of recombination was determined by genetics.

In both white and brown layer chickens, females and males had similar family structure, marker densities, inbreeding coefficients, and haplotype structures, but females had higher recombination rates and genome-wide recombination numbers than males. The same phenomena was observed for *Drosophila* (Tsai et al., 2011), mice (Petkov et al., 2007), and humans (Coop et al., 2008; Kong et al., 2014). Kong et al. (2014) identified distinct genetic variants associated with female and male genome-wide recombination rate separately in humans. The observed sex-specific recombination pattern should be controlled by genetics.

Besides the above factors, the quality of map assembly (Weng et al., 2014), the assumptions regarding the distribution of crossover events on the chromosome, and the interference between adjacent crossover events (Broman and Weber, 2000) may influence detection of genome-wide recombination events. The identification of recombination events in the genome is a specific and delicate task. Understanding the characteristics of genome-wide recombination will expand our knowledge of genetic variation across the genome as well as genetic evolution throughout generations.

Recombination events also offer an opportunity to detect map errors. Simulation studies were performed to check the relationship between map error and recombination in beef cattle. Regions with many instances of dense double crossover events, excessive numbers of crossover events, or with unusually high recombination rates, contain putative SNP location errors. For analysis of real data, regions with putative map errors are corrected (if possible) or removed. Therefore, identifying recombination events provides evidence of map errors and will eventually enable us to improve the map assembly.
Recombination events break haplotype blocks during meiosis, eroding LD between all loci on either side of the recombination spots. Conditional on the probability of observed recombination events, the accuracy of reconstructing haplotype blocks and estimating haplotype alleles could be enhanced. By knowing the locations of recombination hotspots and coldspots along the genome, it is possible to apply different genotype imputation and phasing strategies separately for these regions to increase the prediction accuracy. For example, the size of haplotype block can vary within a chromosome, based on information of the recombination spots. The length of haplotype block is shorter in recombination hotspots compared to coldspots. Different methods have been proposed to define haplotype blocks. Some define a haplotype block as a fixed length of segment (e.g. Sun et al. (2014)) or as a fixed number of SNPs (e.g. Calus et al. (2008)). Haplotype blocks can also be defined as regions with limited haplotype diversity (Wall and Pritchard, 2003). Alternative methods have focused on using pairwise LD information, and haplotype blocks were defined as regions with little or no historical recombination (Gabriel et al., 2002). Haplotype blocks defined by LD may be similar to those defined by recombination rates, both of which have their own merit. However, recombination rate integrates information of historical recombination events with recent recombination events. In some circumstances, recombination-defined haplotype blocks may be superior to LD-defined blocks.

6.1.3 Selection on recombination

Meiotic recombination generates new combinations of alleles, yields novel haplotype alleles, and contributes to genetic diversity. Recombination is thought to promote adaption, and is important for evolution (Otto and Barton, 2001). Recombination can increase the genetic variation and enhance selection response in a population under the favorable environments for the evolution of recombination (Burt, 2000). It has been found that selecting on traits unrelated to recombination leads to an increase in recombination rate in some artificial selection programs (Otto and Barton, 2001).

Genome-wide recombination rate is controlled by both genetic and environmental effects (Jefreys and May, 2004). Since several causal variants affecting genome-wide recombination rate have been identified in humans (Kong et al., 2014), it is possible to select on recombination resulting
in greater genetic variation. Mészáros et al. (2014) investigated the possibility of increasing genetic gain by selecting on recombination rate in a simulated population. In their study, a higher recombination rate reduced the loss in genetic variance and increased genetic gain, compared to a lower recombination rate. Several questions concerning selection on recombination in livestock breeding programs, such as selection intensity of recombination rate, weight of recombination rate in the selection index, and the effect of long-term selection on recombination rate, are up for debate. Correlations between recombination rates with economically important traits are not clear. Intense selection on recombination rate may lead a drop in performances for traits of interest or reduce the fitness of animals. Recombination also erodes LD between markers and QTL, which lowers the accuracy of genomic prediction. Therefore, although selecting on recombination rate presents challenges, it also has potential benefits for animal breeders.

