농학박사학위논문

Deciphering genomic variation and effective population size using NGS and SNP data in mammals

차세대 염기서열 및 단일염기다형성 데이터를 이용한
포유류의 유전체 변이와 유효집단크기 해독

2014년 8월

서울대학교 대학원
농생명공학부 동물생명공학전공
신 동 현

Deciphering genomic variation and effective population size using NGS and SNP data in mammals

By

Donghyun Shin

Supervisor: Professor Heebal Kim, Ph.D

August, 2014

Department of Agricultural Biotechnology

Seoul National University

차세대 염기서열 및 단일염기다형성 데이터를 이용한 포유류의 유전체 변이와 유효집단크기 해독

지도교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함

2014 년 7 월

서울대학교 대학원

농생명공학부 동물생명공학전공

신 동 현

신동현의 농학박사 학위논문을 인준함

2014 년 8 월

위 원 장 이 창 규 (인)

부위원장 김 희 발 (인)

위 원 윤 숙 희 (인)

위 원 김 경 모 (인)

위 원 조 서 애 (인)

Abstract

Deciphering genomic variation and effective population size using NGS and SNP data in mammals

Donghyun Shin

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

This doctoral dissertation consists of five studies related to mammalian genetic variation and effective population size using SNP data or NGS data. Effective population size is essential to measure data size, quality and genetic diversity of animal population. I thus investigated economic trait-associated genetic variation of domesticated animal using SNP data. In addition, I examined copy number variation related to domestication process of cattle using NGS data.

In chapter 1, I introduced the basic background and necessity of the series of worked in this doctoral dissertation.

The effective population size (N_e) is important to assess the genetic diversity of animal populations. In chapter 2, I characterized more accurate linkage disequilibrium in a sample of 96 dairy cattle producing milk in Korea and estimated N_e that is approximately 122. And I inferred historical N_e and I can knew that a rapid increase N_e over the past 10 generations, and increased slowly thereafter. These results can be rationalized using current knowledge of the history of the dairy cattle

breeds producing milk in Korea. In chapter 3, I investigated the common minke whale (*Balaenoptera acutorostrata*) genome using next generation sequencing. After then, I estimated historical effective population size in the minke whale based on coalescent model to know when minke whale population size decreases rapidly. As a result, I guessed that minke whale population diversity downsized to approximately 3.1%. And strong predicted time of minke whale declination during Holocene is approximately between 194 and 902 years ago. These whole-genome sequencing offers a chance to better understand the population history of the largest aquatic mammals on earth.

After knowing population characteristic, I investigated genetic variant related to economic traits of domesticated animal. In chapter 4, I identified SNPs related to horse racing performance. Thoroughbred, a relatively recent horse breed, is best known for its use in horse racing. Although myostatin (MSTN) variants have been reported to be highly associated with horse racing performance, the trait is more likely to be polygenic in nature. I conducted a two-stage genome-wide association study to search for genetic variants associated with the EBV. I identified 28 significant SNPs related to 17 genes. Among these, six genes have a function related to myogenesis and five genes are involved in muscle maintenance. To my knowledge, these genes are newly reported for the genetic association with racing performance of Thoroughbreds. It complements a recent horse GWAS of racing performance that identified other SNPs and genes as the most significant variants. These results will help to expand my knowledge of the polygenic nature of racing performance in Thoroughbreds. In chapter 5, I identified SNPs related to milk production of dairy cattle. Holsteins are known as the world's highest-milk producing dairy cattle. I inferred each EBVs using recent ridge regression BLUP. After then, I conducted multivariate genome-wide association study to search for genetic variants associated with the EBVs for milk production traits using SNP data. I identified 128 significant

SNPs related to 47 genes. These genes were related to cellular component localization, protein localization, intracellular signaling cascade and microtubule. These genes are newly reported for the genetic association with milk production of Holstein. It complements a recent Holstein GWAS that identified other SNPs and genes as the most significant variants. These results will help to expand my knowledge of the polygenic nature of milk production in Holstein.

Finally, I detected cattle copy number variations related to domestication process, as another genetic source except SNP. Copy number variation (CNV), a source of genetic diversity in mammals, has been shown to underlie biological functions related to production traits. Notwithstanding, there have been few studies conducted on CNVs using next generation sequencing at the population level. I used NGS data containing ten Holsteins, a dairy cattle, and 22 Hanwoo, a beef cattle. The sequence data for each of the 32 animals varied from 13.58-fold to almost 20-fold coverage. I detected a total of 6,811 deleted CNVs across the analyzed individuals (average length = 2,732.2 bp) corresponding to 0.74% of the cattle genome (18.6 Mbp of variable sequence). By examining the overlap between CNV deletion regions and genes, I selected 30 genes with the highest deletion scores. These genes were found to be related to the nervous system, more specifically with nervous transmission, neuron motion, and neurogenesis. I regarded these genes as having been effected by the domestication process. Further analysis of the CNV genotyping information revealed 94 putative selected CNVs and 954 breed-specific CNVs. This study provides useful information for assessing the impact of CNVs on cattle traits using NGS data at the population level.

Key words : Effective Population Size, Estimated Breeding value, Genome Wide Association Study, Copy Number Variation

Student number : 2009-21260

Contents

List of Tables

List of Figures

Abbreviation

SNP : Single Nucleotide Polymorphism

NGS : Next Generation Sequencing

N_e : Effective Population Size

GWAS : Genome Wide Association Study

MAF : Minor Allele Frequency

BLUP : Best Linear Unbiased Prediction

rrBLUP : Ridge Regression Best Linear Unbiased Prediction

EBV : Estimated Breeding Value

LD : Linkage Disequilibrium

CNV : Copy Number Variation

QTL : Quantitative Trait Locus

PCR : Polymerase Chain Reaction

FST : Fixation index based on Wright’s F-statistics

FDR : False Discovery Rate

General Introduction

These content were largely divided into three categories (effective population size estimation, genome-wide association study, copy number variation using NGS data) related to animal genetics. And effective population size estimation and genome-wide association study categories contained two subjects.

Effective population size estimation

N_e provides information into the inbreeding level of a population and showed genetic diversity of interesting animal population. And historical N_e showed the impact of selective breeding or rapid population change. So N_e estimation is very important pre-step study to understand animal population.

Firstly, I estimated N_e of domesticated animal population using linkage disequilibrium. The recent availability of SNP genotyping allows for the application of genomic techniques to domestic animals. Genomic techniques such as genome-wide association studies and genomic selection depend on the character of linkage disequilibrium. Therefore, accurate characterization of linkage disequilibrium will assist in planning future studies using genomic techniques. Also, Linkage disequilibrium can provide insights into the evolutionary history of a population through the effective population size (N_e) (Falconer 1960). The strength of linkage disequilibrium at different genetic distances between makers can be used to infer ancestral N_e. The pattern of historical N_e in animal populations can increase my understanding of the impact of selective breeding strategies on the genetic variation within the framework of population genetics. Especially, N_e is an important measure in the global dairy cattle industry as well as the Korean dairy cattle industry in identifying the state of genetic resource. So I estimated the current N_e and inferred

an historical N_e using SNP data of Korea dairy cattle after characterization linkage disequilibrium using novel methods.

Seconds, I inferred N_e change of wild animal population based on NGS data. Cetaceans are a group of secondarily adapted marine mammals with a history of transition from terrestrial to aquatic environments. Although the origin and evolutionary history of cetaceans remains unclear, a widely accepted view is that their terrestrial ancestors returned to the seas around 50 Mya and finally diversified into a group of fully aquatic mammals. Minke whale (Balaenoptera acutorostrata) has been greatly hunted since human civilization. Especially, because whaling technology were developed rapidly in medieval time, I guessed that the minke whale population size declined rapidly after then. So I estimated two parameters: the magnitude of the population decline and the time of its occurrence.

Genome-wide association study

After knowing population character, I want to investigate genetic variants associated with economic traits of domesticated animal. So, I performed genome-wide association studies for horse and cattle. I used estimated breeding value for each traits as phenotype in association studies, because these measure can show characteristic of each economic traits considering well, taking into account several environmental factors. EBV estimation methods are largely divided into two types. One is estimation based on pedigree information and another is estimation based on genetic variant information. I used former EBV in Horse GWAS and latter EBV in cattle GWAS.

First, I performed genome-wide association study of horse racing records based on SNP data. The Thoroughbred which is best known for horse racing is a relatively recent horse breed derived from a small number of Arabian stallions and

native British mares in the 17th and 18th century England (Cunningham et al. 2001, Hill et al. 2002). Race time for each race is the most direct measure of speed and hence, makes it a suitable quantitative measure for evaluating the genetics of racing performance (Moritsu, Funakoshi and Ichikawa 1994, Oki, Sasaki and Willham 1994). In a horse breeding study, race time showed moderate heritability in the range of 0.1–0.3 (Mota, Abrahão and Oliveira 2005), with higher heritability for shorter distance race time. Previously, a study of 12,279 racehorses registered in the Korea Racing Authority, adjusted race time showed a 0.324 heritability (Park et al. 2011). However, as racing Thoroughbreds have multiple records for race time under different conditions and environmental factors, race time alone is not suitable as phenotypic value for GWAS. Because the estimated breeding value is a statistical prediction value that indicates how much each Thoroughbred has gene effects, I decided that EBV from these racing records and pedigree information was suitable for this GWAS as phenotypic value than others measures A candidate gene approach to identify genetic variants associated with racing performance in Thoroughbreds revealed a single-nucleotide polymorphism in the first intron of the equine myostatin (MSTN) gene (Hill et al. 2010c). Several genome-wide association studies have confirmed this finding that SNPs within or near the MSTN gene are strongly associated with racing performance (Hill et al. 2010c, Binns, Boehler and Lambert 2010, Tozaki et al. 2010). Although MSTN variants have been reported to be highly associated with horse racing performance, this complex trait is more likely to be polygenic in nature. In the case of human athletic performance, more than 220 genes were reported to be associated with the phenotype (Bray et al. 2009). Similarly, I speculate that other SNPs not-related to MSTN could be associated with racing performance in Thoroughbreds. So, I used the EBV of race time as the phenotype for GWAS and conducted a joint-analysis of two-stage GWAS to search for significant genetic variants associated with race time. In the first stage of GWAS, a

relatively large number of markers were evaluated in a relatively small number of samples. In the second stage, a relatively small number of markers identified as having large effects in the first stage were evaluated in a relatively large number of samples. This joint analysis of two-stage GWAS has been shown to increase the power to detect genetic association (Skol et al. 2006, Skol et al. 2007, Amos 2007).

Seconds, Holsteins are the world's highest-milk producing dairy cattle that have been continuously selected and genetically evolved into the efficient, high producing black-and-white dairy cattle, Holstein-Friesian. For last several decades, intensive application of traditional animal breeding technologies has significantly improved their milk performances throughout the world. Over the last decades, technology of molecular biology make it possible to identify genome regions or variant underlying complex traits such as milk yield in dairy cattle. Instead of traditional animal breeding program solely relying on phenotype and pedigree information, information by genetic evaluation provides a great potential to enhance selection accuracies, hence expediting the genetic improvement of animal productivity. Meanwhile, QTL mapping using linkage analysis and/or linkage disequilibrium was developed and has provided a great potential to enhance selection accuracies, hence expediting the genetic improvement of productivity in dairy cattle. Recently, the advent of genome-wide panels including hundreds of thousands of single nucleotide polymorphisms has resulted in the development of commercial SNP chips and rapid, large-scale genotyping of common SNP across large populations. These SNPs have proved powerful and useful in identification of casual mutations associated with economically important traits in livestock (Brym, Kamiński and Wójcik 2004, Amills et al. 2005, Georges 2007). At the same time, Genome-wide association studies based on high throughput SNP genotyping technologies open a broad avenue for exploring genes associated with milk production traits in dairy cattle (Jiang et al. 2010). Generally, most of economic traits

in dairy cattle are controlled by many polymorphisms, SNPs of small or large effect. Genetic variances captured by the SNP markers can be used for the calculation of direct genomic breeding values of milk production traits (Erbe et al. 2012). To find the casual variants and determine the distribution of genomic effect and contribution for milk production traits beyond previous studies, I used multivariate GWAS based on EBVs using genetic variant information. In this Multivariate GWAS, I used linear combination traits of three phenotypes related to milk production.

Copy number variation using NGS data

Finally, I investigated copy number variation associated with cattle domestication using NGS data. Since the completion of the bovine genome assembly (Elsik, Tellam and Worley 2009, Liu et al. 2009, Zimin et al. 2009), a large number of genetic variation as single-nucleotide polymorphisms, have become widely known and commercial SNP panels have been developed for cattle (Matukumalli et al. 2009). The continued discovery of SNPs in diverse cattle breeds has been further expanded (Eck et al. 2009, Stothard et al. 2011) by the recent availability of massively parallel sequencing technologies called next-generation sequencing. SNPs and the commercial SNP marker panels have been successfully used to identify genomic regions that potentially underlie the economic traits of cattle (Barendse et al. 2009a, Gibbs et al. 2009, Hayes et al. 2009b). Another source of genetic variation is mammals come from gains and losses of genomic structural sequence variants, copy number variations (CNVs), that occur in more than two individuals (Mills et al. 2011). While SNPs are more frequently used in cattle breeding than CNVs, CNVs occupy a higher percentage of genomic sequence than SNPs. It is possible that CNVs have a potentially greater effect on phenotype, including changing of gene structure and dosage, altering gene regulation and exposing recessive alleles (Zhang et al. 2009). These points are attracting attention to CNV as structural variation that can

account for diverse economically important traits in domestic animals. In particular, the CNV type, deletions, which is the focus of this study has been shown to be one of the five CNV types and one of the two main classes with duplications (Redon et al. 2006). Previous study of cattle using next generation sequencing (NGS) data has reported that CNVs play a crucial role in diverse biological functions as pathogen- and parasite-resistance, lipid transport and metabolism, breed-specific differences in adaptation, health, and production traits (Bickhart et al. 2012). For example, partial deletion of the bovine gene ED1 causes anhidrotic ectodermal dysplasia (Drögemüller et al. 2001). Beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. It is possible that CNVs may be associated with these agriculturally important traits (Bickhart et al. 2012). Until now, CNV screens were routinely performed by comparative genomic hybridization (CGH) and SNP arrays, and many studies have extensively reviewed their performances (Lai et al. 2005, LaFramboise 2009, Winchester et al. 2009, Pinto et al. 2011). However, these methods, which are often affected by low probe density and cross-hybridization of repetitive sequence, were not able to detect CNVs at the whole genome level. A limited number of investigations in cattle CNV has been performed to detect CNVs using methods that include high-density aCGH and the 50K SNP panel (Bae et al. 2010a, Fadista et al. 2010, Liu et al. 2010). The recent advances of NGS and complementary analysis programs have provided better approaches to systematically identify CNVs at a deep genome-wide level than the currently available commercial SNP chip and aCGH methodologies (Stothard et al. 2011, Alkan et al. 2011). These sequence-based approaches, which are becoming more popular due to the ongoing developments and cost decreases in NGS data, allow for CNV reconstruction at a higher effective resolution and sensitivity. In this study, I attempt to detect genome-wide CNVs at the population level based on NGS data of 32 cattle. Using UMD3.1 (Zimin et al. 2009)

as a reference genome, I used Genome STRiP to detect cattle CNVs at the population level using Hanwoo (22 individuals), a Korea beef cattle, and Holstein (10 individuals), a dairy cattle. This study confirmed that CNVs are common, associated with deleted regions, and often occur in gene-rich regions in cattle. I analyzed genes related to CNVs using deletion score in order to explore their potential function and contributions in domestication. In addition, I investigated the selected CNVs using FST and breed-specific CNVs for traits related to beef and milk production. By providing several types of information on cattle CNV at the population level and presenting deleted CNV maps with breed-specific CNVs, I provide the basis for further studies into the role of deleted CNVs in the cattle genome.

Chapter 1. Literature Review

1.1 Effective Population Size

1.1.1 Overview of Effective Population Size

In population genetics, effective population size (N_e) was defined as the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration. As same means, an effective population size defined as the number of individuals in an idealized population that has a value of any given population genetic quantity that is equal to the value of that quantity in the population interest.

Population size is very important value in population genetics, but unfortunately it is difficult or impossible to determine the number of individual or gametes which contribute to the next generation. So I need another way to define the size of population. The definition of the population size in population genetics relies on the dynamics of genetic variation in the population. To make a distinction between the dynamics of genetic variation in a population and the number of individual, scientist suggested really two types of population size. One is the count of individual in a population (N, census population size) and the other is the genetic size of a population (N_e, effective population size) that was determined by comparing the rate of genetic drift in an actual population with the rate of genetic drift in an ideal population meeting the assumptions of the Wright-Fisher model. A simple way to think of the difference between the two population size concepts is that the total number of individuals is census population size and the number of individuals that actually contribute to next generation is effective population size.

In consideration of population size and genetic drift, there are several effects that can affect effective population size. Founder effect related to population

starting point and genetic bottleneck linked to population fluctuation are representative effects associated with effective population size. Founder effect is the establishment of a population by small number of individual which result in small effective population size in a new founded population. Genetic bottleneck is sharp but often transient reduction in the size of a population that increases allele frequency sampling error and has an imbalanced impact on the effective population size in later generations. Moreover, there are other aspects of biological population that have the same impact by increasing the sampling error in allele frequency across generations. Mating patterns can have a considerable impact on effective population size when difference sexes in one population make unequal contribution to next generation by reproduction. Also one of other factors affected effective population size is degree to which adult individuals in the population contribute to the next generation. Because one of the assumption of Wright-Fisher populations is that all individuals contributes an equal number of gametes to the infinite gamete pool. And family size affects effective population size, so variance of family sizes used for estimation of population reduction trace.

I have to consider inbreeding in estimation effective population size as well as genetic drift. In large populations with random mating, probability of biparental inbreeding is very low. But, in small population the chance of mating with a relative is larger, because the number of possible mates is limited. Genetic drift also occurs due to finite population size. Therefore genetic drift and tendency of inbreeding are interrelated phenomena connected to population size and both effects increases the homozygosity in a population over time. One of main conclusions that can be drawn from the interrelationship between autozygosity (by genetic drift and inbreeding in finite population) and the effective population size is that genetic drift make population to become more inbred even when mating is random. Because genetic drift arouse fixation or loss of alleles, heterozygosity in a population decreases.

In domesticated animal breeding system, to fast increase economic trait of animal population I has used small number of outstanding male and many female. So, because of high inbreeding coefficient, effective population size of domesticate animal population mostly is much smaller than that of wild animal population. If effective population size is too small, the population can be exposed to severe risk as contagious disease. Because effective population size can inform genetic diversity of specific population, breeder of domesticated animal and scientist who interested endangered species used effective population size as genetic diversity measurement to protect animal population.

1.1.2 Basic Methods of Effective Population Size Estimation

There are several ways to define the effective population size, because there are several models constructed on different foundations used to explain how genetic variation changes in finite population. During inferring effective population size, I have to consider two important concepts, inbreeding and variance effective population size. Inbreeding effective population size (N_{ie}) is estimation of ideal population that has the sample probability of allele copies being identical by descent (IBD) as an actual population. The chances of sampling two copies of the same allele depend on the size of the population ($1/ 2N_{ie}$). So, I have probability that two alleles are identical by descent, as follows:

$$P(IBD) = \frac{1}{2Ne_{inbreeding}} \qquad (1.1)$$

And I can easily restate as:

$$Ne_{inbreeding} = \frac{1}{2P(IBD)} \qquad (1.2)$$

As shown upper equation, the chance of identical by descent for two allele is low in big population, whereas the chance of identical by descent is high in small population. As such, when the effective population size is defined by reference to autozygosity or in breeding, the value is inbreeding effective population size. Genetic drift causes the change in allele frequencies in many replication populations over generations. The range of change in allele frequency could be expressed as a variance. The variance of the change in allele frequencies from one generation to the next ($\Delta p = p_{t-1} - p_t$) is:

$$Variance(\Delta p) = \frac{p * q}{2Ne_{variance}} \qquad (1.3)$$

I can restate by solving for the effective population size.

$$Ne_{variance} = \frac{p * q}{2Variance(\Delta p)} \qquad (1.4)$$

Here, I can define how big the effective size is by quantifying the variance in the change in allele frequency. This definition provides the variance effective population size (N_{ve}). Additionally I need to think about breeding effective population size. In species, where individuals are more or less continuous distributed over large areas without obvious physical and geographic boundaries that define populations. Instead, populations can be defined by average mating and dispersal patterns among individuals that result in limits to the movement of gametes each generation. Based on the size of the breeding and dispersal area, there is the breeding effective population size, which is suitable for population where individuals may occur relatively uniformly and not form discrete aggregation. Breeding effective population size (N_{be}) was the number of individuals found in a genetic neighborhood defined by the variance in gamete dispersal. Breeding effective population size

depends on the probability distribution of gamete dispersal in space. In several method of estimating effective population size, I can chose method in consideration of data type and species to infer accurate effective population size.

1.1.3 Estimation Effective Population Size of Domesticated Animal Using SNP Chip Data

I can apply the recent genomic techniques as SNP Chip to study of domestic animals. Also genome analysis as genomic selection using large sale genomic information depends on the extent and quality linkage disequilibrium that describes the non-random association of alleles at different loci and can result from processes such as migration, selection and genetic drift in finite populations. Effective population size is closely related to its rate of decline. I can estimate available linkage disequilibrium of my interesting population using SNP chip data to estimate effective population size (Hayes et al. 2003). Linkage disequilibrium can provide insights into the evolutionary history of a population. To avoid bias result of genome analysis, I have to use suitable population which meet a certain level of genetic diversity. If I have complete pedigree informations of the population, I can easily know genetic diversity. However, I have incomplete pedigree information of population in most case. So I have to identify genetic diversity of population in a roundabout way through effective population size based on linkage disequilibrium. Additionally, if I use the method that estimates effective population size based on linkage disequilibrium, I can infer historical effective population size variation according to time series. The pattern of historical effective population size of domesticated animal can increase my understanding of the impact of selective breeding strategies on the genetic variation within the framework of population genetics.

In population genetics, there are equation of homozygosity probability using identical by state with population size and mutation rate, as follows:

$$Probability\ homoaygosity(G) \cong \frac{1}{4N\mu + 1} \qquad (1.5)$$

Under the assumption of an isolated population, Sved (Sved 1971) derived an approximate expression for the expectation of linkage disequilibrium measurement (r2). Linkage disequilibrium measurement, r^2 was one of correlation coefficients between pairs of loci, expressed as:

$$r^2 = \frac{D^2}{P_A P_B P_a P_b} \qquad (1.6)$$

Where, P_A, P_a, P_B, P_b are the respective frequencies of alleles A, a, B, b and D is P_{AB}-$P_A P_B$. And modified equation for effective population size using linkage disequilibrium is:

$$E(r^2) = \frac{1}{1 + 4Nc} \qquad (1.7)$$

Where, N is the effective population size and c is the recombination frequency. In estimation effective population size, usually c was replaced by linkage disequilibrium. This was justified by the approximation of the more precise equation for linkage dis equilibrium by Sved (Sved 1971). Based on this equation, I can predict the effective population size at a given point in time, expressed as generations in the past (Hayes et al. 2003) using these equation:

$$N_T(t) = \frac{1}{4c} * \left(\frac{1}{(r^2{}_c)^{-1}} - 1\right) \qquad (1.8)$$

Where N_T is the effective population size t generation ago, c is the distance between markers in Morgans, r^2_c is the mean value of r^2 for marker pairs c Morgans apart, and $c=(2t)^{-1}$ when assuming linear growth (Hayes et al. 2003). From the decay in linkage disequilibrium, I can infer a constant actual present effective population size and evolutionary story related to population size variation.

1.1.4 Effective Population Size Estimation Using NGS Data

I used coalescent model to estimate historical story of effective population size, if genomic data was composed of small size sample. Coalescent theory is a retrospective model of population genetics by tracing allele shared in population to most recent common ancestor (MRCA). So, I can make model of ancestor-descendant relationships (genealogy) that allow us to predict identity by descent in the past based only on knowledge of the present without pedigree information. With this model, I can look at patterns among the individuals in the present and try to reconstruct versions of events such as inbreeding, gene flow or natural selection in the past. These models are referred as coalescent model. Using rules of random sampling based around the Wright-Fisher model, I can develop a prediction for the number of generations back in time until two lineages find MRCA (coalescence to a single lineage). The probability that two lineages coalesce in the immediately preceding generation is the probability that they share a parental DNA sequence. If data is from diploid population with a constant effective population size (N_e), there are 2 N_e potential parents in the previous generation. The probability that two alleles share a parent is 1/2 N_e, therefore the probability of not coalesce is 1-1/2 N_e. Based on geometric distribution, the probability that coalescent event at t generation (non-coalescence at the t-1 preceding generation) is:

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} * \left(\frac{1}{2N_e}\right) \qquad (1.9)$$

In practice, the probability of coalescence are approximated using an exponential function. Therefore the cumulative probability of a pair of lineages coalescing at or before generation t is:

$$P(T_c \leq t) = 1 - e^{-\frac{1}{2Ne}t} \qquad (1.10)$$

Where Tc is the generation of coalescence and t is the maximum time to coalescence. Based on this equation, I can calculate height of a coalescent tree, so I can identify waiting time that is the mean or expected time back in the past until a single coalescence event in a sample of lineage.

As in the models of genetic drift, effective population size also plays a crucial role in coalescent model. The effective population size determines the chance that two alleles copes descend from the same ancestor when working back in time from the present to the past. In the coalescent model, two randomly sampled alleles have the probability 1/2 N_e of finding their MRCA in the previous generation, as follows:

$$P_c = \frac{1}{2N_e} \qquad (1.11)$$

This equation can be rearranged:

$$Ne = \frac{1}{2P_c} \qquad (1.12)$$

The probability that two randomly sampled alleles copies do not coalescence model, P_{nc}, over some number of generations can be used to show the overall effective population size when the population size is not constant. As the assumption that

modeling coalescent times follows exponential distribution (the continuous time coalescence time), e (-t/2 N_e) is used to approximate 1-1/2 N_e as long as N_e does not get too small. This means that Pnc can be approximated by an exponential function:

$$P_{nc} = \left(1 - \frac{1}{2N_e}\right)^t \approx e^{-\frac{t}{2N_e}} \qquad (1.13)$$

Where t is the number of generation. If population fluctuates in size over three generations and no experiencing a coalescence event over the three generation could be approximated by:

$$P_{nc} \approx e^{-\frac{t}{2N_e}} = e^{\left(-\frac{1}{2N_e(t=1)}\right)\left(-\frac{1}{2N_e(t=2)}\right)\left(-\frac{1}{2N_e(t=3)}\right)} \qquad (1.14)$$

The exponential terms of e can be solved for the effective population size by taking the natural log of both sizes to eliminate e:

$$-\frac{t}{2N_e} = \left(-\frac{1}{2N_e(t=1)}\right) + \left(-\frac{1}{2N_e(t=2)}\right) + \left(-\frac{1}{2N_e(t=3)}\right) \qquad (1.15)$$

And then multiplying both sides by 1/t and then -2 to get:

$$\frac{1}{N_e} = \frac{1}{t}\left(\frac{1}{N_e(t=1)} + \frac{1}{N_e(t=2)} + \frac{1}{N_e(t=3)}\right) \qquad (1.16)$$

The term on the right side is the harmonic means of the effective size. And this Equation also can be induced is from identity by descent.

Interesting subject related to coalescent model in animal genetics using wild animals in crisis of extinction as minke whale is tracing growing or shrinking populations. In a population growing in size over time, the probability of a coalescent event is least near the present because the population is at its largest size. When the size of the population is continually shrinking, the probability of a coalescence event (1/2 N_e) must also be continually increasing as I move back in time toward the MRCA, since N is shrinking. A common way to model growing or shrinking

population is to assume that population size is changing exponentially over time. Under exponential growth, the population size at time t in the past is a function of the initial population size in the present N0 and the rate of population growth r according to:

$$N(t) = N_0 * e^{-rt} \qquad (1.17)$$

The generalizations above regarding coalescent waiting times depend on rapid and sustained changes in population size over time such as under exponential population growth with a constant rate (r). Increases in population size over time cause the probability of coalescence to decrease toward the present. At the same time, the chance of a coalescence event increases toward the present simply due to a larger number of lineages available to coalesce. With these equations and assumptions, I can infer demographic parameter as growth rate from genetic data, based on diffusion approximation.

1.2 Genome-wide Association Study

1.2.1 Overview of Genome-wide Association Study

In genetics, genome-wide association study (GWAS, or Whole genome association study) is an examination of many common genetic variant in different individuals to investigate which variant is associated with an interesting trait. GWAS mainly focus on associations between genetic variant as single-nucleotide polymorphisms (SNPs) or Copy Number Variations (CNVs) and trait as major disease or economic traits of domesticated animal. Genome-wide data sets are used to identify biological pathways and network underlying complex diseases and in drug development process as well as associated gene.

In medicine, these studies mainly compare the DNA of two groups of participants (e.g. people with the disease and normal people without disease). Genetic variants are read using chip or other sequencing method, from sample of each people. If one type of the variant is more frequent in patients, the SNP is said to be associated with the disease. The associated SNPs are then considered to mark a region of the genome which influences the risk of disease. Also in animal science, these studies normally compare the DNA of domesticated animals with economic trait (e.g. meat quality, milk quantity, racing record). Based on linear model to identify significant SNP associated with continuous data, if one type of SNPs is fitting highly, the SNP is said to be associated with the economic trait. Unlike methods which specifically test one or few genetic regions, the GWAS investigate the entire genome. This approach is non-candidate-driven study, not gene-specific candidate-driven study. So GWAS can identify genetic variant in DNA associated with interesting trait, but cannot their own specify which genes are causal without additional validation process.

1.2.2 GWAS Method in Population Data

The most common GWAS is the case-control study which compares two groups of individuals, one healthy control group and one case patients group. All individuals in each group are genotyped for the majority of common known SNPs. If the allele frequency is significantly altered between the case and the control group, significant SNPs will be investigated. In such setups, the fundamental unit of each SNPs is the odds ratio which reports the ratio between two proportions. When the allele frequency in the case group is much higher than in the control group, the odds ratio will be higher than 1, and vice versa for lower allele frequency. Additionally, a p-value for the significance of the odds ratio is typically calculated using a simple chi-squared test. Finding SNPs that have significant odds ratios is the objective of the GWAS, because this shows that a SNP is associated with disease. A common alternative to case-control GWA studies is the analysis with quantitative phenotypic data (e.g. height). In this study, linear model mainly are used to investigate significant SNP related to quantitative trait. Calculations are typically done using bioinformatics software such as PLINK (Purcell et al. 2007), which also include tools for many of these alternative statistics.

A key point in GWAS is the imputation of genotypes at SNPs that are not on the genotype chip (Marchini and Howie 2010). This process increases the number of SNPs that can be tested for association test, increases the power of the analysis and facilitates meta-analysis of GWAS across distinct cohorts. Genotype imputation is performed by statistical methods that combine the GWAS data together with a reference panel of haplotypes. Popular software packages for genotype imputation are IMPUTE2 (Howie, Marchini and Stephens 2011) and MaCH (Li et al. 2010b).

Also genotype imputation without reference panel is possible in case of BEAGLE (Browning and Browning 2011).

In addition to the calculation of association, it is common to take into account variables that could confound the results. Basically, sex and age are common confounding variables. Additionally, it is known that many genetic variations are closely associated with the geographical and historical populations (Novembre et al. 2008). Because of this association, studies must take account of several variables related to the geographical and ethnical background by controlling for population stratification. Especially there is spurious association in GWAS approaches of domesticated animal population. To solve this problem, there are several methods for structured population (Pritchard and Rosenberg 1999). The first approach is transmission disequilibrium test (TDT) which tests for the presence of genetic linkage between a genetic marker and a trait using within-family (Spielman, McGinnis and Ewens 1993). Also QTDT (quantitative TDT) and the family-based association tests extended this within-family comparisons with quantitative traits (Laird and Lange 2008, Abecasis, Cardon and Cookson 2000). These methods are robust to population stratifications. And to correct test statistics from structured population, several methods were proposed. Mainly “genomic control” (GC) and the model including structural effect were widely used (Devlin, Bacanu and Roeder 2004, Price et al. 2006) due to its power and computational efficiency, linear mixed model is widely used. The mixed model has been mainly used in animal breeding (Quaas and Pollak 1980) and recently have received attention in human genetics (Price et al. 2010). In this mixed models, polygenic and residual variances should be estimated separately before the each SNP test. So, association test between only genetic value (like breeding value) of phenotype and SNP is possible. GRAMMAR method analyzes the association between these residuals adjusted for family effect and SNPs using rapid least-squares methods (Amin, van Duijn and Aulchenko 2007).

GRAMMAR-GC (Price et al. 2010) and GRAMMAR-Gamma (Svishcheva et al. 2012) extended the method of GRAMMAR. Efficient Mixed-Model Association eXpedited (EMMAX) efficiently test each SNP by using heritability estimated from the null model fixed instead of repeatedly estimating variance components and can correct stratification and population structure (Kang et al. 2010). Additionally methods based on principal component analysis (PCA) have been developed as EIGENSTRAT (Zeggini et al. 2008) and PCA based in logistic regression (Lee et al. 2010).

After odds ratios and p-values calculation for all SNPs, a common next step is to create a Manhattan plot. In the GWAS, this plot shows the negative logarithm of the p-value as a function of genomic location. The SNPs with the most significant association will stand out on the plot. The p-value threshold for significance is corrected for multiple testing issues. The exact threshold varies (Wittkowski et al. 2014) but usually p-values must be very low ($10^{-7 \text{ or } -8}$) to be considered significant in the face of the millions of tested SNPs. GWAS typically perform the first analysis in a discovery cohort, followed by validation of the most significant SNPs in an independent validation cohort.

1.2.3 Regularization Approach in GWAS

In recent biology, big and high-dimensional data have emerged with the rapid development of sequencing technology. Recent GWAS usually uses over 500K SNP chip in human or 60K SNP chip in pig, or 50K SNP chip in horse. Also there were about 133,000 individuals in GIANT consortium height association study (Allen et al. 2010), but sample size is still much less than number of SNPs. Many studies have performed association analysis of SNPs with diseases or quantitative traits through mainly single-SNP analysis. Additionally, several quantitative traits

that have a strong correlation with each other, multivariate analysis was performed to identify a wide range of significant SNPs. Occasionally, these analyses often identifies only a few significant SNPs, so these results provide limited information. To overcome this limitation, Peng (2009) suggested second-wave analysis as gene based, pathway based, network based and haplotype based GWAS (Peng et al. 2010a, Akula et al. 2011).

Some scientists proposed another approach using multiple regression methods to treat high-dimensional data. The large number of genetic variant as SNP and small sample size can make ill-defined result of multiple regression methods. Also the process that is analysis of several hundreds of thousands of SNPs at the same time is computationally very heavy. In order to solve these statistical and computational challenges for model fitting and variable selection of these multiple regressions, recent shrinkage methods such as the LASSO (Tibshirani 1996), ridge regression (Hoerl and Kennard 1970), elastic-net (Zou and Hastie 2005) and SCAD (Fan and Li 2001) usually are used.

1.2.4 Using Estimated Breeding Value in Animal GWAS

Estimated breeding value (EBV) is an estimate of true breeding value of an individual for a trait based on the performance of the individual or close relatives. Originally breeding value can be defined as the value of an individual as parent. Parents can transfer a random sample of their genes to their offspring. Breeding value gives an estimate of the transmitting ability of the parent. In practical animal breeding, selection is not only based on phenotype but also on estimates of breeding value that are derived from records on the animal itself as well as relatives using Best Linear Unbiased Prediction (BLUP) for animal model (Lynch and Walsh 1998). An

important property of EBV from an animal model is that many and high quality records are available on the individual and its relatives. Also non-genetic factors (e.g. herd-year-season, age, sex) that contribute to total variation observed in a population have to be adjusted to maximize the accuracy of EBV.

The common methods of animal model is the use of linear regression for the prediction of EBV from phenotypic records. Originally, BLUP is used in linear mixed models for the estimation of random effects and is similar to Best Linear Unbiased Estimates (BLUE) of fixed effects in statistics. In the field of animal breeding, phenotypes is conceptually composed of two parts: a genetic component and a non-genetic component. The genetic component was passed down to each individual by its parents. The non-genetic component includes the environment in which each individual was raised and all other aspects of each individual's own existence. In many cases, the non-genetic component can be further decomposed into parts attributable to "fixed" effects (e.g. sex, year of birth, management conditions, and age when measured). Then, an observation on a single animal can be expressed as:

$$\text{phenotypes} = \text{Environmental effect} + \text{Genetic effect} + \text{Residual}$$

In terms of a statistical model, this can be expressed as:

$$y_{ij} = \mu_i + g_i + e_{ij} \qquad (1.18)$$

Where,

y_{ij} = the j'th record observed for the i'th animal

u_i = the non-random (fixed) environmental effects

g_i = the sum of additive, dominance and epistatic genetic values of the i'th animal

e_{ij} = the sum of random environmental effects from the j'th record of the i'th animal
Now, by substituting ga(additive), gd(dominance) and ge(epistatic) for gi, the statistical model becomes:

$$y_{ij} = \mu_i + (g_a)_i + (g_d)_i + (g_e)_i + e_{ij} = \mu_i + (g_a)_i + e^*_{ij} \qquad (1.19)$$

Where, y_{ij} and μ_i are defined as above, but e^*_{ij} is now equal to $(g_d)_i + (g_e)_i + e_{ij}$. Modified equation is the basic equation used to calculate EBV. An implicit assumption in applying equation is that μ_i is known without error. But this is almost never true in fact, so the data in hand are often used to simultaneously estimate μ_i and $(g_a)_i$. The method used to estimate both μ_i and $(g_a)_i$ simultaneously from the same data is BLUP. In BLUP process, EBV are obtained by a process that simultaneously adjusts for differences attributable to fixed effects and accounts for all genetic relationships through the pedigree of the population. And In matrix notation, the basic linear model can be written as:

$$Y = X\beta + Z\mu + e \qquad (1.20)$$

Where,

Y = the vector of observed values,

β = the vector of fixed effects,

u = the vector of random genetic effects,

e = the vector of residual errors

X and Z are design matrices that relate elements in β and u, respectively, to elements in Y. Leaving out many intermediate steps in the derivation, it is stated without complicated proof that the mixed model equations for estimating breeding values for a single trait or characteristic are:

$$\begin{bmatrix} X'X & XZ \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix} \begin{bmatrix} \beta \\ \mu \end{bmatrix} = \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix} \qquad (1.21)$$

Where,

A^{-1} = matrix inverse of A, the numerator relationship matrix,

$\alpha = (1 - h^2) / h^2$, where h^2 is heritability of the trait under analysis.

The powerful ability of BLUP to obtain estimates of breeding values using all genetic relationships is embodied in the numerator relationship matrix. In estimation of traditional breeding value, numerator relationship matrix are constructed from pedigree information. Recently, numerator relationship matrix can be replaced with genetic relationship matrix that are constructed from relationship information based on generic variant information as SNP. Instead of estimation of each individual, animal model can estimate each SNP effect. After that, I can calculated EBV that is sum of each SNP effect of each individuals. This BLUPs which is based on genetic variant is said to GBLUP.

The reason why EBV is used in GWAS is that individuals who originally has not phenotype can have phenotypic information. This is important point, because a certain economic traits of domesticated animal are related to sex (e.g. milk). Bull cannot make milk but is very important due to supply sperm in milk industry. But, I can estimate Bull's phenotype related to milk and evaluate their capacity to perform GWAS or set a price of each individual or semen. Already, several GWA studies using EBV were performed in animal genetics. Emma K. Finlay (2012) performed genome wide analysis of the association between EBV of the tuberculosis susceptibility and each SNP in Holstein-Friesian dairy cattle (Finlay et al. 2012). Heidi Signer-Hasler (2012) analyzed the association of the genotype data of 1,077 horse data with EBV for height (Signer-Hasler et al. 2012). In sheep, Dag I Vage (2013) performed GWAS using EBV for litter size (Våge et al. 2013).

1.3 Copy Number Variation Using Next Generation Sequencing

1.3.1 Next Generation Sequencing

Nucleic acid sequencing is a method for determining the exact order of nucleotides present in DNA or RNA. In the past decade, the use of nucleic acid sequencing has increased exponentially as sequencing technology has become accessible to research and clinical labs. The first major try into DNA sequencing was the Human Genome Project, a $3 billion, 13-year-long endeavor, completed in 2003. The Human Genome Project was accomplished with first-generation sequencing, known as Sanger sequencing. Sanger sequencing developed in 1975 was considered the gold standard for sequencing for the subsequent two and a half decades (Sanger, Nicklen and Coulson 1977).

Since completion of the first human genome sequence, demand for cheaper and faster sequencing methods has increased greatly. Because of this demand, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis over the past four years. Recent sequencing technologies have been directed towards the development of new methods, leaving Sanger sequencing. Since first-generation technology, newer methods are referred to as next-generation sequencing (NGS) composed of various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced.

Template preparation consists of building a library of nucleic acids and amplifying that library. Sequencing libraries are constructed by fragmenting the DNA (or cDNA) sample and ligating adapter sequences onto the ends of the DNA fragments. Once constructed, libraries are clonally amplified in preparation for sequencing. To obtain nucleic acid sequence from the amplified libraries, several NGS platforms rely on sequencing by synthesis. The library fragments act as a template, and a new DNA fragment is synthesized. The sequencing occurs through a cycle of washing and flooding the fragments with the known nucleotides in a sequential order.

After sequencing is complete, raw sequence data must undergo several analysis steps. A generalized data analysis pipeline for NGS data includes preprocessing the data to remove adapter sequences and low-quality reads, mapping of the data to a reference genome or de novo alignment of the sequence reads, and analysis of the compiled sequence. Analysis of the sequence can include a wide variety of bioinformatics assessments, including genetic variant calling for detection of SNPs or indels (insertion or deletion), structural variation (Copy Number variation), detection of novel genes or regulatory elements, and assessment of transcript expression levels. Analysis can also include identification of genetic variation that may contribute to the diagnosis of a disease or genetic condition. Many free online tools and software packages exist to perform the bioinformatics necessary to successfully analyze sequence data (Gogol-Döring and Chen 2012).

The NGS technologies has changed scientific approaches in basic, applied research. Although NGS is similar with initial PCR, the major advance of NGS technology is the capacity to produce large genome data cheaply. So I can sequence the whole genome of organism at reasonable cost. Since NGS era, the whole genome of many related organisms were sequenced and large-scale comparative and evolutionary studies that were unimaginable just a few years ago were performed.

Above this, applications of NGS is almost endless, allowing for rapid advances in many fields. Human re-sequencing of the human genome will be performed to identify genetic variants in pathological processes. NGS has also provided a wealth of knowledge for comparative biology studies through whole genome sequencing of a wide variety of organisms. And NGS is applied in the fields of public health through the sequencing of bacterial and viral species to facilitate the identification of novel virulence factors. Additionally, gene expression studies using RNA-Seq have begun to replace microarray analysis with the ability to visualize RNA expression in sequence form. As NGS grows in popularity, it is inevitable that there will be additional innovative applications.

1.3.2 Overview of Copy Number Variation

Copy number variations (CNVs) as form of structural variation are one of several alterations of the DNA of a genome that results in the cell having an abnormal or a normal variation in the number of copies of one or more sections of the DNA. CNVs correspond to large regions of the genome that have been deleted or duplicated on genome. For example, the chromosome that normally has sections in order as A-B-C-D might instead have sections A-B-C-C-D (a duplication of "C") or A-B-D (a deletion of "C"). CNV accounts for roughly 12% of human genomic DNA and each variation may range from about one kilobase (1,000 nucleotide bases) to several megabases in size (Stankiewicz and Lupski 2010). So Many studies predicted that CNVs have more explanation power than single-nucleotide polymorphisms (SNPs), because physical portion of CNVs on genome relatively is higher than SNP.

CNVs may either be inherited or caused by de novo mutation. JA Lee (2007) proposed mechanism for the cause of some CNVs was fork stalling and template

switching, a replication misstep (Lee, Carvalho and Lupski 2007). Also CNVs can be caused by structural rearrangements of the genome such as deletions, duplications, inversions, and translocations. Low copy repeats, which are region-specific repeat sequences, are susceptible to such genomic rearrangements resulting in CNVs. Factors such as size, orientation, percentage similarity and the distance between the copies influence the susceptibility of CNVs to genomic rearrangement (Lee and Lupski 2006).

Copy number variation can be discovered by cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization, and by virtual karyotyping with SNP arrays. Recent advances in DNA sequencing technology have further enabled the identification of CNVs by next-generation sequencing (Korbel et al. 2007, Sudmant et al. 2010, Mills et al. 2011, Paudel et al. 2013). Through several studies, CNVs can result in either too many or too few of the dosage-sensitive genes, which may be responsible for a substantial amount of complex phenotypic variability (Redon et al. 2006, Freeman et al. 2006). In cases of rapidly growing Escherichia coli cells, the gene copy number can be 4-fold greater than others. Elevating the gene copy number of a particular gene can increase the expression of the protein that it encodes (Atkinson et al. 2003, Perry et al. 2007).

1.3.4 General Study of Copy Number Variation in Animal Genetics

Since the completion of the several animal genome assembly, a large number of genetic variation as single nucleotide polymorphisms (SNPs), have become widely known and commercial SNP panels have been developed for domesticated animal cattle. The continued discovery of SNPs has been further

expanded by NGS data. SNPs and the commercial SNP marker panels have been successfully used to identify genomic regions that potentially underlie the economic traits of cattle (Barendse et al. 2009b). Another source of genetic variation is from gains and losses of genomic structural sequence variants, copy number variations (CNVs), that occur in more than two individuals (Mills et al. 2011).While SNPs are more frequently used in cattle breeding than CNVs, CNVs occupy a higher percentage of genomic sequence than SNPs. Already, many studies have endeavored to understand CNVs in mammals (Graubert et al. 2007, Guryev et al. 2008). It is possible that CNVs have a potentially greater effect on phenotype, including changing of gene structure and dosage, altering gene regulation and exposing recessive alleles (Zhang et al. 2009). These points are attracting attention to CNV as structural variation that can account for diverse economically important traits in domestic animals. So, many CNV studies of domesticated animals were performed using several domesticated animals including dog, goat, cattle, pig and sheep (Chen et al. 2009, Fontanesi et al. 2009b, Ramayo-Caldas et al. 2010b, Fontanesi et al. 2011, Bickhart et al. 2012). Considering the heritability of CNVs and their higher rates of mutation, CNVs may be largely associated with or affect animal health and production traits under recent selection. In the case of cattle, partial deletion of the bovine gene ED1 causes anhidrotic ectodermal dysplasia (Drögemüller, Distl and Leeb 2001). Likewise, beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. It is possible that CNVs may be associated with these agriculturally important traits (Bickhart et al. 2012).

1.3.5 Detection of Copy Number Variation in Population Level using NGS Data

Until now, CNV screens were routinely performed by comparative genomic hybridization (CGH) and SNP arrays, and many studies have extensively reviewed their performances (Lai et al. 2005, Winchester, Yau and Ragoussis 2009, Pinto et al. 2011). However, these methods, which are often affected by low probe density and cross-hybridization of repetitive sequence, were not able to detect CNVs at the whole genome level. A limited number of investigations in cattle CNV has been performed to detect CNVs using methods that include high-density aCGH and the 50 K SNP panel (Bae et al. 2010b, Fadista et al. 2010). The recent advances of NGS and complementary analysis programs have provided better approaches to systematically identify CNVs at a deep genome-wide level than the currently available commercial SNP chip and aCGH methodologies (Stothard et al. 2011). These approaches have several drawbacks, including hybridization noise, limited coverage for genome, low resolution, and difficulty in detecting novel and rare mutations (Snijders et al. 2001, Shendure and Ji 2008). Over the last few years, NGS has evolved into a popular strategy for genotyping and has included comprehensive characterization of CNVs by generating hundreds of millions of short reads in a single run. The advantages of the NGS approach include higher coverage and resolution, more accurate estimation of copy numbers, more precise detection of breakpoints, and higher capability to identify novel CNVs (Meyerson, Gabriel and Getz 2010, Alkan, Coe and Eichler 2011). Many SV-detecting methods can be applied to CNV identification. The NGS based CNV detection methods can be categorized into five different strategies, including: (1) paired-end mapping (PEM), (2) split read (SR), (3) read depth (RD), (4) de novo assembly of a genome (AS), and (5) combination of the above approaches (CB). Taking approach and their advantages into account, a diverse set of tools has been developed to detect CNVs based on different features that can be extracted from NGS data. In NGS CNV, recent approach proposed population-level concepts to reinterpret the technical features of sequence data that reflect structural

variation. This method is based on enable new kinds of analytical approaches. In this approach, True structural alleles has additional evidence in population-scale data than within genome data. Segregating alleles distinguish some genomes from others, substitute for alternative structural alleles, give rise to discrete allelic states in a diploid genome, are often shared across genomes and segregate on haplotypes with other variants (Handsaker et al. 2011). Based on CNV from NGS data, previous study of cattle using next generation sequencing (NGS) data has reported that CNVs play a crucial role in diverse biological functions as pathogen- and parasite-resistance, lipid transport and metabolism, breed-specific differences in adaptation, health, and production traits.