### 6.1.4 Identification of causal QTL

The integration of SNP panels with the genotypes of causal mutations can potentially improve genomic prediction accuracy and improve genetic gain. Hayr et al. (2014) reported that including genotypes for four causal mutations for milk fat yield slightly increased the accuracy of prediction and also decreased bias. In our study, 18 significant QTL associated with growth and body composition traits in Brangus beef cattle were identified. Whether these candidate QTL can improve genomic prediction accuracy for Brangus requires further study.

Three strategies are proposed to investigate the identified candidate genes. The first requires genotypes of candidate genes and may require custom designed SNP chips. The statistical effects of candidate genes may be explored by comparing accuracy of genomic prediction in different cross validation scenarios. Scenario 1 uses genotypes of all markers from the commercial SNP panel with all markers fitted as random covariates in the model. Scenario 2 only includes genotypes of candidate genes from a custom SNP panel in the prediction. Interactions among these candidate genes may also be considered. Scenario 3 utilizes genotypes of all markers from the commercial panel and genotypes of candidate genes from the custom panel. The allelic states of candidate genes could be fitted as either fixed effects or random effects. If more than one candidate gene were
tested, interactions among these candidate genes should be considered. Other scenarios may also be considered to examine the impact of candidate genes on genomic prediction. The second strategy is to assess the biological function of candidate genes by using microarray or RNA sequencing data. Whether the candidate gene has an impact on the trait of interest can be examined by differential expression of transcripts between case animals and control animals. Case animals are defined as animals with poorer or better performance than average. The last one is to validate the effects of causal genes using molecular experiments in animals. Such experiments require the technology of gene knock-out, or gene-editing approach to perform a site-specific modifications of the sequence of candidate genes. Gene-editing is a new technology that will lead an evolution in interpreting genetic variants and modeling disease in plant, animal, human, and other organisms. CRISPR/Cas9 is an efficient gene-editing platform that can be used to introduce a specific sequence to modify genes at a specific location in live animals (Editorial, 2014). The expression of candidate genes can be altered by rewriting the sequence of its promoters or exons. Therefore, the effects of the candidate gene can be related to the change in performance, if any, after gene-editing.

Among these significant QTL, 5 were pleiotropic QTL and segregate in other cattle breeds. It is expected that the across-breed prediction accuracy can be improved by implementing genotypes from the large-effect, common QTL that segregate across breeds. Current research shows that multi-breed genomic prediction may be effective and successful when studied in populations that are closely related (Calus et al., 2014). Saatchi and Garrick (2014) showed that due to the segregation of common QTL, across-breed genomic prediction had some utility for traits when applied to other breeds. More research on the development of prediction model sand the utilization of denser markers and large-effect, common QTL are needed for across-breed genomic prediction.

6.1.5 Training generations

Increasing the size of the reference population improves prediction accuracy (Hayes et al., 2009). Including ancestral generations in the training dataset is a feasible way of enlarging the size of the reference population. However, the prediction accuracy of different traits varied when distant
ancestral generations were included in the training dataset. This was due in part to the number of training generations used, as well as heritability of the trait.

The analysis was conducted based on segregating SNP genotypes from the 42K panel in layer chickens. Results indicate that if the QTL or causal mutation genotype is observed, accuracy of genomic prediction for an additive model will increase by adding more distant generations in training until the prediction accuracy reaches a plateau. When a QTL genotype is not observed, markers in LD with the QTL will capture the effect of the QTL. However, such LD erodes over generations due to recombination between markers and the QTL, and to a greater extent with low density marker panels (Georges, 2007). With a high-density panel, LD is better preserved since markers are closer to the QTL. For some traits, a tag marker for the QTL or the causal mutation will ultimately be included in high-density panels. Thus, it is expected that the accuracy of genomic prediction will not be reduced with high-density panels and sufficiently large datasets when distant generations are used for training. Two strategies might be useful for datasets with low-density panels: truncating the data to only includes close relatives of selection candidates, or decreasing the emphasis on information from distant relatives to maximize accuracy of GEBVs (Wolc et al., 2013).