This chapter was published in *Asian-Australasian Journal of Animal Sciences*
as a partial fulfillment of Dong-Hyun Shin's Ph.D program.

Chapter 2. Accurate estimation of effective population size in the Korean dairy cattle based on linkage disequilibrium corrected by genomic relationship matrix

2.1 Abstract

Linkage disequilibrium between markers or genetic variants underlying interesting traits affects many genomic methodologies. In many genomic methodologies, the effective population size (N_e) is important to assess the genetic diversity of animal populations. In this study, dairy cattle were genotyped using the Illumina BovineHD Genotyping BeadChips for over 777,000 SNPs located across all autosomes, mitochondria and sex chromosomes, and 70,000 autosomal SNPs were selected randomly for the final analysis. I characterized more accurate linkage disequilibrium in a sample of 96 dairy cattle producing milk in Korea. Estimated linkage disequilibrium was relatively high between closely linked markers (>0.6 at 10 kb) and decreased with increasing distance. Using formulae that related the expected linkage disequilibrium to N_e, and assuming a constant actual population size, N_e was estimated to be approximately 122 in this population. Historical N_e, calculated assuming linear population growth, was suggestive of a rapid increase N_e over the past 10 generations, and increased slowly thereafter. Additionally, I corrected the genomic relationship structure per chromosome in calculating r^2 and estimated N_e. The observed N_e based on r^2 corrected by genomics relationship structure can be rationalized using current knowledge of the history of the dairy cattle breeds producing milk in Korea.

2.2 Introduction

The recent availability of SNP genotyping allows for the application of genomic techniques to domestic animals. Genomic techniques such as genome-wide association studies and genomic selection depend on the extent of linkage disequilibrium and its rate of decline with distance between loci within a population. These genomic tools depend on sample size and the quality of linkage disequilibrium estimation. Therefore, accurate characterization of linkage disequilibrium will assist in planning future studies using genomic techniques.

Linkage disequilibrium can provide insights into the evolutionary history of a population through the effective population size (N_e) (Falconer 1960). Using N_e, I can monitor genetic diversity in livestock populations and explain the observed range and pattern of genetic variation with regard to population genetics. In addition, if pedigrees are incomplete or unavailable, N_e provides information into the inbreeding level of a population. The strength of linkage disequilibrium at different genetic distances between makers can be used to infer ancestral N_e. The pattern of historical N_e in animal populations can increase my understanding of the impact of selective breeding strategies on the genetic variation within the framework of population genetics.

Researchers have already applied SNP chips for genome-wide association study (Huang et al. 2010, Jiang et al. 2010, Sahana et al. 2010) and genomic selection (Schaeffer 2006, Hayes et al. 2009a) in dairy cattle. The pattern of linkage disequilibrium in cattle has been characterized, and predictions of N_e estimated using SNP chip data have already been made. Flury (2010), based on 128 Swiss Eringer breed genotyped using the Illumina BovineSNP50 BeadChip, estimated linkage

disequilibrium-based actual N_e and ancestral pedigree-based N_e (Flury et al. 2010). Other studies have evaluated the historical N_e of a variety of cattle breeds, all of which suggested a continuous decrease in N_e since the time of domestication (Thevenon et al. 2007, De Roos et al. 2008).

N_e is an important measure in the global dairy cattle industry as well as the Korean dairy cattle industry in identifying the state of genetic resource. As Korea is a major-semen importing country in the dairy cattle industry (Table 2.4), dairy cattle in Korea have diverse genetic sources. In this study, I used r^2 corrected by the genomic relationship structure based on SNPs, which is more accurate than the existing r^2 to estimate linkage disequilibrium. I then estimated the current N_e and inferred an accurate N_e. These results are compared with other studies and considered in the context of current knowledge for the establishment of genomics methods in dairy cattle in Korea.

2.3 Materials and Methods

2.3.1 Genotypic data

Almost all of the dairy cattle in Korea are Holstein, which have been produced by selected domestic seed bull or imported semen. Therefore, the Korean dairy cattle population in this manuscript is the Holsteins population in Korea. The country of origin for imported semen is shown in Table 2.4. The 96 Holstein samples were collected from four field dairy farms located in Gyeonggi and Gangwon provinces. The samples were collected by the Animal and Plant Quarantine Agency of Korea (a governmental agency) by random sampling for a survey on population disease resistance in Korea. The collected samples for this study represent the Korean Holstein cattle population despite the limitations of the sample size and is treated as a single population sample representing the dairy cattle of Korea. There are few regional differences in the dairy cattle in Korea because the supply and management of dairy cattle semen or seed bull takes place at the national level.

The Illumina BovineHD Genotyping BeadChip, which contains 777kb SNP located across all autosomes, sex chromosomes, and mitochondria was used. These informative SNPs were selected from the Bovine Genome Database. Genotyping data were analyzed using Plink for quality control (GENO > 0.05, MAF < 0.1, HWE test P <= 0.0001), which removed 226,156 SNP from the analysis because of poor genotyping quality (Purcell, Neale et al. 2007). As the Illumina BovineHD Genotyping BeadChip contains numerous SNP at a high density, the distance between adjacent SNPs is very short making the computing time long. To reduce the computational time, I used 70, 000 randomly selected SNPs including only autosomal SNPs after quality control. Interbull (Uppsala, Sweden;

www.interbull.org) uses BovineSNP50 Genotyping BeadChips, which contains 54,609 SNP to estimate genetic parameters and hence, I considered 70,000 SNPs sufficient for my purposes.

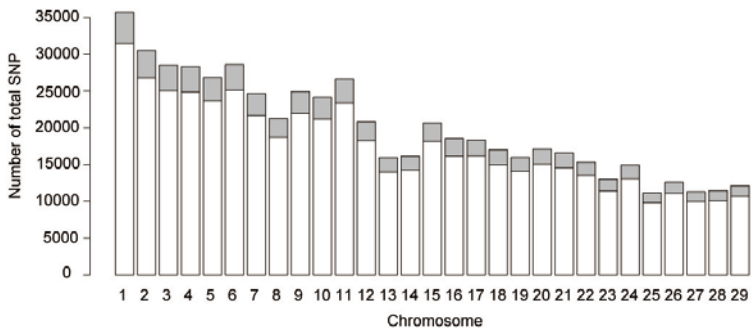

Figure 2.1. Number of SNPs per chromosome. Gray indicates the portion of used SNPs and white denotes the unused portion.

2.3.2 Linkage disequilibrium

I estimated linkage disequilibrium using the "LDcorSV" package implemented in R. This package provides a set of functions to measure the existing r^2 and the r^2 corrected by the sample structure (Mangin et al. 2011). Estimating the r^2 of all pairwise per chromosome using high-density SNP chips is time-consuming. To reduce computing time, the input for each chromosome was split into files containing 100 loci, and r^2 was calculated for all syntenic marker pairs in each file as has been done in previous studies (Flury et al. 2010).

The standard measures of existing r^2 are respectively equivalent to the covariance and the correlation between alleles at two different loci (Hill and Robertson 1968), computed as:

$$r^2 = \frac{D^2}{P_A P_a P_B P_b} \quad (2.1)$$

Where PA, Pa, PB, and Pb are the respective frequencies of alleles A, a, B and b, and D is PAB - PAPB. Sample structure information is required for the corrected r^2. I used the genomic relationship matrix based on SNP data instead of pedigree data. In this study, the genomic relationship value of individual pairs was the number of common SNP between two individuals / number of total allele sites. In this way, I proposed a simple genomic relationship matrix (96 × 96). A process of calculating r^2 corrected by the genomic relationship structure based on SNP was identical to the existing r^2 calculation.

Details on the physical position of the markers can be found in the Illumina product literature. To determine linkage disequilibrium in relation to the physical distance between markers, marker pairs were divided into distance bins. I established two kinds of classes (0 - 0.5Mb, 0 - 5Mb) and subsequently, applicable marker pairs to each class were put into 50 distance bins with bin ranges dependent on the class (see Table 2.1). The two types of mean r^2 for each of the distance bins were then plotted against the median of the distance bin range (Mb). This estimation of r^2 was performed on a chromosome by chromosome basis; the pooled results are presented in Figure 2.2.

Also r^2 was calculated for a random selection of non-syntenic SNP. Across the autosome, 700 SNP were randomly selected and r^2 values were calculated for only non-syntenic marker pairs. This resulted in a total of 250,096 pairwise comparisons that were not corrected by the genomic relationship structure.

Table 2.1. Distance classes and bin ranges for the linkage disequilibrium summary

Class	Minimum Distance (Mb)	Maximum Distance (Mb)	Within class bin distance range (Mb)	No. of bins
1	0	0.5	0.01	50
2	0	5	0.1	50

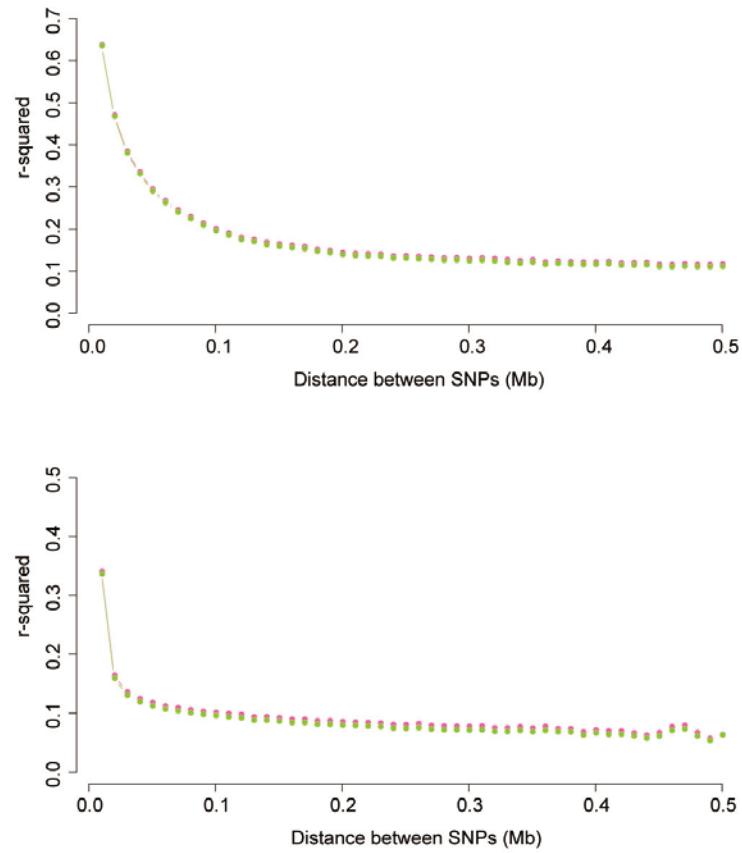

Figure 2.2. Average linkage disequilibrium (solid line) plotted against the median of the distance bin range (Mb). Each hot pink and yellow-green color represents the existing r2 and the r2 corrected by the genomic relationship structure based on SNP, respectively. (a) Distances ranged from 0 to 0.5 Mb. r2 values were averaged using bins of 0.01 Mb and pooled over autosomes. (b) Distances ranged from 0 to 5 Mb. r2 values were averaged using bins of 0.1 Mb and pooled over autosomes.

2.3.2 Construction model of linkage disequilibrium with distance

Under the assumption of an isolated population with random mating, Sved (1971) derived an approximate expression for the expectation of r^2 (Sved 1971):

$$\mathrm{E}(r^2) = \frac{1}{1 + 4Nc} \quad (2.2)$$

where N is the effective population size, and c is the recombination frequency. In this study, as in previous studies, c was replaced by linkage distance in Morgans (Hayes et al. 2003, Tenesa et al. 2007, Thevenon et al. 2007, De Roos et al. 2008, Villa-Angulo et al. 2009, Qanbari et al. 2010, Flury et al. 2010, Corbin et al. 2010b). This was justified by the approximation of the more precise equation for E(r^2) given by Sved (1971) (Sved 1971). Based on this formula, a non-linear least-squares approach to statistically model the observed r^2 was implemented within R (nls function) using this model:

$$y_i = \frac{1}{a + 4bd_i} + e_i \quad (2.3)$$

where y_i is the value of r^2 for marker pair i, at linkage distance d_i in Morgans. Parameters a and b were estimated iteratively using the least-squares method. Chromosome-specific megabase-to-centimorgan conversion rates were calculated based on total physical chromosome length as stated on the UCSC Web site, and total chromosome genetic length derived from the bovine linkage map (Arias et al. 2009) (see Table 2.2). I used marker pairs for which r^2 values were higher than the mean of r^2 for non-syntenic marker pairs. This model was applied to each chromosome in turn and the parameters were estimated. Similar to Corbin (2010), estimated parameters were combined by meta-analysis in R using an inverse variance

method for pooling and random effects method based on the DerSimonian-Laird method (DerSimonian and Laird 1986, Corbin et al. 2010b).

Table 2.2. Chromosome-specific centimorgan to megabase conversion ratios

Chromosome	Length (Mb)	Length (cM)	cM/Mb Ratio
1	158	166	1.05
2	137	148	1.08
3	121	141.8	1.17
4	121	132.5	1.10
5	121	130	1.07
6	119	134.2	1.13
7	113	125.5	1.11
8	113	124.4	1.10
9	106	110.3	1.04
10	104	118.9	1.14
11	107	129.9	1.21
12	91	117.3	1.29
13	84	118.3	1.41
14	85	127.4	1.50
15	85	110.3	1.30
16	82	112.4	1.37
17	75	97	1.29
18	66	103.2	1.56
19	64	100.8	1.58
20	72	73.7	1.02
21	72	90.2	1.25
22	61	91.4	1.50
23	53	90	1.70
24	63	85.8	1.36
25	43	62	1.44
26	52	69.8	1.34
27	45	60.9	1.35
28	46	57.3	1.25
29	52	68	1.31

2.3.4 Ancestral effective population size estimation

Equation (2.2) can predict the effective population size at a given point in time, expressed as generations in the past (Hayes et al. 2003, De Roos et al. 2008).

$$N_T(t) = \frac{1}{4c} * \left[\frac{1}{r_c^2} - 1\right] \quad (2.4)$$

where N_T is the effective population size t generations ago, c is the distance between markers in Morgans, r^2_c is the mean value of r^2 for marker pairs c Morgans apart, and c=(2t)-1 when assuming linear growth (Hayes et al. 2003). As previously mentioned, marker pairs with linkage distances less than the mean of r^2 for non-syntenic marker pairs (< 0.014248) were excluded from this analysis. To estimate N_T, the number of prior generations was selected and a suitable range for c was calculated. The binning process was designed to ensure sufficient marker pairs within each bin and obtain a representative r^2 mean (see Table 2.3). This process was performed for markers pooled across autosomes.

Table 2.3. Description of the generation binning process

Generation range applied to	No. of generations represented by each bin	Example for first bin		
		Generation	Generation range	Corresponding distance range (Morgans)
5 - 10	5	5	4.5 to 5.5	0.11 to 0.09
20 - 100	10	20	15 to 25	0.02 to 0.033
200 - 1,000	100	200	150 to 250	0.002 to 0.0033
2,000 - 10,000	1,000	2,000	1,500 to 2,500	0.0002 to 0.00033
20,000 - 100,000	10,000	20,000	15,000 to 25,000	0.00002 to 0.000033

2.3.5 Estimation of effective population size based on the genomic relationship structure per chromosome

For the 96 genotyped individuals, I were able to establish genomic relationship structure per chromosome using SNP information of each chromosome. I proposed r^2 corrected by the genomic relationship structure per chromosome using "LDcorSV" package implemented in R (Mangin et al. 2011). A process of estimating N_e based on genomic relationship per chromosome and ancestral N_e was identical to that described above.

2.4 Results

2.4.1 Genotypic data

Of the 734,862 genotyped autosomal SNPs, 508,707 (69.2%) remained after quality control processes and 70,000 were selected for analysis. The number of SNPs per autosome remaining after filtering and selection ranged from 1,300 to 4,280 and was closely related to chromosome length and total number of SNP as shown in Figure 2.1. The minor allele frequency of remaining SNP followed a uniform distribution and averaged (±SD) to be 0.31 ± 0.11. The average distance between marker pairs (±standard deviation) for this analysis was 1,202.77 ± 932 kb, with the distance between markers ranging from 0.134 kb to 8,398 kb.

2.4.2 Linkage Disequilibrium

Linkage disequilibrium declined with increasing distances between marker pairs, as shown in Figure 2.2 a and b. The most rapid decrease was seen over the first four bins, with the mean r^2 decreasing by about half. The mean existing r^2 decreased more slowly with increasing distance and was constant after 0.2 Mb of distance. The mean r^2 corrected by the genomic relationship structure based on SNP with distance between marker pairs was slightly less than that of the existing r^2. However, change patterns in the r^2 means with distance for both methods were similar. The decline in linkage disequilibrium showed slight differences with log-transformed distance (Figures 2.6 and 2.7). According to the existing r^2, 23,797 of the 3,385,800 marker pairs were in complete linkage disequilibrium; 12,225 of these were adjacent pairs. For r^2 corrected by the genomic relationship structure based on SNP, 23,794 of the

3,385,800 marker pairs were in complete linkage disequilibrium; 12,223 of these were adjacent pairs. The mean (± standard deviation) r^2 between random non-syntenic marker pairs was 0.014248 ± 0.0197. 835,324 using the existing r^2 measure and 861,676 using r^2 corrected by the genomic relationship structure based on SNP are less than the mean of non-syntenic marker pairs.

2.4.3 Construction model of linkage disequilibrium with distance

The non-linear regression model of the declining linkage disequilibrium with distance resulted in both parameters, a and b, being significantly different from zero. For parameter estimation using the existing r^2, the mean estimate and 95% confidence interval by meta-analysis across autosomes for parameters a and b were 2.95 [2.84; 3.07] and 106.32 [92.95; 119.70], respectively. The parameters a and b for r^2 corrected by the genomic relationship structure based on SNP were 2.87 [2.75; 2.99] and 122.28 [107.21; 137.34], respectively. The predicted r^2 from the non-linear regression equation was similar to the mean observed r^2 of regions with massed bins with discrepancies occurring in other regions for both two types of r^2 (Figures 2.6 and 2.7). Parameter b showed greater variability between chromosomes than parameter a. No such relationship was observed between the estimated parameters (a, b) and chromosome length (cM) as shown in Figure 2.3 and 2.4. This reason for the lack of relationship and interpretation of parameter b in this non-linear regression model as an estimated N_e is demonstrated in the discussion.

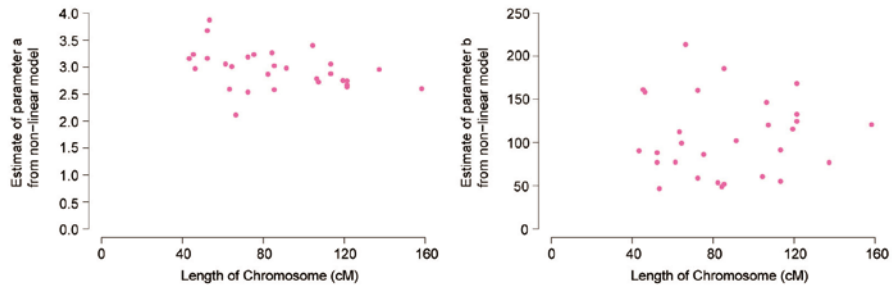

Figure 2.3. Parameter estimates from Equation (3) plotted against chromosome length (cM) according to the bovine linkage map using the existing r2 (Arias, Keehan et al. 2009). (a) Estimates of parameter a plotted against chromosome length (cM). (b) Estimates of parameter b plotted against chromosome length (cM).

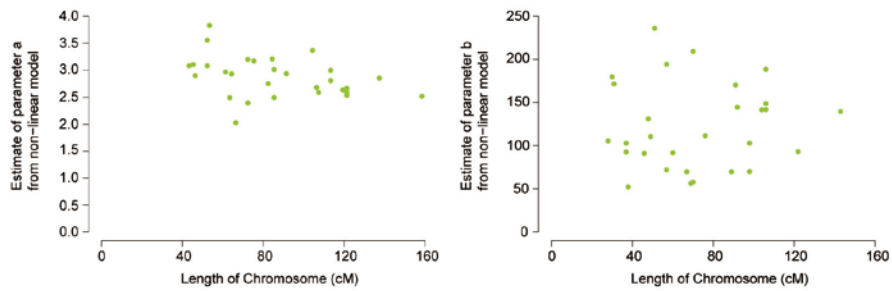

Figure 2.4. Parameter estimates from Equation (3) plotted against chromosome length (cM) according to the bovine linkage map using the r2 corrected by the genomic relationship structure based on single nucleotides (Arias, Keehan et al. 2009). (a) Estimates of parameter a plotted against chromosome length (cM). (b) Estimates of parameter b plotted against chromosome length (cM).

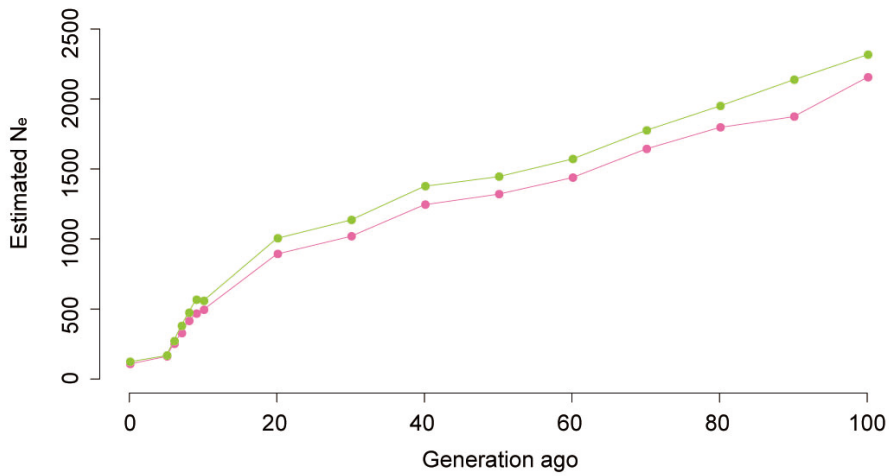

Figure 2.5. Average estimated effective population size plotted against generations in the past, truncated at 100 generations. Each hot pink and yellow-green color dot represents the use of the existing r2 and the r2 corrected by the genomic relationship structure based on SNP, respectively.

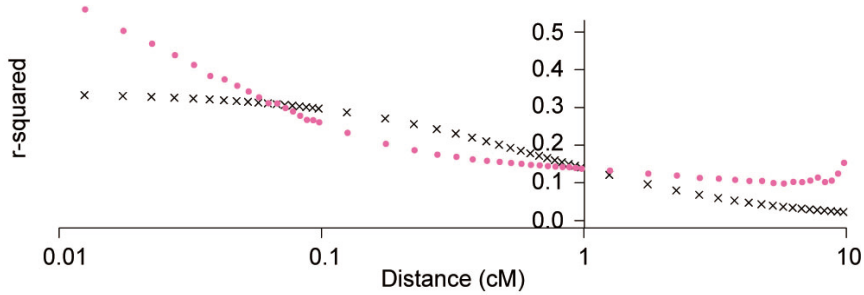

Figure 2.6. Predicted r^2 versus observed r^2 against mean distance between markers (cM, on a log scale) using the typical r^2 in Equation (3)

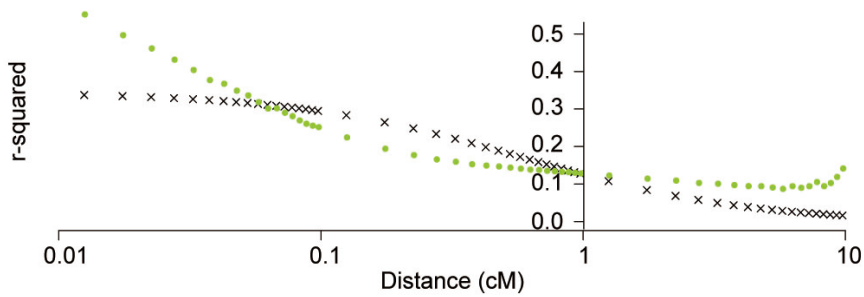

Figure 2.7. Predicted r^2 versus observed r^2 against mean distance between markers (cM, on a log scale) using r^2 corrected by the genetic relationship structure based on total single nucleotides in Equation (3)

2.4.4 Ancestral effective population size estimation

I observed rapid increase in N_e over the past 10 generations with the values increasing fivefold (close to 500) by 10 generations ago (Figure 2.5). From the past 10 generations to the past 100 generations, N_e increased slowly. My results suggest that a continuous increase in N_e occurred over the last 100,000 generations (Figure 2.8). Overall, the tendencies for the two different r^2 were similar, but N_e based on r^2 corrected by the genomic relationship structure based on SNP was slightly higher than N_e based on the existing r^2. In Figure 2.5, although the recent difference between the two estimated measures was small, the difference in value increased with time.

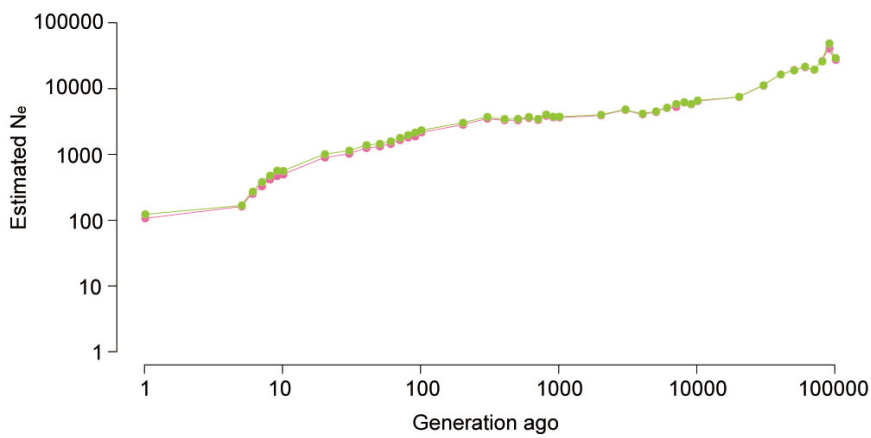

Figure 2.8. Average estimated effective population size plotted against generations in the past, truncated at 100,000 generations. Estimated effective population size and generations in the past plotted on a log scale. (a) hot pink for the typical r2 (b) yellow-green for r2 corrected by the genomic relationship structure based on total single nucleotide polymorphisms.

Table 2.4. Number of registered semen per country in Korea (domestic and imported)

	Estimated number of reliable semen	Estimated number of non-reliable semen	Total	%	Domestic or imported
Korea	38	175	213	12.4	domestic
U.S.A	958	175	1,133	66	imported
Canada	311	48	359	21	imported
Australia	5	0	5	0.3	imported
Japan	5	0	5	0.3	imported
Total	1,317	398	1,715	100	-

2.4.5 Estimation of effective population size based on genomic relationship per chromosome

Linkage disequilibrium corrected by the genomic relationship structure per chromosome declined with increasing distance between marker pairs, as shown in Figure 2.9 a and b. Over the first four bins, the decline was greater than two previous results. The mean r^2 corrected by the genomic relationship structure per chromosome with distance between marker pairs was less than the two previous mean but the change pattern was similar (Figure 2.10). In total, 23,783 of marker pairs were in complete linkage disequilibrium; 12,220 of these were adjacent pairs. For parameter estimation using r^2 corrected by the genomic relationship structure per chromosome, the mean estimate and 95% confidence interval by meta-analysis across autosomes for parameters a and b were 2.11 [2.04; 2.18] and 361.73 [335; 339.47], respectively. Parameter b was approximately three times greater than the two previous values for parameter b. The decline in linkage disequilibrium was almost linear with log-transformed distance and the decline in linkage disequilibrium with log-transformed distance was better fitted with predicted values than that of the other two. No such relationship was observed between parameters (a, b) and chromosome length (cM) in r^2 corrected by the genomic relationship structure per chromosome as shown in Figure 2.11. Similar to previous patterns, I observed a rapid increase in N_e over the past 10 generations and then a slow increase up to 100 generations ago. However, the values increased fivefold (close to 1,500) at 10 generation than two kinds of r^2 (Figure 2.12). From 10 generation ago, my results suggest that continuous increase in N_e has occurs over the 100,000 generations (Figure 2.13). Overall, these tendencies were similar with that of the previous two but N_e based on r^2 corrected by the genomic relationship structure per chromosome was greater than the two previous kinds of r^2 measures. Although recent difference with two estimated measures was big, their difference in value decreased with generations ago.

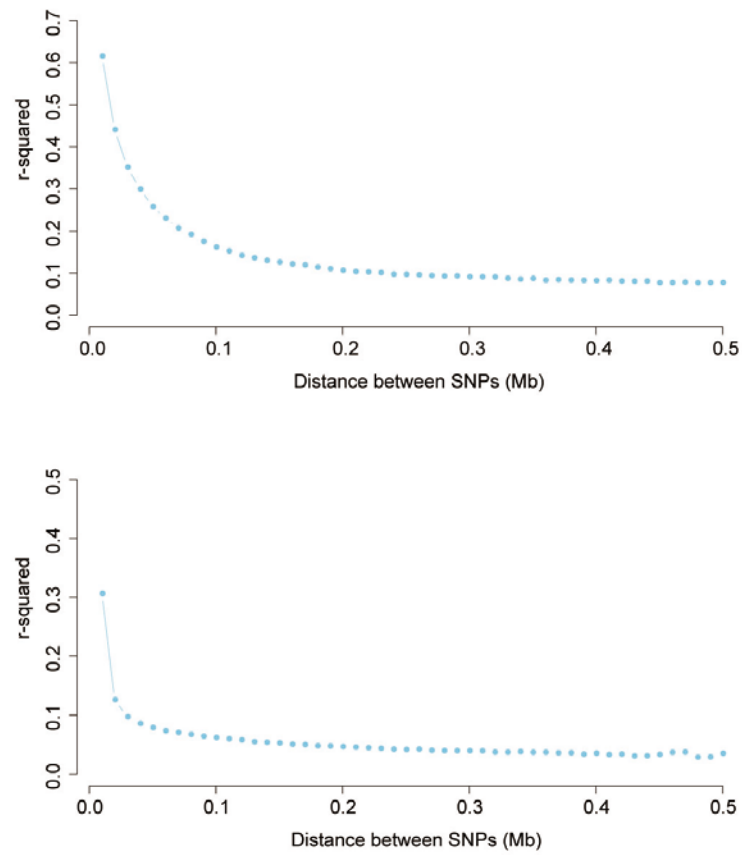

Figure 2.9. Average linkage disequilibrium (solid line) plotted against the median of the distance bin range (Mb) using r2 corrected by the genetic relationship structure per chromosome. (a) Distance range from 0 to 0.5 Mb. r2 values averaged using bins of 0.01 Mb and pooled over autosomes. (b) Distance range from 0 to 5 Mb. r2 values averaged using bins of 0.1 Mb and pooled over autosomes.

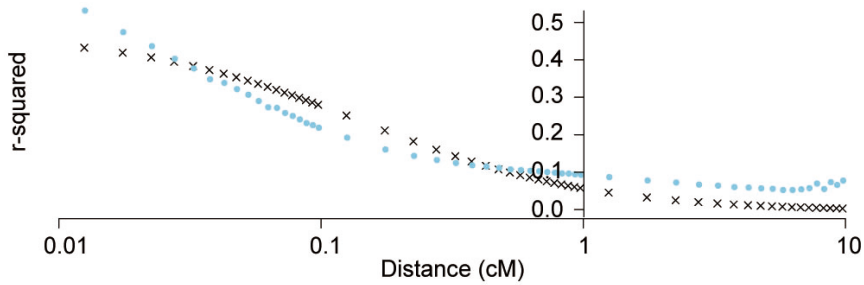

Figure 2.10. Predicted r^2 versus observed r^2 against mean distance between markers (cM, on a log scale) using the r^2 corrected by the genetic relationship structure per chromosome in Equation (3).

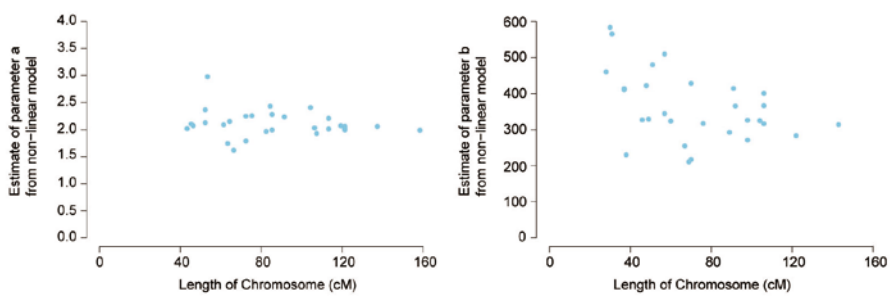

Figure 2.11. Parameter estimates from Equation (3) plotted against chromosome length (cM) according to the bovine linkage map using r2 corrected by the genetic relationship structure per chromosome. (Arias, Keehan et al. 2009). (a) Estimates of parameter a plotted against chromosome length (cM). (b) Estimates of parameter b plotted against chromosome length (cM).

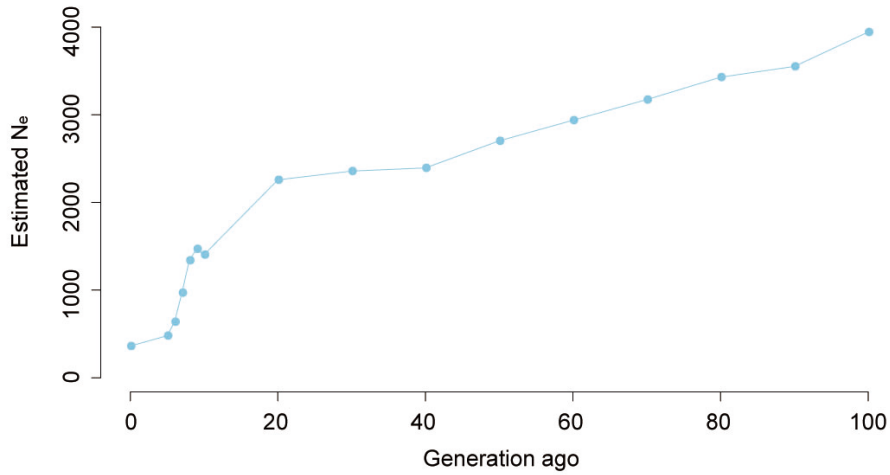

Figure 2.12. Average estimated effective population size plotted against generations in the past, truncated at 100 generations using r^2 corrected by the genomic relationship structure per chromosome.

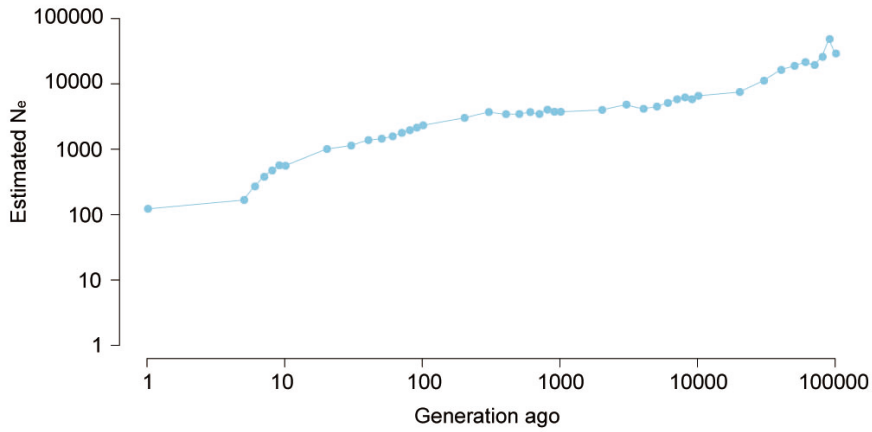

Figure 2.13. Average estimated effective population size plotted against generations in the past, truncated at 100,000 generations using r2 corrected by the genomic relationship structure per chromosome. This study provides an overview of linkage di Estimated effective population size and generations in the past plotted on a log scale.

2.5 Discussion

This study provides an overview of linkage disequilibrium in dairy cattle in Korea using a high density SNP panel. Validation work by Khatkar (2008) on their cattle data suggested that the r^2 correlation between 100 and 1,000 samples is 0.94 (Khatkar et al. 2008). Thus, I can obtain an unbiased picture of linkage disequilibrium (3,385,801 pairs) in my population of 96 dairy cattle. The pattern of decline of linkage disequilibrium in this population is consistent with that reported by Flury (2010) in 128 Swiss cattle (Flury et al. 2010) with existing r^2 remaining higher than approximately 0.3 for distance up to 0.05 Mb. The linkage disequilibrium observed is higher at short distances and more extensive than that observed in human populations (Shifman et al. 2003). Overall, r^2 values corrected by the genomic relationship structure based on SNP were slightly lower than the existing r^2 because the correction process prevented the overestimation of linkage disequilibrium.

My population consisted of individuals from several genetic origins, which produces linkage disequilibrium between unlinked loci because of differences in allele frequencies. Consequently, such a structured sample can lead to a biased estimate of linkage disequilibrium, which may increase false positives rates. Therefore r^2 was corrected for by inferring N_e to take into account the non-independence of loci due to population differentiations (Yu et al. 2005). Not corrected r^2 can be a biased estimation of linkage disequilibrium, which could lead to estimated genetic parameters that reduce the power of analysis (Mangin et al. 2011). I used an R package ("LDcorSV") to calculate r^2 corrected by the genomic relationship structure based on SNP, which was slightly smaller than the existing r^2. This was expected, as I speculated that the calculated r^2 corrected by the genomic

relationship structure based on SNP was more accurate. Thus I can infer an accurate N_e of dairy cattle in Korea.

The mean value of r^2 between non-syntenic markers was 0.014248, which provides an approximation of the linkage disequilibrium that can be expected by chance. The value observed here is higher than the value observed by Khatkar (2008) in a sample of over 1,500 cattle (0.0032) (Khatkar et al. 2008). Sample size and genetic sampling (drift) affect the mean of non-syntenic r^2 values, and hence the mean may be expected to decrease with an increase in sample size. The larger non-syntenic value observed by Khatkar (2008) may be more affected by large populations. Since the sample size was smaller in this study than the other studies, the larger non-syntenic values of my dataset are reasonable. I used marker pairs with more than mean of r^2 for non-syntenic marker pairs as a standard for estimating linkage disequilibrium. Many low r^2 can cause overestimated N_e more than expected. Therefore I decided to use marker pairs more than mean of non-syntenic marker pairs for inferring N_e.

Using Sved's (1971) formula for the expected r^2, a non-linear regression model was fitted to the data to describe the relationship between linkage distance and linkage disequilibrium. My estimate of parameter a supports an alternative version of Sved's (1971) equation (Sved 1971), derived by Tenesa (2007), which accounts for mutation and puts a = 2 (Tenesa et al. 2007). While estimating parameters, the initial value of parameter a was two with this approach. The estimated parameter a ranged from 2 to 3. For the heterogeneity of variance of the observed r^2, variance of r^2 declined with increasing distances between markers, which may have impacted my results. A significant negative relationship between chromosome length (cM) and estimates of parameter b from the non-linear model have been observed (Corbin et al. 2010b), while others have observed a positive

relationship in domestic livestock species (Khatkar et al. 2008, Muir et al. 2008). In this study, all maker pairs were calculated in each bin so r^2 was not affected by chromosome length. Thus, I could not observe a relationship between chromosome length (cM) and estimates of b.

My estimate of b represents an estimated N_e assuming a constant population size. However, this assumption is difficult to maintain, b represents a conceptual average of N_e e over the period inferred from the range of marker pairs distances (Toosi, Fernando and Dekkers 2010). Two measures of r^2 resulted in two different estimates of N_e. N_e based on the existing measure of r^2 is about 106 and N_e based on r^2 corrected by the genomic relationship structure based on SNP was about 122. Assuming that r^2 corrected by the genomic relationship structure based on SNP was more accurate than the existing r^2, I predict that N_e of dairy cattle in Korea is about 122. Figure 2.5, 2.8, 2.12 and 2.13 show historical N_e assuming a linear population following Hayes (2003) (Hayes et al. 2003). The observed pattern shows a rapid increase in N_e up to around 10 generations ago. Several explanations exist for this pattern including bottlenecks associated with domestication, selection and breed formation, and endangerment of the breed. Therefore, it is useful to consider my results in the context of the demographic history of the dairy cattle in Korea. The reliability of this method depends both on the technical implementation and approached used in a previous study approach (Corbin et al. 2010a).

To the best of my knowledge, this is a novel study on N_e estimation based on r^2 corrected by the genomic relationship structure per chromosome resulting in an underestimation of inbreeding and thus an overestimation of the current population size. The estimates for recent N_e for dairy cattle in Korea based on r^2 corrected by the genomic relationship structure based on SNP was around 120 individuals in contrast to estimates in the range of 361 using r^2 corrected by the

genomic relationship structure per chromosome. Interesting thing is that recent N_e for Swiss Eringer breed using pedigree information covering 15 generations (the range of 110–321) is three times higher than the linkage disequilibrium-based estimates for recent N_e (around 100 individuals). Thus, I guess that the number of genomic relationship matrix that covers generation depends on how I establish the genomic relationship structure. These differences are important for inferring N_e.

The extent of linkage disequilibrium in a population can be used to estimate the SNP density required for genome-wide association studies to be effective, as well as providing some indication of genomic selection. This has generated thresholds for useful linkage disequilibrium described as the proportion of QTL (quantitative trait locus) variance explained by a marker (Zhao et al. 2005) . The consensus is that an average $r^2 > 0.3$ will permit reasonable sample sizes for genome-wide association studies (Ardlie, Kruglyak and Seielstad 2002, Du, Clutter and Lohuis 2007, Khatkar et al. 2008). In this data set, markers 200 kb apart achieved an average linkage disequilibrium of $r^2 = 0.303$ excluding marker pairs less than the mean of non-syntenic marker pairs. However, marker pairs with $r^2 = 1$, which represent the high variability of r^2 values at small distances are typically excluded in genomic selection. This is likely due to underestimation of the actual number of SNP needed. Genomic selection appears to be effective at lower average r^2 with simulation results demonstrating accuracies of up to 0.65 with an average r^2 between adjacent markers as low as 0.2 and a trait heritability of 0.1 (Calus, de Roos and Veerkamp 2008). Deterministic equations demonstrates that the accuracy of genomic selection can be expressed as a function of the effective number of loci in a population (Daetwyler et al. 2010). The effective number of loci in a population relates to the number of independent chromosome segments and assumes a random mating population. My dataset covered an effective number of loci. However, because the estimated N_e was

greater than my sample size, I required more preparation to predict the potential accuracy of genomic selection in dairy cattle populations.

I used dense SNP genotype data to characterize linkage disequilibrium and infer the ancestral N_e for a sample of dairy cattle. In the population studied, linkage disequilibrium extended for long distances, reaching baseline levels at more than 5 Mb. From the decay in linkage disequilibrium with genetic distance, I inferred the ancestral N_e and observed a recent rapid increase in N_e which reached approximately 500 10 generations ago followed by a decrease until the present time. The final results were that I used correction by the genomic relationship structure to ensure accurate derivation of N_e, resulting in 122 individuals.

This chapter is a part of paper which will be published in elsewhere as a partial fulfillment of Dong-Hyun Shin's Ph.D program.

Chapter 3. Estimation of historical effective population size in the Minke whale based on coalescent model

3.1 Abstract

Whales have captivated the human imagination for millennia. These incredible cetaceans (whales, dolphins and porpoises) are the only aquatic mammals that have adapted to life in the open oceans and have been a source of human food, fuel and tools around the globe. Yet, the genetic story of whale populations is not fully explored. So, I investigated the common minke whale (Balaenoptera acutorostrata) genome, one of the most abundant baleen whales using next generation sequencing. After that I estimated historical effective population size in the Minke Whale based on coalescent model to know when minke whale population size decreases rapidly. This whole-genome sequencing offers a chance to better understand the population history of the largest aquatic mammals on earth.

3.2 Introduction

Cetaceans are a group of secondarily adapted marine mammals with a history of transition from terrestrial to aquatic environments. Although the origin and evolutionary history of cetaceans remains unclear, a widely accepted view is that their terrestrial ancestors returned to the seas around 50 Mya and finally diversified into a group of fully aquatic mammals. Cetaceans include nearly 85 species that consist of two suborders, one is the Mysticeti (baleen whales such as right whale, blue whale, humpback whale, and minke whale) and another is the Odontoceti (toothed whales such as sperm whales and dolphins), which arose from a common Eocene ancestor around 34 Mya. In spite of their variation in body size, all cetaceans are mostly similar in shape (Thewissen et al. 2009, Uhen 2010, Shen et al. 2012).

Recently, Yim (Yim et al. 2014) reported the whole genome sequencing and de novo assembly of the minke whale genome to assemble the draft genome using a high-depth male minke whale sequence (128× average depth of coverage). Also they supported the hypotheses regarding adaptation to hypoxic resistance, metabolism under limited oxygen conditions and the development of unique morphological traits.

In this paper, I use “dadi” (Gutenkunst et al. 2009) to estimate historical effective population size based on resequencing data from four common minke whales. So I can reveal when minke whale population size decreases rapidly.

3.3 Materials and Methods

3.3.1 Minke whale genome sequencing and assembly

Four individual minke whale samples (S30, S34, S35, and S37) were collected from the Whale Research Institute at the National Fisheries Research & Development Institute (NFRDI), Korea. The samples were caught incidentally by fishing net in Hupo, Ganggu, Pohang, and off the east coast of Korea, and were donated to NFRDI for research purposes. DNA was extracted from the muscle tissue of each minke whale, and paired-end libraries were constructed with insert sizes of about 270 bp and 480 bp. Then 101 cycle paired-end sequencing was conducted using the Illumina HiSeq 2000 sequencer. The data are listed in Table 3.1.

FastQC (Andrews 2010) was used to check the quality of the raw read data, and sequencing errors were discarded using the error-correction module of Allpaths-LG (Gnerre et al. 2011). Fq2fa was used to merge error-corrected paired-end reads of each sample into one shuffled-form fasta file, with a filter option for filtering N bases in the reads. I assembled error-corrected paired-end reads using IDBA_UD (Peng et al. 2012) with the option of pre-correction and kmin = 40. Gaps (N bases) in assembled sequences were filled using Gapcloser (Luo et al. 2012) with parameter k value=31. I carried out a 10 genome assembly for the S30 sample using CLC Assembly with minimum contig lengths = 2000, similarity = 0.85, length fraction = 0.5, insert cost = 3, deletion cost = 3, and mismatch cost = 2.

The reads were mapped to the combined assembled genome using Bowtie2 (Langmead and Salzberg 2012) with the default option. Alignment of the SAM file and removal of duplicated reads were conducted using Picard and SAMtools (Li et

al. 2009). Local realignment was conducted using the Genome Analysis Toolkit (GATK) (McKenna et al. 2010) and SNPs were extracted from the reads alignment file using UnifiedGenotyper, based on multi sample calling. Detected variants (QUAL<30, QD<5, FS>200, MQ0 > 4, MQ0/DP > 0.1) and missing variants (which were found in one sample) were discarded from further analysis.