In this study, the optimal number of training generations was determined for each trait separately. Some traits required more than ten training generations, while others only require one or two to achieve the greatest prediction accuracy. Different strategies were proposed to identify the optimal number of training generations for a multi-traits selection program. One strategy is to choose the maximum number of training generations among all traits. This strategy applies one optimal training generation, but introduces unnecessary computational burdens. An alternative approach is to vary the number of training generations for different traits is another option. Lastly, another option is to select an optimal number of training generations for the majority of the traits. The latter option is used to obtain a weighted optimal number based on the importance of traits in the specific breeding program. Future study is recommended to test the feasibility of these strategies.

Only the effect of the number of training generations and the trait heritability on genomic prediction were discussed in this study. Besides factors described below, other factors, such as interactions between genotype and environment and genetic relationships between the training and
validation populations should be further investigated using simulation studies and also real data.  

1) Impact of selection. Selection can introduce the Bulmer Effect, which results in a negative correlation between pairs of loci, thereby reducing the additive genetic variance and response to selection (Bulmer, 1971). This decline in genetic variance reduces the accuracy of selection as a result of the reduction in signal to noise ratio. At the same time, selection and random drift will change allele frequencies, which may result in fixation of the QTL alleles, reduce genetic variance, and create LD over short distance, all of which lead to a reduction of accuracy.  

2) Impact of genomic selection. Dekkers et al. (2009) showed that selecting on genomic estimated breeding values (GEBV) will lead to a decay in accuracy because of the fixation of QTL alleles determined by the extent of LD between marker and QTL and the size of the QTL effect.  

3) Effect of recombination. In their simulated study, Habier et al. (2007) showed that markers can affect the accuracy of GEBV by capturing genetic relationships between genotyped animals, as well as LD between markers and QTL. Recombination erodes haplotype blocks, generating new haplotype alleles. As a consequence, recombination reduces the LD between haplotype alleles and QTL, thereby decreasing the prediction accuracy of GEBV.  

4) Model assumptions. The genomic models applied in the study only take into account the additive effect of markers. If dominance or epistatic effects are significant, or there is a significant interaction between genotype and environment, the estimation of marker effects would be biased.

6.2 Conclusions

In conclusion, 1) genome-wide recombination is regulated in a species-specific, sex-specific, and polygenic manner. Recombination has an impact on haplotype phasing and genotype imputation. Family structure and marker density can influence the identification of recombination events. The difference in recombination rates and genome-wide recombination numbers between breeds may be caused by distinct inbreeding levels, and/or haplotype structures (e.g. the number of haplotype alleles and the proportion of homozygous haplotype alleles). The sex feature of recombination is controlled by genetics. Genome-wide recombination patterns relate to chromosome size, GC
content, and CpG island density. The mapping results of genome-wide recombination numbers in two beef cattle breeds (Angus and Limousin) and two purebred layer lines (white and brown layer chickens) detected previously identified genes (e.g. SPO11, RNF212), as well as several novel candidate genes that are involved in the meiotic crossover procedure. 2) A genome-wide association study of growth and body composition traits in Brangus beef cattle identified 18 significant QTL using Bovine770K SNP genotypes, of which 7 QTL with pleiotropic effects and 11 QTL were trait-specific. The identified candidate genes will enable us better to understand the biological processes of growth and body composition traits. It is expected that the accuracy of genomic prediction can be enhanced by including large-effect causal QTL in the model. 3) The impact of including distant ancestral generations in the training dataset on prediction accuracy varied between traits. The optimal number of training generations is determined by heritability and genetic structure of the trait. Given the data used in this study, compared to traits with higher heritability, lowly heritable traits had a smaller optimal number of training generations. In a multi-trait selection population, the optimal number of training generations should be defined based on the importance

6.3 Bibliography