Table 3.1. Sequencing results of the four minke whale samples

Sample Name	Insert Size	Total Base (bp)	Depth (X)	Read Count	N (%)	GC (%)	Q20 ratio(%) /depth(X)	Q30 ratio(%) /depth(X)
S30	270bp	51,959,636,648	17.32	514,451,848	1.98	39.89	92 / 15.9	87 / 15.0
S34	270bp	40,610,199,180	13.54	402,081,180	2.54	39.72	92 / 12.3	87 / 11.8
S35	480bp	47,488,854,074	15.83	470,186,674	1.82	42.4	90 / 14.2	83 / 13.1
S37	480bp	40,666,301,650	13.56	402,636,650	2.17	40.98	89 / 12.1	82 / 11.1
Total		180,724,991,552	60.25	1,789,356,352	2.13	40.75	90 / 54.5	85 / 51.0

* Estimated minke whale genome size : 3 Gb
* Fastq Quality Encoding : Sanger Quality (ASCII Character Code = Phred Quality Value + 33)

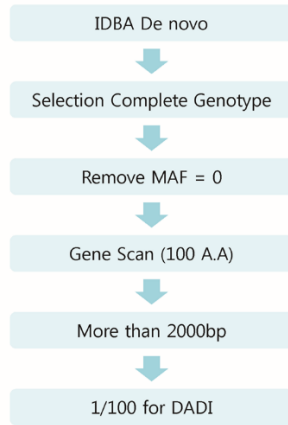

Figure 3.1. Genomic data (SNP) quality control processes for estimating the demographic model using DaDi.

3.3.2 Estimation of the recent demographic model using SNPs

To demonstrate the changes in the minke whale population, I used SNP data from NGS data of four minke whales. To accurately estimate the population history, I performed a quality control procedure (Figure 3.1). I detected 628,081 contigs (2.22 GB) containing 4,099,489 SNPs using IDBA de novo assembler (Peng et al. 2010b) (Figure 3.2, Figure 3.3). I extracted the completely genotyped SNPs (3,730,122) and contigs (259,871, 2.10 GB). I excluded SNPs with a minor allele frequency of 0, retaining 3,610,057 SNPs (352,771 contigs, 2.09 GB). Then, using RepeatMasker (Version 4.0.1) (Smit, Hubley and Green 2012), Augustus (Version 2.5.5) (Stanke and Morgenstern 2005), and blastall (Version 2.2.26) (Altschul et al. 1997), I extracted 26,372 contigs (457,150 SNPs, 0.28 GB) containing 100 amino acids which was estimated to be an entire gene. Then I extracted 24,688 contigs (451,034 SNPs, 0.28 GB) that were greater than 2,000 pb in length (as high-quality contigs). I downsized the SNP data to 1/100 to make a DaDi input file (4,510 SNPs, 3,823 contigs, 0.05 GB). Using the Watterson estimator, I estimated the effective population size before the decline as follows:

$$\theta = 4 * N_e * \mu = \frac{\mathrm{K}(\#\text{segregation site } \cong \#\mathrm{SNP})}{\sum \frac{1}{i}(\mathrm{i} = 1,2,\dots,\mathrm{n}-1)}$$

where N_e is effective population size, n is sample population size, and μ is the mutation rate for minke whale, which is 4.54 × 10-10 (per pb per year) (ref). I used the total genome size and the average generation time (17.65 years) of the minke whale to calculate the mutation rate (RUEGG et al. 2010). K represents the segregation sites in the Watterson estimator, so the number of SNPs were also considered as the number of segregation sites.

To estimate the decline in the historical effective population, I used DaDi software, which can be used to infer population genetic parameters based on coalescent analysis using a single whale population (Gutenkunst et al. 2010). Using the estimated decline in the minke whale population, I assumed a simple decline demographic model (Figure 3.4). I wanted to estimate two parameters using DaDi: the magnitude of the population decline and the time of its occurrence. The process of estimating a demographic model using DaDi was divided into two steps. First, I identified the proper range for each parameter using the ML method in DaDi. Initial values of the two parameters were based on the effective population size before the decline. Then the estimated parameters were used as the new parameters in the next run. After repeating this process 50 times, I excluded outliers and estimated the proper range for each of the two parameters (Figure 3.5). In the second step, I identified each of the two parameters. I selected 100 random parameter sets from the first step. Thereafter, I repeated this parameter estimation process 100 times, and inferred two accurate parameters (Figure 3.6 and Figure 3.7).

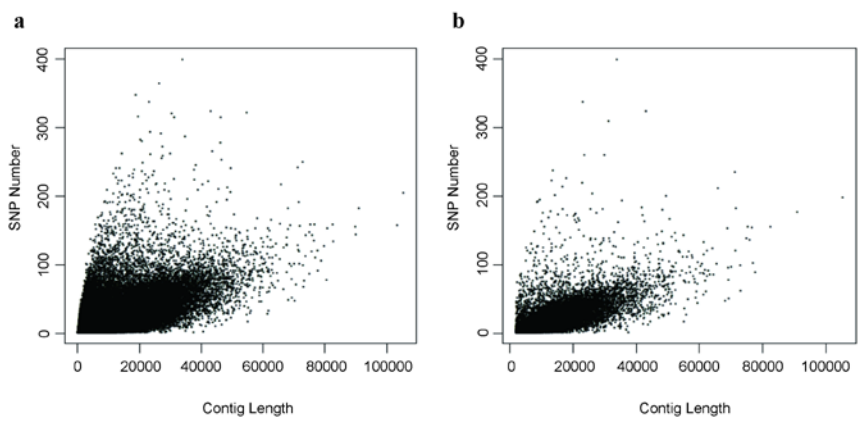

Figure 3.2. Plots identifying the relationship between contig length and number of SNPs: (a) using total contigs before quality control and (b) using contigs after quality control.

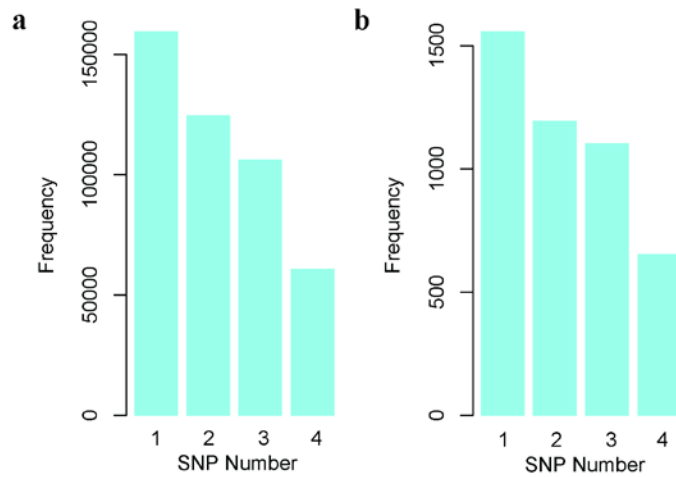

Figure 3.3. Plots identifying the relationship between contig length and number of SNPs: (a) using total contigs before quality control and (b) using contigs after quality control.

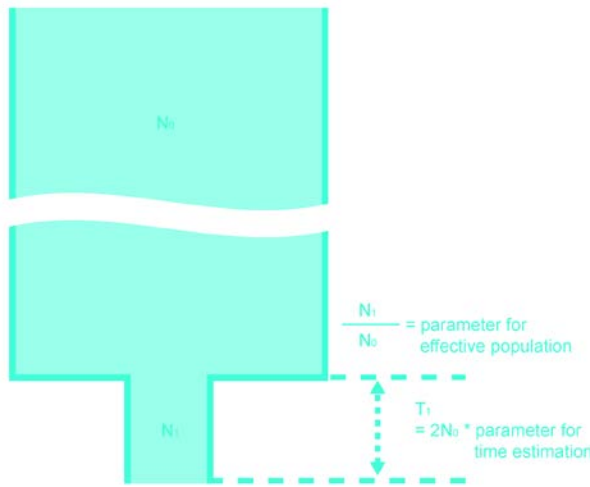

Figure 3.4. Assumed demographic model of minke whales using SNPs from next generation sequencing, based on decline in the population size of minke whales during the Holocene

3.4 Results

I made DADI input data using 4,510 SNPs (0.05 GB). Because coalescent analysis used relation between SNP number and frequency, clear input data is very important. Relation between SNP number and contig length must be positive and I confirmed this relation in total SNPs (Figure 3.2). Also DADI infers parameter based on SNP number frequency per site. I could investigate expected negative relation between Frequency and SNP number per site. I estimated effective population size of minke whales before decline using Watterson estimator. Estimated effective population size before decline is 26938.49 and effective population size using total SNPs and 1/100 downsized SNPs are almost similar. Initial values of two parameters in the first step of DADI in this study are based on effective population size of previous analysis and papers. Initial value of parameter 1 (related to how much minke whale population reduced) is based on paper of Joe Roman and Stephen R. Palumbi (2003) (Roman and Palumbi 2003). Joe Roman and Stephen R. Palumbi reported that prewhaling populations is 6 to 20 times higher than present-day population estimates. So I determined 1/13 (thirteen is average six and 20) as parameter 1. Initial value of parameter 2 (related to when minke whale population reduced) is based on historical facts and estimated effective population size before deline. Before modern whaling era, slow whales were caught by men hurling harpoons from small open boats. Since 1860s, Norway invented many new techniques (steam-powered catcher boats, harpoon gun) and disseminated them worldwide. So I determined minke whale generations since 1860 / 2*N0 (effective population size before deline) as parameter 2. In boxplot of estimation parameter 1 of first step, I could not identify outliers and determined range 0.0007859 to

0.007860 as proper range of parameter 1 (Figure 3.5). In a case of parameter 2 of first step, I identified six outliers and exclude them, and determined range 0.000204 to 0.001969 as proper range of parameter 2 (Figure 3.5). In second step, I identified no outlier and normal distribution of parameter 1 of which mean is 0.003187 and median is 0.002467 (Figure 3.6-a). So I guessed that minke whale population diversity downsized to approximately 3.1%. Also I identified six outlier and skewed distribution of parameter 2 of which mean is 0.001584 and median is 0.007731 (Figure 3.6-b). Except six outliers, parameter 2 is in between 0.000204 and 0.001969. And in Figure 3.6-b, first interval (0-0.001) are most frequent. Based on this range, strong predicted time of minke whale declination during Holocene is between 194.36 and 902.31 years ago.

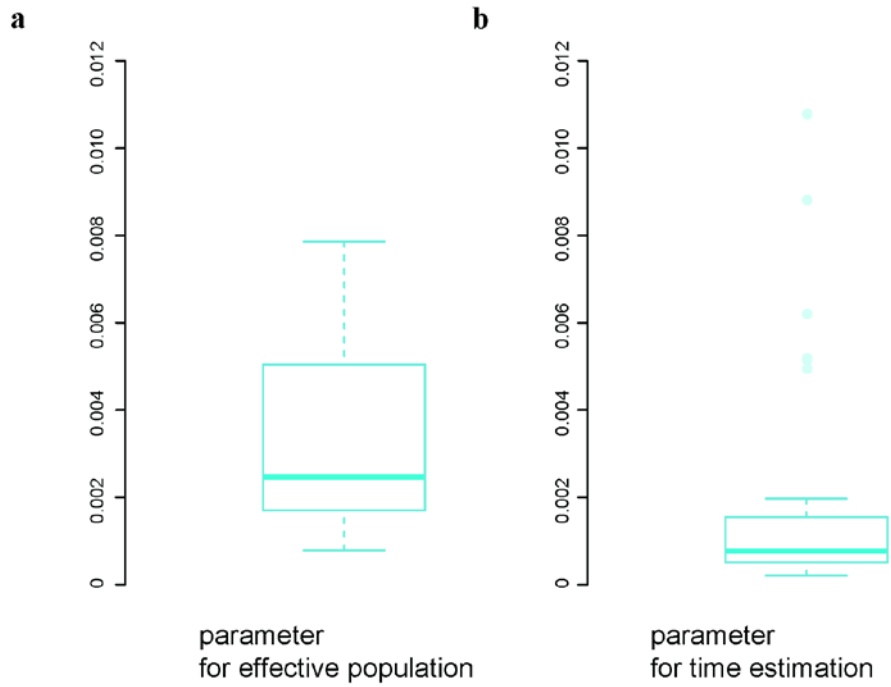

Figure 3.5. Boxplot of two parameters in the first step using DaDi to find proper ranges for the second step: (a) parameter for effective population and (b) parameter for time estimation.

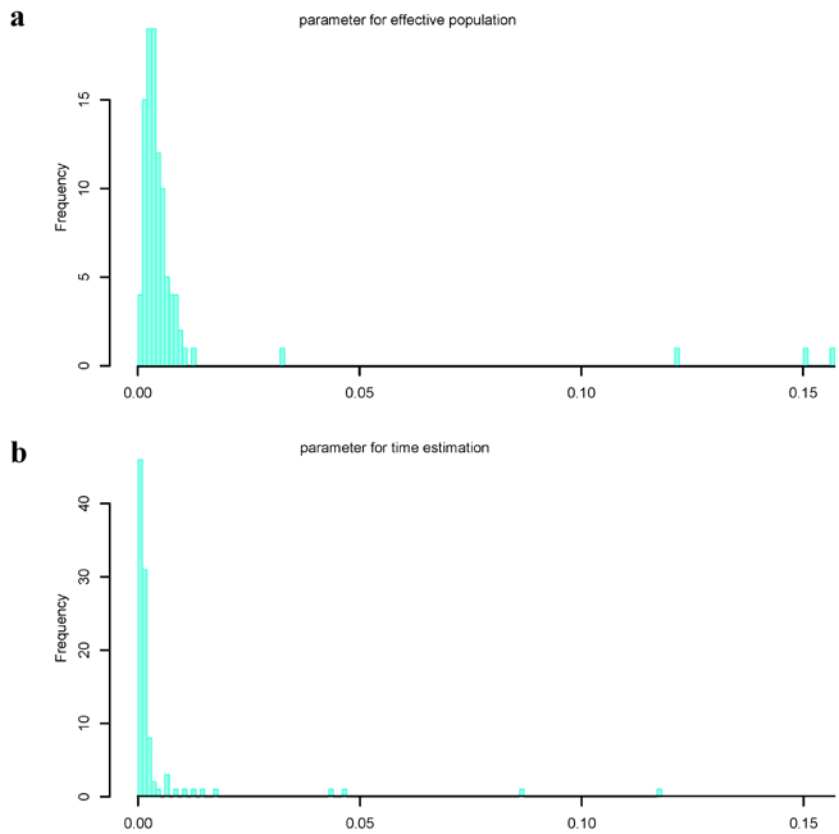

Figure 3.6. Distribution of two parameters in the second step using DaDi to infer the parameters: (a) parameter for effective population and (b) parameter for time estimation.

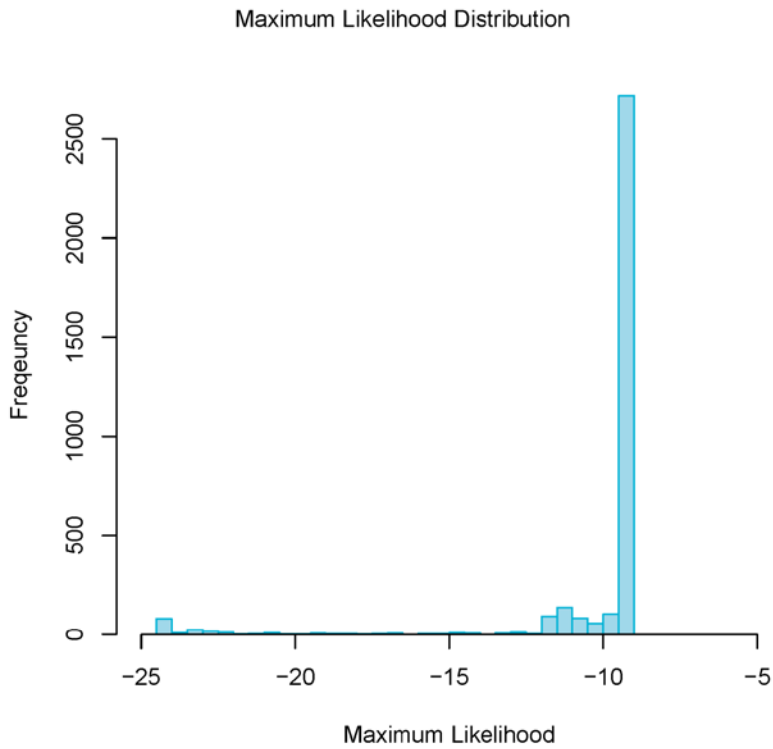

Figure 3.7 Distribution of maximum likelihood estimators in the second step using DaDi for inferring the two parameters.

3.5 Discussion

Minke whale (Balaenoptera acutorostrata) has been greatly hunted and in 1986 the International Whaling Commission (IWC) declared a moratorium on whaling commercially. The minke whale were harvested greatly after development of technology related to whaling and declined. So the preservationists would resolve by banning the even ‘aboriginal’ and ‘small-scale coastal’ whaling. The historical effective population size has declined to 3.1%, because of bottlenecks of minke whale. A population bottlenecks is a sharp reduction in size of a population owing to the environmental stochastic events such as scanty of foods or human activities. So I guess that full-scale decline of mink whale related to biological diversity during Holocene happened between 194.36 and 902.31 years ago.

Some of marine mammals went to extinction and over the last 500 years (or more), minke whale’s population size undoubtedly have decreased. The scale of world whaling has been global, continental and spanning bays and gulfs and pelagic waters and island shelves. Whaling began in antiquity (more than a thousand years ago) and continues to the current time. It supports the historical effective population size declining from antiquity to the present time. In fact, rather than the extreme climate change, the main cause of declining of minke whale effective population size can be regarded as the scanty of foods or human activities, as one might know. In medieval times, developed method for whaling was introduced and full-scale whaling had started. In Basque whaling, the whaling was made in 1059 and was spread to the Spanish Basque Country in 1150 and. By the 14th century they were making "seasonal trips" for whaling to the English Channel and southern Ireland. In Greenland whaling, beginning in the 1630s, whaling of the Dutch expanded into the open sea for regular and intensive whaling. In Japanese open-boat whaling, between 1500s and 1600s, whaling by the use of harpoons was spreads. I guessed that full-

scale whaling in medieval times is one of main reason of bottlenecks of minke whale with modern whaling.

Even though the human's whaling have been prohibited, I might think the cause of blocking the increment of minke whale's population size or their effective population size is the scanty of foods or current global climate change. The global climate change can lead to the migration of minke whale to the new spots, but the scanty of foods caused by migration of new environments or current spots' environmental change can lead to complicated and hard situation to the minke whale population. If possible, I think that the endeavors of protecting and increasing current effective population size of minke whale up to the proper level must be executed by human's own capability and I expect the increments of population size of minke whales up to now.

This chapter is paper which will be published in elsewhere as a partial fulfillment of Dong-Hyun Shin's Ph.D program.

Chapter 4. Multiple genes related to muscle identified through a joint analysis of a two-stage genome-wide association study for racing performance of 1,156 Thoroughbreds

4.1 Abstract

Thoroughbred, a relatively recent horse breed, is best known for its use in horse racing. Although myostatin (MSTN) variants have been reported to be highly associated with horse racing performance, the trait is more likely to be polygenic in nature. The purpose of this study was to identify genetic variants strongly associated with racing performance by using estimated breeding value (EBV) for race time as a phenotype. I conducted a two-stage genome-wide association study to search for genetic variants associated with the EBV. In the first stage of genome-wide association study, a relatively large number of markers (~54,000 single-nucleotide polymorphisms; SNPs) were evaluated in a small number of samples (240 horses). In the second stage, a relatively small number of markers identified to have large effects (170 SNPs) were evaluated in a much larger number of samples (1,156 horses). I also validated the SNPs related to MSTN known to have large effects on racing performance and found significant associations in the stage two analysis, but not in stage one. I identified 28 significant SNPs related to 17 genes. Among these, six genes have a function related to myogenesis and five genes are involved in muscle maintenance. To my knowledge, these genes are newly reported for the genetic association with racing performance of Thoroughbreds. It complements a recent horse GWAS of racing performance that identified other SNPs and genes as the most significant variants. These results will help to expand my knowledge of the polygenic nature of racing performance in Thoroughbreds.

4.2 Introduction

The Thoroughbred which is best known for horse racing is a relatively recent horse breed derived from a small number of Arabian stallions and native British mares in the 17th and 18th century England (Cunningham et al. 2001, Hill et al. 2002). To measure horse racing performance, various phenotypic value are used including race time, best race time, rank, position rates, annual earnings, and earnings per start (Ricard 1998). In particular, race time for each race is the most direct measure of speed and hence, makes it a suitable quantitative measure for evaluating the genetics of racing performance (Moritsu, Funakoshi and Ichikawa 1994, Oki, Sasaki and Willham 1994). In a horse breeding study, race time showed moderate heritability in the range of 0.1–0.3 (Mota, Abrahão and Oliveira 2005), with higher heritability for shorter distance race time. Previously, a study of 12,279 racehorses registered in the Korea Racing Authority, adjusted race time showed a 0.324 heritability (Park et al. 2011). However, as racing Thoroughbreds have multiple records for race time under different conditions and environmental factors, race time alone is not suitable as phenotypic value for GWAS. However, the estimated breeding value is a statistical prediction value that indicates how much each Thoroughbred has gene effects, so EBV was suitable for this GWAS as phenotypic value than others. Measures genetic variance and does not take into account other variable environmental factors.

A candidate gene approach to identify genetic variants associated with racing performance in Thoroughbreds revealed a single-nucleotide polymorphism (SNP; ECA18 g.66493737C/T) in the first intron of the equine myostatin gene (MSTN gene) (Hill et al. 2010c). Several genome-wide association studies (GWAS)

have confirmed this finding that SNPs within or near the MSTN gene are strongly associated with racing performance (Hill et al. 2010c, Binns, Boehler and Lambert 2010, Tozaki et al. 2010). Although MSTN variants have been reported to be highly associated with horse racing performance, this complex trait is more likely to be polygenic in nature. In the case of human athletic performance, more than 220 genes were reported to be associated with the phenotype (Bray et al. 2009). Similarly, I speculate that other SNPs not-related to MSTN could be associated with racing performance in Thoroughbreds.

To identify the genetic basis of horse racing performance, I used the EBV of race time as the phenotype for GWAS and conducted a joint-analysis of two-stage GWAS to search for significant genetic variants associated with race time. EBV was used as the phenotype as it only considers the genetic component of phenotypic variance, increasing the statistical power of the analysis. In the first stage of GWAS, a relatively large number of markers were evaluated in a relatively small number of samples. In the second stage, a relatively small number of markers identified as having large effects in the first stage were evaluated in a relatively large number of samples. This joint analysis of two-stage GWAS has been shown to increase the power to detect genetic association (Skol et al. 2006, Skol et al. 2007, Amos 2007). Using this approach, I identified 28 SNPs to be associated with the Thoroughbred racing performance. The SNPs were related to 17 genes including genes for myogenesis and muscle maintenance.

4.3 Materials and Methods

4.3.1 Ethics and blood collection

Korea Racing Authority (KRA) has established an animal experimentation ethics committee according to the Animals Protection Act 14 of Korea. This committee, titled Korea Racing Authority Institutional Animal Care and Use Committee (KRA IACUC) is composed of two external members and three internal members. One external member is a research veterinarian with experience in experimental animals (Veterinarian Act 2, paragraph 1, in Korea) and the other member is from an animal protection organization (Animals Protection Act 14, paragraph 2, in Korea). Three internal members are composed of the general manager (Chairman of KRA IACUC) and senior managers of the Equine Health & Welfare Section and the Disease Control & Prevention Section of the veterinary Center of KRA. KRA IACUC is under the auspices of the Equine Health & Welfare section of veterinary Center of KRA. The committee operates on a regular basis rather than approving each blood collection as blood collection of the race horses are performed routinely before every race. KRA operates experimental procedures including drug testing and ethics problem according to international guidelines, which is guaranteed by an affiliate association of the Korean government (Korea Racing Authority Act, Article 44) and is a member of the Association of Official Racing Chemist (AORC). In addition, the owners of the horses in KRA have granted permission for blood extraction for research and development purposes (Korea Racing Authority Act, Article 11, 12, 36).

Genomic DNA of the Thoroughbreds was isolated from blood collected for drug testing, health care and horse bloodlines management by the KRA. Legally, 25ml of blood, divided into three heparin tubes must be collected from the carotid

artery of all race horses participating in the race 2 to 3 hours before the race. The samples are stored at KRA, and two samples are used in drug testing, while the third sample is stored for either DNA identification or for additional drug testing. After the race, urines is collected from horses with high standing in the race for primary drug testing. If prohibited drugs are discovered in the urine, the third sample of blood collected before the race is used for secondary drug testing. From an animal welfare point of view, drug testing protects the racehorse from use of prohibited drug for enhanced racing performance.

The DNA information from the collected blood is used to preserve horse bloodlines and used in the genetic improvement of horse by genetic method. With the development of genomics, horse breeding and selection strategy is moving towards the use of SNP information derived from genomic DNA. So KRA uses archived blood samples of racehorses that passed the dope test both before and after the race.

For imported stallions and retired racehorses of KRA that does not participate in races, 10ml of blood was collected for DNA extraction. The collection of blood from these horses was approved as routine procedures for DNA information storage and horse bloodlines management. The genotyping SNPs in this study was conducted for the dual purpose of academic achievement and horse preservation within the legal and ethical framework. All blood-collection was performed by KRA veterinarians.

4.3.2 Estimated Breeding Value of race time

To improve the accuracy of the EBV, I simplified the animal model by reducing the factors deemed unnecessary (racing year, racing type, and type of

weight carried). A multiple-record animal model was used to estimate the genetic parameters as breeding value. The animal model used in this study is as follows:

$$\mathrm{Y} = \mathrm{Xb} + W_1\mathrm{v} + W_2\mathrm{pe} + \mathrm{Za} + \mathrm{e}$$

where Y = the vector of observations, b = the vector of fixed effects, v = the vector of random effects, pe = the vector of permanent environmental effect: common environment, a = the vector of individual additive genetic effect, Z = relationship matrix and e = the vector of residual error. X, W1, W2, and Z are coefficient matrices for b, v, pe and a, respectively.

Observations comprised a total of 262,326 race time records from 14,752 Thoroughbreds between 1994 and 2010 in Seoul Racecourse, Busan-Gyeongnam Racecourse of the KRA. The fixed effects used were racecourse, racing distance, country of foaling, sex, and age. Racing distances in the KRA were 1000, 1200, 1300, 1400, 1600, 1700, 1800, 1900, 2000, 2200, and 2300 m. Countries of foaling was divided into Korea or others with sex divided into female, male, or gelding. Random effects were moisture, jockey, weight of handicap, and trainer. Different performances in common environment represented the difference between the repeatability and heritability. Repeatability was explained by the permanent environment of the horse during the rearing period. The relationship matrix Z included all animal racers, nonracers, reproducers, and non-reproducers. The vector of the individual genetic effect, EBV, is the coefficients vector of Z matrix. Residual error, e, represent inexplicable factors such as temporary environmental effects (e.g., racing strategy, race condition, etc.).

All parameters were estimated using the ASREML program (Gilmour, 2000) facilitated by the derivative-free restricted maximum likelihood (DFREML) method for a single-trait animal model. Using this model, I calculated the EBV for race time.

4.3.3 Initial genome-wide scan: stage 1

DNA samples were obtained from a total of 240 Thoroughbreds registered in the KRA. The 180 racehorses and 60 stallions were genotyped for the initial genome-wide scan using EquineSNP50 Genotyping BeadChips (Illumina, San Diego, CA). The chip includes 54,602 SNPs that are uniformly distributed on the 31 equine autosomes and X chromosome (average density of 1 SNP per approximately 43 kb) from the EquCab2 SNP database of the horse genome. All samples were genotyped in the National Instrumentation Center for Environmental Management (NICEM) at Seoul National University. I excluded SNPs with a missing rate of >0.05, minor allele frequency (MAF) of <0.05, and Hardy–Weinberg equilibrium (HWE) test P-value of <0.001. SNPs on the X chromosome were also excluded, retaining 41,371 autosomal SNPs for analysis. Association analysis of stage 1 was conducted on the basis of linear regression using the software PLINK (Purcell et al. 2007).

4.3.4 Joint analysis in replication study: stage 2

For stage 2 association analysis, 190 SNPs selected in stage 1 and the two SNPs (BIEC2-417495 and BIEC2-417274) related to MSTN, were included. For the replicate study, 916 Thoroughbreds with more than four race records in Korea were genotyped using the customized Equine BeadXpress SNP Chips (Illumina) at NICEM. A total of 172 SNPs were successfully genotyped on the Equine SNP Chips I applied the same quality control criteria as in stage 1 and retained 158 SNPs for further analysis. For the joint analysis, I combined the data from 158 SNPs from 240 Thoroughbreds from stage 1 and 916 Thoroughbreds from stage 2. The joint association analysis was based on linear regression implemented in PLINK. The

same analyses were also conducted on the 2 SNPs related to MSTN gene. Haploview program was used to visualize several linkage disequilibrium (LD) blocks of interest regions (Barrett et al. 2005).

4.4 Results

4.4.1 Initial genome-wide scan of 240 Thoroughbred SNPs: stage 1

In this GWAS, I compared the genotypes of Thoroughbreds with the EBV as a phenotypic value. I conducted the initial genome-wide scan in a population of 240 Thoroughbreds. The chromosome sorting is displayed as a Manhattan plot (Figure 4.1). Initial analysis of my data found 3,919 SNPs that were associated with the EBV for race time (unadjusted P-value, <0.05), though none of these SNPs exceeded the threshold of multiple tests ($P < 2.41*10^{-06}$, equivalent to P = 0.05 after Bonferroni correction). At this stage, speculation on the plausibility and biological significance of these candidate SNPs is not mature because of the inevitably high false-positive rate from several tens of thousands of tests are performed on the same data set. Replication of these findings in a larger population was required to identify significant SNPs associated with racing performance. I selected 190 SNPs that were evenly distributed on the equine chromosomes for the second stage.

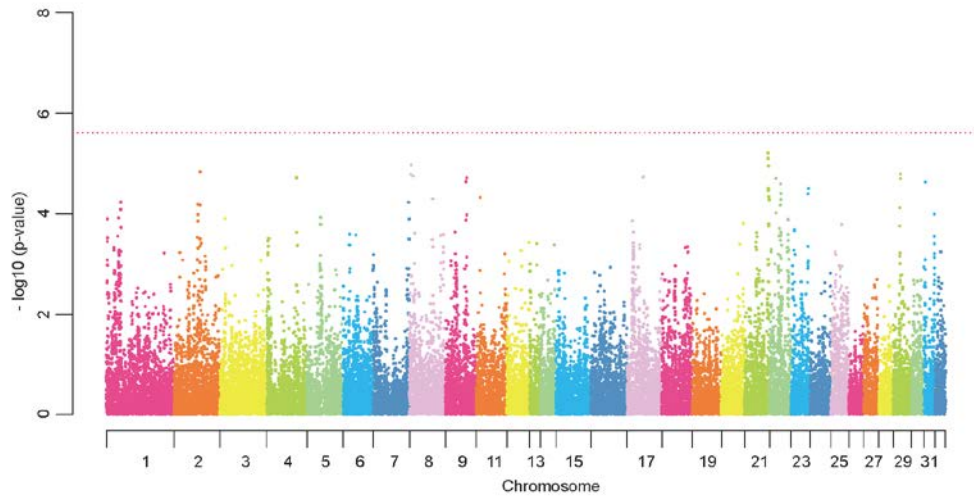

Figure 4.1. Manhattan plot of genome-wide association with the EBV for race time (Stage 1). In this plot, results are plotted as negative log-transformed P-values from the genotypic association test (observed –log10 P-values by position); the red horizontal dotted line indicates $P = 2.41*10^{-06}$ which means P = 0.05 after Bonferroni correction.

4.4.2 Joint analysis of 1,156 Thoroughbreds for replication study: stage 2

In the second stage, 916 additional samples were randomly selected from the Thoroughbred population of the KRA not included in stage one. To further test the association with the EBV, I genotyped 190 of the most associated SNPs from stage 1 in the 916 Thoroughbreds. After data quality control procedures, 158 SNPs were available for stage 2 for a total of 1,156 Thoroughbreds. These SNPs covered 119 distinct chromosomal regions defined by a maximal distance between two SNPs of <100 kb. Out of the 119 regions, 92 contained only one SNP, and 27 contained more than two SNPs. In joint analysis using the combined population data, 28 SNPs (17.7% of the 158 SNPs) achieved genome-wide significance criteria (P-value = 0.000632911, equivalent to P = 0.1 after Bonferroni correction) (Figure4.2, Table 4.1).

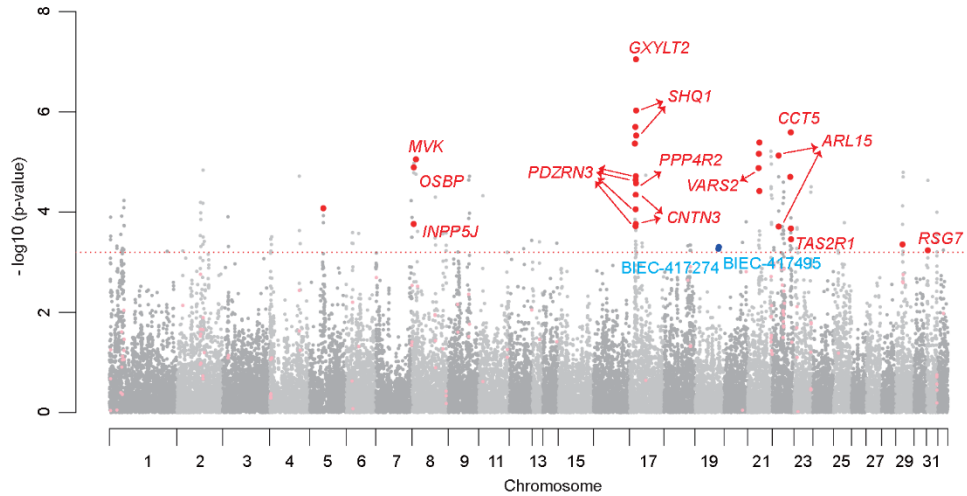

Figure 4.2. Manhattan plot of genome-wide association with the EBV for race time (Stage 1). In this plot, results are plotted as negative log-transformed P-values from the genotypic association test (observed –log 10 P-values by position); Result of stage 1 is shown in two different gray colors. Odd chromosome numbers are in dark grey, and even chromosome numbers in light grey. The red horizontal dotted line indicates stage 2 threshold. SNPs in stage 2 are shown in two red colors. Significant SNPs with a P-value less than threshold are in red, and others are in light red. Reported SNPs that associated with racing performance of Thoroughbred are in blue.

Table 4.1. List of SNPs with a P-value of < 0.000632911 (equivalent to P = 0.1 after Bonferroni correction) based on the linear regression model of stage 2 data (n=1,156)

	Chr	BP	Neareat Gene (Ensemble Gene ID)	SNP Type	Minor Allele	Major Allele	Stage 1			Stage 2		
							P-value	MAF	SNP effect	P-value	MAF	SNP effect
BIEC2_906777	5	40747218	- (ENSECAG00000019204)	InterGenic	A	G	1.63.E-04	0.356	0.278	8.38.E-05	0.330	0.141
BIEC2_1026019	8	5906904	INPP5J (ENSECAG00000006664)	InterGenic	A	G	1.64.E-05	0.159	-0.4315	1.73.E-04	0.190	-0.163
BIEC2_1026200	8	6126833	OSBP2 (ENSECAG00000015443)	InterGenic	G	A	1.06.E-05	0.107	-0.5102	1.28.E-05	0.142	-0.215
BIEC2_1029757	8	11590300	MVK (ENSECAG00000023067)	InterGenic	C	A	1.75.E-05	0.119	-0.4715	8.85.E-06	0.137	-0.219
BIEC2_330101	16	14185461	-	InterGenic	C	A	1.39.E-04	0.254	0.3328	4.29.E-06	0.216	0.195
BIEC2_330360	16	15550375	-	InterGenic	A	G	8.12.E-04	0.388	0.2536	2.00.E-06	0.371	0.167
BIEC2_330495	16	16289366	CNTN3 (ENSECAG00000013575)	Genic	A	G	3.81.E-04	0.298	0.2847	1.91.E-04	0.271	0.143
BIEC2_330509	16	16334663		InterGenic	G	A	4.43.E-04	0.297	0.2818	8.79.E-05	0.274	0.150
BIEC2_330558	16	16948655	PDZRN3 (ENSECAG00000014864)	Genic	G	A	6.02.E-04	0.377	0.2565	2.28.E-05	0.362	0.150
BIEC2_330572	16	16993987		Genic	A	G	7.93.E-04	0.367	0.2484	4.53.E-05	0.358	0.145
BIEC2_330575	16	16999220		Genic	A	G	2.32.E-04	0.354	0.2734	1.73.E-04	0.339	0.134
BIEC2_330578	16	17001997		Genic	G	A	6.02.E-04	0.377	0.2565	1.92.E-05	0.363	0.152
BIEC2_330677	16	17421511	PPP4R2 (ENSECAG00000000689)	Genic	A	G	6.13.E-04	0.377	0.2562	2.66.E-05	0.359	0.149

BIEC2_330691	16	17546589	GXYLT2 (ENSECAG00000008483)	Genic	G	A	7.05.E-04	0.371	0.2611	8.94.E-08	0.355	0.191
BIEC2_330725	16	17763943	SHQ1 (ENSECAG00000015673)	InterGenic	G	A	4.49.E-04	0.325	0.2802	9.40.E-07	0.316	0.180
BIEC2_330739	16	17790340		InterGenic	G	A	6.71.E-04	0.315	0.2672	2.97.E-06	0.310	0.172
BIEC2_527753	20	29897398	VARS2 (ENSECAG00000018202)	Genic	A	G	4.60.E-04	0.219	0.3011	1.32.E-05	0.172	0.195
BIEC2_527879	20	30126579	- (ENSECAG00000015285)	InterGenic	A	G	4.60.E-04	0.219	0.3011	6.86.E-06	0.172	0.202
BIEC2_529755	20	32127869	- (ENSECAG00000017401)	InterGenic	A	C	5.23.E-04	0.256	0.2969	3.80.E-05	0.210	0.171
BIEC2_529760	20	32131071		InterGenic	C	A	2.39.E-04	0.296	0.2909	4.09.E-06	0.239	0.180
BIEC2_554645	21	18154976	ARL15 (ENSECAG00000014970)	Genic	A	G	1.98.E-05	0.218	-0.3696	7.44.E-06	0.213	-0.188
BIEC2_554739	21	18348602		InterGenic	C	A	9.53.E-05	0.265	-0.3267	1.94.E-04	0.256	-0.146
BIEC2_568963	21	47502588	-	InterGenic	A	G	4.70.E-04	0.471	-0.2539	1.98.E-05	0.474	-0.149
BIEC2_569862	21	48967346	CCT5 (ENSECAG00000019192)	InterGenic	G	A	2.59.E-04	0.373	-0.2739	2.57.E-06	0.381	-0.165
BIEC2_570062	21	49384234		Genic	A	C	1.30.E-04	0.404	0.2819	2.12.E-04	0.386	0.127
BIEC2_570485	21	49894191	TAS2R1 (ENSECAG00000005160)	InterGenic	A	G	3.40.E-04	0.215	0.3012	3.48.E-04	0.209	0.148
BIEC2_732151	28	17405077	-	InterGenic	A	C	7.52.E-05	0.290	0.3151	4.42.E-04	0.274	0.134
BIEC2_814518	30	3869336	RGS7 (ENSECAG00000009422)	InterGenic	A	G	2.33.E-05	0.388	0.3152	5.79.E-04	0.390	0.119

Table 4.2. List of two reported SNPs that associated with racing performance of Thoroughbreds.

	Chr	BP	Neareat Gene (Ensemble Gene ID)	SNP Type	Minor Allele	Major Allele	Stage 1			Stage 2		
							P-value	MAF	SNP effect	P-value	MAF	SNP effect
BIEC2_417274	18	65868604	-	InterGenic	C	A	0.801	0.498	-0.02	5.38.E-04	0.491	-0.124
BIEC2_417495	18	67186093	-	InterGenic	G	A	0.526	0.483	-0.05	4.97.E-04	0.480	-0.123

The most significant SNP (P = $8.941*10^{-08}$ in joint analysis), BIEC2_330691, identified by my two-stage GWAS was found on chromosome 16 with the other 11 most significant SNPs. Out of the 11 most significant SNPs, three were top-ranking and located in the sixth intron of the gene GXYLT2 (included in block 10 of Figure 4.3). Twelve of the top 27 SNPs were located together, (contained seven of 119 distinct chromosomal regions), spanning a 3.61 Mb region on chromosome 16 (chr16: 14.18–17.79 Mb). This regions have not been previously reported as being related to the racing performance of Thoroughbreds (Figure 4.2 and Table 4.1). Four of these 12 significant SNPs were in PDZRN3 genes and other four containing BIEC2_330691 was in a 0.37-Mb interval (chr16: 17.74–17.79 Mb). LD calculations were performed within the 3.6-Mb region on chromosome 16 (chr16: 14.18–17.79 Mb). Thirteen discrete LD blocks were identified in the 3.6-Mb peak of association on chromosome 16. (Figure 4.3)

Six significant SNPs including the fourth most significant SNPs, BIEC2_569862, were located on chromosome 21. These SNPs were not located on a distinct chromosomal region but on two regions. One is chr21: 18.15-18.34 (2 SNPs) and another is chr21:47.50-49.89 (4 SNPs). Four significant SNPs were on chromosome 20. Two SNPs were on one distinct chromosomal regions (chr20: 32.12-32.13, 2 SNPs) and other two were mostly near each other (chr20: 29.89-32.13). Three significant SNPs were dispersed in chromosome 3. Chromosome 5, 28, 30 each had one significant SNP.

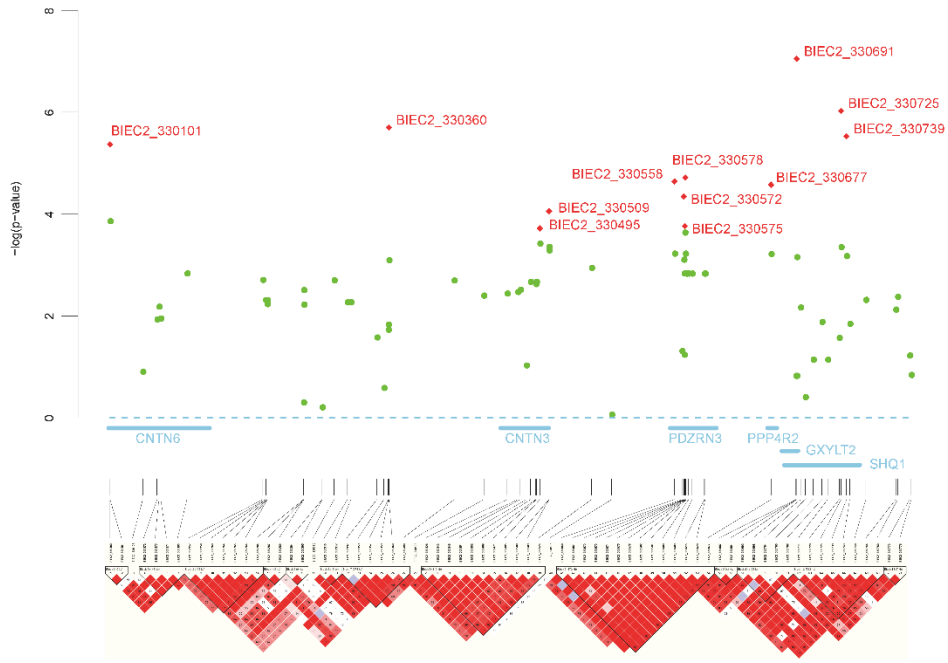

Figure 4.3. Location of the association signal and pairwise linkage disequilibrium (LD) surrounding the most associated SNPs on chromosome 16:14.18-17.79 Mb. LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has six nearest genes related to significant SNPs. The most associated SNP, BIEC2-330691, lies in the gene GXYLT2 and are in small LD. Four SNPs lies in the gene PDZRN3 and are in almost complete LD.

4.4.3. Candidate genes associated with racing performance of Thoroughbreds

The transcriptional content of 28 significant SNPs was assessed using the EquCab2.0 assembly and annotation of the horse genome. The genes were annotated using the gene IDs from ENSEMBL genome browser EquCab2 (http://www.ensembl.org/). I collected the genes whose entire information was located within or near (±10kb) the SNPs. Twenty-five SNPs associated with EBV for race time belonged to 17 genes (Ensemble Gene ID) which were mainly related to muscle terms. (Table 4.1) Twenty SNPs of 25 SNPs were located in 13 protein-coding genes. Five of these identified genes contained more than one SNP. I found that 11 of the 13 genes associated with EBV were associated with muscle. Of the 11 genes related to muscle, six play a role in myogenesis. These six genes are glucoside xylosyltransferase 2 (GXYLT2), SHQ1 homolog, Saccharomyces cerevisiae (SHQ1), PDZ domain-containing ring finger 3 (PDZRN3), ADP-ribosylation factor-like 15 (ARL15), oxysterol binding protein 2 (OSBP2), and mevalonate kinase (MVK).

GXYLT2 encodes the 37.4-kDa proteoglycan core protein, glucoside xylosyltransferase 2. Proteoglycans, one of the macromolecule groups in the extracellular matrix, impact the regulation of muscle cell proliferation and differentiation during myogenesis (Velleman et al. 2012). The fact that the SNP with the most significant p-value was located in the intron of GXYLT2 suggests that this gene is a likely to be associated with racing performance (Figure 4.3). SHQ1 is an SHQ1 is an assembly factor required for the assembly of telomerase RNPs (Grozdanov et al. 2009). O'Connor et al. (2009) showed that the telomerase activity in muscle stem cells is retained in old and age-specific telomere shortening and is not detected in the old differentiated muscle fibers in other mammals (O'Connor, Carlson and Conboy 2009). In addition, the second and fifth most significant SNPs, BIEC2_330725 and BIEC2_330739, respectively were located near the SHQ1 gene

(Figure 4.3). Ko (2006) suggested that PDZRN3 plays an crucial role in the differentiation of myoblasts into myotubes by acting of myogenin (Ko et al. 2006). PDZRN3 contains four significant SNPs in the 3.6-Mb region on chromosome 16 (Figure 4.3). Two significant SNPs were found in the ARL15 gene on chromosome 21, which encodes the ADP-ribosylation factor-like 15. ADP-ribosylation factor is reported to be critical regulator of myoblast fusion (Bach et al. 2010) (Figure 4.7). OSBP2 encodes oxysterol binding protein 2 and is highly regulated during transitional-phase post-differentiation induction during myogenic differentiation (Szustakowski et al. 2006) (Figure 4.10). Therefore, for before oxyterol operation, mevalonate kinase (encoded by MVK gene) which is an intermediate substance of the oxysterol synthesis pathway may be important in myogenesis (Liscurn 2002) (Figure 4.11).

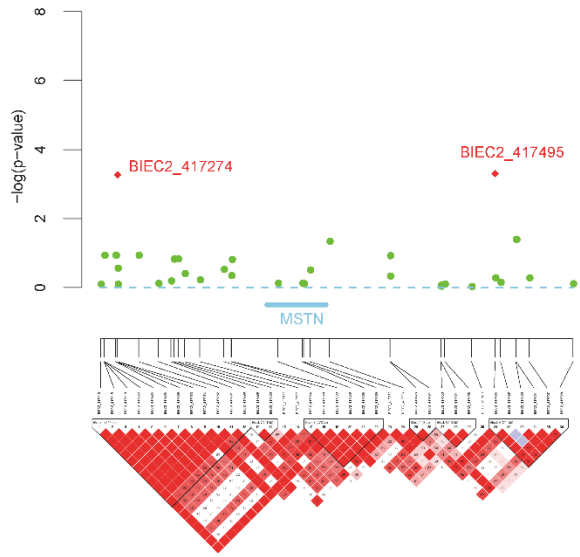

Figure 4.4. Location of the association signal and pairwise linkage disequilibrium (LD) surrounding two reported SNPs related to gene MSTN. LD pattern is depicted using stage 1 data (n = 240). Association signals of two SNPs are shown in combined dataset of stage 2 (red diamond, n = 1,156).

Five other genes identified in this study are related to muscle maintenance. The five genes include contactin 3 (CNTN3); Chaperonin Containing TCP1, Subunit 5 (CCT5); valyl-tRNA synthetase 2, mitochondrial (VARS2) and inositol polyphosphate-5-phosphatase J (INPP5J), Protein Phosphatase 4, Regulatory Subunit 2 (PPP4R2). CNTN3 encodes contactin 1 protein, which is a member of the immunoglobulin superfamily. Jelinsky (2010) defined a set of tendon-selective genes present in both adult rat and human tendons that contained CNTN3 (Jelinsky et al. 2010). Tendon connects muscle with bone, so play a crucial role in muscle functions. Two SNPs were located in the CNTN3 gene on chromosome 16 (Figure 4.3). Two SNPs were also found in the CCT5 gene on chromosome 21. TCP-1 Ring Complex encoded CCT is known to play a synergistic role in the process of actin folding (Kim, Löwe and Hoppe 2008). VARS2 encodes valyl-tRNA synthetase 2. Phosphorylation of aminoacyl-tRNA synthetases could play a role in the regulation of protein synthesis in a similar manner as insulin regulates muscle (Kimball, Vary and Jefferson 1994). INPP5J encodes inositol polyphosphate 5-phosphatase J; its role as a second messenger in signal transduction has been well established in many cell types involved in skeletal muscle signaling (Moschella et al. 1995). PPP4R2 is important partner of SMN (Survival of Motor Neuron) protein which affect modulation of skeletal muscle (Bosio et al. 2012).

4.4.4 Comparison of previously reported polymorphisms associated with racing performance

Hill (2010) and Binns (2010) previously reported a peak of association with best race distance on chromosome 18, and each identified the most important SNPs as BIEC2-417495 (Hill et al. 2010d) and BIEC2-417274 (Binns et al. 2010), respectively. These two SNPs did not exceed the threshold genome-wide

significance criteria (P-value = 0.000632911, equivalent to P = 0.1 after Bonferroni correction) in stage 1 of this study. However, I tested the relationship of these SNPs with EBV once more for comparison with my result in the second stage. Two reported SNPs reached the threshold in the joint analysis of this study (Figure 4.2 and Table 4.1). Moreover, LD blocks across a 1.7-Mb region on chromosomes in a study by Hill (2010) were almost identical to the LD blocks using my stage 1 data (Hill et al. 2010c). (Figure 4.4)

4.4.5 Evaluation of SNP effects on EBV

Using the results of the linear regression model, I wanted to identify 28 SNP effects in this study. To evaluate the effects of 28 significant SNPs and compared them with the two SNPs of MSTN, I made 30 plots for the effect allele score with EBV (Figure 4.5, Figure 4.6, Figure 4.13-15). For each subject in this population, each minor allele was scored for the EBV. Each box plot shows the relationship between the effect allele number and EBV. Each allele had either a positive or negative effect on the racing performance of Thoroughbreds. In chromosome 16, all 12 SNPs had a positive effect on racing performance (Figure 4.6). In other chromosomes, nine out of 15 SNPs had positive effects while the remaining seven had negative effects. (Figure 4.13-15). Both SNPs related to MSTN showed a negative effect on the EBV (Figure 4.5).

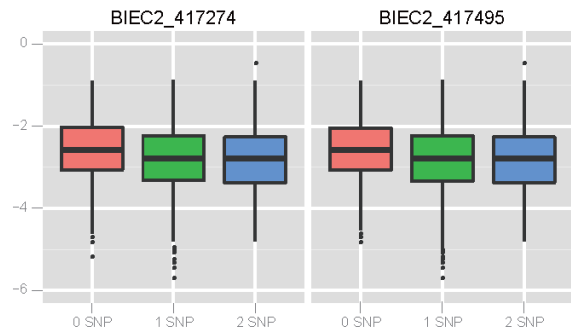

Figure 4.5. Boxplots show cumulative effect for EBV of the effect allele number of significant 2 SNPs near MSTN. These two SNPs have negative effects.

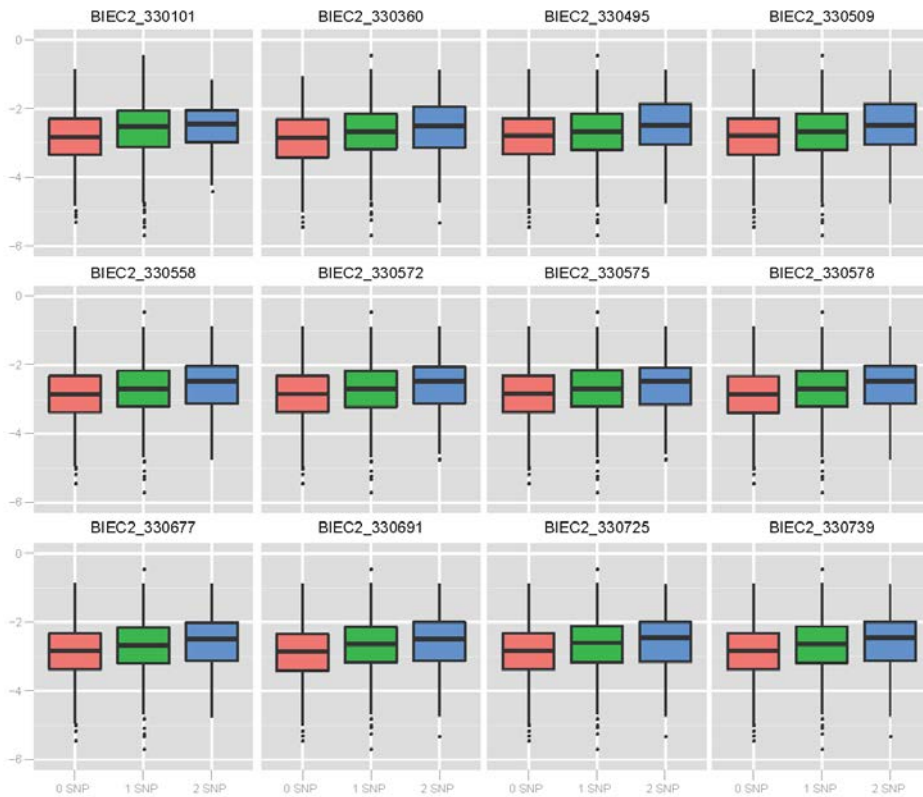

Figure 4.6. Boxplots show cumulative effect for EBV of the effect allele number of significant 12 SNPs on chromosome 16. All 12 SNPs on chromosome 16 have positive effects.

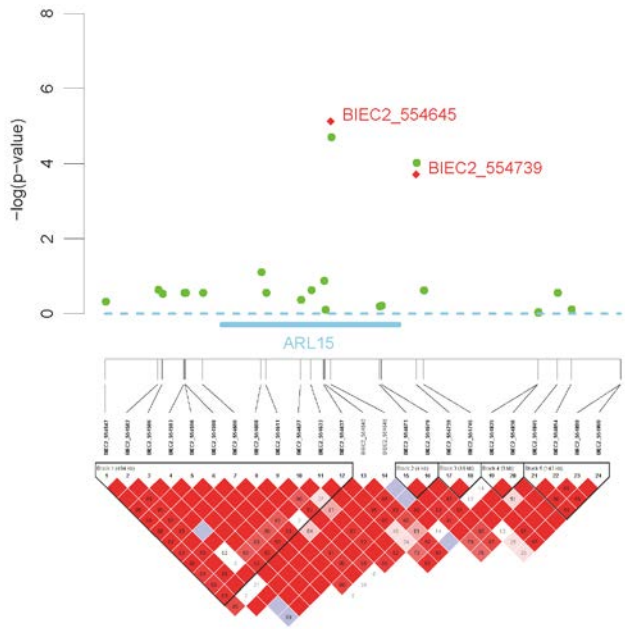

Figure 4.7. Location of the association signal and pairwise linkage disequilibrium LD) surrounding two on chromosome 21. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *ARL15* genes related to two significant SNPs.

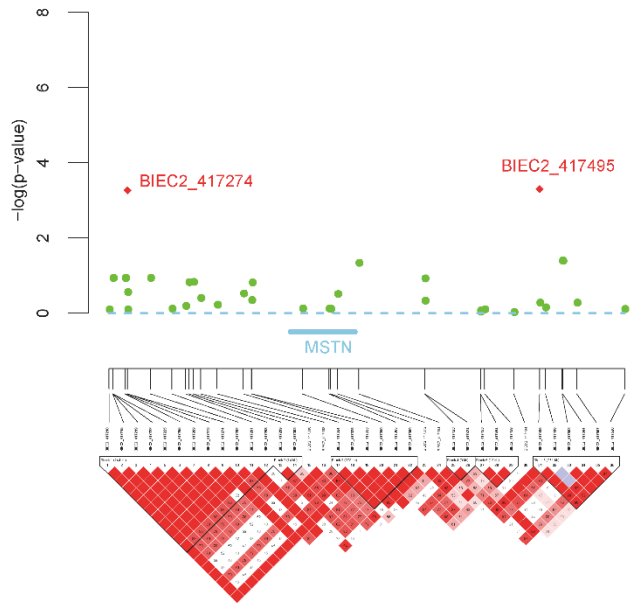

Figure 4.8. Location of the association signal and pairwise linkage disequilibrium LD) surrounding four on chromosome 21. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *CCT5*, *TAS2R1* genes related to each two, one significant SNPs.

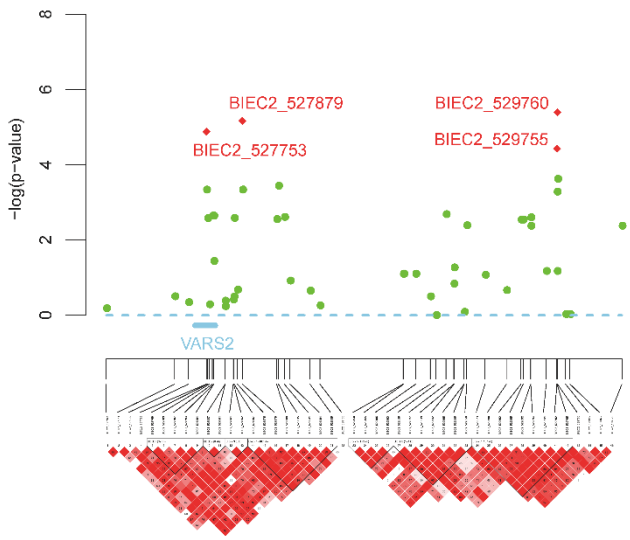

Figure 4.9. Location of the association signal and pairwise linkage disequilibrium LD) surrounding four on chromosome 20. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *VARS2* genes related to one significant SNPs.

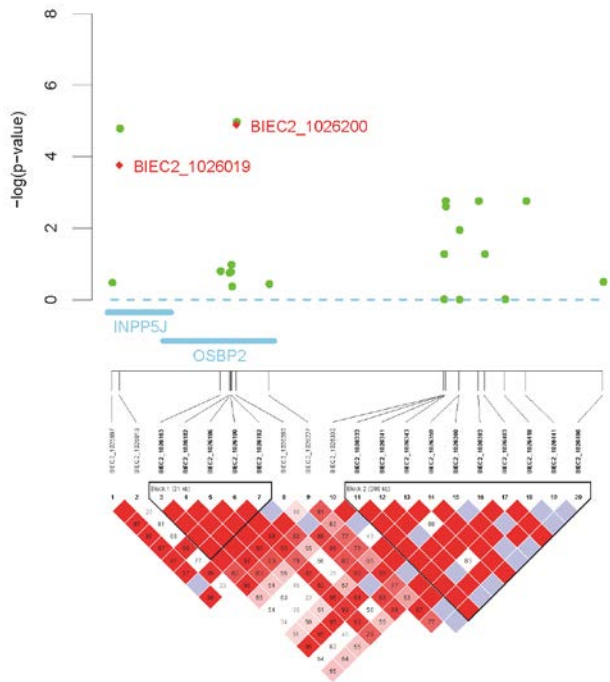

Figure 4.10. Location of the association signal and pairwise linkage disequilibrium LD) surrounding four on chromosome 8. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *INPP5J*, *OSBP2* genes related to each one significant SNP.

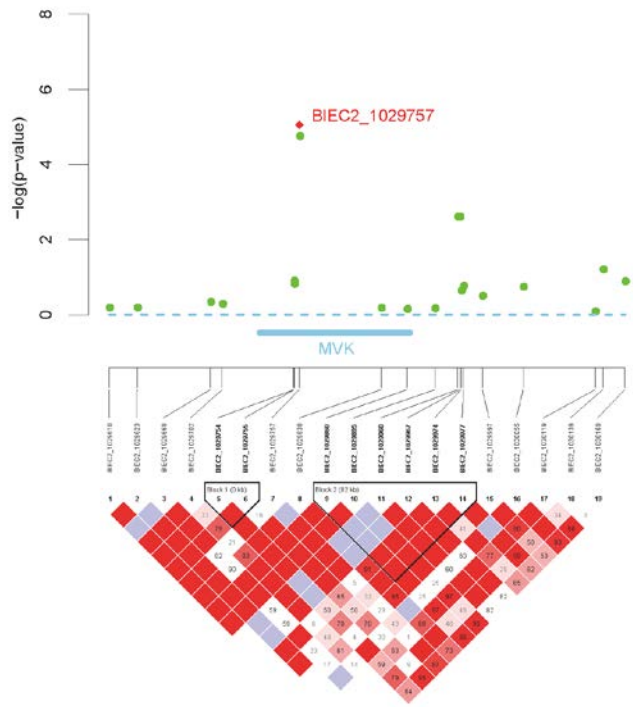

Figure 4.11. Location of the association signal and pairwise linkage disequilibrium LD) surrounding four on chromosome 8. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *MVK* genes related to one significant SNP.

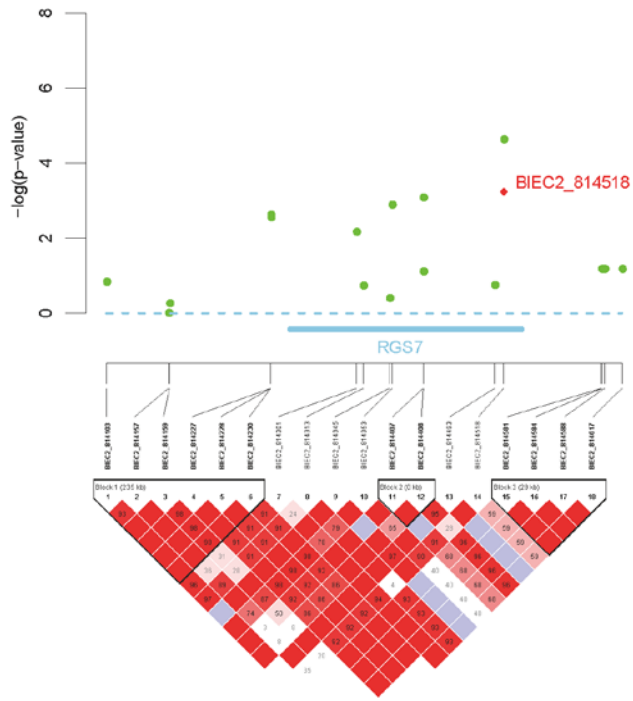

Figure 4.12. Location of the association signal and pairwise linkage disequilibrium LD) surrounding four on chromosome 30. This LD pattern is depicted using stage 1 data. Association signals are shown for all SNPs genotyped in stage 1 samples (green circles, n = 240); significant SNPs in combined dataset of stage 2 (red diamond, n = 1,156). This region has *RGS7* genes related to one significant SNP.

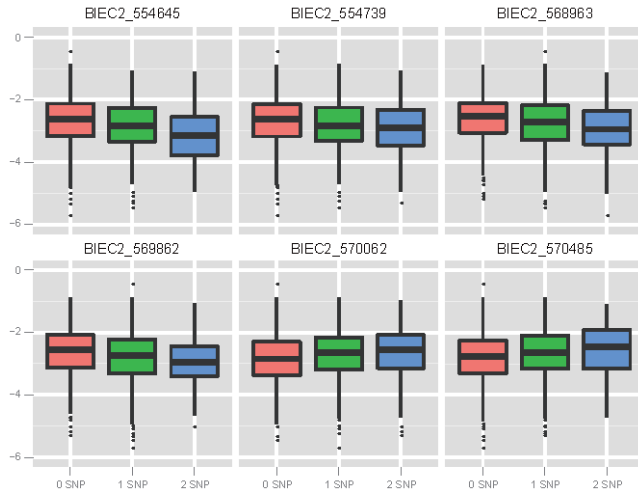

Figure 4.13. Boxplots show cumulative effect for EBV of the effect allele number of significant 6 SNPs on chromosome 21.

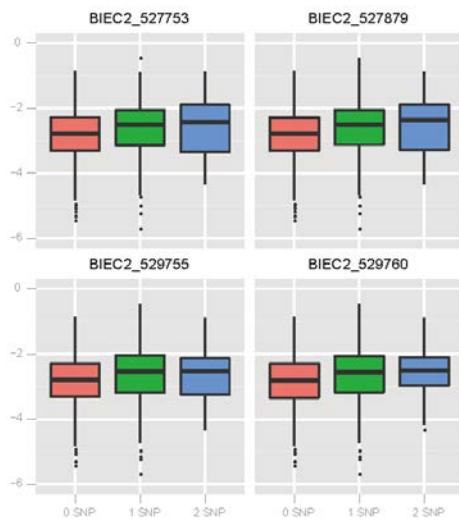

Figure 4.14. Boxplots show cumulative effect for EBV of the effect allele number of significant 4 SNPs on chromosome 20.

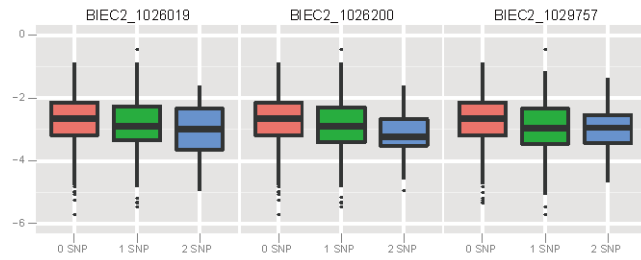

Figure 4.15. Boxplots show cumulative effect for EBV of the effect allele number of significant 3 SNPs on chromosome 8.

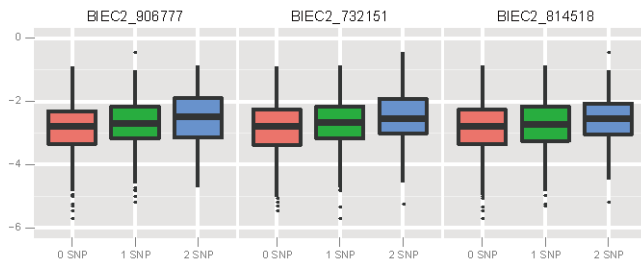

Figure 4.16. Boxplots show cumulative effect for EBV of the effect allele number of significant each one SNPs on chromosome 5, 28, 30.

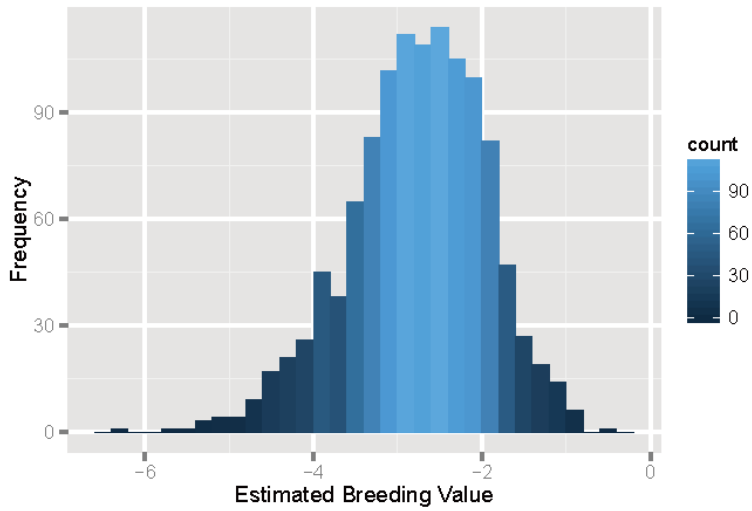

Figure 4.17. Histogram of EBVs for 1,156 Thoroughbreds.

4.5 Discussion

To identify SNPs underlying the racing performance in Thoroughbred, I applied a two-stage GWAS to search for genetic variants associated with the EBV for race time. I present the results of a joint-analysis of two-stage GWAS involving 1,156 Thoroughbreds. In the first stage of my GWAS, a relatively large number of markers were evaluated in a relatively small number of samples. In the second stage, a relatively small number of markers with large effects were evaluated in a much larger number of samples. In data set of the first and second stage were combined in a joint analysis to increase the power to detect genetic associations (Skol et al. 2006, Amos 2007, Skol et al. 2007). Using this approach, I identified 28 SNPs associated with racing performance. The SNPs were in genes related to myogenesis and muscle maintenance that have not been previously reported.

Various racing traits such as race time, best race time, rank, starting position, and annual earnings are used to measure racing performance (Ricard 1998). Out of these, race time of each race is the most direct measure of speed and a suitable quantitative measure for evaluating the racing performance of horses (Moritsu et al. 1994). However, racing performance is a complex phenotype affected by a variety of different factors such as management of Thoroughbreds, climate, geographical region, age, and reproductive status. In addition, management of exercise conditioning and nutrition has been shown to account for about 65% of racing capacity in the development of elite Thoroughbreds. Even considering these factors, previous studies show that race time had a heritability in the range of 0.1–0.3. Also, a significant proportion of variation in athletic ability has been shown to be heritable (Gaffney and Cunningham 1988).

In animal breeding, EBV is used to rank breeding stock for selection as it only considers the genetic effect on phenotype and predicts the genetic value of an individual based on the phenotypes measured in their relatives. Breeding value is the sum of gene effects of an animal as measured by the performance of its progeny. To exclude other effects and increase the power of the analysis, I calculated a composite phenotype for genetic merit (i.e., EBV) of race time. Each Thoroughbred racehorse has multiple records for race time and each record was achieved under different conditions and environment factors. This makes using race time as a phenotypic value for GWAS difficult. As EBV is a single numerical prediction value that indicates how each Thoroughbred contributes to its progeny, it is suitable as a phenotype for GWAS. However, with the use of EBV as a phenotype in GWAS, inflation becomes a problem.

To investigate the population stratification in stage 1 data, I calculated the lambda value by the statistical package R. The lambda value of stage 1 data was 1.38, which is much higher than that of other studies. The inflation of significant association signals most likely resulted from the relatedness of the horses studied, which contain a massively structured population and high LD. In addition, because EBV comprises only genetic factors, the inflation value increases more than expected. However, inflation is a normal phenomenon in animal GWASs and not a problem for detecting significant SNPs. I excluded relatedness from race time as much as possible by using the EBV and conducted stringent multiple testing (Bonferroni correction) to reduce false-positive errors. Moreover, I performed association analysis using a combined dataset in stage 2 to detecting significant SNPs.

I carried out a two-stage GWAS to search for common variants associated with EBV for race time. GWAS is a promising method to discover common genetic variants that could explain diseases or economic phenotypes of animals and plants.,

Due to the high cost of genotyping, often a two stage design is used, in which a portion of the sample is genotyped on a large number of markers in stage 1 and a certain number of these markers are followed-up by genotyping the remaining samples in stage 2. Compared with one-stage designs that genotype all samples on all markers, well-constructed two-stage association designs maintain power and reduce cost (Thomas, Xie and Gebregziabher 2004). Most often, two-stage design is used to include a replication study in the second stage to define the findings that reach statistical thresholds. I used an alternative strategy, in which the second stage analysis was a joint analysis that combines the data from both stages. This increases the power to detect genetic association (Skol et al. 2006). Skol (2006) reported that joint analysis is far more powerful than replication-based analysis (Skol et al. 2006) and recommended this strategy for two-stage GWAS studies that genotype a large portion of the samples in stage 1 and genotype a large portion of SNPs in stage 2. In this study, the number of selected markers in stage 2 was smaller than expected. However, as Thoroughbreds have stronger and larger LD than other species because of selective breeding over the three centuries, a small portion of the markers still retained power in the joint analysis stage. This indicated that a low number of markers could explain large regions of the genome. My study was sufficiently powered to reliably detect SNPs moderately associated with the racing performance of Thoroughbreds.

Twenty-eight loci of the studied clearly achieved a genome-wide level of significance. This contrasts with GWASs in other phenotypes linked to racing performance such as racing distance, where the most associated SNP related to MSTN had a P-value of $<1.61*10^{-9}$ (Hill et al. 2010c). The most significant SNP in my study had a P-value of $< 8.9*10^{-8}$, and 28 SNPs exceeded the threshold after multiple testing. However, no overlap was observed between the results of this study and those of previously published horse GWAS studies (Hill et al. 2010c, Gu et al.

2010, Hill et al. 2010a, Hill et al. 2010b). In previous studies, rare mutations with large effects on the racing performance of Thoroughbreds have been identified in genes such as MSTN. Several possible explanations exist for this discrepancy in the results. Tens of thousands of tests performed in GWASs increases the likelihood that associated SNPs in the initial genome-wide scan represent false-positives arising by chance. Another possible explanation for this lack of overlap may be that EBV is a more directly related to phenotype (Finlay et al. 2012), which would significantly expand the power of GWASs. The identification of multiple SNPs involved in disparate biological pathways supports this notion.

In humans, numerous GWAS and candidate gene studies have revealed more than 220 gene loci (Bray et al. 2009) that influence athletic performance. This hints at the fact that racing performance has the potential to be a polygenic in nature. However in Thoroughbreds only a small number of racing performance-associated sequence variants have been reported (Hill et al. 2010c, Gu et al. 2010, Hill et al. 2010a, Hill et al. 2010b).

In this study, SNPs were identified in genes with diverse functions related to those previously identified as being selected for the Thoroughbred. Gu (2009) indicated that genomic regions containing genes responsible for other biological functions such as insulin signaling, fatty acid metabolism, steroid metabolic processes, and muscle strength have been selected for during the development of the Thoroughbred (Gu et al. 2009). INPP5J in chromosome 8 revealed that these enzymes regulate insulin signaling (Ooms et al. 2009). In addition, VARS2 could play a role in the regulation of protein synthesis such as that in muscle by insulin. MVK, encoding mevalonate kinase, is a part of the mevalonate pathway, which is involved in steroid biosynthesis (Goldstein and Brown 1990). A GWAS revealed

variants in ARL15 that influence adiponectin levels involved in fatty acid breakdown (Richards et al. 2009).

Due to the a priori hypothesis of the candidate gene approach, I did not perform direct experimental evaluation of genes with racing performance-associated SNPs (Jorgensen et al. 2009). The results of this study is important in that it reveals many SNPs that are associated with racing performance. Assuming that the significant variants identified in this study are truly associated with the EBV for race time, I investigated the LD block that includes the genes identified in this study. Next, assuming that LD is significant with racing performance of Thoroughbreds, I speculated that these genes are strong candidates for racing performance of Thoroughbreds. However, these biological hypotheses should be interpreted cautiously as the genes identified as simply the genes that are closest to the significant SNPs. In my study significant SNPs may affect the expression of cis genes up to 10 kb away or act in trans to alter gene expression on other chromosomes (Myers et al. 2007). Also the SNPs could alter the function or tissue-specific expression of a previously unidentified microRNA or genetic element. Therefore I used two indirect evaluations for my study. The first was the addition of two reported SNPs related to MSTN in the joint analysis. Based on previous candidate gene approach, SNPs, BIEC2-417495 and BIEC2-417274, near the equine MSTN gene were found to have a highly significant association with racing performance in Thoroughbreds (Binns et al. 2010, Hill et al. 2010d). Several MSTN variants are likely to be responsible for large amounts of the phenotypic variance. In domestic animals' traits, few quantitative trait loci (QTL) have been shown to have large effects on the phenotype. For example, most of the morphological traits across domestic dog breeds (Boyko et al. 2010) and the double-muscling gene in cattle (McPherron and Lee 1997) are based on few SNPs. In addition, several GWASs confirmed that SNPs within or near the MSTN gene are strongly associated with

racing performance in Thoroughbreds (Binns et al. 2010, Hill et al. 2010c, Tozaki et al. 2010, Tozaki et al. 2011). This region contained BIEC2-417495, BIEC2-417274, and the top SNP (g.66493737C>T) associated with optimum race distance according to Hill (2010) (Hill et al. 2010c). These SNPs are related to myostatin MSTN, which is a member of the transforming growth factor β family expressed in skeletal muscle and acts as a negative regulator of the proliferation and differentiation of myocytes (Hill et al. 2010c).

However racing performance is more likely to be polygenic in nature. In this study all P-values of the 28 selected SNPs were more significant than the two previously identified SNPs related to MSTN. Therefore I consider that the 28 SNPs of this study are as important as or more so than those two reported SNPs. The second indirect evaluation was a display of relationships between EBV and effect allele number. I noted that the two SNPs related to MSTN have a negative effect. Thoroughbreds with one or two of these SNPs could produce about 0.5-second faster progeny than those with no SNPs (Figure 4.6). The 28 SNPs identified in this study had similar or steeper slopes for racing performance than MSTN SNPs. Thus, I speculated that the 28 SNPs identified in this study influences the racing capacity of Thoroughbreds.

The two-stage GWAS of this study conducted in a large population of Thoroughbreds suggested the presence of 17 protein coding genes related to muscle based on 28 SNPs associated with EBV for race time. My results strongly support a major involvement of myogenesis in the genetic predisposition for high race performance and suggest several genes as genetic factors for muscle maintenance. These candidate genes may provide insights into the genetic secrets underlying the racing performance of Thoroughbreds.

This chapter is paper which and will be published in elsewhere as a partial fulfillment of Dong-Hyun Shin's Ph.D program.

Chapter 5. Multivariate GWAS of milk production traits using genomic estimated breeding value

5.1 Abstract

Holsteins are known as the world's highest-milk producing dairy cattle. The purpose of this study was to identify genetic variants strongly associated with milk production by using estimated breeding value (EBV) as a phenotype. I inferred each EBVs using R packages “rrBLUP”. After then, I conducted multivariate genome-wide association study to search for genetic variants associated with the EBVs for milk production traits using Illumina BovineSNP50 Beadchip (~54,000 single-nucleotide polymorphisms; SNPs). I identified 128 significant SNPs related to 47 genes. These genes were related to cellular component localization, protein localization, intracellular signaling cascade and microtubule. To my knowledge, these genes are newly reported for the genetic association with milk production of Holstein. It complements a recent Holstein GWAS that identified other SNPs and genes as the most significant variants. These results will help to expand my knowledge of the polygenic nature of milk production in Holstein.

5.2 Introduction

Holsteins are the world's highest-milk producing dairy cattle. Since the black Batavians and white Friesians cows were bred to produce better ones 2,000 years ago, they have been continuously selected and genetically evolved into the efficient, high producing black-and-white dairy cattle, Holstein-Friesian. For last several decades, intensive application of traditional animal breeding technologies has significantly improved their milk performances throughout the world.

Over the last decades, technology of molecular biology make it possible to identify genome regions or variant underlying complex traits such as milk yield in dairy cattle. Instead of traditional animal breeding program solely relying on phenotype and pedigree information, information by genetic evaluation provides a great potential to enhance selection accuracies, hence expediting the genetic improvement of animal productivity. Meanwhile, QTL mapping using linkage analysis and/or linkage disequilibrium was developed and has provided a great potential to enhance selection accuracies, hence expediting the genetic improvement of productivity in dairy cattle. Since the seminal work on QTL mapping by Georges el al (Georges et al. 1995), a large number of articles including detection of QTLs for milk production traits have been published. So far a total number of 1,137 QTL for milk production traits have been reported via genome scan based on marker-QTL linkage analyses. The limitations of QTL mapping using linkage analysis (LA) and/or linkage disequilibrium (LD) based on panels of low to moderate density markers have been well documented previously (Meuwissen and Goddard 2001). Additionally, in the past decades merely few strong candidate genes with potential

effects on milk production traits using QTL information (Blott et al. 2003, Grisart et al. 2004).

The advent of genome-wide panels including hundreds of thousands of single nucleotide polymorphisms (SNPs) has resulted in the development of commercial SNP chips and rapid, large-scale genotyping of common SNP across large populations. These SNPs have been widely used for the detection and localization of QTL for complex traits in many species (Daw, Heath and Lu 2005), and have proved powerful and useful in identification of casual mutations associated with economically important traits in livestock (Brym, Kamiński and Wójcik 2004, Amills et al. 2005, Georges 2007) as well as human diseases (Craig and Stephan 2005, Coon et al. 2007). At the same time, Genome-wide association studies (GWAS) based on high throughput SNP genotyping technologies open a broad avenue for exploring genes associated with milk production traits in dairy cattle (Jiang et al. 2010). Most recently, along with maturing of genome sequencing and high throughput SNP genotyping technologies, genome-wide association studies (GWAS) are becoming practical for exploring genes associated with complex traits. GWAS has been widely accepted as a primary approach for gene finding and achieved huge success in identifying genes conferring modest disease risks in human.

Several studies focusing on identifying genes for milk production traits have been performed. Associations between milk traits and polymorphisms in candidate genes have been suggested producing long list of potential markers with some significant effects reported in regional Holstein cattle population (Fontanesi et al. 2014). Generally, most of economic traits in dairy cattle are controlled by many polymorphisms, SNPs of small or large effect. Genetic variances captured by the SNP markers can be used for the calculation of direct genomic breeding values of milk production traits (Erbe et al. 2012). Thus, it is important to determine the

distribution of genomic effect and contribution for milk traits, use this information in genomic prediction and plan future breeding program accordingly (Hayes and Goddard 2010). To find the casual variants and determine the distribution of genomic effect and contribution for milk production traits beyond previous studies, I used multivariate GWAS based on EBVs. EBV was used as the phenotype as it only considers the genetic component of phenotypic variance, increasing the statistical power of the analysis.

In the Multivariate GWAS, I used lineat combination traits of three phenotypes related to milk production. Using this approach, I identified 128 SNPs to be associated with milk production of Holstein. And the identified SNP loci may be considered as preliminary foundation for further replication studies and eventually unraveling the causal mutations for milk production traits in dairy cattle. Additionally, the identified SNP loci may be used as potential candidate markers for selection in Korean dairy cattle breeding programs and provide unprecedented insight into the structure of Holstein cattle populations in Korean and eventually unraveled the causal genes and their pathways for milk production control.

5.3 Materials & Methods

5.3.1 Animals and Data

A total of 456 Holstein in Korea (including bulls and cows) were used in the study. 301 of 456 Holsteins are cow and 290 of 301 cows had records related to milk production traits of one parity. I used three traits that were milk quantity, milk fat and milk proteins of and two environmental factor that are year and season.

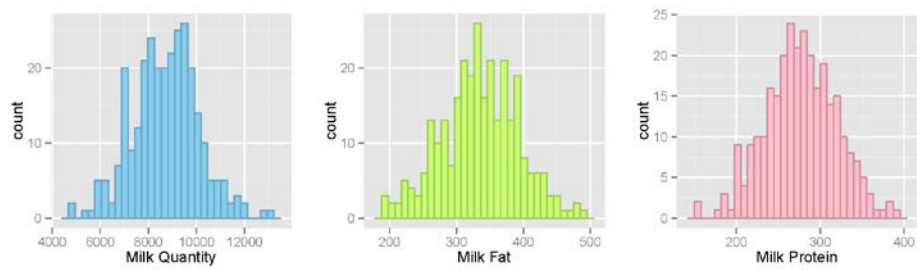

Figure 5.1. Phenotype distribution of the three traits related to Milk.

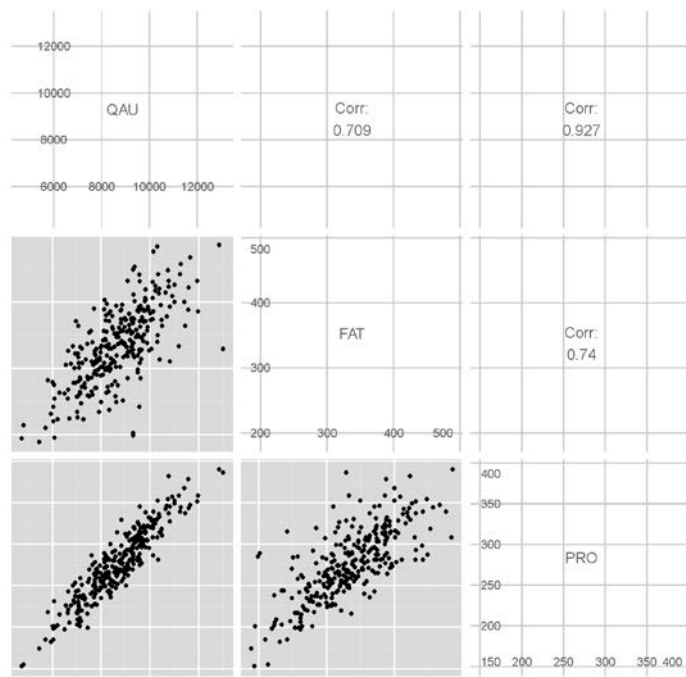

Figure 5.2. All pairwise phenotypic correlation of the three traits related to Milk production.

5.3.2 Genotyping

DNA was extracted from nasal discharge sample of the some seed bull and daughters using the nasal collection kit and semen sample of the sires which were in Korea. DNA was quantified and genotyped using the Illumina BovineSNP50 BeadChip containing 54,609 SNPs. Features of the Illumina BovineSNP50 BeadChip have been detailed previously (Matukumalli et al. 2009). All samples were genotyped using BEADSTUDIO (Illumina).

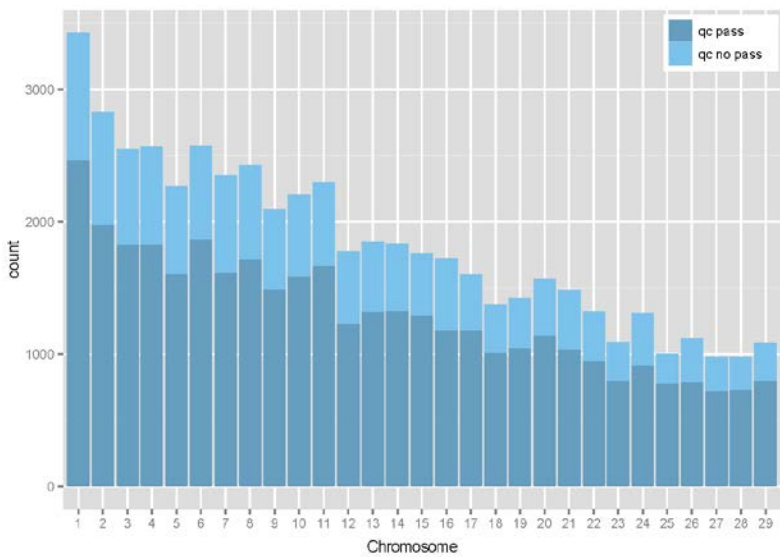

Figure 5.3. 38,656 SNPs after data quality control. Skyblue and blue indicates SNP that did not pass and passed SNP, respectively.

5.3.3 Genotype Quality Control and Imputaion

The chip includes 54,609 SNPs that are distributed on the 29 bovine autosomes, X and Y chromosome (average density of 1 SNP per approximately 49 kb) from the horse genome, UMD 3.1. All samples were genotyped in the National Instrumentation Center for Environmental Management (NICEM) at Seoul National University. I excluded SNPs with Hardy–Weinberg equilibrium (HWE) test P-value of <0.001, a missing rate of >0.05 and minor allele frequency (MAF) of <0.05. Additionally, SNPs on the X and Y chromosome were also excluded retaining finally 38,656 autosomal SNPs. These quality control process was performed using the software PLINK (Purcell, Neale et al. 2007). The 38,656 autosomal SNP data of Holsteins after quality control was imputed without no panel using BEAGLE (Browning and Browning 2011).

5.3.4 Estimation Heritability using SNP data

I estimated the genetic relationship matrix (GRM) between all pairs of 290 individuals having phenotypic records from all the genotyped SNPs after quality control. For each traits, I estimated the variance that can be captured by all SNPs using the restricted maximum likelihood (REML) approach in mixed linear model:

$$y = Xb + g_G + e$$

Where y is a vector of phenotypes, b is a vector of fixed effects with its incidence matrix X, g_G is a vector of aggregate effects of all SNPs:

$$var(g_G) = A_G * {\sigma^2}_G$$

Where A_G is the SNP-derived GRM and ${\sigma^2}_G$ is the additive genetic variance. The proportion of variance explained by all SNPs (heritability) is defined as:

$$h^2{}_G = A_G * \sigma^2{}_G$$

Where σ^2_P is phenotypic variance. Method details have been described in Yang's Papers (Yang et al. 2011a, Yang et al. 2011b). I estimated variance explained by SNPs for milk quantity, milk fat and milk proteins using GCTA (Yang et al. 2011a).

5.3.5 Estimated Breeding Value

I used animal model (BLUP) to infer estimated breeding values, as follows:

$$y = Xb + Za + e$$

Where y is a vector of phenotype, b is a vector of fixed effects, a is a vector of random effects, e is a vector of residual error. X and Z are coefficient matrices for b and a respectively. Phenotype data was comprised of records (including milk quantity, fat, protein) of one parity from 290 cows between 2007 and 2013. The fixed effects in this inferring were year and season of milking. Instead of relationship with individuals, Z matrix included all SNP after quality control. So the coefficients vector of Z matrix is each SNP effects. After that, estimated breeding value is sum of SNP effects considering each genotype for each individual

All parameters were estimated using the "rrBLUP" R packages (Endelman 2011) facilitated by the ridge regression for a single-trait animal model. Using this model, I calculated the EBV for each three traits related to milk production of all Holstein in this study:

$$EBV = \sum_{i=1}^{n} SNP_i \; effect * genotype \; code \; (0, 1, 2)$$

Where n is total SNP number, i is SNP order and genotype code was composed of 0, 1, 2 that means SNP number.

5.3.6 Genome-Wide Association Study

I performed Association analysis on the basis of multivariate test using PLINK (Ferreira and Purcell 2009), as follows:

$$y = g + e$$

Where y is a vector of n phenotype values and g is a random effect and e is residual. In this study, y contains three EBVs for milk quantity, fat and protein and g is SNP data of Holstein. I performed this association analysis using PLINK (Ferreira and Purcell 2009). In multivariate test of PLINK, canonical correlation analysis (CCA) which is a multivariate generalization of the Pearson product-moment correlation was used to measure the association between the two sets of variables. CCA extracts the linear combination of traits that explain the largest possible amount of the covariation between the marker and all traits. This approach is most appropriate for the analysis of normally distributed traits.

5.4 Results

Phenotypes used in this study were three trait related to milk production (milk quantity, fat and protein) of 290 cows of one parity. Milk quantity values were in range 4,747 kg to 13,007 kg. Mean and standard deviation were 8706.17 kg and 1367.58 kg respectively. Milk fat values were in range 189 kg to 486 kg. Mean and standard deviation of Milk fat were 334.97 kg and 57.68 kg respectively. In case of milk protein, values were in range 153 kg to 390 kg. Mean was 273.71 kg and standard deviation is 42.55 kg. Additionally, total three traits followed normal distribution (Figure 5.1). All pairwise phenotypic correlation of three traits are higher than 0.7 (Figure 5.2). Especially, correlation values between milk quantity and protein was 0.927 that was more than expected.

I used three criteria to perform data quality control to reduce false positive result. 1,978 SNPs were excluded in first criteria Hardy-Weinberg equilibrium test. 2,779 SNPs were not passed second criteria as missing rate <0.05. In third criteria which was minor allele frequency >0.05, 11,892 failed to pass. Additionally, because I used cow and bull in association analysis, 729 SNPs on sex chromosome were excluded. Remained 37,927 SNPs were distributed evenly on autosome (Figure 5.3).

After imputation, I estimated the proportion of variance explained by fitting all the SNPs in a mixed linear model for each three traits (Table 5.1). Three proportions were higher than 0.6 and proportion of milk quantity was the highest of them. And p-value of all three proportions in significant test were less than 0.05 that means three proportion were significant. Because of that, I decided that GWAS test using these phenotype can find important SNPs which can explain the phenotype.

To infer phenotypic value related to milk production of bull that had no records, I estimated marker effect of 37,927 SNPs for each traits using “rrBLUP” R packages. Marker effect of Milk quantity in range -2.474 kg to 2.323 kg. Mean and standard deviation were -0.018 kg and 0.578 kg respectively. Those of Milk fat were in range -0.103 kg to 0.122 kg. Mean and standard deviation of Milk fat were -0.00011 kg and 0.024 kg respectively. In case of milk protein, marker effects were in range -0.09 kg to 0.08 kg. Mean was -0.00069 kg and standard deviation is 0.02 kg. Additionally, marker effects of total three traits followed normal distribution (Figure 5.4). Using these marker effect, EBVs of 456 Holstein including cow and bull for three traits were calculated and their distributions were shown in Figure 5.5. And Correlation coefficient of real phenotype and EBV of 290 cows having records for milk quantity, fat and protein were 0.616, 0.653 and 0.624, respectively.

Table 5.1. Estimates of variance explained by all SNPs for the milk production traits.

Trait	*n*	$*h^2_G$	*P*
Milk Quantity	290	0.606 (0.148)	2.68E-05
Milk Fat	290	0.601 (0.147)	1.27E-05
Milk Protein	290	0.677 (0.140)	3.77E-07

* Estimate of variance explained by all SNPs with its standard error given in the parentheses

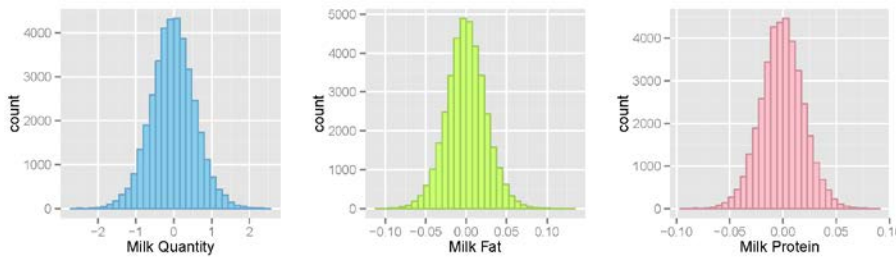

Figure 5.4. Marker effect distribution of the three traits related to Milk.

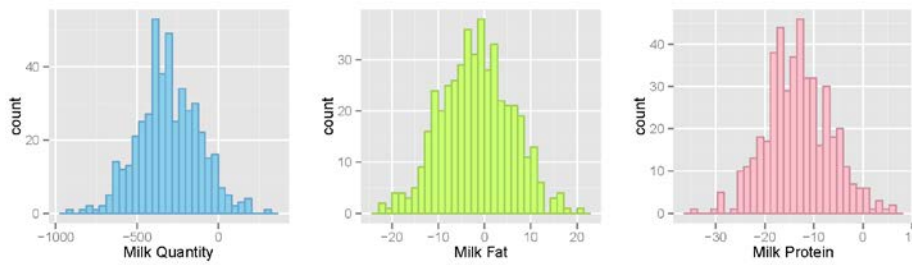

Figure 5.5. EBV distribution of the three traits related to Milk.

In this multivariate GWAS, I compared the genotypes of 456 Holsteins with the three EBVs as a phenotypic value at the same time. Result of GWAS after chromosome sorting was displayed in Manhattan plot (Figure 5.6). Multivariate GWAS found 128 SNPs that were associated with the three EBVs for milk production. Those 128 SNPs exceeded the threshold of multiple tests (p-value = $1.318 * 10^{-6}$, equivalent to adjust p-value after Bonferroni correction). Nine (distributed five chromosome) of chromosomal regions covered by 128 significant SNPs which were defined by a maximal distance between two SNPs of <100 kb contained more than two SNPs. Chromosome 6 had the largest region (2 SNPs or more than 2 SNPs) containing six SNPs (ARS BFGL NGS 34509, ARS BFGL NGS 75181, ARS BFGL NGS 94020, ARS BFGL NGS 54787, Hapmap41517 BTA 35994, Hapmap48747 BTA 40164). And chromosome 10 of five chromosome had the highest number of regions (2 SNPs or more than 2 SNPs).

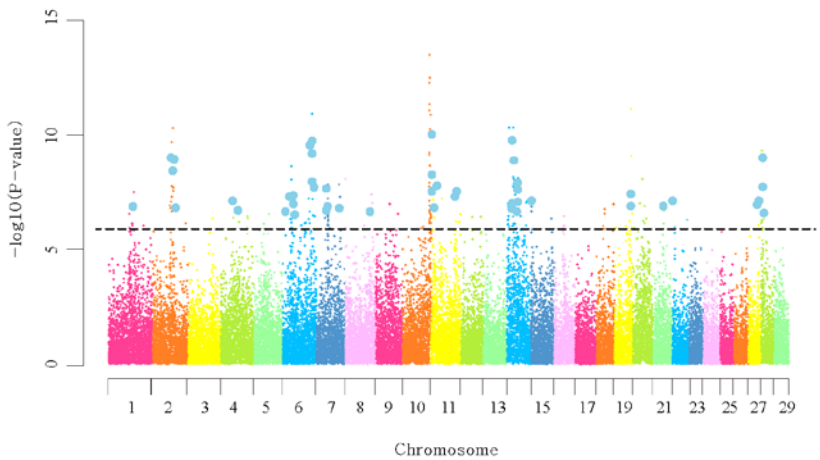

Figure 5.6. Manhattan plot of multivariate GWAS result of the three traits related to Milk. Skyblue circles mean that those SNP belongs to protein coding genes.

The most significant SNP ($P = 3.26 * 10^{-14}$), BTB_01901596 was found on chromosome 10 with the other 18 significant SNPs. Three of 18 SNPs were top ranking (ARS BFGL NGS 112800, BTB 01255583, ARS BFGL NGS 117433) and two of them were in one of Nine regions (chromosome 10: 99037336-99132260). But relationship between detailed biological function and these four top SNPs was concealed, because they were in inter-genic regions. Except these four SNPs on chromosome 10, for the most significant SNP was ARS_BFGL_NGS_71052 on chromosome 19 with the other three significant SNPs and two of them on one of nine regions (chromosome 19: 57714094-57770336, ARS BFGL NGS 71052, UA IFASA 6546). This region included two genes related to significant SNPs which were BTBD17 and DNAI2. BTBD17 encodes BTB/POZ Domain-Containing Protein which was component of Myc-interacting zinc finger protein in mammary gland affecting lactation (Sanz-Moreno et al. 2014). DNAI encodes Dynein intermediate chain 2 which associated with secretory activation of Bovine mammary epithelial cells (Stiening et al. 2006). Chromosome containing the highest number of significant SNPs was chromosome 14 including 27 SNPs. Thirteen SNPs of 27 were related to Ensemble Gene and Eight SNPs of 13 were related to protein coding gene.

The transcriptional content of 128 significant SNPs was assessed using the UMD 3.1 assembly and annotation of the bovine genome. The genes were annotated using the gene IDs from ENSEMBL genome browser and I collected the genes whose entire information was located within or near (±50kb) the SNPs (Table 5.2). Sixty five of 128 SNPs had nearby Ensemble Gene and Forty seven of 65 SNPs were related to protein coding gene. Result of gene ontology analysis (biological process) was in Figure 5.7. Eleven genes of cluster 1 in GO analysis were related to cellular component localization (ABCA1, ARID4B, CLASP1, COG1, DERL1, DNAI2, HOOK3, ROCK2, RPS6KA5, RYR2, VCAN) which means that eleven genes were involved in transporting or maintaining of cellular entity such as a protein complex

or organelle. Previous study reported that ABCA1 protein was localized in mammary gland epithelial cells and showed differential activity between dry period and lactation suggesting a role of ABCA1 in the removal of excess cellular cholesterol (Mani et al. 2009). The protein encoded by COG1 is one of eight proteins (COG1-8) which form a Golgi-localized complex (COG) required for normal Golgi morphology and function which mature milk protein droplets (Wellings and Philp 1964). Five genes of cluster 2 in GO analysis were related to protein localization which means that five genes were involved in a certain process, such as transported to or maintained in a specific location (DERL1, RAB37, ABCA1, COG1, HOOK3). DERL1 encodes a member of the derlin family which recognizes substrate in the ER and works in a complex to retrotranslocate it across the ER membrane into the cytosol and directly associated with the regulation of lactogenesis in bovine mammary epithelial cell differentiation (Stiening et al. 2008). Six genes of cluster 3 in GO analysis were related to intracellular signaling cascade which means that six genes were associated with a certain signal passed on to downstream components within the cell (RPS6KA5, RAB37, ROCK2, RB1CC1, ABCA1, TNIP2). TNIP2 was involved in activation of the MAP kinase pathway related to oxytocin that was essential hormone for mammalian lactation (Tomizawa et al. 2003). Nine genes of cluster 4 in GO analysis were related to microtubule (ABCA1, ARID4B, CLASP1, COG1, DERL1, HOOK3, ROCK2, RPS6KA5, RYR2). Microtubules are known to play a major role in organelle transport and secretory vesicles or granules have been shown to be transported on microtubules. So microtubules was important cellular component to mammalian in milk production.

To compare Multivariate GWAS with single GWAS for each trait, I compared the genotypes of 456 Holsteins with the three EBVs, respectively (Figure 5.8, 5.9, 5.10). 22 of 33 significant SNPs in milk quantity GWAS, 20 of 34 significant SNPs in milk fat GWAS and 20 of 37 significant SNPs in milk fat GWAS were

overlapped with 128 significant SNPs Multivariate GWAS. Intersection of three single GWAS had only 2 SNPs and those were significant in Multivariate GWAS. And union of three single GWAS included 81 SNPs and 45 SNPs were in results of Multivariate GWAS, too.

Table 5.2. The most significant SNPs in Multivariate GWAS with milk production traits

SNP	CHR	BP	Minor/Major	MAF	P-value	Adjusted P-value	Nearby Gene1	Type	Distance
BTB_01901596	10	99234390	C/T	0.462	3.26E-14	1.24.E-09			
ARS_BFGL_NGS_112800	10	100544731	T/G	0.366	3.12E-13	1.18.E-08			
BTB_01255583	10	99037336	C/T	0.374	5.42E-13	2.06.E-08			
ARS_BFGL_NGS_117433	10	99132260	A/G	0.432	4.63E-12	1.76.E-07			
ARS_BFGL_NGS_71052	19	57714094	T/C	0.287	7.44E-12	2.82.E-07	ENSBTAG00000030166 (BTBD17)	InterGenic	-3401
ARS_BFGL_NGS_82902	10	98138943	C/T	0.327	8.74E-12	3.31.E-07			
ARS_BFGL_NGS_4767	6	107186270	C/T	0.316	1.19E-11	4.51.E-07	ENSBTAG00000005711 (NSG1)	Genic	Genic
ARS_BFGL_BAC_15488	10	103048273	T/C	0.296	1.34E-11	5.08.E-07	ENSBTAG00000045016	InterGenic	4023
ARS_BFGL_BAC_10345	14	6133529	C/A	0.359	4.75E-11	1.80.E-06			
Hapmap50356_BTA_42148	14	17401030	A/G	0.252	4.75E-11	1.80.E-06			
ARS_BFGL_NGS_106694	2	74101924	G/A	0.159	5.11E-11	1.94.E-06			
Hapmap36326_SCAFFOLD266420_4418	10	101130681	T/C	0.366	5.64E-11	2.14.E-06			
ARS_BFGL_NGS_27366	10	103470001	A/G	0.315	8.27E-11	3.14.E-06	ENSBTAG00000010906 (RPS6KA5)	InterGenic	-38677
UA_IFASA_6693	14	18296407	G/A	0.231	1.85E-10	7.02.E-06	ENSBTAG00000020693 (DERL1)	InterGenic	-10733
ARS_BFGL_NGS_107266	2	70200269	G/A	0.165	2.10E-10	7.96.E-06			
ARS_BFGL_NGS_36089	14	25698286	A/G	0.480	2.24E-10	8.50.E-06			
ARS_BFGL_NGS_112814	6	108041362	T/C	0.493	2.44E-10	9.25.E-06	ENSBTAG00000013996 (SH3BP2)	InterGenic	7131

ARS_BFGL_NGS_104112	6	108076099	G/A	0.443	2.98E-10	1.13.E-05	ENSBTAG00000015126 (TNIP2)	InterGenic	-48437
ARS_BFGL_NGS_60834	28	2047488	C/T	0.197	5.01E-10	1.90.E-05			
ARS_BFGL_NGS_94895	6	108225240	T/C	0.452	5.45E-10	2.07.E-05			
BTA_09585_no_rs	6	108135281	A/G	0.393	5.62E-10	2.13.E-05	ENSBTAG00000015126 (TNIP2)	InterGenic	7107
Hapmap60701_rs29010356	10	98245112	A/C	0.367	6.05E-10	2.29.E-05			
Hapmap51063_BTA_82431	10	97753107	A/G	0.337	7.30E-10	2.77.E-05			
UA_IFASA_6546	19	57770336	C/T	0.475	8.72E-10	3.31.E-05	ENSBTAG00000016732 (DNAI2)	Genic	Genic
ARS_BFGL_NGS_18597	28	5290625	G/A	0.121	9.09E-10	3.45.E-05	ENSBTAG00000000942 (SIPA1L2)	Genic	Genic
ARS_BFGL_NGS_22157	2	71906695	T/G	0.186	1.11E-09	4.21.E-05	ENSBTAG00000009058 (TMEM177)	InterGenic	11830
Hapmap25319_BTA_156716	2	73509732	G/A	0.283	1.25E-09	4.74.E-05	ENSBTAG00000019781 (CLASP1)	Genic	Genic
ARS_BFGL_NGS_42006	14	23122719	A/G	0.266	1.27E-09	4.82.E-05	ENSBTAG00000000878 (RB1CC1)	InterGenic	-25273
BTB_01308993	10	100917539	A/G	0.375	2.27E-09	8.61.E-05			
BTA_119035_no_rs	6	32293744	C/G	0.190	2.38E-09	9.03.E-05			
ARS_BFGL_NGS_101411	2	72163562	A/G	0.186	3.00E-09	1.14.E-04	ENSBTAG00000018236 (EPB41L5)	InterGenic	-2988
ARS_BFGL_NGS_34509	10	101474077	T/C	0.420	5.99E-09	2.27.E-04	ENSBTAG00000009994 (EML5)	Genic	Genic
Hapmap41517_BTA_35994	14	18175312	C/T	0.345	6.93E-09	2.63.E-04	ENSBTAG00000015319	InterGenic	2024
Hapmap48747_BTA_40164	14	35151359	A/G	0.338	7.62E-09	2.89.E-04			
ARS_BFGL_NGS_75181	8	2357514	C/T	0.162	8.44E-09	3.20.E-04			
ARS_BFGL_NGS_94020	20	34158865	T/C	0.416	8.51E-09	3.23.E-04			
ARS_BFGL_NGS_54787	14	20606100	G/A	0.240	8.63E-09	3.27.E-04	ENSBTAG00000009474	Genic	Genic

Hapmap27136_BTC_073585	6	32927979	C/T	0.434	9.15E-09	3.47.E-04	ENSBTAG00000047255	InterGenic	-14284
ARS_BFGL_NGS_102418	14	35744766	A/G	0.442	1.18E-08	4.48.E-04	ENSBTAG00000044870 (U3)	InterGenic	49844
ARS_USMARC_Parent_DQ984825 _rs29012457	10	98230479	C/T	0.496	1.21E-08	4.59.E-04			
ARS_BFGL_NGS_101630	14	19220744	G/C	0.248	1.38E-08	5.23.E-04			
ARS_BFGL_NGS_67190	7	85579651	G/T	0.226	1.45E-08	5.50.E-04			
Hapmap43685_BTA_77920	6	108026956	A/G	0.500	1.46E-08	5.54.E-04	ENSBTAG00000013996 (SH3BP2)	Genic	Genic
BTB_01964039	7	43479777	C/T	0.185	1.47E-08	5.58.E-04	ENSBTAG00000038722	InterGenic	-2938
Hapmap58976_rs29019573	2	68877969	T/C	0.194	1.75E-08	6.64.E-04			
Hapmap11534_BTA_82438	10	97410796	T/C	0.436	1.82E-08	6.90.E-04			
ARS_BFGL_NGS_110191	11	6758495	A/G	0.349	1.83E-08	6.94.E-04	ENSBTAG00000006343 (IL1R2)	InterGenic	718
BTA_103106_no_rs	6	111321399	G/A	0.387	1.85E-08	7.02.E-04	ENSBTAG00000007449 (HS3ST1)	InterGenic	-48515
Hapmap58921_rs29010046	14	16902500	G/A	0.207	1.85E-08	7.02.E-04			
Hapmap47360_BTA_63966	28	8002146	A/G	0.366	1.86E-08	7.05.E-04	ENSBTAG00000000222 (ARID4B)	Genic	Genic
ARS_BFGL_NGS_100845	7	42460306	G/A	0.254	1.91E-08	7.24.E-04	ENSBTAG00000031412 (OR6F1)	InterGenic	-19374
ARS_BFGL_NGS_69160	2	76136460	C/T	0.181	2.03E-08	7.70.E-04			
Hapmap41742_BTA_107716	14	36837792	G/A	0.285	2.42E-08	9.18.E-04	ENSBTAG00000043212 (SNORD12)	InterGenic	18325
UA_IFASA_4619	2	76309053	C/T	0.283	2.74E-08	1.04.E-03			
BTB_00562506	14	30113351	C/T	0.406	2.82E-08	1.07.E-03			
BTA_05225_no_rs	10	103908013	C/G	0.413	2.91E-08	1.10.E-03	ENSBTAG00000015385 (CTDSPL2)	InterGenic	-16162

ARS_BFGL_NGS_106939	1	94730717	C/T	0.363	3.10E-08	1.18.E-03			
ARS_BFGL_NGS_105912	11	86655500	T/C	0.336	3.57E-08	1.35.E-03	ENSBTAG00000001903 (C11H2orf50)	InterGenic	34743
ARS_BFGL_NGS_81636	19	58687075	T/C	0.456	3.84E-08	1.46.E-03	ENSBTAG00000006087 (COG1)	Genic	Genic
BTB_01066770	8	97834727	A/C	0.400	4.11E-08	1.56.E-03			
ARS_BFGL_NGS_111395	14	25731992	T/C	0.428	4.15E-08	1.57.E-03			
Hapmap58847_rs29023123	7	90239425	A/G	0.204	4.98E-08	1.89.E-03			
Hapmap47485_BTA_107591	2	69883074	A/G	0.229	4.99E-08	1.89.E-03	ENSBTAG00000015077	InterGenic	8198
BTB_01536946	2	68429456	G/C	0.206	5.10E-08	1.93.E-03			
ARS_BFGL_NGS_104379	11	86497486	T/C	0.129	5.26E-08	1.99.E-03	ENSBTAG00000005847 (ROCK2)	InterGenic	-4091
BTA_00906_rs29014057	6	96260709	A/C	0.114	5.94E-08	2.25.E-03			
BTA_75768_no_rs	6	35147153	T/C	0.366	6.02E-08	2.28.E-03	ENSBTAG00000019808 (CCSER1)	Genic	Genic
BTA_99659_no_rs	11	30945111	C/T	0.461	6.08E-08	2.31.E-03	ENSBTAG00000027015	InterGenic	-17239
BTB_00308411	7	43808593	T/C	0.175	6.17E-08	2.34.E-03	ENSBTAG00000016065	InterGenic	-860
Hapmap50091_BTA_75608	6	27433375	T/C	0.424	6.36E-08	2.41.E-03	ENSBTAG00000012343 (TSPAN5)	Genic	Genic
ARS_BFGL_NGS_8311	6	108257168	C/T	0.441	6.46E-08	2.45.E-03			
Hapmap49856_BTA_108815	28	3998395	T/C	0.225	6.48E-08	2.46.E-03	ENSBTAG00000039845	InterGenic	10120
ARS_BFGL_NGS_85936	15	6234482	G/A	0.444	6.83E-08	2.59.E-03	ENSBTAG00000040567 (MMP27)	InterGenic	25319
ARS_BFGL_NGS_13472	11	4696808	G/A	0.427	6.99E-08	2.65.E-03			
BTA_54025_no_rs	22	2687429	G/A	0.211	7.38E-08	2.80.E-03	ENSBTAG00000035286 (CMC1)	Genic	Genic
ARS_BFGL_NGS_105598	4	47195467	A/G	0.344	7.69E-08	2.92.E-03	ENSBTAG00000002037 (ATXN7L1)	Genic	Genic

ARS_BFGL_NGS_2937	10	98680972	G/A	0.474	7.74E-08	2.94.E-03			
ARS_BFGL_NGS_82976	14	68962221	A/G	0.389	7.76E-08	2.94.E-03	ENSBTAG00000032432	InterGenic	23678
ARS_BFGL_NGS_59506	2	66860284	G/A	0.313	8.28E-08	3.14.E-03			
BTB_00554463	14	18568023	A/G	0.285	8.56E-08	3.25.E-03			
ARS_BFGL_NGS_63803	27	38631236	T/C	0.082	8.63E-08	3.27.E-03	ENSBTAG00000033137 (PSD3)	Genic	Genic
Hapmap26883_BTC_055211	6	44441744	G/A	0.295	8.64E-08	3.28.E-03			
Hapmap41415_BTA_107703	14	37535038	C/T	0.499	9.20E-08	3.49.E-03	ENSBTAG00000005404 (MSC)	InterGenic	-14301
ARS_BFGL_NGS_113655	14	66118162	C/T	0.396	9.26E-08	3.51.E-03			
Hapmap58471_rs29026091	8	100967241	A/G	0.374	9.41E-08	3.57.E-03			
BTB_00581487	15	13226449	G/A	0.372	9.65E-08	3.66.E-03			
ARS_BFGL_BAC_35584	20	22542386	T/C	0.339	9.93E-08	3.77.E-03			
BTB_01346626	9	56522192	T/C	0.222	1.00E-07	3.79.E-03			
ARS_BFGL_NGS_110993	14	4808166	A/C	0.385	1.01E-07	3.83.E-03			
BTB_00672453	14	17288504	T/C	0.375	1.05E-07	3.98.E-03	ENSBTAG00000022114 (TMEM65)	Genic	Genic
ARS_BFGL_NGS_111782	18	63029071	T/C	0.280	1.06E-07	4.02.E-03	ENSBTAG00000038797	InterGenic	239
UA_IFASA_4968	9	57559232	C/T	0.275	1.07E-07	4.06.E-03	ENSBTAG00000003743	InterGenic	41926
ARS_BFGL_NGS_112812	6	38627070	A/G	0.348	1.09E-07	4.13.E-03	ENSBTAG00000005932 (FAM184B)	Genic	Genic
ARS_BFGL_NGS_37799	27	37315792	A/G	0.180	1.14E-07	4.32.E-03	ENSBTAG00000007634 (HOOK3)	Genic	Genic
ARS_BFGL_NGS_31820	19	57330068	T/C	0.383	1.17E-07	4.44.E-03	ENSBTAG00000008092 (RAB37)	Genic	Genic
ARS_BFGL_NGS_114507	2	61080508	C/T	0.295	1.21E-07	4.59.E-03			

BTB_01646116	10	99103087	T/C	0.372	1.24E-07	4.70.E-03			
BTB_01687158	7	85712075	C/T	0.178	1.29E-07	4.89.E-03	ENSBTAG00000014906 (VCAN)	Genic	Genic
ARS_BFGL_NGS_19469	20	42937884	T/C	0.083	1.30E-07	4.93.E-03			
ARS_BFGL_NGS_69831	7	41956061	T/C	0.206	1.32E-07	5.01.E-03	ENSBTAG00000038284 (ZNF496)	Genic	Genic
BTB_00325087	7	88317307	C/T	0.218	1.32E-07	5.01.E-03			
ARS_BFGL_NGS_96559	11	7851919	T/C	0.140	1.32E-07	5.01.E-03	ENSBTAG00000042285 (U6)	InterGenic	47010
ARS_BFGL_NGS_117336	21	41947454	T/G	0.311	1.32E-07	5.01.E-03	ENSBTAG00000021844 (COCH)	InterGenic	-585
BTB_01086804	1	94952782	G/A	0.231	1.35E-07	5.12.E-03	ENSBTAG00000031802 (SPATA16)	InterGenic	-12416
Hapmap23302_BTC_052123	14	4848750	T/G	0.229	1.38E-07	5.23.E-03			
Hapmap27294_BTC_032117	6	32850666	G/A	0.465	1.41E-07	5.35.E-03			
ARS_BFGL_BAC_31451	15	3811254	G/A	0.152	1.41E-07	5.35.E-03			
Hapmap53388_rs29010903	2	63581955	G/A	0.162	1.59E-07	6.03.E-03			
ARS_BFGL_NGS_113949	14	25501417	A/C	0.291	1.59E-07	6.03.E-03	ENSBTAG00000043923 (U6)	InterGenic	9233
Hapmap54618_rs29021334	14	25612510	C/T	0.286	1.61E-07	6.11.E-03			
BTB_00248463	14	19441969	G/A	0.431	1.70E-07	6.45.E-03			
ARS_BFGL_NGS_83348	2	76884059	C/A	0.378	1.74E-07	6.60.E-03	ENSBTAG00000031898 (CNTNAP5)	Genic	Genic
ARS_BFGL_NGS_108426	18	28324801	T/G	0.443	1.76E-07	6.68.E-03			
BTB_00194107	4	66583105	T/C	0.151	1.81E-07	6.86.E-03	ENSBTAG00000039231 (C4H7orf41)	InterGenic	-27850
Hapmap59709_rs29021868	14	31014368	A/G	0.328	1.94E-07	7.36.E-03	ENSBTAG00000001299 (CYP7B1)	Genic	Genic
Hapmap58172_rs29027092	10	98548526	G/A	0.412	1.95E-07	7.40.E-03			

ARS_BFGL_NGS_65709	21	39496121	C/T	0.136	2.02E-07	7.66.E-03			
Hapmap44568_BTA_77505	6	17563402	T/C	0.196	2.05E-07	7.78.E-03	ENSBTAG00000017282 (COL25A1)	Genic	Genic
BTB_01364730	7	38967158	C/T	0.219	2.09E-07	7.93.E-03	ENSBTAG00000019071 (COMMD10)	Genic	Genic
BTA_82304_no_rs	8	96366646	G/T	0.132	2.13E-07	8.08.E-03	ENSBTAG00000020661 (ABCA1)	Genic	Genic
BTA_21857_no_rs	2	69917549	G/T	0.317	2.15E-07	8.15.E-03	ENSBTAG00000016785	InterGenic	279
BTB_02021101	15	3416019	T/G	0.321	2.19E-07	8.31.E-03			
BTA_77536_no_rs	6	100921725	T/C	0.123	2.20E-07	8.34.E-03			
Hapmap31719_BTA_125984	10	98627766	C/T	0.384	2.23E-07	8.46.E-03			
ARS_BFGL_NGS_61189	28	9964623	A/G	0.128	2.30E-07	8.72.E-03	ENSBTAG00000022886 (RYR2)	InterGenic	-1293
UA_IFASA_8554	14	28381520	G/T	0.169	2.35E-07	8.91.E-03	ENSBTAG00000045005	InterGenic	43185
ARS_BFGL_NGS_104268	14	24057354	C/T	0.485	2.50E-07	9.48.E-03	ENSBTAG00000047303	Genic	Genic
Hapmap54823_rs29021261	6	44545708	G/A	0.463	2.53E-07	9.60.E-03	ENSBTAG00000042453 (U6)	InterGenic	-8102

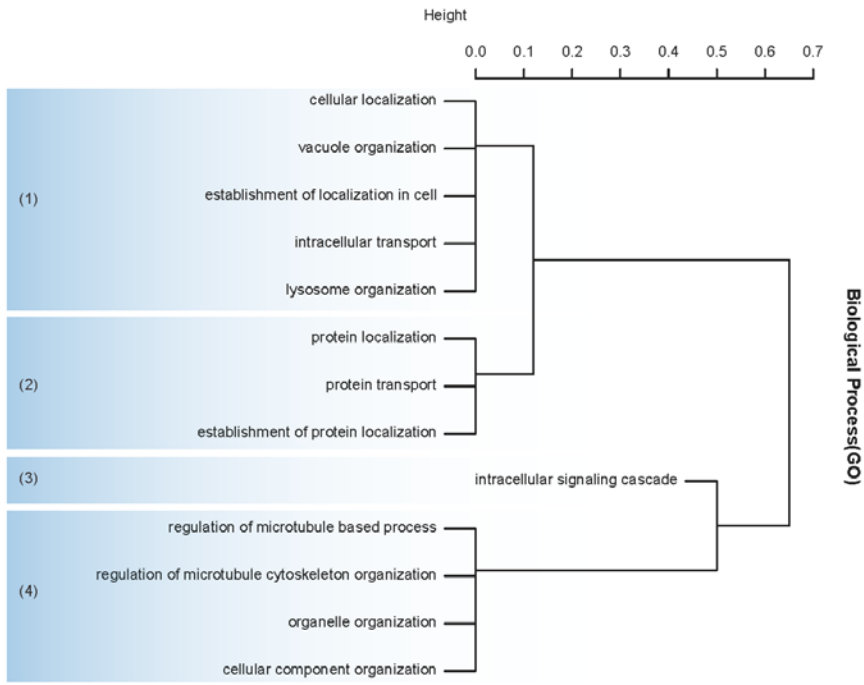

Figure 5.7. Gene Ontology analysis of 47 genes associated with milk production after multivariate GWAS

5.5 Discussion

To identify SNPs underlying milk production of Holstein, I applied a multivariate GWAS to search for genetic variants associated with the linear combination of three EBVs for milk production. I present the results of a multivariate GWAS involving 456 Holsteins. Before association study, I inferred marker effects for three phenotypes that were milk quantity, fat and protein using 290 cows and I calculated genomic estimated breeding value for 456 Holsteins including cows and bulls. In association study, I performed a multivariate GWAS using PLINK whish used linear combination of several traits as dependent variable (Ferreira and Purcell 2009). Using this approach, I identified 128 SNPs associated with milk production of Holstein. The SNPs were in genes related to lactation that mostly have not been previously reported.

Several measurements were used to estimate good milk production. Milk quantity, fat and protein are the most direct measure for evaluating good milk of Holsteins. In this study, I used records of Milk quantity, fat and protein of first lactation in Holstein. And heritability of each traits were reported as 0.40 to 0.59 for milk yield, 0.34 to 0.68 for fat yield and 0.33 to 0.69 for protein yield. Those were mostly similar with heritability based on data of this study using GCTA. So, I thought that SNP associated with these trait can explain many things about milk production.

In animal breeding, EBV is used to rank breeding stock for selection as it only considers the genetic effect on phenotype and predicts the genetic value of an individual based on the phenotypes measured in their relatives. Breeding value is the sum of gene effects of an animal as measured by the performance of its progeny. Traits relater to milk production were for only cow after delivery and bull can not

have these records related to milk production. So I infer EBVs for three traits associated with milk production, because I want to assign bulls phenotypic value related to milk production.

GWAS is a promising method to discover common genetic variants that could explain diseases or economic phenotypes of animals and plants. So far, many single-analysis of Holstein were performed to investigate SNPs related to milk production. In Generally, the statistical power of detecting associations using multiple-trait GWAS was as good as or better than that of the best single-trait GWAS. Therefore, I carried out multivariate GWAS to search for genetic variants associated with EBV for milk production traits. Holsteins have distinctive markings and outstanding milk production with milk fat and protein. Already, I reported that correlation coefficient of all pairwise among milk quantity, fat and protein. When correlated traits are analyzed, the sampling errors tend to be correlated and this makes the interpretation of the results difficult. To overcome this problem, methods that generate uncorrelated traits are useful and have been used in this study. So multivariate GWAS which generated linear combination trait for several trait was very appropriate method.

However, inflation becomes a problem with the use of EBV as a phenotype in GWAS. I calculated the lambda value by the statistical package R. The lambda value of multi GWAS was 3.3462, which is much higher than expected. There are several potential reasons for this higher lambda value. First, the linkage disequilibrium in the Holstein population is extended to a very long range and this effective population size was around only 122 (Shin et al. 2013). Therefore, in a single marker analysis, all the SNPs in linkage disequilibrium will have an effect on the analyses. Second, the milk production traits analyzed in this study are under intense directional artificial selection. This can result in genome-wide inflation of p-

values. In addition, because EBV comprises only genetic factors, the inflation value can increase more than expected. For these reasons, inflation was a normal phenomenon in animal GWASs and not a problem for detecting significant SNPs. Additionally, I excluded false positive significant SNPs as much as possible by stringent multiple testing (Bonferroni correction).

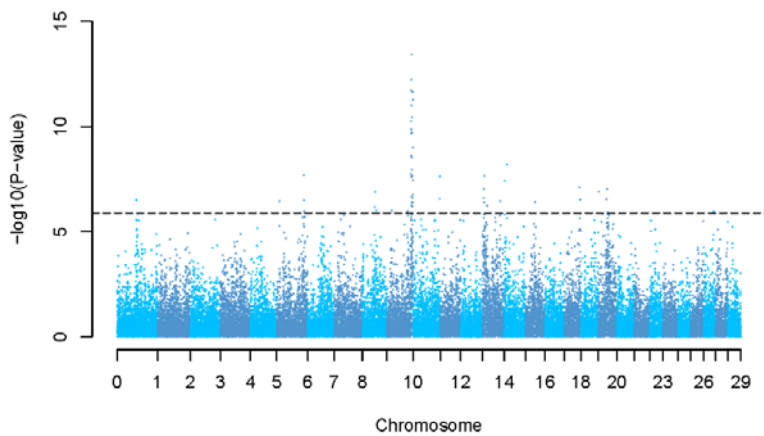

Figure 5.8. Manhattan plot of single GWAS result of milk quantity.

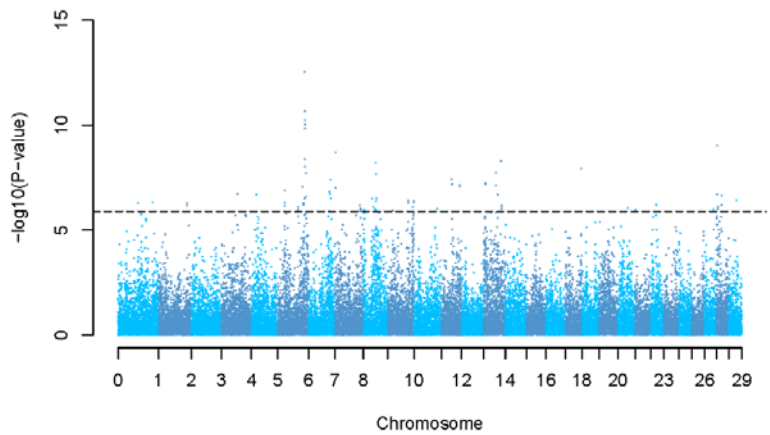

Figure 5.9. Manhattan plot of single GWAS result of milk fat.

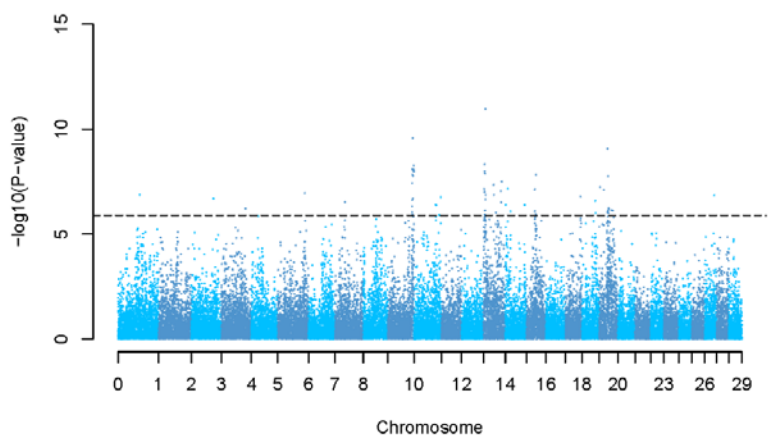

Figure 5.10. Manhattan plot of single GWAS result of milk protein.

This chapter was published in *BioMedCentral Genomics*
as a partial fulfillment of Dong-Hyun Shin's Ph.D program.

Chapter 6. Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level

6.1 Abstract

Copy number variation (CNV), a source of genetic diversity in mammals, has been shown to underlie biological functions related to production traits. Notwithstanding, there have been few studies conducted on CNVs using next generation sequencing at the population level.

Illumina NGS data was obtained for ten Holsteins, a dairy cattle, and 22 Hanwoo, a beef cattle. The sequence data for each of the 32 animals varied from 13.58-fold to almost 20-fold coverage. I detected a total of 6,811 deleted CNVs across the analyzed individuals (average length = 2732.2 bp) corresponding to 0.74% of the cattle genome (18.6 Mbp of variable sequence). By examining the overlap between CNV deletion regions and genes, I selected 30 genes with the highest deletion scores. These genes were found to be related to the nervous system, more specifically with nervous transmission, neuron motion, and neurogenesis. I regarded these genes as having been effected by the domestication process. Further analysis of the CNV genotyping information revealed 94 putative selected CNVs and 954 breed-specific CNVs.

This study provides useful information for assessing the impact of CNVs on cattle traits using NGS data at the population level.

6.2 Introduction

Since the completion of the bovine genome assembly (Elsik, Tellam and Worley 2009, Liu et al. 2009, Zimin et al. 2009), a large number of genetic variation as single-nucleotide polymorphisms (SNPs), have become widely known and commercial SNP panels have been developed for cattle (Matukumalli et al. 2009). The continued discovery of SNPs in diverse cattle breeds has been further expanded (Eck et al. 2009, Stothard et al. 2011) by the recent availability of massively parallel sequencing technologies called next-generation sequencing (NGS). SNPs and the commercial SNP marker panels have been successfully used to identify genomic regions that potentially underlie the economic traits of cattle (Barendse et al. 2009a, Gibbs et al. 2009, Hayes et al. 2009b). Another source of genetic variation is mammals come from gains and losses of genomic structural sequence variants, copy number variations (CNVs), that occur in more than two individuals (Mills et al. 2011). While SNPs are more frequently used in cattle breeding than CNVs, CNVs occupy a higher percentage of genomic sequence than SNPs.

Many studies have endeavored to understand CNVs in mammals, especially in humans (Redon et al. 2006, Conrad et al. 2009, Altshuler et al. 2010, Mills et al. 2011) and rodents (Graubert et al. 2007, Guryev et al. 2008, She et al. 2008, Yalcin et al. 2011). In particular, several CNVs were shown to be important in both normal phenotypic variability and disease susceptibility in human (Aitman et al. 2006, Fellermann et al. 2006, Le Maréchal et al. 2006, Yang et al. 2007, Stankiewicz and Lupski 2010). It is possible that CNVs have a potentially greater effect on phenotype, including changing of gene structure and dosage, altering gene regulation and exposing recessive alleles (Zhang et al. 2009). These points are

attracting attention to CNV as structural variation that can account for diverse economically important traits in domestic animals. In particular, the CNV type, deletions, which is the focus of this study has been shown to be one of the five CNV types and one of the two main classes with duplications (Redon et al. 2006). Previous study of cattle using next generation sequencing (NGS) data has reported that CNVs play a crucial role in diverse biological functions as pathogen- and parasite-resistance, lipid transport and metabolism, breed-specific differences in adaptation, health, and production traits (Bickhart et al. 2012).

The focus of CNV studies has also extended into other domesticated animals including dog, goat, cattle, pig, and sheep (Chen et al. 2009, Fontanesi et al. 2009a, Nicholas et al. 2009, Bae et al. 2010a, Fadista et al. 2010, Liu et al. 2010, Ramayo-Caldas et al. 2010a, Kijas et al. 2011, Fontanesi et al. 2011, Bickhart et al. 2012). Considering the heritability of CNVs and their higher rates of mutation, CNVs may be largely associated with or affect animal health and production traits under recent selection. In the case of cattle, partial deletion of the bovine gene ED1 causes anhidrotic ectodermal dysplasia (Drögemüller et al. 2001). Bos taurus indicus has the capacity to adapt to warm climates and superior resistance to tick infestation than Bos taurus taurus breeds (Porto Neto et al. 2011). Likewise, beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. It is possible that CNVs may be associated with these agriculturally important traits (Bickhart et al. 2012).

Until now, CNV screens were routinely performed by comparative genomic hybridization (CGH) and SNP arrays, and many studies have extensively reviewed their performances (Lai et al. 2005, LaFramboise 2009, Winchester et al. 2009, Pinto et al. 2011). However, these methods, which are often affected by low probe density and cross-hybridization of repetitive sequence, were not able to detect CNVs at the

whole genome level. A limited number of investigations in cattle CNV has been performed to detect CNVs using methods that include high-density aCGH and the 50K SNP panel (Bae et al. 2010a, Fadista et al. 2010, Liu et al. 2010). The recent advances of NGS and complementary analysis programs have provided better approaches to systematically identify CNVs at a deep genome-wide level than the currently available commercial SNP chip and aCGH methodologies (Stothard et al. 2011, Alkan et al. 2011). These sequence-based approaches, which are becoming more popular due to the ongoing developments and cost decreases in NGS, allow for CNV reconstruction at a higher effective resolution and sensitivity.

In this study, I attempt to detect genome-wide CNVs at the population level based on NGS data of 32 cattle. Using UMD3.1 (Zimin et al. 2009) as a reference genome, I used Genome STRiP to detect cattle CNVs at the population level using Hanwoo (22 individuals), a Korea beef cattle, and Holstein (10 individuals), a dairy cattle. I discovered 18.6 Mbp of deleted sequence in the reference genome. However, using Genome STRiP, I could only extract deleted CNVs from the population data (Mills et al. 2011). This study confirmed that CNVs are common, associated with deleted regions, and often occur in gene-rich regions in cattle. I analyzed genes related to CNVs using deletion score in order to explore their potential function and contributions in domestication. In addition, I investigated the selected CNVs using FST and breed-specific CNVs for traits related to beef and milk production. By providing several types of information on cattle CNV at the population level and presenting deleted CNV maps with breed-specific CNVs, I provide the basis for further studies into the role of deleted CNVs in the cattle genome.

6.3 Materials & Methods

6.3.1 Ethics statement

All experimental procedures on animals in this study were performed in strict accordance with good animal practice as defined by the relevant national and/or local welfare bodies. In addition, all animal experiments were approved by the Institutional Animal Care and Use Committee of the National Institute of Animal Science (No. 2012-C-005, CNU-00300).

6.3.2 DNA Sampling & Resequencing Process

Based on the breed history and breed-specific information, I obtained 22 Hanwoo and ten Holsteins for whole-genome resequencing. Individuals were selected as representatives of its breed. Out of the 22 Hanwoo, 11 individuals were from the Hanwoo Experiment Station, National Institute of Animal Science, Rural Development Administration, Korea, and the other 11 individuals were from Kyungpook National University, Korea. Ten Holsteins were obtained from National Institute of Animal Science, Rural Development Administration, Korea. Blood was collected from each animal and treated with heparin to prevent clotting. Manufacturers' instructions were followed to create a paired library. Pair-end sequence data was generated using HiSeq 2000 (Illumina, Inc). Pair-end sequence reads were mapped to the reference cattle genome UMD 3.1 with aligner based on the Burrows-Wheeler transform and the FM-index (Bowtie2; version 2.1.0) using default setting (Langmead and Salzberg 2012). Three open-source packages were used for downstream processing and variant calling: Picard Tools, SAMtools, and

Genome analysis toolkit (GATK) (McKenna et al. 2010) (Figure 6.1). All calls with a Phred-scaled quality of less than 20 were filtered out. The origin, features, and general sequencing information of the individual animals are summarized in Table 6.1.

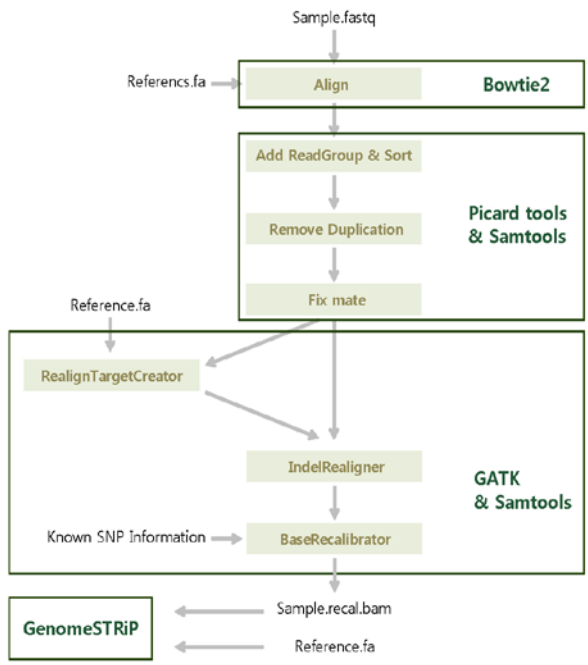

Figure 6.1. Resequencing NGS data process pipeline before Genome STRiP for CNV extraction.

Table 6.1. Sample information and NGS quality score

Sample	Breed	Region	Number of Reads	Align Accuracy	Raw X coverage
HW_1	HanWoo	RDA in Suwon	234,569,860	98.04	18.68
HW_2	HanWoo	RDA in Suwon	187,253,355	98.08	14.91
HW_3	HanWoo	RDA in Suwon	239,492,976	98.05	19.07
HW_4	HanWoo	RDA in Suwon	237,065,133	98.08	18.87
HW_5	HanWoo	RDA in Suwon	217,926,984	98.09	17.35
HW_6	HanWoo	RDA in Suwon	203,717,705	98.01	16.22
HW_7	HanWoo	RDA in Suwon	219,348,590	98.18	17.46
HW_8	HanWoo	RDA in Suwon	212,530,861	98.03	16.92
HW_9	HanWoo	RDA in Suwon	246,479,541	97.70	19.62
HW_10	HanWoo	RDA in Suwon	233,226,278	98.05	18.57
HW_11	HanWoo	RDA in Suwon	217,612,130	98.01	17.33
HW_12	HanWoo	Kyungpook Nat'l University	204,631,996	97.46	16.29
HW_13	HanWoo	Kyungpook Nat'l University	218,317,802	97.34	17.38
HW_14	HanWoo	Kyungpook Nat'l University	214,567,000	97.07	17.08
HW_15	HanWoo	Kyungpook Nat'l University	197,016,954	97.39	15.69
HW_16	HanWoo	Kyungpook Nat'l University	216,161,654	97.41	17.21
HW_17	HanWoo	Kyungpook Nat'l University	213,761,749	97.47	17.02
HW_18	HanWoo	Kyungpook Nat'l University	192,688,294	96.53	15.34
HW_19	HanWoo	Kyungpook Nat'l University	196,881,793	97.32	15.67
HW_20	HanWoo	Kyungpook Nat'l University	176,882,706	97.21	14.08
HW_21	HanWoo	Kyungpook Nat'l University	176,265,378	97.20	14.03
HW_22	HanWoo	Kyungpook Nat'l University	183,510,511	97.27	14.61
HS_1	Holstein	RDA in Suwon	170,511,249	97.31	13.58
HS_2	Holstein	RDA in Suwon	242,875,799	97.35	19.34
HS_3	Holstein	RDA in Suwon	240,641,118	97.48	19.16
HS_4	Holstein	RDA in Suwon	215,999,610	97.15	17.20
HS_5	Holstein	RDA in Suwon	245,048,441	97.08	19.51
HS_6	Holstein	RDA in Suwon	225,557,863	97.77	17.96
HS_7	Holstein	RDA in Suwon	239,365,182	97.45	19.06
HS_8	Holstein	RDA in Suwon	239,776,485	97.17	19.09
HS_9	Holstein	RDA in Suwon	230,316,320	97.75	18.34
HS_10	Holstein	RDA in Suwon	238,847,676	94.42	19.02

6.3.2 Copy Number Variations Extraction

The re-sequencing data of the 32 cows were aligned and CNVs were extracted from the combined dataset. The CNV extraction tool Genome STRucture in Population (Genome STRiP) was used to retrieve deletion calls of CNVs at the population level (Mills et al. 2011). Each CNV was genotyped, and the genotype quality was estimated based on the measurement of genotype likelihoods. To ensure that only highly plausible variants are retained, I selected CNVs that passed all genotype quality thresholds in Genome STRiP. Genome STRiP has four filtering criteria for defining deleted CNVs. The definition and default values of the four criteria in Genome STRiP are as follows: COHERENCE (incoherence metric > 0.01), COVERAGE (median normalized read depth of samples with observed evidentiary pairs < 1.0, this filter was used to remove calls in regions of unusually high sequence coverage across many samples), DEPTH (depth ratio <= 0.63 or depth ratio <= 0.8 and heterogeneity P value < 0.01), DEPTHPVAL (Depth p-value using chi-squared test < 0.01). When Genome STRiP defines CNVs, each CNV must pass the four criteria. The number of CNVs decreased from 44,388 to 9,732 CNVs following the filtering criteria. After this step, I applied a secondary criterion to check individual quality for each CNV. In this filtering process, Genome STRiP used genotype likelihoods test. If all individual did not pass through this filtering, I could not obtain the CNV genotype information. After removing the low quality CNVs, 6,811 deleted CNVs remained. I regarded these 6,811 CNVs as the cattle CNVs in this study for additional analyses (Figure 6.2).

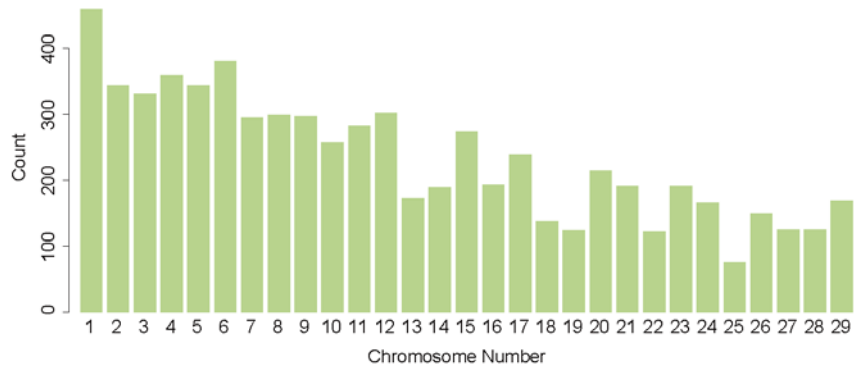

Figure 6.2. Distribution per chromosome of the deleted CNV on the cattle genome

6.3.3 Gene Content of Deleted Copy Number Variations

The gene content of each CNV was assessed by searching each CNV sequence against the Ensemble gene database (Flicek et al. 2012). I used BioMart in the Ensemble database to obtain the Ensemble gene IDs for the genes that overlapped with a CNV of the 32 cows (Smedley et al. 2009). The total number of Ensemble gene IDs was 23,431 and 1,508 CNVs were related to 1,228 Ensemble genes. Deletion score was defined as the number of total deletions in a gene region, as follows:

$$Deletion\ Score\ per\ gene = \sum_{n=1}^{l} (\#Deletion\ per\ CNV\ in\ Gene) \quad (6.1)$$

Where,

#Deletion = deletion number in 32 individuals in each CNV (range 0 to 32)

l = number of CNVs in each gene

I assigned a deletion score to each gene that overlapped with CNVs (Figure 6.3).

To discover significant genes that overlap with CNVs that may be affected by the deleted CNV, I calculated empirical p-values for each CNV overlapping gene. I assumed that the distribution of total deletion score values of the 1,228 CNV overlapping genes was a normal distribution. The empirical p-value of each CNV overlapping gene was derived from this normal distribution. Then I selected genes with the top deletion scores (p-value < 0.01) as the representative genes related to cattle domestication. These genes were used to perform Gene Ontology (GO) analysis and pathway analysis in Database for Annotation, Visualization and Integrated Discovery (DAVID; version 6.7) (Da Wei Huang and Lempicki 2008).

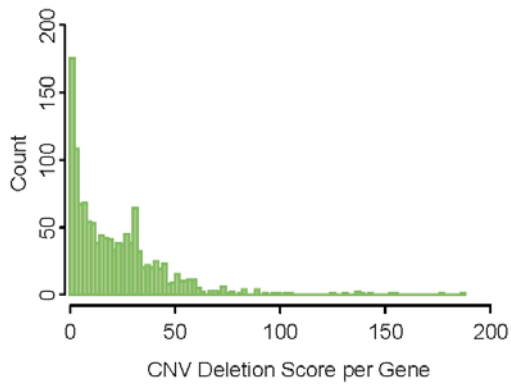

Figure 6.3. Distribution of the deletion score for the bovine genes

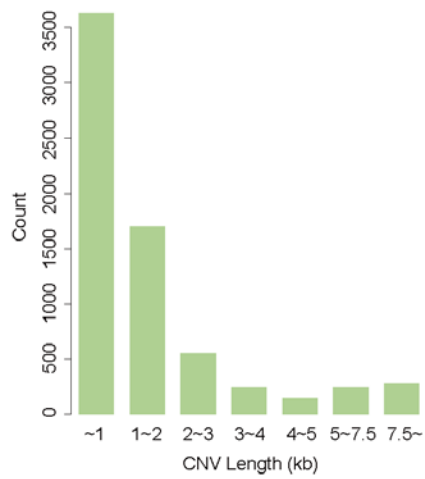

Figure 6.4. Histogram of the bovine deleted CNV length

6.3.3 Quantitative Trait of Deleted Copy Number Variations

I compared CNV regions with the cattle QTL regions to explain the role of extracted CNVs in a quantitative trait. The quantitative trait content of each CNV was assessed by selecting QTL regions that overlapped with CNV regions in the 32 cows. The Animal QTL database was used to obtain all QTL region information (Hu, Fritz and Reecy 2007). QTL traits of cattle can be largely divided into 12 traits with 3,605 loci. The length of QTL was found to be highly variable (minimum: 1000 bp; maximum: 134,956,528 bp; median: 208,803 bp; average score: 7,738,095bp) (Figure 6.4). The average distance between deletions calculated as the QTL length divided by the deletion score was defined as the CNV density within a QTL. The CNV density was calculated for all QTL related to cattle CNV and the top 30 QTL were selected as being representative QTL related to cattle CNV.

6.3.4 Population Structure Analysis & Phylogenetic Inference

Two preliminary analyses were performed to infer the population structure of the 32 cows used in this study. The program STRUCTURE was used to evaluate the extent of substructuring between Holstein and Hanwoo (Pritchard, Stephens and Donnelly 2000). I determined that an initial burn-in of 10,000 iterations followed by 10,000 iterations for parameter estimation was sufficient to ensure convergence of parameter estimates. To estimate the number of populations (the K parameter of STRUCTURE), the dataset was analyzed by allowing for the values of K = 2 and 3. PCA was conducted for the CNV genotypes in the 32 cows using the statistical program R. For further identification of the evolutionary history of the samples, I constructed a phylogenetic tree using Bayesian inferences (BI) approaches. Bayesian phylogenetic inference is based on Bayes's rule. The first characteristic of Bayesian

inferences is the use of distribution referred to as the prior that specifies the prior probability of different parameter values. Additionally, this method uses the likelihood function that describes the probability of the data under different parameter values and the total probability of the data summed and integrated over the parameter space to infer a phylogenetic tree. As a result, Bayesian inference is based on the so-called posterior distribution. Phylogenetic analysis in this study was carried out using BI analytical method executed in MrBayes 3.1.2. (Ronquist and Huelsenbeck 2003) with the following options: nst: 6, rates: gamma, number of generations: 2,000,000, sample frequency: 100, number of chains: 4, and burn-in generation: 20,000. To estimate the reliability of the nodes, the Bayesian posterior probability (BPP) values were calculated as shown on the BI tree.

6.3.5 Identifying Selection Signal using FST

Wright (Wright 1949) defined several F coefficients that describe evolutionary processes. His definition was in terms of correlations among gamete: so I used Nei's equivalent definitions in terms of deviations from expected heterozygosities.

$$F = \frac{H_{exp} - H_{obs}}{H_{exp}} \; and \; H_{obs} = \sum c_j(2p_j q_j), H_{exp} = \sum 2E(p)E(q) \quad (6.2)$$

Where,

H_{obs} = the observed frequency of heterozygotes

c_j = relative size (proportion) of j^{th} subpopulation

p_j = frequency of deletion in j^{th} subpopulation

q_j = frequency of allele in j^{th} subpopulation

6.3.6 Identification of Breed-specific Copy Number Variations

Each CNV that passed the applied filtering criteria was labeled as a putative breed-specific CNV if the allele was present in only one of the two breeds. Among the putative breed-specific CNVs, CNVs with a deletion frequency of more than 0.1 in each population were selected as breed-specific CNVs. Gene related to the breed-specific CNV were selected and Gene Ontology (GO) analysis was performed in Database for Annotation, Visualization and Integrated Discovery (DAVID; version 6.7) (Da Wei Huang and Lempicki 2008).

6.3.7 CNV Validation

I selected seven putative genes (TTN, SLIT3, KLHL1, NCAM2, MDGA2, EFNA5 and PRKG1) that contain the impact of cattle domestication and performed PCR to confirm the 19 CNVs within these genes. Originally, there were 25 CNVs in the seven genes but six CNVs with a length of greater than 1.5 Kb were excluded in this validation. Genomic DNA (gDNA) samples from ten Holstein and 22 Hanwoo were used to validate the CNV region selected by Genome STRiP (Mills et al. 2011) and determine if they were genuine CNV regions. The primer pairs were designed to be located outside of the predicted CNV region or inside and outside of the CNV region for cases where only the deleted allele was detected. Fifty nanograms of gDNA was used for PCR amplification and the reaction was performed by using a 2x PCR master mix solution (iNtRON Bio Technology, Seongnam, Gyeonggi, Korea) with 0.5 μM of each primer set. The amplification was performed under the following conditions: 1 cycle of 95°C for 5 min; 35 cycles of 95°C for 30 sec, annealing at the 58~66 °C for 30 sec, and 72°C for 1 min or 1min 30 sec; and 1 cycle

of 72°C for 10 min. All PCR products were visualized on 1% ethidium bromide stained gels run for 25min.

6.4 Results

Illumina NGS data were obtained from 10 Holsteins, a dairy cattle, and 22 Hanwoo, a beef cattle. The sequence data for each individual yielded approximately 13.58-fold to 20-fold coverage (Table 6.1). To provide a complete and accurate estimate of CNV at the population level, I used Genome STRiP which combines several technical features including breakpoint-spanning reads, paired-end sequences, and local variation in read depth of coverage (Mills et al. 2011). This method had sufficient power to detect deleted CNVs across the autosomes but not enough power to discover inserted events. In this analysis, I focused on the characterization of high-confidence deleted CNVs from known autosomes in UMD 3.1. A total of 6,811 deleted CNVs were detected among the analyzed animals (average length = 2732.2 bp) corresponding to 18.6 Mbp of variable sequence or 0.74% of the entire cattle genome. Using this information, I constructed deleted CNV maps for the cattle genome, which encompassed 1,228 Ensemble cattle reference genes and 2,220 quantitative trait loci (QTL). A full CNV call is shown in the deleted CNV map with breed-specific CNVs (Figure 6.5). Out of the 6,811 CNVs, 4,407 (9.9 of 18.6 Mbp; 53.1%) were shared between Holstein and Hanwoo and only 2 CNVs (BovineCNV5631, BovineCNV5701) were monomorphic.

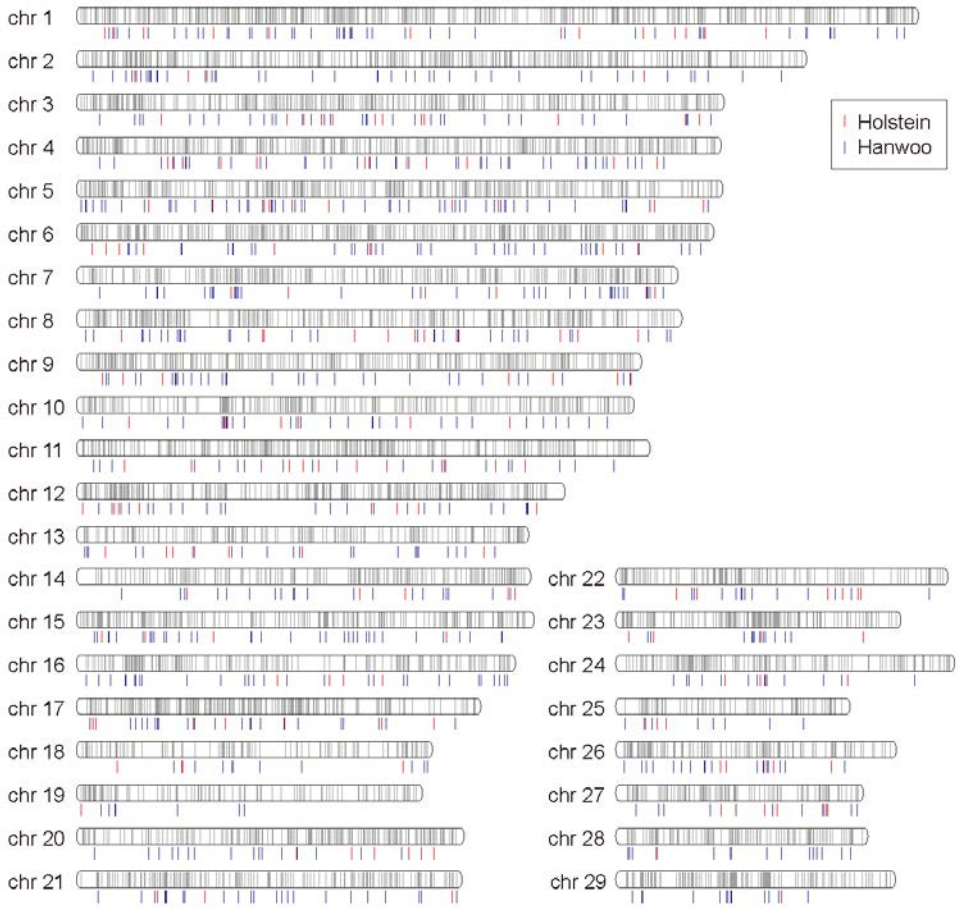

Figure 6.5. Cattle deleted CNV map with breed-specific CNVs. Gray bar represents the cattle deleted CNVs using the entire population of this study. Holstein and Hanwoo-specific CNVs are represented as red and blue bars, respectively. The position of the bars is based on the physical position information of UMD 3.1.

Using the cattle reference gene annotations, I identified CNVs that overlap with genes and then assigned deletion scores to each gene. Out of the 23,431 cattle Ensemble reference genes, 1,228 genes (5.24%) overlapped with the deleted CNVs in this study. The overlapping genes showed high variation in the deletion score with a minimum score of 1, maximum score of 187, median score of 14 and an average score of 21.95. Among the overlapping genes, 33 Ensemble genes had an empirical p-value of less than 0.01 and were considered as being significant in cattle domestication (Table 6.2). While 10 of the 33 Ensemble genes did not have a defined function, Gene Ontology analysis revealed that the remaining 23 genes were related to the nervous system, more specifically nervous transmission, neuron motion, and neurogenesis (Figure 6.6). Ten genes (cluster 1 of GO analysis, Figure 6.6) were found to be related to nervous transmission (NCAM2, PIK3C2G, EFNA5, RASGRF2, UNC13C, GUCY1A2, ACCN1, GRM7, DCDC2, and PCDH15). Of these 10 genes, five genes (NCAM2, EFNA5, UNC13C, GRM7, and PCDH15) have been previously reported to be related to nervous transmission (Collingridge and Lester 1989, Xu et al. 1998, McIntyre, Meldrum and Garthwaite 1990, Bliss and Collingridge 1993, Cartmell and Schoepp 2000, Yagi and Takeichi 2000, Titlow and McClintock 2010, Winther, Berezin and Walmod 2012) (Table 6.3). Six genes (cluster 2 of GO analysis, Figure 6.6) were found to be related to neuron motion (EFNA5, KLHL1, DNAH5, SLIT3, DCDC2, and PRKG1). Five of these genes in cluster 2 (EFNA5, DNAH5, SLIT3, DCDC2, and PRKG1) were reported to be related to neuron motion in previous studies (Gleeson et al. 1999, Davy et al. 1999, Sasaki et al. 2000, Brose and Tessier-Lavigne 2000, Bilimoria and Bonni 2013) (Table 6.3). Eight genes (cluster 3 of GO analysis, Figure 6.6) were found to be related to neurogenesis (NCAM2, EFNA5, MDGA2, KLHL1, SLIT3, PRKG1, PCDH15, and FAT3). I identified that seven of these eight genes in cluster 3 (NCAM2, EFNA5, MDGA2, KLHL1, SLIT3, PRKG1, FAT3) have previously been

reported to be related to neurogenesis (Itoh et al. 1998, Rønn, Hartz and Bock 1998, Nemes, Benzow and Koob 2000, Litwack et al. 2004, Nagae, Tanoue and Takeichi 2007, Hara et al. 2010, Yoneyama et al. 2011) (Table 6.3). Also, the pathway analysis using 33 significant Ensemble gene IDs based on deletion scores showed that only the pathway related to axon guidance is significant. Three genes, EPHA6, EFNA5, and SLIT3, were associated with this pathway.

Table 6.2. Deletion score top 1 % (p-value < 0.01) genes.

Ensemble Gene	Gene Symbol	Chr	Gene Start	Gene End	#CNV	CNV Deletion Score	Empirical P-value
ENSBTAG00000045905	PCDH15	26	5,017,714	5,578,654	8	187	2.03E-14
ENSBTAG00000044144	GUCY1A2	15	16,604,381	17,100,655	10	177	6.20E-13
ENSBTAG00000018996	BT.61689	9	98,421,510	99,568,077	6	155	5.53E-10
ENSBTAG00000035007	GALNTL6	8	3,816,704	5,330,369	8	154	7.35E-10
ENSBTAG00000025522	UNC13C	10	55,717,362	56,389,601	4	143	1.48E-08
ENSBTAG00000005697	MDGA2	10	39,914,871	40,445,335	7	140	3.22E-08
ENSBTAG00000001133	KIAA0564	12	11,705,780	12,059,605	5	137	6.86E-08
ENSBTAG00000045699	BT.102825	28	22,419,203	24,270,401	5	137	6.86E-08
ENSBTAG00000013047	GRM7	22	18,740,484	19,647,747	6	132	2.33E-07
ENSBTAG00000016515	EFNA5	7	109,049,590	109,217,439	4	125	1.18E-06
ENSBTAG00000002966	BT.20044	1	138,139,496	138,305,752	3	105	7.14E-05
ENSBTAG00000018404	PRKG1	26	6,906,081	8,343,629	5	104	8.58E-05
ENSBTAG00000006392	TTC7B	10	103,185,307	103,381,020	3	99	0.00021
ENSBTAG00000008708	BT.93891	12	66,292,489	67,324,791	5	97	0.00029
ENSBTAG00000025200	ACCN1	19	16,353,233	17,562,209	7	93	0.00057
ENSBTAG00000027899	-	5	67,852,917	67,930,472	3	89	0.00107
ENSBTAG00000021969	BT.40893	11	75,290,644	75,623,195	3	89	0.00107
ENSBTAG00000020715	PIK3C2G	5	91,835,146	92,276,939	2	89	0.00107

ENSBTAG00000008647	KLHL1	12	44,295,888	44,616,940	4	89	0.00107
ENSBTAG00000000655	MIPOL1	21	47,852,815	48,169,316	2	84	0.00224
ENSBTAG00000009798	DCDC2	23	33,102,946	33,246,914	2	84	0.00224
ENSBTAG00000021972	DNAH5	20	59,285,274	59,560,606	4	84	0.00224
ENSBTAG00000005014	BT.63292	11	69,204,181	69,321,396	2	83	0.00259
ENSBTAG00000004081	FAT3	29	1,965,869	2,605,125	4	82	0.00298
ENSBTAG00000021291	BT.88441	21	35,415,163	35,655,955	2	78	0.00513
ENSBTAG00000017746	SLIT3	20	324,518	507,045	2	77	0.00585
ENSBTAG00000019823	ADAMTS17	21	6,514,360	6,924,114	2	76	0.00666
ENSBTAG00000030259	RASGRF2	7	83,362,788	83,546,393	3	74	0.00857
ENSBTAG00000003301	NCAM2	1	14,791,090	15,026,555	2	74	0.00857
ENSBTAG00000031358	BT.66173	3	45,563,732	46,487,165	4	74	0.00857
ENSBTAG00000044111	EPHA6	1	40,608,729	41,623,778	3	74	0.00857
ENSBTAG00000000432	BT.104278	10	22,106,909	23,304,334	2	74	0.00857
ENSBTAG00000000939	KIF16B	13	10,238,640	10,519,339	3	73	0.00970

Table 6.3. Deletion score top 1 % (p-value < 0.01) genes identified in this study and in previous studies.

Clustering Name	Gene	Chr	Reference	Gene Description
Clustering 1 : nervous transmission	NCAM2	chr1	Winther, Berezin et al. 2012	The protein encoded by NCAM2 may play important roles in selective fasciculation and zone-to-zone projection of the primary olfactory axons.
	EFNA5	chr7	McIntyre, Titlow et al. 2010	EFNA5 Receptors that mediate axonal inhibition or repulsion tended to be expressed in olfactory sensory neurons.
	UNC13C	chr10	Xu, Wes et al. 1998	UNC13C encodes protein unc-13 homolog, which has been shown to function in synaptic transmission
	GRM7	chr22	Collingridge and Lester 1989, Meldrum and Garthwaite 1990, Bliss and Collingridge 1993, Cartmell and Schoepp 2000	Metabotropic glutamate receptors of GRM7 are present in varying degree at various synapses and regulate transmitter release. Glutamate receptors mediate most of the excitatory synaptic transmission in the mammalian central nervous system and play crucial roles in synaptic plasticity, learning and memory, and in some neuropathological disorders.
	PCDH15	chr26	Yagi and Takeichi 2000	Cadherins encoded by PCDH15 have been identified as synaptic components, and their suggested roles include neuronal circuitry, synaptic junction formation, and synaptic plasticity.
Clustering2 :neuron motion	EFNA5	chr7	Davy, Gale et al. 1999	Cell surface GPI-bound ligands for Eph receptors are crucial for migration, repulsion and adhesion during neuronal, vascular and epithelial development.
	DNAH5	chr20	Sasaki, Shionoya et al. 2000	This DNAH5 encodes a dynein protein, which is part of a microtubule-associated motor protein complex consisting of heavy, light, and intermediate chains.
	SLIT3	chr20	Brose and Tessier-Lavigne 2000	SLIT3 encodes slit homolog 3, which may act as molecular guidance cue in cellular migration and SLIT proteins were identified as being both negative and positive regulators, repelling various axonal and cell migrations.
	DCDC2	chr23	Gleeson, Lin et al. 1999	DCDC2 encodes a member of the doublecortin family which is a microtubule-associated protein expressed by neuronal precursor cells and immature neurons.

	PRKG1	chr26	Bilimoria and Bonni 2013	The soluble I alpha and I beta isoforms of PRKG by alternative transcript splicing, which play a central role in axon branching.
Clustering3 :neurogenesis	NCAM2	chr1	Rønn, Hartz et al. 1998	NCAM have been shown to be crucial for the formation of the olfactory bulb and the mossy fiber system in the hippocampus.
	EFNA5	chr7	Hara, Nomura et al. 2010	Ephrin-A5, a ligand for Eph receptor tyrosine kinases, plays multiple roles in both neurogenesis and vascular formation in the adult hippocampus.
	MDGA2	chr10	Litwack, Babey et al. 2004	MDGA2 protein play a role in neural development, including axon guidance
	KLHL1	chr12	Nemes, Benzow et al. 2000	Protein KLHL1 belongs to a family of actin-organizing proteins and may play a role in organizing the actin cytoskeleton of the brain cells
	SLIT3	chr20	Itoh, Miyabayashi et al. 1998	The slit proteins may participate in the formation and maintenance of the nervous and endocrine systems by protein–protein interactions
	PRKG1	chr26	Yoneyama, Kawada et al. 2011	The PRKG1 proteins play a role in proliferation of neural stem/progenitor cells
	FAT3	chr29	Nagae, Tanoue et al. 2007	Classic Fat is known to regulate cell proliferation and planar cell polarity And Fat3 plays a role in the interactions between neurites derived from specific subsets of neurons during development.

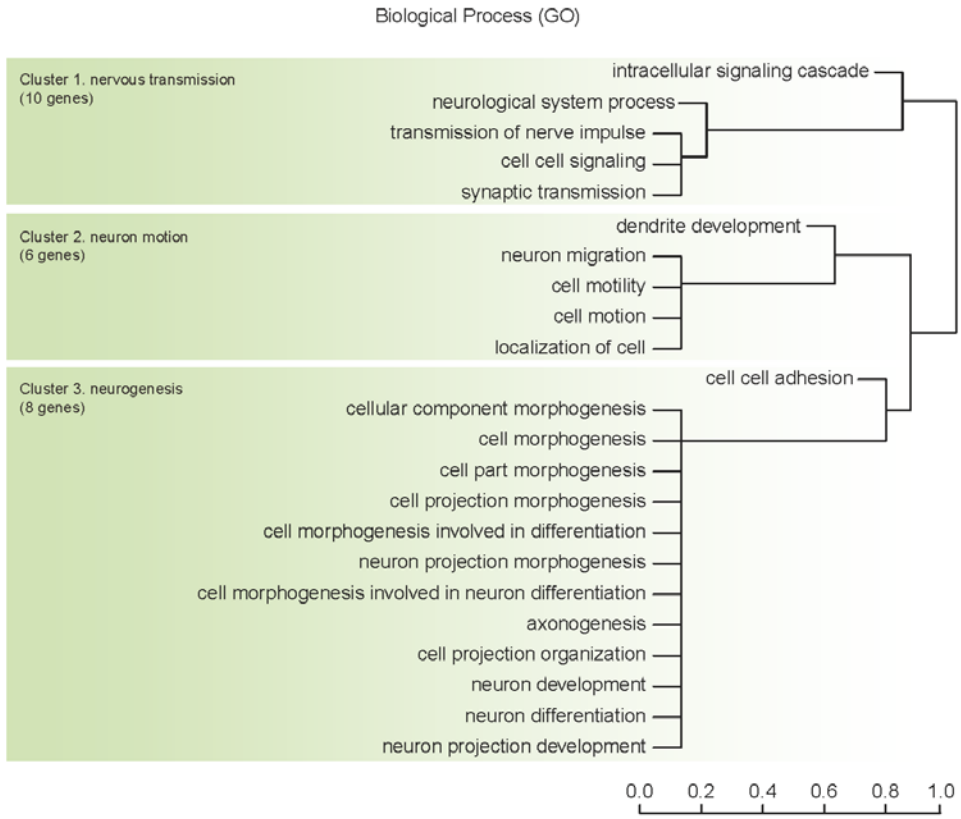

Figure 6.6. Hierarchical clustering of significant GO terms for genes with top 1% deletion scores. Significant results of GO analysis using genes with top 1% deletion scores run with default criteria in DAVID GO analysis (COUNT=2, EASE=0.1). These results are largely divided into three clusters: cluster 1 with 10 genes related to nervous transmission. cluster 2 with six genes associated with neuron motion, and cluster 3 with eight genes linked to neurogenesis.

QTL related to CNV regions were identified using the Animal QTL database [62]. I found that 2,220 out of 3,605 (61.58%) cattle QTL overlapped with 6,623 putative deleted CNVs. The index used to measure deletion density, the average distance between deletions, showed large variations (minimum: 1069.86 bp; maximum: 3,728,838 bp; median: 20693.06 bp; average score: 31433.17 bp). The top 30 QTL overlapping with CNVs are listed in Table 6.4. CNV deletion scores of the top 30 QTL were also highly variable (between 50 and 142). Six of the top QTL were directly related to meat production while eight of the top QTL were associated with milk production. I also propose genes that overlap with the top 30 QTL (Table 6.4), which are mainly related to sensory perception as olfactory receptor.

Table 6.4. Top 30 QTLs using average distance between deletions

QTL Name	QTL ID	Chr	#CNV	#CNV Deletion Score	QTL Length (bp)	average distance between deletions	QTL Trait	Overlapping Ensemble Gene (Gene Symbol)
chr_2_5293683_5505175	20298	2	2	82	211,492	2579.17	Meat Association	ENSBTAG00000019177 (BT.54746) ENSBTAG00000045010 (bta-mir-2350) ENSBTAG00000034949 (HIST1H2AE)
chr_3_18099270_18291063	13113	3	2	77	191,793	2490.82	Health Association Milk Association	ENSBTAG00000008700 (CRCT1) ENSBTAG00000045563 ENSBTAG00000046138 (C1ORF68) ENSBTAG00000032446
chr_4_6695822_6899873	1499	4	2	77	204,051	2650.01	Health QTL	-
chr_5_58828203_59001914	15407	5	5	94	173,711	1847.99	Reproduction Association	ENSBTAG00000024607 (OR6C75) ENSBTAG00000046645 ENSBTAG00000006313 (OR6C76) ENSBTAG00000048224 ENSBTAG00000045684
chr_5_59557792_59731504	14027	5	3	81	173,712	2144.59	Production Association Meat Association	ENSBTAG00000039756 (OR10A7) ENSBTAG00000047259 ENSBTAG00000047967 ENSBTAG00000031097 ENSBTAG00000031096 ENSBTAG00000047619 ENSBTAG00000026078 ENSBTAG00000037629

QTL Name	QTL ID	Chr	#CNV	#CNV Deletion Score	QTL Length (bp)	average distance between deletions	QTL Trait	Overlapping Ensemble Gene (Gene Symbol)
chr_5_59592534_59766246	4412	5	3	81	173,712	2144.59	Production QTL	ENSBTAG00000039756 (OR10A7) ENSBTAG00000047259 ENSBTAG00000047967 ENSBTAG00000031097 ENSBTAG00000031096 ENSBTAG00000047619 ENSBTAG00000037629
chr_5_7933409_8107121	20385	5	2	65	173,712	2672.49	Meat Association	-
chr_5_99090659_99264371	5068	5	3	74	173,712	2347.46	Reproduction QTL	ENSBTAG00000030468 (BT.76064) ENSBTAG00000023258 (BT.76070) ENSBTAG00000030466 (BT.76067) ENSBTAG00000030461 (BT.76068) ENSBTAG00000030463 (BT.76065)
chr_6_10548494_10723996	10146	6	4	92	175,502	1907.63	Milk QTL	-
chr_6_28262771_28438274	16296	6	3	88	175,503	1994.35	Health Association	-
chr_6_32148739_32324241	4541	6	2	66	175,502	2659.12	Production QTL	-
chr_6_33640512_33816015	5382	6	2	110	175,503	1595.48	Production QTL	-
chr_6_33658063_33833565	14705	6	2	110	175,502	1595.47	Reproduction Association	-
chr_6_92328701_92504203	9914	6	2	71	175,502	2471.86	Milk QTL	ENSBTAG00000032074 (U1) ENSBTAG00000007692 ENSBTAG00000017028 (USO1) ENSBTAG00000015449 (PPEF2)

QTL Name	QTL ID	Chr	#CNV	#CNV Deletion Score	QTL Length (bp)	average distance between deletions	QTL Trait	Overlapping Ensemble Gene (Gene Symbol)
chr_7_31966928_32132336	14036	7	2	81	165,408	2042.07	Milk Association	ENSBTAG00000020578 (PRDM6) ENSBTAG00000001568 (PPIC)
chr_9_17969953_18152603	18466	9	4	78	182,650	2341.67	Meat Association	-
chr_10_26933288_27109657	10052	10	5	91	176,369	1938.12	Meat Association	ENSBTAG00000047483 ENSBTAG00000039315 ENSBTAG00000048109 ENSBTAG00000038485 ENSBTAG00000037959 ENSBTAG00000006198 ENSBTAG00000038868 ENSBTAG00000038227 ENSBTAG00000045521 ENSBTAG00000032798
chr_11_99886324_100048919	10463	11	2	82	162,595	1982.87	Milk Association	ENSBTAG00000019513 (C9ORF50) ENSBTAG00000015437 (BT.29263)
chr_12_49559394_49720836	3385	12	3	99	161,442	1630.73	Reproduction Association Production Association	ENSBTAG00000042618 (7SK) ENSBTAG00000005760 (TBC1D4)
chr_12_50520004_50681445	5047	12	3	67	161,441	2409.57	Milk QTL	-
chr_15_3685265_3828842	5063	15	2	82	143,577	1750.94	Reproduction QTL	-
chr_17_23527916_23679836	4444	17	4	142	151,920	1069.86	Production QTL	-
chr_17_24530877_24682797	14865	17	2	74	151,920	2052.97	Health Association	-
chr_18_65270255_65421839	6114	18	3	73	151,584	2076.49	Milk QTL	-

QTL Name	QTL ID	Chr	#CNV	#CNV Deletion Score	QTL Length (bp)	average distance between deletions	QTL Trait	Overlapping Ensemble Gene (Gene Symbol)
chr_19_7061002_7168138	4935	19	1	50	107,136	2142.72	Production QTL	-
chr_23_20221868_20346736	20494	23	2	73	124,868	1710.52	Health QTL	ENSBTAG00000021609 (GPR110)
chr_23_29418035_29542903	12178	23	4	87	124,868	1435.26	Meat QTL	ENSBTAG00000038562 ENSBTAG00000040582 ENSBTAG00000027955 ENSBTAG00000040280 (OR2J3) ENSBTAG00000038928 ENSBTAG00000047558
chr_27_21868197_21976582	15284	27	2	53	108,385	2045.00	Reproduction Association	-
chr_28_10873331_11010357	6140	28	3	93	137,026	1473.40	Milk QTL	ENSBTAG00000046453
chr_28_17021860_17158886	15537	28	2	74	137,026	1851.70	Reproduction Association	-

I identified selective signals between Hanwoo and Holstein populations from CNV based FST to annotate regions of selection. Differences in the frequencies of deleted CNVs for each breed were used to characterize signatures of selection. The CNVs selected based on FST exhibited evidence of evolutionary selection in genomic regions that were considered to have been under positive selection in meat and dairy cattle. Ninety-four deleted CNVs were identified as putatively harboring selective sweep signals with FDR multiple test corrected empirical p-values (less than 0.01) of FST. Seventeen Ensemble genes overlapped with CNVs and gene function was defined in 14 of the genes (Figure 6.7). Seven (TTN, MATN3, DST, HDAC4, TSHR, CCDC141, GALK2) of the 14 genes were reported to be related to representative economic traits of each breed [63-68] (Table 6.5). Two (MATN3, DST) of these seven genes had deleted CNVs mainly in Holstein while the other five (TTN, HDAC4, TSHR, CCDC141, GALK2) had deleted CNVs mainly in Hanwoo (Figure 6.8).

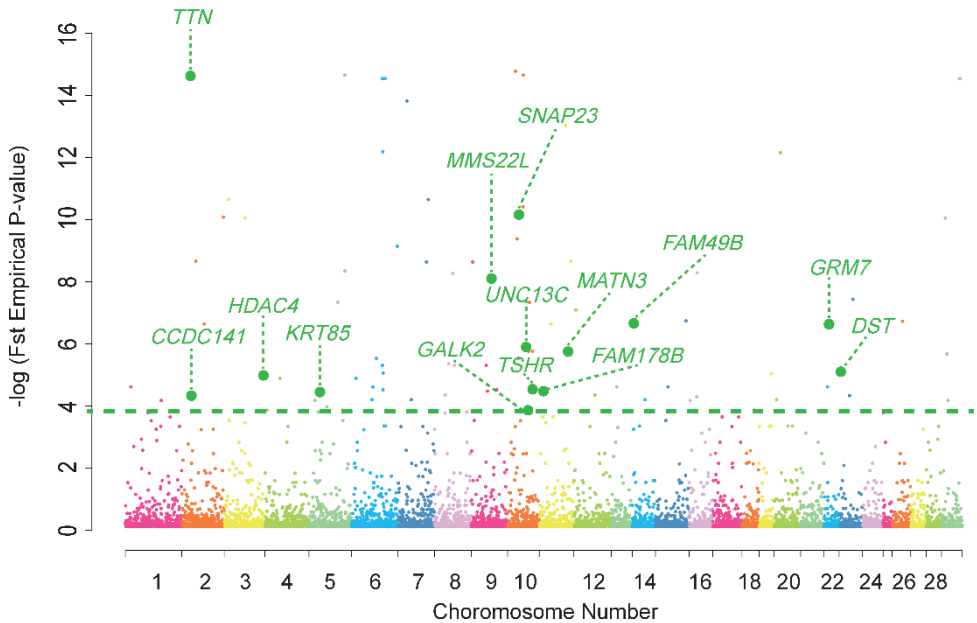

Figure 6.7. Manhattan plot of FST empirical p-value. Results plotted as negative log-transformed empirical p-values of FST. The green horizontal line indicates the threshold (p-value after FDR correction <=0.01). The green circle represents the cattle CNVs that are significant and belongs to genic regions.

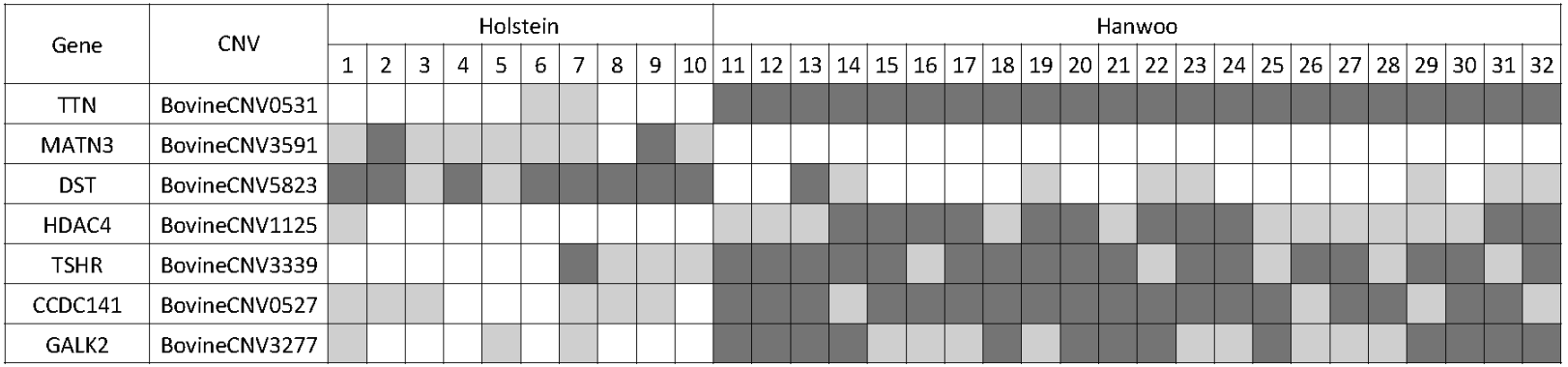

Figure 6.8. Genotype information of the top seven cattle CNVs using FST, which may impact the differences between Hanwoo and Holstein.

Table 6.5. Gene description and reference of the top seven cattle CNVs using Fst, which may impact the differences between Hanwoo and Holstein.

CNV	Gene	Chr	Reference	Gene Description
BovineCNV0531	TTN	chr2	Yamada, Sasaki et al. 2009	TTN is involved in myofibrillogenesis and through association study of a single nucleotide polymorphism (SNP) in Japanese Black beef cattle.
BovineCNV3591	MATN3	chr11	Yucesoy, Charles et al. 2013	MATN3 is related to genetic risk factors for osteoarthritis which related to dairy production.
BovineCNV5823	DST	chr23	Cole, Wiggans et al. 2011	The second most significant SNP effect in US holstein GWAS study for daughter stillbirth was the dystonin gene (DST) on BTA23.
BovineCNV1125	HDAC4	chr3	Youn, Grozinger et al. 2000	HDAC4 constitute a family of calcium-sensitive transcriptional repressors of myocyte enhancer factor 2.
BovineCNV3339	TSHR	chr10	Pipes, Bauman et al. 1963	TSHR encodes thyroid stimulating hormone receptor and it has been shown that in beef cattle there is a significantly lower thyroxine secretion rate than in dairy cattle.
BovineCNV0527	CCDC141	chr2	Fukuda, Sugita et al. 2010	CCDC141 encodes Coiled-Coil Protein Associated With Myosin II.
BovineCNV3277	GALK2	chr10	Mohammad, Hadsell et al. 2012	GALK2 has been shown to be upregulated during the secretory activation in initiation of milk production.

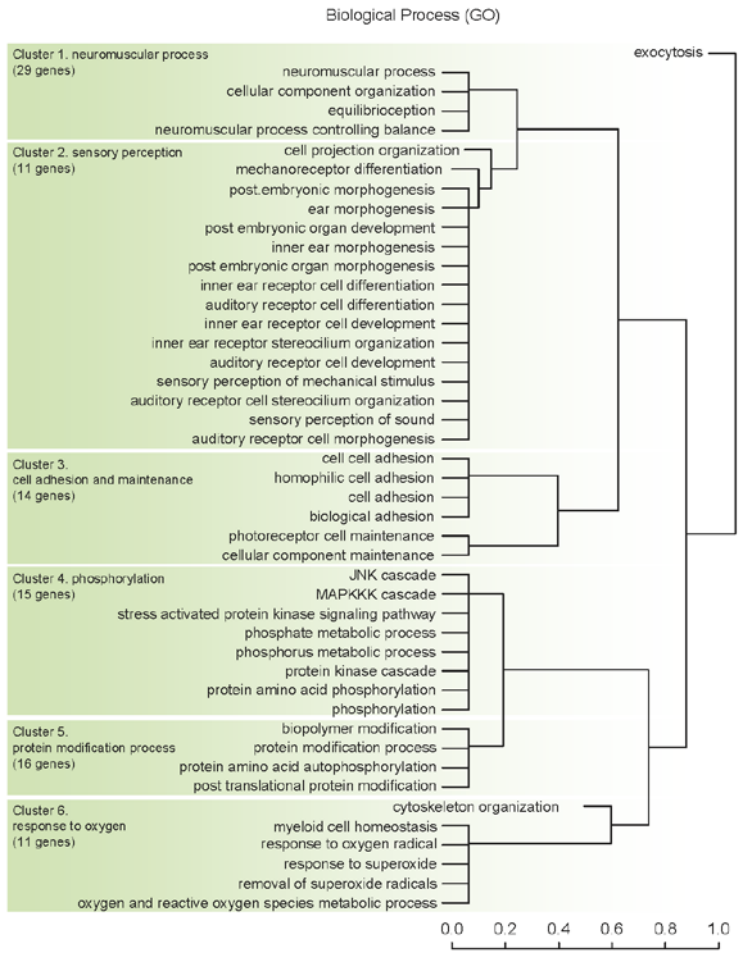

Figure 6.9. Hierarchical clustering of significant GO terms for genes that overlap with Hanwoo breed-specific CNVs. Significant result of GO analysis using genes with top 1% deletion scores run with default criteria in DAVID GO analysis (COUNT=2, EASE=0.1). These results are largely divided into six clusters: cluster 1 with 29 genes related to neuromuscular process, cluster 2 with 11 genes associated with sensory perception, cluster 3 with 14 genes linked to cell adhesion and maintenance, cluster 4 with 15 genes related to phosphorylation, cluster 5 with 16 genes associated with protein modification process and cluster 6 with 11 genes linked to response to oxygen and maintenance.

Breed-specific CNVs were identified to investigate their ability to explain breed-specific traits. Although substantial portions of the total CNVs were shared between Holstein and Hanwoo, I found putative breed-specific CNVs for each breed. A total of 2,404 CNVs corresponding to 8.73 Mbp of sequence indicated that deletion was present in only one of the two breeds. After filtering, 767 Hanwoo-specific CNVs and 187 Holstein-specific CNVs were identified (Table 6.6, Table 6.7). Hanwoo-specific CNVs were more abundant than Holstein-specific CNVs. I assigned all breed-specific deleted CNVs to a nearby Ensemble gene ID. For the Hanwoo-specific CNVs, 177 of 767 CNVs were related to 173 Ensemble genes, of which 137 had gene symbol for biological interpretation (Table 6.6). Gene Ontology analysis showed that these genes were related to neuromuscular process, sensory perception, cell adhesion and maintenance, phosphorylation, protein modification process, and response to oxygen (Figure 6.9). Cluster 1 of GO analysis result includes 29 genes (Figure 6.9) associated with neuromuscular process (ARHGAP10, ARID2, CADPS2, CDH23, CHD9, DNAH9, DSG1, DYNC2H1, EPB41L2, EXOC4, FANCC, GORASP2, ITGAV, KLHL1, LMX1A, MPP6, MYO7A, PALLD, PCDH15, RAPGEF4, RIN3, SMYD3, SOD1, STXBP5L, TLN2, TRPM7, TTF2, USH2A, and UTRN). The second cluster of GO analysis result (Figure 6.9) includes 11 genes related to sensory perception (CDH23, DNAH9, DYNC2H1, GRM7, KLHL1, LMX1A, MYO7A, NTRK3, PCDH15, SOD1, and USH2A). Cluster 3, which contained 14 genes, was associated with cell adhesion and maintenance (CDH23, CNTN6, COL28A1, DSG1, FAT3, FER, ITGAV, LAMB3, PCDH15, PTPRC, PTPRT, TLN2, TRPM7, and USH2A). Cluster 4 with 15 genes was related to phosphorylation (DAPK1, EPHA5, FER, GAB1, LRRK2, MAP4K3, MAPK10, NDUFA10, NTRK3, PTPRC, PTPRT, RPS6KA2, SOD1, TRPM7, and WNK1) and 16 genes in cluster 5 were linked to protein modification (DAPK1, EPHA5, FBXW2, FER, GAB1, LRRK2, MAP4K3, MAPK10, NTRK3, PTPRC, PTPRT, RPS6KA2,

SOD1, TPST1, TRPM7, and WNK1). The final cluster, cluster 6, which includes 11 genes, was connected with response to oxygen (ARHGAP10, CDH23, EPB41L2, FANCC, KLHL1, MYO7A, PALLD, PLCB1, SOD1, TLN2, and TRPM7). Seventeen genes were associated with phosphorylation (cluster 4 in Figure 6.9) or protein modification process (cluster 5 in Figure 6.9). I speculate that many of the genes are related to cell growth in phosphorylation and protein modification process, which are needed for the production of meat in the muscle mechanism. Previous studies reported relationships between 10 (NDUFA10, WNK1, MAPK10, FER, RPS6KA2, MAP4K3, PTPRT, PTPRC, GAB1, NTRK3) of the 17 genes related to phosphorylation and protein modification process, and cell growth (Sun and Tonks 1994, Aspenström 1997, Moore et al. 2000, Rodrigues et al. 2000, Lannon and Sorensen 2005, Fluckey et al. 2006, Yin et al. 2007, Perez, Cañón and Dunner 2010) (Table 6.7). Fourteen genes were related to cell adhesion and maintenance (cluster 3 in Figure 6.9). Out of these, nine genes (ITGAV, COL28A1, FER, TLN2, LAMB3, DSG1, PCDH15, CDH23 and FAT) have been shown to be directly linked to cell adhesion and maintenance (Vidal et al. 1995, Rosato et al. 1998, Runswick et al. 2001, Marthiens et al. 2002, Mitsui et al. 2002, Veit et al. 2006, Halbleib and Nelson 2006, Takada, Ye and Simon 2007, Senetar, Moncman and McCann 2007) (Table 6.9). Additionally, it was found that only the neurotrophin signaling pathway was significant when 137 significant genes that overlapped with Hanwoo breed-specific CNVs were analyzed. Five genes, NTRK3, YWHAG, RPS6KA2, GAB1, and MAPK10, were associated with this pathway. For Holstein, 31 out of 187 breed-specific CNVs were related to 26 Ensemble genes (PNKD, PKLR, HCN3, SLC30A7, KIAA1324L, ADCYAP1R1, ZNF804B, NELL2, CNTN1, CRY1, SYNPO2, EFNA5, BAI3, PDE10A, AP3B1, CDAN1, GALM, MATN3, SUGT1, ZMYND8, CUX2, C6ORF10, BRUNOL4, C10ORF28, PSD3, and SLC35F3) (Table 6.6). I predicted that these genes might be linked to dairy production. Previous studies

reported relationships between ten (PKLR, ADCYAP1R1, NELL2, CRY1, EFNA5, PDE10A, AP3B1, GALM, MATN3, and C6ORF10) of the 26 genes and dairy production supporting the results of the analysis (Connor et al. 2008, Baik et al. 2009, Dostaler-Touchette et al. 2009, Sadkowski et al. 2009, Li et al. 2010a, Winters and Moore 2011, Zolla and Scaloni 2011, Casey and Plaut 2012, Bionaz et al. 2012, D'Alessandro, Yucesoy et al. 2013) (Table 6.10).

Table 6.6. Genes that overlapped with Hanwoo breed-specific CNVs

Ensemble Gene	Gene Symbol	Chr	Gene Start	Gene End	# Breed Specific CNV	# CNV
ENSBTAG00000021396	HEG1	1	70,027,330	70,106,415	1	2
ENSBTAG00000021703	MED12L	1	117,548,538	117,917,463	2	2
ENSBTAG00000030913	MX1	1	143,176,083	143,204,865	1	1
ENSBTAG00000020990	P2RY14	1	117,785,180	117,817,225	1	1
ENSBTAG00000001918	STXBP5L	1	66,213,612	66,537,723	1	1
ENSBTAG00000004891	OXNAD1	1	155,017,757	155,158,936	1	1
ENSBTAG00000013663	C1H3orf26	1	43,730,028	44,124,279	1	1
ENSBTAG00000026994	C2H2orf88	2	6,038,403	6,113,008	1	2
ENSBTAG00000044009	PPP1R1C	2	14,502,890	14,623,643	1	2
ENSBTAG00000020984	RAPGEF4	2	23,644,876	23,974,945	1	1
ENSBTAG00000013218	GORASP2	2	25,566,351	25,599,241	1	1
ENSBTAG00000011649	FARSB	2	111,504,465	111,578,408	1	1
ENSBTAG00000012217	PLA2G2F	2	133,111,283	133,124,161	1	1
ENSBTAG00000023963	RHBDD1	2	115,826,528	115,960,780	1	1
ENSBTAG00000019929	ITGAV	2	9,651,631	9,760,100	1	1
ENSBTAG00000000937	SSFA2	2	14,683,868	14,751,322	1	1
ENSBTAG00000025621	HNRNPA1	3	54,445,872	54,446,858	4	4
ENSBTAG00000012025	LMX1A	3	3,692,343	3,865,786	1	3
ENSBTAG00000043876	U2	3	54,610,767	54,610,931	2	2
ENSBTAG00000015392	TTF2	3	26,244,074	26,297,632	1	1
ENSBTAG00000003279	NDUFA10	3	119,779,215	119,794,481	1	1
ENSBTAG00000005439	FAM102B	3	34,839,059	34,877,721	1	1
ENSBTAG00000046773	MCOLN2	3	59,303,476	59,351,124	1	1
ENSBTAG00000032121	C7orf10	4	80,866,977	81,642,046	2	3
ENSBTAG00000010437	MKLN1	4	95,807,761	96,016,756	1	2
ENSBTAG00000014112	EXOC4	4	97,791,626	98,594,890	1	3
ENSBTAG00000024420	COL28A1	4	15,332,616	15,511,692	1	1
ENSBTAG00000005110	CADPS2	4	87,522,151	88,100,974	1	4
ENSBTAG00000015303	MPP6	4	71,625,584	71,690,219	1	1
ENSBTAG00000032650	DPY19L2	4	62,259,290	62,349,341	1	1
ENSBTAG00000042539	U6	4	47,483,218	47,483,324	1	1
ENSBTAG00000032148	TMEM117	5	36,197,272	36,807,308	2	4
ENSBTAG00000026880	KRT85	5	27,711,064	27,729,751	1	2
ENSBTAG00000027064	BTBD11	5	70,923,456	71,257,389	1	2
ENSBTAG00000021287	SLC16A7	5	53,987,909	54,214,799	1	1
ENSBTAG00000013333	GYS2	5	89,020,801	89,077,024	1	1
ENSBTAG00000011087	ARID2	5	34,488,354	34,583,315	1	1

ENSBTAG00000020914	CPNE8	5	42,421,554	42,698,411	1	1
ENSBTAG00000016274	CCDC38	5	60,564,497	60,591,379	1	1
ENSBTAG00000004569	GLIPR1L1	5	4,824,769	4,866,969	1	1
ENSBTAG00000016260	LRRK2	5	40,703,505	40,916,225	1	1
ENSBTAG00000016204	C1RL	5	103,633,960	103,644,323	1	1
ENSBTAG00000008595	PPHLN1	5	38,397,360	38,571,901	1	1
ENSBTAG00000013912	TXNRD1	5	68,239,611	68,302,678	1	1
ENSBTAG00000005221	WNK1	5	108,079,510	108,207,083	1	1
ENSBTAG00000006156	BST1	6	115,687,877	115,716,694	1	2
ENSBTAG00000035776	C6H4orf22	6	96,794,894	97,503,220	1	3
ENSBTAG00000009438	EPHA5	6	82,560,093	82,962,887	1	1
ENSBTAG00000020048	MAPK10	6	102,687,307	103,063,481	1	1
ENSBTAG00000002348	SLC4A4	6	88,182,303	88,541,046	1	1
ENSBTAG00000039275	ERAP2	7	98,715,725	98,767,729	1	2
ENSBTAG00000009975	PBX4	7	3,631,055	3,685,738	1	2
ENSBTAG00000038117	MGC138057	7	84,242,450	84,505,756	1	1
ENSBTAG00000003051	FER	7	110,465,382	110,864,228	1	1
ENSBTAG00000030210	SLCO6A1	7	103,818,199	103,932,551	1	2
ENSBTAG00000014661	CHSY3	7	25,178,592	25,466,387	1	1
ENSBTAG00000013810	GABBR2	8	63,840,064	64,089,259	1	2
ENSBTAG00000000738	DAPK1	8	82,207,240	82,309,507	1	1
ENSBTAG00000017195	FANCC	8	83,023,629	83,270,596	1	2
ENSBTAG00000000712	FBXW2	8	112,057,391	112,089,639	1	1
ENSBTAG00000005247	FRMD3	8	77,785,287	77,910,313	1	1
ENSBTAG00000001081	PALLD	8	571,565	923,258	1	1
ENSBTAG00000008836	ZNF782	8	84,850,159	84,937,203	1	1
ENSBTAG00000021741	RPS6KA2	9	102,918,982	103,074,109	1	1
ENSBTAG00000018634	SH3BGRL2	9	19,526,852	19,547,939	1	1
ENSBTAG00000005960	EPB41L2	9	69,916,605	70,027,346	1	1
ENSBTAG00000009665	UTRN	9	82,762,003	83,311,756	1	2
ENSBTAG00000031165	TRPM7	10	59,853,461	59,943,298	1	1
ENSBTAG00000003667	TLN2	10	47,303,876	47,796,327	1	1
ENSBTAG00000021876	WDR72	10	56,678,901	56,805,084	1	1
ENSBTAG00000018947	SYT16	10	74,384,705	74,497,702	1	1
ENSBTAG00000008466	CCNB1IP1	10	26,830,647	26,838,482	1	1
ENSBTAG00000025642	RYR3	10	28,789,149	29,100,901	1	1
ENSBTAG00000044173	EHBP1	11	61,117,376	61,472,058	1	1
ENSBTAG00000016442	MAP4K3	11	21,579,432	21,761,348	1	1
ENSBTAG00000008647	KLHL1	12	44,295,888	44,616,940	1	4
ENSBTAG00000004165	CLYBL	12	80,409,751	80,638,113	1	2
ENSBTAG00000000939	KIF16B	13	10,238,640	10,519,339	1	3
ENSBTAG00000008338	PLCB1	13	789,380	1,695,139	3	5
ENSBTAG00000000309	PTPRT	13	71,397,829	71,678,072	1	2
ENSBTAG00000008279	FRMD4A	13	28,631,455	28,961,637	1	2

ENSBTAG00000009475	PLXDC2	13	21,595,581	22,015,323	1	1
ENSBTAG00000027412	SOD1	13	51,930,067	51,930,888	1	1
ENSBTAG00000008969	SLC9A8	13	78,628,908	78,698,768	1	1
ENSBTAG00000011908	CPQ	14	69,287,302	69,893,052	2	2
ENSBTAG00000004518	GRHL2	14	64,905,620	65,031,370	1	1
ENSBTAG00000015229	DNAJC5B	14	32,050,488	32,143,515	1	1
ENSBTAG00000013033	BTBD10	15	39,778,550	39,840,041	1	1
ENSBTAG00000002382	DDX10	15	18,644,692	18,952,411	1	1
ENSBTAG00000003955	MYO7A	15	57,332,143	57,419,714	1	1
ENSBTAG00000037384	OR10AB1P	15	45,478,430	45,479,365	1	1
ENSBTAG00000000727	RNF169	15	54,840,266	54,923,335	1	1
ENSBTAG00000047461	DYNC2H1	15	5,313,517	5,463,876	1	1
ENSBTAG00000020046	ASAM	15	34,226,369	34,332,243	1	1
ENSBTAG00000037661	DENND1B	16	78,480,806	78,590,234	1	2
ENSBTAG00000002164	AXDND1	16	62,138,403	62,214,479	1	2
ENSBTAG00000033180	SMYD3	16	31,589,396	32,333,109	1	2
ENSBTAG00000023144	PTPRC	16	79,522,522	79,592,696	2	2
ENSBTAG00000004407	KCNK2	16	69,865,018	70,097,416	1	1
ENSBTAG00000016542	LAMB3	16	75,567,714	75,610,920	1	1
ENSBTAG00000006188	USH2A	16	19,573,856	20,502,175	1	2
ENSBTAG00000017561	HHIPL2	16	26,713,225	26,741,364	1	1
ENSBTAG00000024555	EFCAB2	16	32,961,375	33,072,297	1	1
ENSBTAG00000046256	TMEM132C	17	49,445,324	49,739,129	1	2
ENSBTAG00000012738	ZNF827	17	12,587,397	12,739,481	1	1
ENSBTAG00000002531	ARHGAP10	17	10,182,664	10,560,361	1	1
ENSBTAG00000002813	GAB1	17	14,625,987	14,753,591	1	1
ENSBTAG00000003447	ZNF551	18	65,291,985	65,319,947	1	2
ENSBTAG00000002287	CHD9	18	21,726,409	21,857,238	1	1
ENSBTAG00000003334	ADAP2	19	18,353,810	18,388,312	1	1
ENSBTAG00000044618	SNORA31	19	3,899,300	3,899,416	1	1
ENSBTAG00000022509	DNAH9	19	30,963,518	31,248,936	1	1
ENSBTAG00000027074	SV2B	21	16,214,670	16,460,490	2	3
ENSBTAG00000047543	NTRK3	21	19,493,058	19,520,421	1	1
ENSBTAG00000007382	SCAPER	21	32,118,828	32,512,606	1	1
ENSBTAG00000010416	RIN3	21	57,859,148	57,953,844	1	1
ENSBTAG00000013047	GRM7	22	18,740,484	19,647,747	1	6
ENSBTAG00000012073	VOPP1	22	495,506	677,104	1	1
ENSBTAG00000003592	CNTN6	22	25,053,636	25,233,677	1	1
ENSBTAG00000034496	SHFM1	22	22,098,224	22,098,436	1	1
ENSBTAG00000007360	TMEM43	22	58,726,302	58,747,763	1	1
ENSBTAG00000003359	ELOVL5	23	25,155,743	25,228,997	1	1
ENSBTAG00000013831	DSG1	24	26,089,489	26,133,405	1	1
ENSBTAG00000009822	PPP4R1	24	42,093,001	42,121,847	1	1
ENSBTAG00000000390	TPST1	25	28,339,035	28,445,471	1	1

ENSBTAG00000016244	VWA3A	25	19,992,453	20,049,454	1	1
ENSBTAG00000004077	YWHAG	25	34,884,283	34,906,639	1	1
ENSBTAG00000045905	PCDH15	26	5,017,714	5,578,654	1	8
ENSBTAG00000007948	SORCS1	26	27,810,198	28,389,391	3	6
ENSBTAG00000022715	DMBT1	26	42,782,329	42,813,472	1	1
ENSBTAG00000037795	CYP2C87	26	16,030,292	16,065,395	1	1
ENSBTAG00000004830	ADAM18	27	34,517,618	34,625,508	1	3
ENSBTAG00000033137	PSD3	27	38,483,892	38,797,972	1	2
ENSBTAG00000020361	SLC35F3	28	6,762,322	7,195,661	1	5
ENSBTAG00000011072	ADK	28	30,215,525	30,732,444	1	2
ENSBTAG00000021497	CDH23	28	27,729,164	28,123,449	1	1
ENSBTAG00000004081	FAT3	29	1,965,869	2,605,125	1	4
ENSBTAG00000001043	MGC157332	29	20,259,769	20,557,376	1	1
ENSBTAG00000016506	ST3GAL-IV	29	30,073,042	30,112,563	1	1

Table 6.7. Genes that overlapped with Holstein breed-specific CNVs

Ensemble Gene	Gene Symbol	Chr	Gene Start	Gene End	# Breed Specific CNV	# CNV
ENSBTAG00000003936	PNKD	2	107,020,657	107,093,550	1	1
ENSBTAG00000017056	PKLR	3	15,399,755	15,408,994	1	1
ENSBTAG00000017055	HCN3	3	15,409,813	15,418,775	1	1
ENSBTAG00000019027	SLC30A7	3	42,465,567	42,559,234	1	1
ENSBTAG00000004023	KIAA1324L	4	33,518,953	33,763,007	1	2
ENSBTAG00000020247	ADCYAP1R1	4	65,670,543	65,729,276	1	1
ENSBTAG00000046430	ZNF804B	4	73,326,980	73,897,041	1	1
ENSBTAG00000032183	NELL2	5	35,657,329	36,042,694	1	2
ENSBTAG00000020679	CNTN1	5	39,998,363	40,264,645	1	1
ENSBTAG00000010149	CRY1	5	70,606,115	70,701,030	1	1
ENSBTAG00000006434	SYNPO2	6	7,388,728	7,590,933	1	3
ENSBTAG00000016515	EFNA5	7	109,049,590	109,217,439	1	4
ENSBTAG00000015335	BAI3	9	7,914,952	8,455,993	1	2
ENSBTAG00000007758	PDE10A	9	101,987,619	102,068,801	1	2
ENSBTAG00000005016	AP3B1	10	9,040,253	9,300,567	1	1
ENSBTAG00000005751	CDAN1	10	38,138,863	38,151,656	1	1
ENSBTAG00000021102	GALM	11	21,040,425	21,095,033	1	1
ENSBTAG00000020893	MATN3	11	78,889,151	78,907,349	1	1
ENSBTAG00000002137	SUGT1	12	11,131,393	11,172,149	1	1
ENSBTAG00000013114	ZMYND8	13	76,501,073	76,625,012	1	1
ENSBTAG00000008571	CUX2	17	57,065,816	57,343,548	1	1
ENSBTAG00000023541	C6ORF10	23	26,726,417	26,794,259	1	2
ENSBTAG00000004940	BRUNOL4	24	19,819,065	20,129,055	1	1
ENSBTAG00000021856	C10ORF28	26	19,164,410	19,197,680	1	1
ENSBTAG00000033137	PSD3	27	38,483,892	38,797,972	1	2
ENSBTAG00000020361	SLC35F3	28	6,762,322	7,195,661	1	5

Table 6.8. Gene description and references for genes related to phosphorylation or protein modification process in Hanwoo

Gene	Chr	Reference	Description
NDUFA10	chr3	Perez, Cañón et al. 2010	*NDUFA10* is associated with long-chain omega-3 fatty acids in bovine skeletal muscle.
WNK1	chr5	Moore, Garg et al. 2000	WNK1 encodes Serine/threonine-protein kinase which plays an important role in cell proliferation and in actin cytoskeletal reorganization.
MAPK10	chr6	Fluckey, Knox et al. 2006	MAPK10 are related to the Mitogen-Activated Protein Kinase (MAPK) system, which is a major growth signaling pathway that controls skeletal muscle growth.
FER	chr7	Aspenström 1997	FER encodes FER tyrosine kinase, which may acts downstream of cell surface receptors for growth factors and plays a role in the regulation of the actin cytoskeleton.
RPS6KA2	chr9	Yin, Kim et al. 2007	RPS6KA2 encodes a member of the RSK (ribosomal S6 kinase) family of serine/threonine kinases which has been implicated in controlling cell growth and differentiation.
MAP4K3	chr11	Fluckey, Knox et al. 2006	MAP4K3 are related to the Mitogen-Activated Protein Kinase (MAPK) system, which is a major growth signaling pathway that controls skeletal muscle growth.
PTPRT	chr13	Sun and Tonks 1994	The proteins encoded by PTPRT are members of the protein tyrosine phosphatase (PTP) family that are known to be signaling molecules that regulate cell growth.
PTPRC	chr16	Sun and Tonks 1994	The proteins encoded by PTPRC are members of the protein tyrosine phosphatase (PTP) family that are known to be signaling molecules that regulate cell growth.
GAB1	chr17	Rodrigues, Falasca et al. 2000	GAB1 encodes a bound protein 2-associated protein that plays a central role in cellular growth.
NTRK3	chr21	Lannon and Sorensen 2005	NTRK3 encodes a member of the neurotrophic tyrosine receptor kinase (NTRK) family which plays a role in cell growth, development, and cell survival.

Table 6.9. Gene description and references for genes related to cell adhesion and maintenance in Hanwoo

Gene	Chr	Reference	Gene Description
ITGAV	chr2	Takada, Ye et al. 2007	ITGAV encodes a protein that is a member of the integrin superfamily which interacts with several extracellular matrix proteins to mediate cell adhesion.
COL28A1	chr4	Veit, Kobbe et al. 2006	COL28A1 belongs to a class of collagens containing von Willebrand factor.
FER	chr7	Rosato, Veltmaat et al. 1998	Fer protein regulates cell-cell adhesion.
TLN2	chr10	Senetar, Moncman et al. 2007	This gene encodes a protein related to talin 1, a cytoskeletal protein that plays a significant role in the assembly of actin filaments and may play an important role in cell adhesion.
LAMB3	chr16	Vidal, Baudoin et al. 1995	In biological process related to collagen, LAMB3 product mediates the attachment, migration and organization of cells into tissues by interacting with other extracellular matrix components.
DSG1	chr24	Runswick, O'Hare et al. 2001	DSG1 encodes desmosomal glycoprotein and desmosomal adhesion as these genes regulate intercellular junctions of epithelia.
PCDH15	chr26	Halbleib and Nelson 2006	PCDH15 encodes protocadherin 15, a member of the cadherin superfamily which related to cell adhesion.
CDH23	chr28	Marthiens, Gavard et al. 2002	CDH23 encode Cadherin-23, a calcium dependent cell-cell adhesion glycoprotein and previous study implicated this cadherin in myogenesis.
FAT3	chr29	Mitsui, Nakajima et al. 2002	Fat3 protein l may be involved in cell adhesion

Table 6.10. Gene description and references for genes related to dairy production in Holstein

Gene	Chr	Reference	Gene Description
PKLR	chr3	Baik, Etchebarne et al. 2009	PKLR encoding pyruvate kinase was found to have expression in mammary tissue of lactating dairy cow.
ADCYAP1R1	chr4	Winters and Moore 2011	ADCYAP1R1 encodes type I adenylate cyclase activating polypeptide receptor, which may regulate the release of prolactin.
NELL2	chr5	Connor, Siferd et al. 2008	NELL2 encodes Neural Epidermal Growth Factor-Like 2, which is down regulated in the bovine mammary gland and affects milking frequency.
CRY1	chr5	Casey and Plaut 2012	CRY1 encodes a flavin adenine dinucleotide-binding protein that is a key component of the circadian core that affected development of the mammary gland and lactation.
EFNA5	chr7	Li, Wang et al. 2010	EFNA5 is a member of the ephrin gene family and was previously reported to be the top milk production trait SNP in Canadian Holstein cattle.
PDE10A	chr9	Dostaler-Touchette et al. 2009	The protein encoded by PDE10A belongs to the cyclic nucleotide phosphodiesterase family and PDE10 appear to be functional in the bovine mammary gland.
AP3B1	chr10	Bionaz, Periasamy et al. 2012	The protein encoded by AP3B1 interacts with the scaffolding protein clathrin, which is important for lactating mammary glands as lactation secretes milk components through vesicles in dairy cattle.
GALM	chr11	D'Alessandro, Zolla et al. 2011	The protein encoded by GALM is expressed in the cytoplasm and has a preference for galactose and contained in networks of bovine milk proteins.
MATN3	chr11	Yucesoy, Charles et al. 2013	MATN3 is related to genetic risk factors for osteoarthritis which related to dairy production.
C6ORF10	chr23	Sadkowski, Jank et al. 2009	Transcriptional profiles of dairy and beef breeds bulls showed that C10ORF28 is down-regulated in Holstein

Table 6.11. Top cattle CNV (p-value after FDR correction < 0.01) using Fst.

CNV	CHR	Fst	Fst p-value	Fst fdr	Ensemble Gene ID	Gene Symbol
BovineCNV0531	chr2	0.861	2.22E-15	2.18E-12	ENSBTAG00000026986	TTN
BovineCNV1785	chr5	0.730	2.22E-15	2.18E-12	-	-
BovineCNV2092	chr6	0.727	2.89E-15	2.18E-12	-	-
BovineCNV2093	chr6	0.727	2.89E-15	2.18E-12	-	-
BovineCNV2094	chr6	0.727	2.89E-15	2.18E-12	-	-
BovineCNV2112	chr6	0.727	2.89E-15	2.18E-12	-	-
BovineCNV3174	chr10	0.733	1.67E-15	2.18E-12	-	-
BovineCNV3237	chr10	0.730	2.22E-15	2.18E-12	-	-
BovineCNV6791	chr29	0.727	2.89E-15	2.18E-12	-	-
BovineCNV2291	chr7	0.709	1.52E-14	1.04E-11	-	-
BovineCNV3583	chr11	0.689	9.36E-14	5.80E-11	ENSBTAG00000021969	BT.40893
BovineCNV2095	chr6	0.666	6.49E-13	3.65E-10	-	-
BovineCNV5332	chr20	0.665	6.97E-13	3.65E-10	-	-
BovineCNV0838	chr3	0.623	2.24E-11	1.02E-08	-	-
BovineCNV2465	chr7	0.623	2.24E-11	1.02E-08	-	-
BovineCNV3235	chr10	0.616	3.84E-11	1.63E-08	-	-
BovineCNV0798	chr2	0.606	8.27E-11	3.03E-08	-	-
BovineCNV0972	chr3	0.605	8.89E-11	3.03E-08	-	-
BovineCNV3213	chr10	0.606	8.27E-11	3.03E-08	ENSBTAG00000005661	SNAP23
BovineCNV6673	chr29	0.605	8.89E-11	3.03E-08	-	-
BovineCNV3187	chr10	0.584	4.18E-10	1.36E-07	-	-
BovineCNV2212	chr6	0.577	7.15E-10	2.21E-07	-	-
BovineCNV0573	chr2	0.562	2.18E-09	6.03E-07	-	-
BovineCNV2450	chr7	0.561	2.30E-09	6.03E-07	-	-
BovineCNV2826	chr9	0.561	2.30E-09	6.03E-07	-	-
BovineCNV3623	chr11	0.562	2.18E-09	6.03E-07	-	-
BovineCNV1786	chr5	0.551	4.48E-09	1.13E-06	-	-
BovineCNV2663	chr8	0.549	5.35E-09	1.26E-06	-	-
BovineCNV4652	chr16	0.549	5.16E-09	1.26E-06	-	-
BovineCNV2969	chr9	0.547	6.12E-09	1.39E-06	ENSBTAG00000000629	MMS22L
BovineCNV3817	chr12	0.534	1.40E-08	3.08E-06	-	-
BovineCNV5921	chr23	0.520	3.68E-08	7.84E-06	-	-
BovineCNV1728	chr5	0.516	4.58E-08	9.17E-06	-	-
BovineCNV3287	chr10	0.516	4.58E-08	9.17E-06	-	-
BovineCNV3665	chr12	0.508	8.03E-08	1.56E-05	-	-
BovineCNV4561	chr15	0.495	1.82E-07	3.44E-05	-	-
BovineCNV6325	chr26	0.494	1.88E-07	3.46E-05	-	-
BovineCNV0642	chr2	0.491	2.28E-07	3.70E-05	-	-
BovineCNV3463	chr11	0.491	2.28E-07	3.70E-05	-	-
BovineCNV4137	chr14	0.491	2.28E-07	3.70E-05	ENSBTAG00000020801	FAM49B

BovineCNV4138	chr14	0.491	2.28E-07	3.70E-05	-	-
BovineCNV5731	chr22	0.491	2.28E-07	3.70E-05	ENSBTAG00000013047	GRM7
BovineCNV3268	chr10	0.459	1.49E-06	0.000230	ENSBTAG00000025522	UNC13C
BovineCNV4306	chr14	0.459	1.49E-06	0.000230	-	-
BovineCNV3280	chr10	0.457	1.72E-06	0.000244	-	-
BovineCNV3313	chr10	0.457	1.72E-06	0.000244	-	-
BovineCNV3574	chr11	0.457	1.72E-06	0.000244	-	-
BovineCNV3591	chr11	0.457	1.72E-06	0.000244	ENSBTAG00000020893	MATN3
BovineCNV1122	chr3	0.455	1.93E-06	0.000268	ENSBTAG00000000634	BT.29245
BovineCNV6687	chr29	0.453	2.13E-06	0.000290	-	-
BovineCNV2042	chr6	0.447	2.93E-06	0.000384	-	-
BovineCNV3197	chr10	0.447	2.93E-06	0.000384	-	-
BovineCNV2630	chr8	0.440	4.41E-06	0.000567	-	-
BovineCNV2091	chr6	0.438	4.94E-06	0.000601	-	-
BovineCNV2677	chr8	0.438	4.94E-06	0.000601	-	-
BovineCNV2936	chr9	0.438	4.94E-06	0.000601	-	-
BovineCNV2097	chr6	0.427	8.72E-06	0.001042	-	-
BovineCNV5258	chr19	0.426	9.11E-06	0.001052	-	-
BovineCNV5823	chr23	0.426	9.11E-06	0.001052	ENSBTAG00000021237	DST
BovineCNV1125	chr3	0.423	1.09E-05	0.001233	ENSBTAG00000017764	HDAC4
BovineCNV1257	chr4	0.419	1.29E-05	0.001421	-	-
BovineCNV1876	chr6	0.419	1.29E-05	0.001421	-	-
BovineCNV0045	chr1	0.407	2.38E-05	0.002421	-	-
BovineCNV2014	chr6	0.407	2.38E-05	0.002421	-	-
BovineCNV3997	chr13	0.407	2.38E-05	0.002421	-	-
BovineCNV4415	chr15	0.407	2.38E-05	0.002421	-	-
BovineCNV5714	chr22	0.407	2.38E-05	0.002421	-	-
BovineCNV3339	chr10	0.405	2.68E-05	0.002688	ENSBTAG00000017489	TSHR
BovineCNV2096	chr6	0.403	2.97E-05	0.002848	-	-
BovineCNV3021	chr9	0.403	2.97E-05	0.002848	-	-
BovineCNV3377	chr11	0.403	2.97E-05	0.002848	ENSBTAG00000006019	FAM178B
BovineCNV1583	chr5	0.401	3.27E-05	0.003049	ENSBTAG00000026880	KRT85
BovineCNV2946	chr9	0.401	3.27E-05	0.003049	-	-
BovineCNV0527	chr2	0.394	4.55E-05	0.003972	ENSBTAG00000027875	CCDC141
BovineCNV2602	chr8	0.395	4.38E-05	0.003972	ENSBTAG00000023143	-
BovineCNV3820	chr12	0.395	4.38E-05	0.003972	-	-
BovineCNV5894	chr23	0.394	4.55E-05	0.003972	-	-
BovineCNV5895	chr23	0.394	4.55E-05	0.003972	-	-
BovineCNV4648	chr16	0.392	5.06E-05	0.004366	-	-
BovineCNV2007	chr6	0.388	6.19E-05	0.005078	-	-
BovineCNV2326	chr7	0.388	6.19E-05	0.005078	-	-
BovineCNV4229	chr14	0.388	6.19E-05	0.005078	-	-
BovineCNV5300	chr20	0.388	6.19E-05	0.005078	-	-
BovineCNV0292	chr1	0.387	6.57E-05	0.005147	-	-

BovineCNV1541	chr5	0.387	6.57E-05	0.005147	-	-
BovineCNV4595	chr16	0.387	6.57E-05	0.005147	-	-
BovineCNV6696	chr29	0.387	6.57E-05	0.005147	-	-
BovineCNV1640	chr5	0.377	0.000105	0.008117	-	-
BovineCNV1152	chr4	0.371	0.000135	0.009777	-	-
BovineCNV1768	chr5	0.371	0.000135	0.009777	-	-
BovineCNV2124	chr6	0.371	0.000135	0.009777	-	-
BovineCNV3277	chr10	0.371	0.000135	0.009777	ENSBTAG00000004011	GALK2
BovineCNV4633	chr16	0.371	0.000135	0.009777	-	-
BovineCNV5121	chr18	0.371	0.000135	0.009777	-	-

To confirm the CNV genotype within some of the putative genes containing the impact of the domestication of cattle, I performed PCR. I selected seven putative genes (TTN, SLIT3, KLHL1, NCAM2, MDGA2, EFNA5 and PRKG1) which had 25 CNVs. However, due to limitations of PCR, I excluded six CNVs that were longer than 1.5 Kb. When the genotype of the examined 19 CNV regions in 10 Holstein and 22 Hanwoo were compared to the expected genotype, various matching rates were discovered (37.19% to 100%, Figure 6.10, Figure 6.11). Almost all of the CNV regions examined by PCR showed similar lengths to the expected CNV lengths (< 200bp) and these CNVs were considered validated. Taken together, the CNV accuracy of this study was determined to be about 80% from the validation experiment (Figure 6.10).

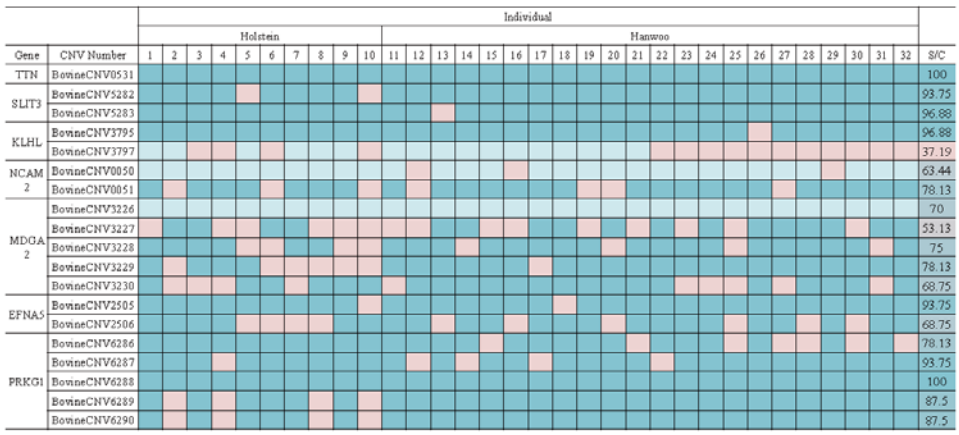

Figure 6.10. CNV validation and its accuracy measure by gDNA PCR. The CNV pattern comparison in 32 individuals was represented by a heat map. The colors of the boxes represent whether the genotypes resulting from PCR matched the predicted genotypes by Genome STRiP. Dark blue indicates that that the two matched while pink indicates that the two did not match. Sky-blue boxes represent restricted matching with the CNV showing only deleted or non-deleted allele in the PCR validation. The average matching rate of the CNVs, 80.02%, was considered to be the CNV accuracy.

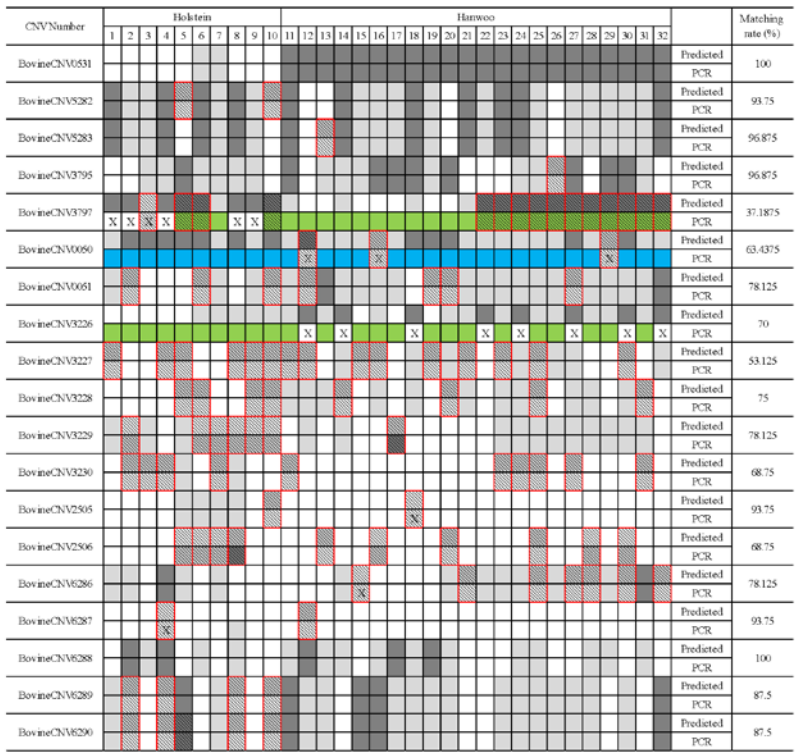

Figure 6.11. Genotype comparison between result from genomic DNA amplification and GenomeSTRiP. The genotype of each CNV and individual are summarized by a heat map. The genotype of the examined 19 CNV regions by PCR was compared to that of the GenomeSTRiP result. Matching score was calculated by dividing the number of individual, whose predicted genotype was the same as the PCR result, to the total number of individuals examined (n = 32). As the CNVs detecting only deleted or non-deleted alleles (BovineCNV0050, BovineCNV3226, and BovineCNV3797) showed limited accuracy, a lower score of 0.7 was used as the matching score.

Table 6.12. Information of the primer pairs used for zygosity validation

Gene Symbol	CNV Number	Primer (5´ to 3´)*		Predicted CNV length (bp)**	Amplicon Size (bp)		PCR condition	
					CNV deletion	CNV Non-deletion	Anealing Temperature	Extention time in 72 °C
TTN	BovineCNV0531	F	TGGGGGAAACCATCATAAAC	1043	319	1362	60	1 min
		R	CCACACAGGATTTGAACCATC					
SLIT3	BovineCNV5282	F	TGAGGGACAGAGACAGAGCA	609	570	1179	66	1 min
		R	TCCAGTTGAGCTGAGTTGAGG					
	BovineCNV5283	F	GACATTCGCTTGGTTGGCTG	249	214	463	60	1 min
		R	GCTTTGAGATTGCTTCATTTCCC					
KLHL1	BovineCNV3795	F	TCATGCTTAGAACTTTCCCCGTT	1106	360	1466	64	1 min
		R	CGCTTGCTTCAGCTTAGCCTT					
	BovineCNV3797	F	CAATCAATGGGGTCACAAAG	212	609	821	64	1 min
		R	AAAGGCTGGGAAAGGAAGAG					
NCAM2	BovineCNV0050	F	ATTCATGGAGAAAAATGCTTGCC	1326	342	1668	65	1min 30sec
		R	GCTGGTTGGTCATAGCCTGAGTT					
	BovineCNV0050_In†	F	ATTCATGGAGAAAAATGCTTGCC	1326	N.D	1019	64	1 min
		R	GCTGGTTGGTCATAGCCTGAGTT					
	BovineCNV0051	F	GCCTCCAGCAAACTTACAGACAT	604	287	891	60	1 min
		R	TTTTCACAAAGAGAACCAGAAGCA					
MDGA2	BovineCNV3226	F	CCCATCCTCAGAAATCCTTTA	700	582	1282	58	1 min
		R	GGTAAATGGGATTGATTCCTTG					
	BovineCNV3227	F	CTACCATCTGGCCCTTCAAC	234	357	591	64	1 min

		R	CAAACATGGAAAGGAATCCAA					
	BovineCNV3228	F	ATGATGTCTTCTGGGCAAGT	1304	374	1678	58	1 min
		R	TTTCGTCTGAGTGCTCCATC					
	BovineCNV3229	F	TTAGTGCCCCTCTCCTTTCC	385	482	867	60	1 min
		R	GCCTTCCTTTCCAACATCAC					
	BovineCNV3230	F	CCCCAGGCTCTTCTGTTCT	1221	302	1523	64	1 min
		R	TGTCAGTTTGTGATGAAAGTTGG					
EFNA5	BovineCNV2505	F	GGAACACAGACAACAGGCAGA	229	440	669	64	1 min
		R	AGGGGAAAGAAGGAGTGGAA					
	BovineCNV2506	F	AAGAGATTCGGGAAGGGACT	289	369	658	64	1 min
		R	AAGAACGACACCTTGCTGCT					
PRKG1	BovineCNV6286	F	TCTCTTTCCCCCAATCTCAA	475	404	879	60	1 min
		R	CACAACATCACCACATCAAGG					
	BovineCNV6287	F	GCAGCAAAAGAAGGGAAAGA	961	434	1395	64	1 min
		R	TGAAGCAACTGAAACCCAGA					
	BovineCNV6288	F	GACACACAAAGGGAAATAGAGGA	1290	718	2008	64	1min 30sec
		R	GACAGTCTGATTGGCTGTTTG					
	BovineCNV6289	F	ATGCTATGGAAACCGAGAGG	572	399	971	60	1 min
		R	AACTATGATGCCCAACTTCACA					
	BovineCNV6290	F	ATGCTATGGAAACCGAGAGG	491	478	969	60	1 min
		R	CTATGATGCCCAACTTCACA					

6.5 Discussion

In this study, I used 32 individual of two cattle breeds, Hanwoo and Holstein, to detect CNVs at the population level. Hanwoo, Bos taurus coreanae, is a breed of cattle raised in Korea, which may be a hybrid of Bos taurus and Bos indicus. Hanwoo migrated and settled in the Korean Peninsula around 5,000 BC. It has been used both as a draft animal and a source of meat but over the past 40 years, the main role of Hanwoo has changed to beef cattle. Since the first official genetic breeding program for Hanwoo by the Korean government started in 1979, the productivity of Hanwoo has improved substantially. In contrast, Holstein is a breed of cattle that has been strongly selected for milk-production and currently has the highest-production of dairy. Genetic resource of Holstein is shared throughout the world by trading in semen and seed bull. I used 22 Hanwoo and 10 Holstein for NGS CNV detection. Holstein individuals were selected using common global criteria while Hanwoo was selected from two different regions to capture the complete genetic picture of the breed. The genetic difference between the Hanwoo individuals of the two populations was identified to be small, and so the 22 individuals were regarded as a single population in this study.

I showed that genes with higher deletion score are more likely to be under genetic drift. Through the CNV deletion score of 32 individuals, I wanted to find out which genes have been affected by cattle domestication. Humans have applied strong selective pressure on each cattle breed through elaborate breeding strategies to form breeds that can provide products such as milk and meat. Animal breeding by humans has been performed during a short period and the cattle population is usually produced by artificial insemination using a small number of seed bull and many cows

to both maintain product quality and manage bloodlines. From a genetics point of view, breeding can be regarded as a genetic diversity reduction event much like a population bottleneck. I predicted in this study that deletion regions with beneficial adaptations might have arisen after this genetic diversity reduction event. The loss of variation leaves a surviving population that is favorable with regard to the selective pressures put on it such as the production of milk or meat. The breeding strategies of each cattle breed share a common domestication process, so I wanted to capture the genic regions affected by the general cattle domestication using deleted CNV. Based on this assumption, I selected 33 significant Ensemble genes, which were strongly affected by deleted CNVs using deletion scores. I regarded the genes with higher deletion score as being under neutral or diversifying selection in the absence of additional information.

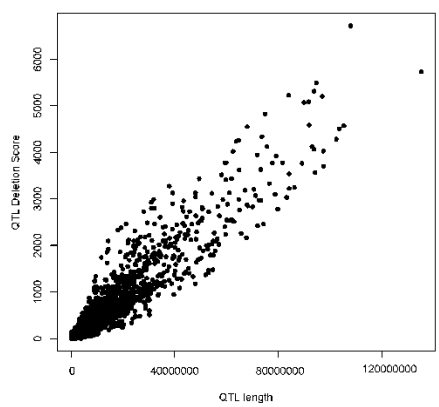

Figure 6.12. Relationships between QTL length and QTL deletion score

I want to discover QTL related to the deleted CNVs. I wanted to suggest novel genetic regions over genic regions, which are affected by deleted CNV by

using QTL that contain information about the region related to each of the economic phenotype of cattle. So, I used QTL information of Animal QTL to detect a wider region of the genome affected by deleted CNV that contains meaningful information. However, QTL mapping is a step prior to gene definition and QTL region information is roughly defined based on phenotype information. As the variance of QTL length was very large and longer QTL tended to have higher deletion scores (Figure 6.12), I could not determine whether high deletion scores of QTL were due to containing many actual deletion regions or simply from the length of the QTL. Therefore, I could not use the deletion scores of QTL to discover QTL affected by deleted CNVs. To overcome this problem, I used a new measurement, average distance between deletions (QTL length/deletion score), to discover QTL affected by deleted CNV. However, the average distance between deletions in QTL was still very variable. It was not possible to create a proper distribution of average distance between deletions in QTL because there were so many QTL regions considering the total number of CNVs. The empirical p-value of QTL had very short average distance between deletions and did not reach the commonly used criteria. However, considering previous studies that discovered important QTL regions that overlapped with SNP in GWAS or selective sweep study, I supposed that QTL containing very short average distance between deletions must be meaningful. So I proposed the top 30 QTL that were selected by the average distance between deletions and regarded these QTL as QTL affected by deleted CNV. I had guessed that QTL types related to CNV in this study would be highly variable, because QTL are roughly defined based on phenotype information and meat and milk traits are complex traits. As expected, the top 30 identified QTL from the QTL analysis had diverse traits. Additionally, as QTL is a region related to economic traits, I focused on the relationship between the region and their traits. I predicted that the gene content of QTL affected by deleted

CNV was very important and that this information would supplement information on the genes selected by deletion score.

The domestication and subsequent selection by humans to create breeds have had an impact on the variation within the cattle genome. Strong selection for breed characteristics or productivity has created regions that have lost variation due to the fixation of advantageous mutations, or selective sweep regions. I identified selective sweep regions in the cattle genome but no study has yet to explore these regions using CNVs. In this study, FST based on the CNV frequency spectra was used to identify and characterize regions of the cattle genome under selective sweep. Additionally, as mentioned earlier, deletion score was used to estimate the genes affected by deleted CNV in cattle domestication by understanding the number of CNVs in each gene and the frequency of each CNV within the population. Selective sweep signal based on FST of deleted CNV was used to estimate how each deleted CNV affects the trait difference between Holstein (for milk) and Hanwoo (for meat). Between the two examined cattle breeds, 94 putative sweep regions were identified. I assumed that economic traits including beef and milk production have historically been under strong selection. Based on this assumption, I wanted to explore CNVs under selective sweep for economic traits. The results were then used as foundation for the selective sweep section of this study. The most significant deleted CNV (BovineCNV0531) was within the titin gene (TTN). Takahisa Yamada et al. (2009) reported that TTN is involved in myofibrillogenesis through a SNP association study in Japanese Black beef cattle. TTN was reported as the gene which is a positional functional candidate responsible for marbling in beef (Yamada et al. 2009). A comparison of Japanese Black breed with Holstein and Brown Swiss breed showed that SNP in TTN has strong selection pressure for high marbling (Watanabe et al. 2011). Therefore, even though the deletion was in the intron, I predict that BovineCNV0531 has had strong impacts from selection during breed formation.

Recently, NGS data have been used to discover breed-specific SNP of domesticated animals. In a previous study on pigs, breed-specific SNPs were selected from NGS resequencing data and then filtered by data validation using SNP chip data of many individual to apply assignment test (Ramos et al. 2011). However, in the case of CNVs it is difficult to validate them, because CNVs in this study are structural variation at the population level that is dependent on the nature of the population and there is no back-up data such as SNP chip data. However, STRUCTURE analysis using 6,811 CNVs could classify individuals into the two breeds, Hanwoo and Holstein (Figure 6.13). Therefore, I wanted to know which CNVs were breed-specific and understand the biological meaning of these breed-specific CNVs. I selected CNVs that belonged to only one breed and regarded these CNVs as breed-specific CNV candidates. And then I only selected CNVs with a frequency of higher than 0.1 in each breed to minimize the false positive breed-specific CNV calls instead of validation using back-up data.

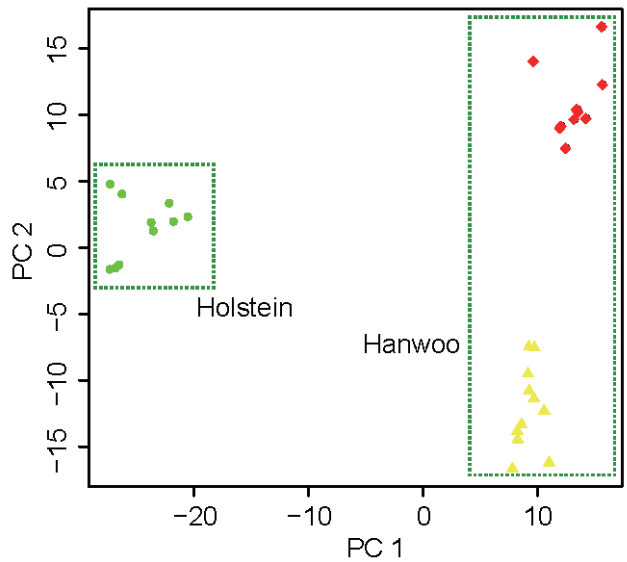

Figure 6.13. PCA using all deleted cattle CNV as markers. Green circle represents Holstein and other two colors represent the two different Hanwoo populations. Red diamond represents Hanwoo from RDA in Suwon and yellow represents Hanwoo from Kyungpook National University.

If deletions occur within coding regions of the listed genes, the missing functional domains of the translated proteins resulting from that gene may be inferred, I were careful in making such inferences as I did not have the phenotype information that is needed to conduct an association study of the relationship between genetic variants and the traits. Therefore, I could not perform additional analysis or experiment to directly investigate the biological phenomenon affected by deleted CNV. Though there are many limitations, I could discover some key points regarding the missing functional domains of translated proteins resulting from genes largely affected by deleted CNVs.

First, as I only identified deleted CNVs, the results only cover a portion of the genes involved in cattle domestication. In Gene Ontology analysis using the top deletion score genes (23), many genes related to nervous system, more specifically nervous transmission, neuron motion, and neurogenesis were identified (Figure 6.6). The 23 genes identified as being related to the nervous system may have played a role in the behavioral changes that occurred in cattle due to domestication. During domestication, humans selected for docility in cattle leading to the loss of cattle's wild nature. Although these genes do not directly code for behavior, they may encode molecular products that govern the functioning of the brain, which then controls character and behavior. A previous study reported that these variations in behavior shape the evolution of genomic elements that influence social behavior through the feedback of selection [97]. The number of genes with a top 1% deletion score was 33 and the number of CNV overlapping with a gene was 135. After comparing the CNV region with the exon region information of 33 genes, only one of 135 overlapped CNVs (BovineCNV3796, chr2: 44486266-44830807) was in exon region. BovineCNV3796 overlapped with 4 exon regions (ENSBTAE00000348420,

ENSBTAE00000348416, ENSBTAE00000092579, ENSBTAE00000246220). The remaining 134 CNVs were in intron regions of the 33 genes with a top 1% deletion score I assumed that the extra structure was needed to produce diverse genes during the evolutionary process and CNV affecting these processes remains as an evolutionary trace. So, through the 33 genes with the highest number of CNV deletions, I can observe evolutionary evidence of changes in important cattle character and behavior during domestication by the potential missing functional domains of translated proteins resulting from genes affected by deleted CNVs.

Additionally, I found that 16 protein-coding genes overlapped with the top 30 QTL identified using average distance between deletions. These QTL were largely affected by deleted CNVs. Out of the 16 genes, four (OR10A7, OR2J3, OR6C75, OR6C76) were related to sensory perception as olfactory receptor. Studies of evolutionary changes of a number of ORs in other mammalian species reported that cow has fewer gene in specific OR gene cluster (Niimura and Nei 2007). And a Holstein CNV study reported that there were many CNV losses in several OR genes (Seroussi et al. 2010). The result, which showed that several top QTL overlapped with some of the OR genes, may be supported by these studies. Based on this result, I guessed that OR genes have been affected by domestication process. Moreover, the rearing time for cattle was longer than that of other domesticated mammals. I guessed that the difference among domesticated animals could remain in the OR genes. Previous study of cattle olfactory receptor gene reported that there was significant variation in the genetic component of olfactory receptor systems among artiodactyl species, indicating that the selection pressure for maintaining the integrity of olfactory receptor genes was lower in cattle compared to pigs (Lee et al. 2013). These results supported that some CNVs in the selected QTL have been reflected in the evolutionary process during domestication by the missing functional domains of translated proteins resulting from genes affected by deleted CNVs.

In selective sweep signal based on FST of deleted CNV, 14 protein coding genes overlapped with CNV containing strong selective sweep. Out of these, seven genes (TTN, MATN3, DST, HDAC4, TSHR, CCDC141, and GALK2) were reported as being related to meat or milk production (Pipes et al. 1963, Youn, Grozinger and Liu 2000, Yamada et al. 2009, Cole et al. 2011, Fukuda et al. 2010, Mohammad, Hadsell and Haymond 2012, Yucesoy et al. 2013). In these seven CNVs genotype information, five CNVs had higher deletion frequency in Hanwoo than in Holstein (TTN, HDAC4, TSHR, CCDC141, GALK2) and two CNVs (MATN3, DST) had higher deletion frequency in Holstein. Specially, TTN which encodes an abundant protein of striated muscle is famous as gene related to marbling SNP in Japanese Black beef cattle (Yamada et al. 2009). I predict that the marbling SNP may give a certain negative effect to muscle production mechanism by TTN gene for intramuscular fat. In this study, all Hanwoo had double deletion CNVs related to TTN genes (Figure 6.8). Though CNV extraction in Hanwoo and FST calculation based on deleted CNV genotype information are trial procedures and not a widely used method, the result matched up with my expectation. Three (HDAV4, TSHR, CCDC141) of other four genes with CNVs that were mainly deleted in Hanwoo were strongly related to muscle (Table 6.5). The last gene, GALK2, has been shown to be up-regulated during the secretory activation in initiation of milk production (Mohammad et al. 2012). In this study, 13 individual had double deletions and 9 individual had single deletions in Hanwoo, but in Holstein only 3 had single deletions and the remaining 7 individuals had no deletions. In the case of CNV that were mainly deleted in Holstein, MATN3 (BovineCNV3277) is related to genetic risk factors for osteoarthritis which is related to dairy production (Yucesoy et al. 2013). Nine of the 10 Holstein individuals in this study had more than one deletion, but none of the Hanwoo individuals had deletions. These facts supported that these

CNVs have contributed to breed differentiation, perhaps, by missing functional domains of translated proteins resulting from genes affected by deleted CNVs.

In the case of breed-specific CNVs, I selected CNVs in one breed, so there were a higher number of breed-specific CNVs than CNVs found for breed differentiation and it was difficult to discover the biological meaning behind them. In Hanwoo, through GO analysis of genes overlapping with Hanwoo-specific CNVs, two clusters were found. Cluster 1 contains 29 genes and cluster 2 contains 11genes that are related to neuromuscular process and sensory perception, respectively (Figure 6.9). These terms are similar to genes and QTL strongly affected by deleted CNV. Therefore, I suggest that Hanwoo-specific CNVs reflected the evolutionary process, which occurred during domestication. Additionally, in the case of beef cattle such as Hanwoo, humans have limited the allowed space for the cows to induce better marbling of the meat. Based on these facts, I predict that individuals that are less sensitive may have had more advantages than sensitive individuals in enduring this breeding environment in captivity. Therefore, I supposed that due to this breeding history, genes related to sensory perception and response to oxygen had many deletions. In Holstein, two genes (NELL2, C6ORF10) related to Holstein-specific CNV were down regulated in milk production and one gene (MATN3), related to dairy production, was reported to be a genetic risk factor for osteoarthritis (Table 6.10) (Connor et al. 2008, Sadkowski et al. 2009, Yucesoy et al. 2013). MATN3 was also selected in the analysis of selective sweep signals based on FST of deleted CNVs. I predict that Holstein-specific deleted CNV may control some biological process, which gives rise to negative effects on the dairy cattle. These results support the hypothesis that CNVs contribute to breed establishment by the missing functional domains of translated proteins resulting from genes affected by deleted CNVs.

Almost all of the CNV regions examined by PCR in the validation experiment were similar to the CNV regions from Genome STRiP (< 200bp). However, three CNV regions, BovineCNV 3797, BovineCNV3226 and BovineCNV0050 were not fully validated by PCR assays across both breeds and all surveyed individuals. In BovineCNV 3797 and BovineCNV3226, the deletion alleles were not successfully amplified (case 3 in Figure 6.14). This is probably due to the fact that the extracted CNV regions ranged over the primer locations. Although the deleted allele was not confirmed, the wild-type allele was well defined in this case and in the case of BovineCNV3226, individuals considered to have only a deletion allele did not produce any amplicons. Interestingly, the opposite case, case 4 in Figure 6.14, was also present. PCR amplification detected only BovineCNV0050-deleted allele. So I carried out PCR again using primer pairs amplifying the CNV and its outside region to confirm the presence of the CNV containing allele. However, it did not work and no non-deleted allele was amplified (data not shown). BovineCNV0050 region contains undefined gap sequence, so this could be the reason for the failed amplification. Similar to case 3, the results showing deleted alleles were well defined. When I calculated the CNV accuracy, these two cases were scored lower than case 1 and 2 (0.7 vs. 1.0). The CNV accuracy examined in this study was about 80 % (Figure 6.10).

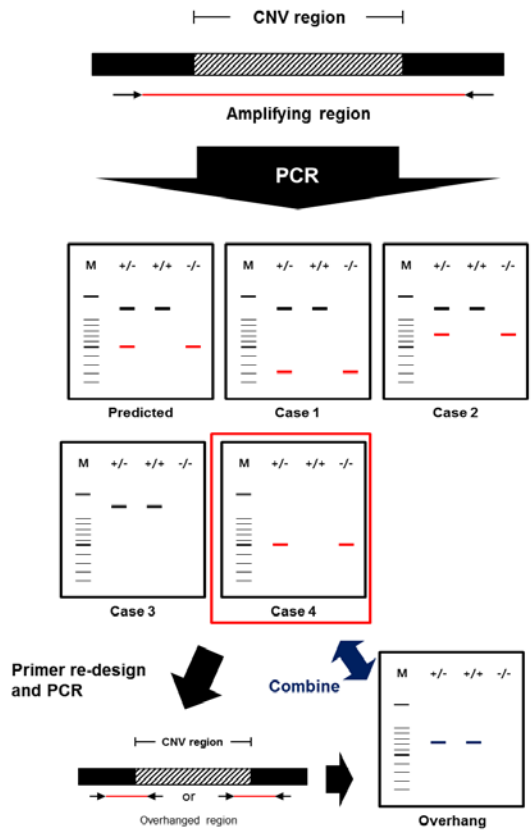

Figure 6.14. CNV validation scheme by genomic DNA PCR. To validate the CNV extracted by GenomeSTRiP, genomic DNA PCR was performed. Primer pairs spanning the extracted CNV were used and each amplicon was visualized by gel electrophoresis. Four patterns of PCR product was detected: the deleted allele being larger or smaller compared to the prediction (case 1 and 2, respectively), absence of deleted allele (case 3) or non-deleted allele (case 4). In case 4, PCR was carried out again with primer pairs which amplify overhanging region between CNV and its outer region. Red and black lines in the diagram representing gel images indicate deleted and non-deleted allele, respectively.

My study presents description of deleted CNVs of cattle by analyzing NGS data of 32 individuals from two breeds. A total of 6,811 deleted CNVs were identified in 22 Hanwoo, and 10 Holsteins individuals. I selected the top 33 genes that had high deletion scores and regarded them as being significantly involved in the domestication process. Their genetic functions were related to nervous system, in particular nervous transmission, neuron motion and neurogenesis. The relationship between these 33 genes and the nervous system may be associated with the changes in behavior due to domestication. The top 30 QTL based on deleted CNVs were associated with diverse quantitative traits including meat and milk production. The genes within top QTL were related to olfactory receptor genes, which reported lower pressure in cattle. I also discovered selective signals in 94 CNVs based on FST values. The top CNVs that were under selection included the TTN gene that has a SNP strongly associated with myofibrillogenesis for marbling in Japanese Black beef cattle. In total, I detected 954 breed-specific CNVs, and 767 of 954 CNVs were Hanwoo-specific and related to several biological processes including phosphorylation, protein modification process, cell adhesion and maintenance, neuromuscular process, sensory perception, and response to oxygen. The other 187 CNVs were Holstein-specific and related to dairy production. Additionally, to confirm the CNV genotype within some putative genes containing the impact on the domestication of cattle, I performed PCR assays. The validation experiment showed that the CNV accuracy of this study is about 80 %.

This study provides information on deleted CNVs across the cattle genome at the population level and suggests their possible roles in both domestication and recent breed selection. This study using deleted CNV at the population level is a trial step towards exploring the underlying genetics of economically important traits in cattle and understanding the genetic changes that occurred during domestication. However, further research into the genes related to CNVs and a comprehensive study

on inserted CNVs is needed to form a more complete picture of the genetic structure variation in the bovine genome. Additionally, when the associations between CNV and economic traits in cows are identified, it will be possible to incorporate them into breeding programs for production enhancement in cattle.

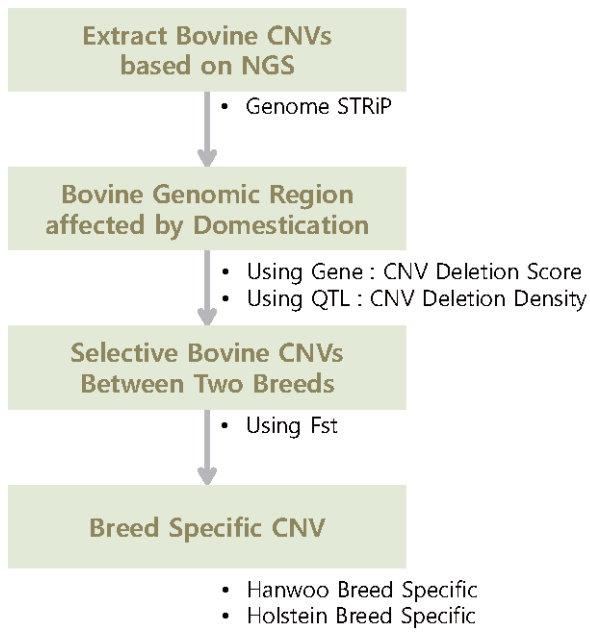

Figure 6.15. Research flow of the study

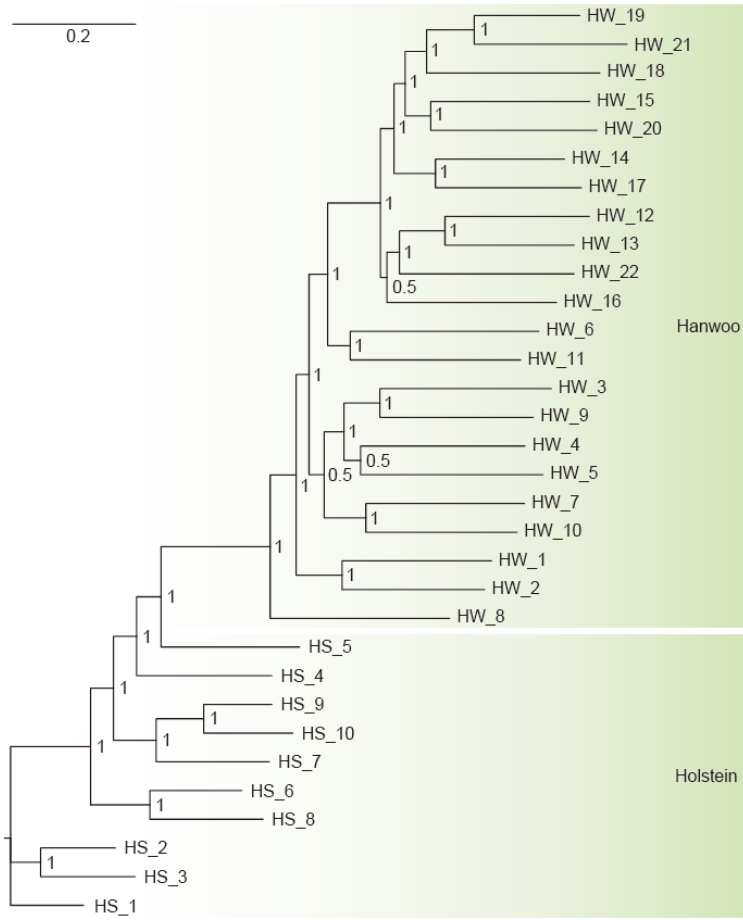

Figure 6.16. Phylogenetic analysis using Bayesian Inference. Sample ID for each branch is in Table 6.1.

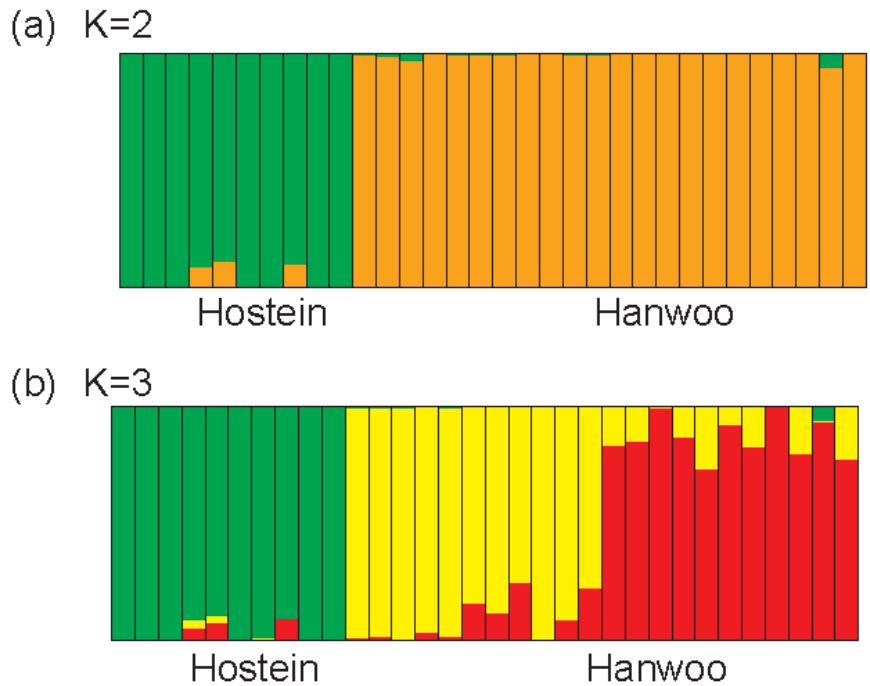

Figure 6.17. Population structure analysis using STRUCTURE. Each individual is represented by a vertical bar, and the length of each colored segment in each of the vertical bars represents the proportion contributed by ancestral populations. (a) Two colors (K=2) mostly represent population structure of 32 individuals. (b) Three colors (K=3) represent population structure of 32 individuals.

General Discussion

By using genomic information from SNP and NGS data, many biological and evolutionary meaning could be obtained to develop animal breeding.

Before genetics variant study, measuring effective population size is important to comprehend interesting animal populations. In chapter 2, I characterized more accurate linkage disequilibrium using dense SNP genotype data. Based on this information, current N_e was estimated and the ancestral N_e was inferred in Korean dairy cattle population. I could know that current N_e of Korean dairy cattle was approximately 122 individuals. And we inferred the ancestral Ne and observed a recent rapid increase in N_e which reached approximately 500, 10 generations ago followed by a decrease until the present time. These result can be rationalized using current knowledge of the history of the dairy cattle breeds producing milk in Korea and be used to other animal breeding based on genomics data as background knowledge. I chapter 3, I estimated historical effective population size in the minke whale based on coalescent model to know when and how much minke whale population size decreases rapidly. As a result, strong predicted time of minke whale declination during Holocene is approximately between 194 and 902 years ago and minke whale population diversity was estimated to downsize to approximately 3.1%. These study offered a chance to better understand the population history of the largest aquatic mammals on earth.

I identified genetic variants associated with specific trait using genotype chips which gives broad regions associated with specific trait in animal. In chapter 4, genome-wide association study was conducted to search for genetic variants of Thoroughbred associated with the EBV for racing record. 28 significant SNPs as GWAS result were related to 17 genes and these genes were related to myogenesis and muscle maintenance. Because those SNPs were newly reported, they will help

to expand my knowledge of the polygenic nature of racing performance in Thoroughbreds. In chapter 5, multivariate GWAS based on EBVs was conducted to search for genetic variants associated with the milk production in Holstein. Novel 128 SNPs related to milk production were identified and they will help to develop animal breeding based on genomics data.

Copy number variation (CNV) is one of several genetic variant types and occupy a higher percentage of genomic sequence than SNP. So, many researcher guessed that CNV has many underlie biological functions related to animal production traits in previous studies. In chapter 6, I performed CNV study using NGS data to explore their potential function and contributions in cattle domestication. This CNV analysis attempted to detect CNVs associated with cattle domestication, CNVs related to differentiation between beef cattle and milk cattle and breed-specific CNVs. This studies had many novel trial in CNV study using NGS data. By providing several types of information on cattle CNV, this study provided the basis for further studies into the role of deleted CNVs in the cattle genome.

Reference

Abecasis, G., L. Cardon & W. Cookson (2000) A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics,* 66**,** 279-292.

Aitman, T. J., R. Dong, T. J. Vyse, P. J. Norsworthy, M. D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A. J. Marshall & E. Petretto (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature,* 439**,** 851-855.

Akula, N., A. Baranova, D. Seto, J. Solka, M. A. Nalls, A. Singleton, L. Ferrucci, T. Tanaka, S. Bandinelli & Y. S. Cho (2011) A network-based approach to prioritize results from genome-wide association studies. *PloS One,* 6**,** e24220.

Alkan, C., B. P. Coe & E. E. Eichler (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics,* 12**,** 363-376.

Allen, H. L., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam & S. Raychaudhuri (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-838.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

Altshuler, D. M., R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, P. De Bakker, P. Deloukas & S. B. Gabriel (2010) Integrating common and rare genetic variation in diverse human populations. *Nature,* 467**,** 52-58.

Amills, M., O. Vidal, L. Varona, A. Tomas, M. Gil, A. Sanchez & J. Noguera (2005) Polymorphism of the pig 2, 4-dienoyl CoA reductase 1 gene (DECR1) and its association with carcass and meat quality traits. *Journal Of Animal Science,* 83**,** 493-498.

Amin, N., C. M. van Duijn & Y. S. Aulchenko (2007) A genomic background based method for association analysis in related individuals. *PloS One,* 2**,** e1274.

Amos, C. I. (2007) Successful design and conduct of genome-wide association studies. *Human Molecular Genetics,* 16**,** R220-R225.

Andrews, S. (2010) FASTQC. A quality control tool for high throughput sequence data. *URL* http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ardlie, K. G., L. Kruglyak & M. Seielstad (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics,* 3**,** 299-309.

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters & R. Spelman (2009) A high density linkage map of the bovine genome. *BMC Genetics,* 10**,** 18.

Aspenström, P. (1997) A Cdc42 target protein with homology to the non-kinase domain of FER has a potential role in regulating the actin cytoskeleton. *Current Biology,* 7**,** 479-487.

Atkinson, M. R., M. A. Savageau, J. T. Myers & A. J. Ninfa (2003) Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell,* 113**,** 597-607.

Bach, A.-S., S. Enjalbert, F. Comunale, S. Bodin, N. Vitale, S. Charrasse & C. Gauthier-Rouvière (2010) ADP-ribosylation factor 6 regulates mammalian myoblast fusion through phospholipase D1 and phosphatidylinositol 4, 5-bisphosphate signaling pathways. *Molecular Biology Of The Cell,* 21**,** 2412-2424.

Bae, J., H. Cheong, L. Kim, S. NamGung, T. Park, J.-Y. Chun, J. Kim, C. Pasaje, J. Lee & H. Shin (2010a) Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics,* 11**,** 232.

Bae, J. S., H. S. Cheong, L. H. Kim, S. NamGung, T. J. Park, J.-Y. Chun, J. Y. Kim, C. F. Pasaje, J. S. Lee & H. D. Shin (2010b) Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics,* 11**,** 232.

Baik, M., B. Etchebarne, J. Bong & M. VandeHaar (2009) Gene expression profiling of liver and mammary tissues of lactating dairy cows. *Asian-Australasian Journal Of Animal Sciences,* 6**,** 871-884.

Barendse, W., B. Harrison, R. Bunch, M. Thomas & L. Turner (2009a) Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. *BMC Genomics,* 10**,** 178.

Barendse, W., B. E. Harrison, R. J. Bunch, M. B. Thomas & L. B. Turner (2009b) Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. *BMC Genomics,* 10**,** 178.

Barrett, J., B. Fry, J. Maller & M. Daly (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-265.

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

Bilimoria, P. M. & A. Bonni (2013) Molecular control of axon branching. *The Neuroscientist,* 19**,** 16-24.

Binns, M., D. Boehler & D. Lambert (2010) Identification of the myostatin locus (MSTN) as having a major effect on optimum racing distance in the Thoroughbred horse in the USA. *Animal Genetics,* 41**,** 154-158.

Bionaz, M., K. Periasamy, S. L. Rodriguez-Zas, W. L. Hurley & J. J. Loor (2012) A novel dynamic impact approach (DIA) for functional analysis of time-course omics studies: validation using the bovine mammary transcriptome. *PloS One,* 7**,** e32455.

Bliss, T. V. & G. L. Collingridge (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature,* 361**,** 31-39.

Blott, S., J.-J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart & D. Johnson (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics,* 163**,** 253-266.

Bosio, Y., G. Berto, P. Camera, F. Bianchi, C. Ambrogio, P. Claus & F. Di Cunto (2012) PPP4R2 regulates neuronal cell differentiation and survival, functionally cooperating with SMN. *European Journal Of Cell Biology,* 91**,** 662-674.

Boyko, A. R., P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt, K. E. Lohmueller, K. Zhao, A. Brisbin, H. G. Parker & M. Cargill (2010) A simple genetic architecture underlies morphological variation in dogs. *PLoS Biology,* 8**,** e1000451.

Bray, M. S., J. M. Hagberg, L. Pérusse, T. Rankinen, S. M. Roth, B. Wolfarth & C. Bouchard (2009) The human gene map for performance and health-related fitness phenotypes: the 2006-2007 update. *Medicine And Science In Sports And Exercise,* 41**,** 35.

Brose, K. & M. Tessier-Lavigne (2000) Slit proteins: key regulators of axon guidance, axonal branching, and cell migration. *Current Opinion In Neurobiology,* 10**,** 95-102.

Browning, B. L. & S. R. Browning (2011) A fast, powerful method for detecting identity by descent. *The American Journal Of Human Genetics,* 88**,** 173-182.

Brym, P., S. Kamiński & E. Wójcik (2004) Nucleotide sequence polymorphism within exon 4 of the bovine prolactin gene and its associations with milk performance traits. *Journal Of Applied Genetics,* 46**,** 179-185.

Calus, M., A. de Roos & R. Veerkamp (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics,* 178**,** 553-561.

Cartmell, J. & D. D. Schoepp (2000) Regulation of neurotransmitter release by metabotropic glutamate receptors. *Journal Of Neurochemistry,* 75**,** 889-907.

Casey, T. & K. Plaut (2012) LACTATION BIOLOGY SYMPOSIUM: Circadian clocks as mediators of the homeorhetic response to lactation. *Journal Of Animal Science,* 90**,** 744-754.

Chen, W.-K., J. D. Swartz, L. J. Rush & C. E. Alvarez (2009) Mapping DNA structural variation in dogs. *Genome Research,* 19**,** 500-509.

Cole, J., G. Wiggans, L. Ma, T. Sonstegard, T. Lawlor, B. Crooker, C. Van Tassell, J. Yang, S. Wang & L. Matukumalli (2011) Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC Genomics,* 12**,** 408.

Collingridge, G. L. & R. A. Lester (1989) Excitatory amino acid receptors in the vertebrate central nervous system. *Pharmacological Reviews,* 41**,** 143-210.

Connor, E., S. Siferd, T. Elsasser, C. Evock-Clover, C. Van Tassell, T. Sonstegard, V. Fernandes & A. Capuco (2008) Effects of increased milking frequency on gene expression in the bovine mammary gland. *BMC Genomics,* 9**,** 362.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes & P. Campbell (2009) Origins and functional impact of copy number variation in the human genome. *Nature,* 464**,** 704-712.

Coon, K. D., A. J. Myers, D. W. Craig, J. A. Webster, J. V. Pearson, D. H. Lince, V. L. Zismann, T. G. Beach, D. Leung & L. Bryden (2007) A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *Journal Of Clinical Psychiatry,* 68**,** 613.

Corbin, L., S. Bishop, J. Swinburne, M. Vaudin, S. Blott & J. Woolliams. 2010a. The impact of method on the estimated effective population size of a Thoroughbred population using genotype data. In *Proceedings Of The 9th World Congress On Genetics Applied To Livestock Production*.

Corbin, L., S. Blott, J. Swinburne, M. Vaudin, S. Bishop & J. Woolliams (2010b) Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Animal Genetics,* 41**,** 8-15.

Craig, D. W. & D. A. Stephan (2005) Applications of whole-genome high-density SNP genotyping.

Cunningham, E., J. Dooley, R. Splan & D. Bradley (2001) Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. *Animal Genetics,* 32**,** 360-364.

D'Alessandro, A., L. Zolla & A. Scaloni (2011) The bovine milk proteome: cherishing, nourishing and fostering molecular complexity. An interactomics and functional overview. *Molecular BioSystems,* 7**,** 579-597.

Da Wei Huang, B. T. S. & R. A. Lempicki (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* 4**,** 44-57.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva & J. A. Woolliams (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics,* 185**,** 1021-1031.

Davy, A., N. W. Gale, E. W. Murray, R. A. Klinghoffer, P. Soriano, C. Feuerstein & S. M. Robbins (1999) Compartmentalized signaling by GPI-anchored ephrin-A5 requires the Fyn tyrosine kinase to regulate cellular adhesion. *Genes & Development,* 13**,** 3125-3135.

Daw, E. W., S. C. Heath & Y. Lu (2005) Single-nucleotide polymorphism versus microsatellite markers in a combined linkage and segregation analysis of a quantitative trait. *BMC Genetics,* 6**,** S32.

De Roos, A., B. J. Hayes, R. Spelman & M. E. Goddard (2008) Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics,* 179**,** 1503-1512.

DerSimonian, R. & N. Laird (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7**,** 177-188.

Devlin, B., S.-A. Bacanu & K. Roeder (2004) Genomic control to the extreme. *Nature Genetics,* 36**,** 1129-1130.

Dostaler-Touchette, V., F. Bédard, C. Guillemette, F. Pothier, P. Chouinard & F. Richard (2009) Cyclic adenosine monophosphate (cAMP)-specific phosphodiesterase is functional in bovine mammary gland. *Journal Of Dairy Science,* 92**,** 3757-3765.

Drögemüller, C., O. Distl & T. Leeb (2001) Partial deletion of the bovine ED1 gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Research,* 11**,** 1699-1705.

Du, F.-X., A. C. Clutter & M. M. Lohuis (2007) Characterizing linkage disequilibrium in pig populations. *International Journal Of Biological Sciences,* 3**,** 166.

Eck, S. H., A. Benet-Pagès, K. Flisikowski, T. Meitinger, R. Fries & T. M. Strom (2009) Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. *Genome Biology,* 10**,** R82.

Elsik, C. G., R. L. Tellam & K. C. Worley (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-528.

Endelman, J. B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome,* 4**,** 250-255.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason & M. Goddard (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal Of Dairy Science,* 95**,** 4114-4129.

Fadista, J., B. Thomsen, L.-E. Holm & C. Bendixen (2010) Copy number variation in the bovine genome. *BMC Genomics,* 11**,** 284.

Falconer, D. S. (1960) Introduction To Quantitative Genetics.

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal Of The American Statistical Association,* 96**,** 1348-1360.

Fellermann, K., D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, C. L. Bevins, W. Reinisch, A. Teml, M. Schwab & P. Lichter (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *The American Journal Of Human Genetics,* 79**,** 439-448.

Ferreira, M. A. & S. M. Purcell (2009) A multivariate test of association. *Bioinformatics,* 25**,** 132-133.

Finlay, E. K., D. P. Berry, B. Wickham, E. P. Gormley & D. G. Bradley (2012) A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle. *PloS One,* 7**,** e30545.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley & S. Fitzgerald (2012) Ensembl 2012. *Nucleic Acids Research,* 40**,** D84-D90.

Fluckey, J. D., M. Knox, L. Smith, E. E. Dupont-Versteegden, D. Gaddy, P. A. Tesch & C. A. Peterson (2006) Insulin-facilitated increase of muscle protein synthesis after resistance exercise involves a MAP kinase pathway. *American Journal Of Physiology-Endocrinology And Metabolism,* 290**,** E1205-E1211.

Flury, C., M. Tapio, T. Sonstegard, C. Drögemüller, T. Leeb, H. Simianer, O. Hanotte & S. Rieder (2010) Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *Journal Of Animal Breeding And Genetics,* 127**,** 339-347.

Fontanesi, L., F. Beretti, P. Martelli, M. Colombo, S. Dall'Olio, M. Occidente, B. Portolano, R. Casadio, D. Matassino & V. Russo (2011) A first comparative map of copy number variations in the sheep genome. *Genomics,* 97**,** 158-165.

Fontanesi, L., F. Beretti, V. Riggio, S. Dall'Olio, R. Davoli, V. Russo & B. Portolano (2009a) Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenetic And Genome Research,* 126**,** 333-347.

Fontanesi, L., F. Beretti, V. Riggio, E. Gómez González, S. Dall'Olio, R. Davoli, V. Russo & B. Portolano (2009b) Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenetic And Genome Research,* 126**,** 333-347.

Fontanesi, L., D. Calò, G. Galimberti, R. Negrini, R. Marino, A. Nardone, P. Ajmone-Marsan & V. Russo (2014) A candidate gene association study for nine economically important traits in Italian Holstein cattle. *Animal Genetics*.

Freeman, J. L., G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith & M. E. Hurles (2006) Copy number variation: new insights in genome diversity. *Genome Research,* 16**,** 949-961.

Fukuda, T., S. Sugita, R. Inatome & S. Yanagi (2010) CAMDI, a novel disrupted in schizophrenia 1 (DISC1)-binding protein, is required for radial migration. *Journal Of Biological Chemistry,* 285**,** 40554-40561.

Gaffney, B. & E. Cunningham (1988) Estimation of genetic trend in racing performance of thoroughbred horses. *Nature,* 332**,** 722-724.

Georges, M. (2007) Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annual Review Genomics And Human Genetics,* 8**,** 131-162.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto, A. T. Pasquino, L. S. Sargeant, A. Sorensen, M. R. Steele & X. Zhao (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics,* 139**,** 907-920.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes & S. Lien (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science (New York, NY),* 324**,** 528-532.

Gleeson, J. G., P. T. Lin, L. A. Flanagan & C. A. Walsh (1999) Doublecortin is a microtubule-associated protein and is expressed widely by migrating neurons. *Neuron,* 23**,** 257-271.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea & S. Sykes (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings Of The National Academy Of Sciences,* 108**,** 1513-1518.

Gogol-Döring, A. & W. Chen. 2012. An overview of the analysis of next generation sequencing data. In *Next Generation Microarray Bioinformatics*, 249-257. Springer.

Goldstein, J. L. & M. S. Brown (1990) Regulation of the mevalonate pathway. *Nature,* 343**,** 425.

Graubert, T. A., P. Cahan, D. Edwin, R. R. Selzer, T. A. Richmond, P. S. Eis, W. D. Shannon, X. Li, H. L. McLeod & J. M. Cheverud (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genetics,* 3**,** e3.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J.-J. Kim, A. Kvasz, M. Mni, P. Simon, J.-M. Frère & W. Coppieters (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield

and composition. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 2398-2403.

Grozdanov, P. N., S. Roy, N. Kittur & U. T. Meier (2009) SHQ1 is required prior to NAF1 for assembly of H/ACA small nucleolar and telomerase RNPs. *Rna,* 15**,** 1188-1197.

Gu, J., D. MacHugh, B. McGivney, S. Park, L. Katz & E. Hill (2010) Association of sequence variants in CKM (creatine kinase, muscle) and COX4I2 (cytochrome c oxidase, subunit 4, isoform 2) genes with racing performance in Thoroughbred horses. *Equine Veterinary Journal,* 42**,** 569-575.

Gu, J., N. Orr, S. D. Park, L. M. Katz, G. Sulimova, D. E. MacHugh & E. W. Hill (2009) A genome scan for positive selection in thoroughbred horses. *PloS One,* 4**,** e5767.

Guryev, V., K. Saar, T. Adamovic, M. Verheul, S. A. Van Heesch, S. Cook, M. Pravenec, T. Aitman, H. Jacob & J. D. Shull (2008) Distribution and functional impact of DNA copy number variation in the rat. *Nature Genetics,* 40**,** 538-545.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson & C. D. Bustamante (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics,* 5**,** e1000695.

Halbleib, J. M. & W. J. Nelson (2006) Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes & Development,* 20**,** 3199-3214.

Handsaker, R. E., J. M. Korn, J. Nemesh & S. A. McCarroll (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics,* 43**,** 269-276.

Hara, Y., T. Nomura, K. Yoshizaki, J. Frisén & N. Osumi (2010) Impaired Hippocampal Neurogenesis and Vascular Formation in Ephrin-A5-Deficient Mice. *Stem Cells,* 28**,** 974-983.

Hayes, B., P. Bowman, A. Chamberlain & M. Goddard (2009a) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal Of Dairy Science,* 92**,** 433-443.

Hayes, B., A. Chamberlain, S. Maceachern, K. Savin, H. McPartlan, I. MacLeod, L. Sethuraman & M. Goddard (2009b) A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. *Animal Genetics,* 40**,** 176-184.

Hayes, B. & M. Goddard (2010) Genome-wide association and genomic selection in animal breeding. *Genome,* 53**,** 876-883.

Hayes, B. J., P. M. Visscher, H. C. McPartlan & M. E. Goddard (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research,* 13**,** 635-643.

Hill, E., D. Bradley, M. Al-Barody, O. Ertugrul, R. Splan, I. Zakharov & E. Cunningham (2002) History and integrity of thoroughbred dam lines revealed in equine mtDNA variation. *Animal Genetics,* 33**,** 287-294.

Hill, E., S. Eivers, B. McGivney, R. Fonseca, J. Gu, N. Smith, J. Browne, D. MacHugh & L. Katz (2010a) Moderate and high intensity sprint exercise induce differential responses in COX4I2 and PDK4 gene expression in Thoroughbred horse skeletal muscle. *Equine Veterinary Journal,* 42**,** 576-581.

Hill, E., J. Gu, B. McGivney & D. MacHugh (2010b) Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. *Animal Genetics,* 41**,** 56-63.

Hill, E. W., J. Gu, S. S. Eivers, R. G. Fonseca, B. A. McGivney, P. Govindarajan, N. Orr, L. M. Katz & D. MacHugh (2010c) A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. *PLoS One,* 5**,** e8645.

Hill, E. W., B. A. McGivney, J. Gu, R. Whiston & D. E. MacHugh (2010d) A genome-wide SNP-association study confirms a sequence variant (g. 66493737C> T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. *BMC Genomics,* 11**,** 552.

Hill, W. & A. Robertson (1968) Linkage disequilibrium in finite populations. *TAG Theoretical And Applied Genetics,* 38**,** 226-231.

Hoerl, A. E. & R. W. Kennard (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics,* 12**,** 55-67.

Howie, B., J. Marchini & M. Stephens (2011) Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics,* 1**,** 457-470.

Hu, Z.-L., E. R. Fritz & J. M. Reecy (2007) AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Research,* 35**,** D604-D609.

Huang, W., B. Kirkpatrick, G. Rosa & H. Khatib (2010) A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Animal Genetics,* 41**,** 570-578.

Itoh, A., T. Miyabayashi, M. Ohno & S. Sakano (1998) Cloning and expressions of three mammalian homologues of Drosophila slit suggest possible roles for Slit in the formation and maintenance of the nervous system. *Molecular Brain Research,* 62**,** 175-186.

Jelinsky, S. A., J. Archambault, L. Li & H. Seeherman (2010) Tendon-selective genes identified from rat and human musculoskeletal tissues. *Journal Of Orthopaedic Research,* 28**,** 289-297.

Jiang, L., J. Liu, D. Sun, P. Ma, X. Ding, Y. Yu & Q. Zhang (2010) Genome wide association studies for milk production traits in Chinese Holstein population. *PloS One,* 5**,** e13661.

Jorgensen, T. J., I. Ruczinski, B. Kessing, M. W. Smith, Y. Y. Shugart & A. J. Alberg (2009) Hypothesis-driven candidate gene association studies: practical design and analytical considerations. *American Journal Of Epidemiology,* 170**,** 986-993.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti & E. Eskin (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics,* 42**,** 348-354.

Khatkar, M. S., F. W. Nicholas, A. R. Collins, K. R. Zenger, J. A. Cavanagh, W. Barris, R. D. Schnabel, J. F. Taylor & H. W. Raadsma (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics,* 9**,** 187.

Kijas, J. W., W. Barendse, W. Barris, B. Harrison, R. McCulloch, S. McWilliam & V. Whan (2011) Analysis of copy number variants in the cattle genome. *Gene,* 482**,** 73-77.

Kim, J., T. Löwe & T. Hoppe (2008) Protein quality control gets muscle into shape. *Trends In Cell Biology,* 18**,** 264-272.

Kimball, S. R., T. C. Vary & L. S. Jefferson (1994) Regulation of protein synthesis by insulin. *Annual Review Of Physiology,* 56**,** 321-348.

Ko, J.-A., Y. Kimura, K. Matsuura, H. Yamamoto, T. Gondo & M. Inui (2006) PDZRN3 (LNX3, SEMCAP3) is required for the differentiation of C2C12 myoblasts into myotubes. *Journal Of Cell Science,* 119**,** 5106-5113.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero & L. Du (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science,* 318**,** 420-426.

LaFramboise, T. (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research,* 37**,** 4181-4193.

Lai, W. R., M. D. Johnson, R. Kucherlapati & P. J. Park (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics,* 21**,** 3763-3770.

Laird, N. M. & C. Lange (2008) Family-based methods for linkage and association analysis. *Advances In Genetics,* 60**,** 219-252.

Langmead, B. & S. L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9**,** 357-359.

Lannon, C. L. & P. H. Sorensen. 2005. ETV6–NTRK3: a chimeric protein tyrosine kinase with transformation activity in multiple cell lineages. In *Seminars in cancer biology*, 215-223. Elsevier.

Le Maréchal, C., E. Masson, J.-M. Chen, F. Morel, P. Ruszniewski, P. Levy & C. Férec (2006) Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nature Genetics,* 38**,** 1372-1374.

Lee, A. B., D. Luca, L. Klei, B. Devlin & K. Roeder (2010) Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology,* 34**,** 51-59.

Lee, J. A., C. Carvalho & J. R. Lupski (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* 131**,** 1235-1247.

Lee, J. A. & J. R. Lupski (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron,* 52**,** 103-121.

Lee, K., D. T. Nguyen, M. Choi, S.-Y. Cha, J.-H. Kim, H. Dadi, H. G. Seo, K. Seo, T. Chun & C. Park (2013) Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. *BMC Genomics,* 14**,** 596.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis & R. Durbin (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Li, H., Z. Wang, S. Moore, F. Schenkel & P. Stothard (2010a) Genome-wide scan for positional and functional candidate genes affecting milk production traits in Canadian Holstein cattle. *Proc. 9th WCGALP, Leipzig, Germany. http://www. kongressband. de/wcgalp2010/assets/pdf/0535. pdf Accessed Nov,* 26**,** 2010.

Li, Y., C. J. Willer, J. Ding, P. Scheet & G. R. Abecasis (2010b) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology,* 34**,** 816-834.

Liscurn, L. (2002) Cholesterol biosynthesis. *New Comprehensive Biochemistry,* 36**,** 409-431.

Litwack, E. D., R. Babey, R. Buser, M. Gesemann & D. D. O'Leary (2004) Identification and characterization of two novel brain-derived immunoglobulin superfamily members with a unique structural organization. *Molecular And Cellular Neuroscience,* 25**,** 263-274.

Liu, G. E., Y. Hou, B. Zhu, M. F. Cardone, L. Jiang, A. Cellamare, A. Mitra, L. J. Alexander, L. L. Coutinho & M. E. Dell'Aquila (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Research,* 20**,** 693-703.

Liu, Y., X. Qin, X.-Z. H. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. P. Kent, M. Sodeland & Y. Ren (2009) Bos taurus genome assembly. *Bmc Genomics,* 10**,** 180.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan & Y. Liu (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1**,** 1-6.

Lynch, M. & B. Walsh (1998) Genetics And Analysis Of Quantitative Traits.

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This & C. Cierco-Ayrolles (2011) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity,* 108**,** 285-291.

Mani, O., M. Sorensen, K. Sejrsen, R. Bruckmaier & C. Albrecht (2009) Differential expression and localization of lipid transporters in the bovine mammary gland during the pregnancy-lactation cycle. *Journal Of Dairy Science,* 92**,** 3744-3756.

Marchini, J. & B. Howie (2010) Genotype imputation for genome-wide association studies. *Nature Reviews Genetics,* 11**,** 499-511.

Marthiens, V., J. Gavard, M. Lambert & R. M. Mège (2002) Cadherin-based cell adhesion in neuromuscular development. *Biology Of The Cell,* 94**,** 315-326.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith & T. S. Sonstegard (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PloS One,* 4**,** e5350.

McIntyre, J. C., W. B. Titlow & T. S. McClintock (2010) Axon growth and guidance genes identify nascent, immature, and mature olfactory sensory neurons. *Journal Of Neuroscience Research,* 88**,** 3243-3256.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel & M. Daly (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

McPherron, A. C. & S.-J. Lee (1997) Double muscling in cattle due to mutations in the myostatin gene. *Proceedings Of The National Academy Of Sciences,* 94**,** 12457-12461.

Meldrum, B. & J. Garthwaite (1990) Excitatory amino acid neurotoxicity and neurodegenerative disease. *Trends In Pharmacological Sciences,* 11**,** 379-387.

Meuwissen, T. & M. E. Goddard (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution,* 33**,** 605-634.

Meyerson, M., S. Gabriel & G. Getz (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics,* 11**,** 685-696.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye & R. K. Cheetham (2011) Mapping copy number variation by population-scale genome sequencing. *Nature,* 470**,** 59-65.

Mitsui, K., D. Nakajima, O. Ohara & M. Nakayama (2002) Mammalian fat3: a large protein that contains multiple cadherin and EGF-like motifs. *Biochemical And Biophysical Research Communications,* 290**,** 1260-1266.

Mohammad, M. A., D. L. Hadsell & M. W. Haymond (2012) Gene regulation of UDP-galactose synthesis and transport: potential rate-limiting processes in initiation of milk production in humans. *American Journal Of Physiology-Endocrinology And Metabolism,* 303**,** E365-E376.

Moore, T. M., R. Garg, C. Johnson, M. J. Coptcoat, A. J. Ridley & J. D. Morris (2000) PSK, a novel STE20-like kinase derived from prostatic carcinoma that activates the

c-Jun N-terminal kinase mitogen-activated protein kinase pathway and regulates actin cytoskeletal organization. *Journal Of Biological Chemistry,* 275**,** 4311-4322.

Moritsu, Y., H. Funakoshi & S. Ichikawa (1994) Genetic evaluation of sires and environmental factors influencing best racing times of Thoroughbred horses in Japan. *Journal Of Equine Science,* 5**,** 53-58.

Moschella, M. C., J. Watras, T. Jayaraman & A. R. Marks (1995) Inositol 1, 4, 5-trisphosphate receptor in skeletal muscle: differential expression in myofibres. *Journal Of Muscle Research And Cell Motility,* 16**,** 390-400.

Mota, M., A. Abrahão & H. Oliveira (2005) Genetic and environmental parameters for racing time at different distances in Brazilian Thoroughbreds. *Journal Of Animal Breeding And Genetics,* 122**,** 393-399.

Muir, W., G. Wong, Y. Zhang, J. Wang, M. Groenen, R. Crooijmans, H.-J. Megens, H. Zhang, J. McKay & S. McLeod (2008) Review of the initial validation and characterization of a 3K chicken SNP array. *World's Poultry Science Journal,* 64**,** 219-226.

Myers, A. J., J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem, D. Leung, L. Bryden & P. Nath (2007) A survey of genetic human cortical gene expression. *Nature Genetics,* 39**,** 1494-1499.

Nagae, S., T. Tanoue & M. Takeichi (2007) Temporal and spatial expression profiles of the Fat3 protein, a giant cadherin molecule, during mouse development. *Developmental Dynamics,* 236**,** 534-543.

Nemes, J. P., K. A. Benzow & M. D. Koob (2000) The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). *Human Molecular Genetics,* 9**,** 1543-1551.

Nicholas, T. J., Z. Cheng, M. Ventura, K. Mealey, E. E. Eichler & J. M. Akey (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research,* 19**,** 491-499.

Niimura, Y. & M. Nei (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One,* 2**,** e708.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann & M. R. Nelson (2008) Genes mirror geography within Europe. *Nature,* 456**,** 98-101.

O'Connor, M. S., M. E. Carlson & I. M. Conboy (2009) Differentiation rather than aging of muscle stem cells abolishes their telomerase activity. *Biotechnology Progress,* 25**,** 1130-1137.

Oki, H., Y. Sasaki & R. Willham (1994) Genetics of racing performance in the Japanese Thoroughbred horse. *Journal Of Animal Breeding And Genetics,* 111**,** 128-137.

Ooms, L., K. Horan, P. Rahman, G. Seaton, R. Gurung, D. Kethesparan & C. Mitchell (2009) The role of the inositol polyphosphate 5-phosphatases in cellular function and human disease. *Biochemical Journal,* 419**,** 29-49.

Park, J.-E., J.-R. Lee, S. Oh, J. W. Lee, H.-S. Oh & H. Kim (2011) Principal components analysis applied to genetic evaluation of racing performance of Thoroughbred race horses in Korea. *Livestock Science,* 135**,** 293-299.

Paudel, Y., O. Madsen, H.-J. Megens, L. A. Frantz, M. Bosse, J. W. Bastiaansen, R. P. Crooijmans & M. A. Groenen (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics,* 14**,** 449.

Peng, G., L. Luo, H. Siu, Y. Zhu, P. Hu, S. Hong, J. Zhao, X. Zhou, J. D. Reveille & L. Jin (2010a) Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal Of Human Genetics,* 18**,** 111-117.

Peng, Y., H. C. Leung, S. Yiu & F. Y. Chin. 2010b. IDBA–a practical iterative de Bruijn graph de novo assembler. In *Research In Computational Molecular Biology*, 426-440. Springer.

Perez, R., J. Cañón & S. Dunner (2010) Genes associated with long-chain omega-3 fatty acids in bovine skeletal muscle. *Journal Of Applied Genetics,* 51**,** 479-487.

Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain & R. Misra (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics,* 39**,** 1256-1260.

Pinto, D., K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. MacDonald & R. Mills (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology,* 29**,** 512-520.

Pipes, G., T. Bauman, J. Brooks, J. Comfort & C. Turner (1963) Effect of season, sex and breed on the thyroxine secretion rate of beef cattle and a comparison with dairy cattle. *Journal Of Animal Science,* 22**,** 476-480.

Porto Neto, L. R., N. N. Jonsson, M. J. D'Occhio & W. Barendse (2011) Molecular genetic approaches for identifying the basis of variation in resistance to tick infestation in cattle. *Veterinary Parasitology,* 180**,** 165-172.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick & D. Reich (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics,* 38**,** 904-909.

Price, A. L., N. A. Zaitlen, D. Reich & N. Patterson (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics,* 11**,** 459-463.

Pritchard, J. K. & N. A. Rosenberg (1999) Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal Of Human Genetics,* 65**,** 220-228.

Pritchard, J. K., M. Stephens & P. Donnelly (2000) Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker & M. J. Daly (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal Of Human Genetics,* 81**,** 559-575.

Qanbari, S., E. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. Sharifi & H. Simianer (2010) The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics,* 41**,** 346-356.

Quaas, R. L. & E. Pollak (1980) Mixed model methodology for farm and ranch beef cattle testing programs. *Journal Of Animal Science,* 51**,** 1277-1287.

Rønn, L. C. B., B. Hartz & E. Bock (1998) The neural cell adhesion molecule (NCAM) in development and plasticity of the nervous system. *Experimental Gerontology,* 33**,** 853-864.

Ramayo-Caldas, Y., A. Castelló, R. Pena, E. Alves, A. Mercadé, C. Souza, A. Fernández, M. Perez-Enciso & J. Folch (2010a) Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics,* 11**,** 593.

Ramayo-Caldas, Y., A. Castelló, R. N. Pena, E. Alves, A. Mercadé, C. A. Souza, A. I. Fernández, M. Perez-Enciso & J. M. Folch (2010b) Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics,* 11**,** 593.

Ramos, A., H. Megens, R. Crooijmans, L. Schook & M. Groenen (2011) Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. *Animal Genetics,* 42**,** 613-620.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson & W. Chen (2006) Global variation in copy number in the human genome. *Nature,* 444**,** 444-454.

Ricard, A. 1998. Developments in the genetic evaluation of performance traits in horses. In *Proceedings of the 6th World congress on genetics applied to Livestock Production, Armidale, NSW, Australia,* 11-16.

Richards, J. B., D. Waterworth, S. O'Rahilly, M.-F. Hivert, R. J. Loos, J. R. Perry, T. Tanaka, N. J. Timpson, R. K. Semple & N. Soranzo (2009) A genome-wide association study reveals variants in ARL15 that influence adiponectin levels. *PLoS Genetics,* 5**,** e1000768.

Rodrigues, G. A., M. Falasca, Z. Zhang, S. H. Ong & J. Schlessinger (2000) A novel positive feedback loop mediated by the docking protein Gab1 and phosphatidylinositol 3-kinase in epidermal growth factor receptor signaling. *Molecular And Cellular Biology,* 20**,** 1448-1459.

Roman, J. & S. R. Palumbi (2003) Whales before whaling in the North Atlantic. *Science,* 301**,** 508-510.

Ronquist, F. & J. P. Huelsenbeck (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574.

Rosato, R., J. M. Veltmaat, J. Groffen & N. Heisterkamp (1998) Involvement of the tyrosine kinase fer in cell adhesion. *Molecular And Cellular Biology,* 18**,** 5762-5770.

RUEGG, K. C., E. C. ANDERSON, C. Scott Baker, M. VANT, J. A. JACKSON & S. R. PALUMBI (2010) Are Antarctic minke whales unusually abundant because of 20th century whaling? *Molecular Ecology,* 19**,** 281-291.

Runswick, S. K., M. J. O'Hare, L. Jones, C. H. Streuli & D. R. Garrod (2001) Desmosomal adhesion regulates epithelial morphogenesis and cell positioning. *Nature Cell Biology,* 3**,** 823-830.

Sadkowski, T., M. Jank, L. Zwierzchowski, J. Oprządek & T. Motyl (2009) Comparison of skeletal muscle transcriptional profiles in dairy and beef breeds bulls. *Journal Of Applied Genetics,* 50**,** 109-123.

Sahana, G., B. Guldbrandtsen, C. Bendixen & M. Lund (2010) Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. *Animal Genetics,* 41**,** 579-588.

Sanger, F., S. Nicklen & A. R. Coulson (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings Of The National Academy Of Sciences,* 74**,** 5463-5467.

Sanz-Moreno, A., D. Fuhrmann, E. Wolf, B. von Eyss, M. Eilers & H.-P. Elsässer (2014) Miz1 Deficiency in the Mammary Gland Causes a Lactation Defect by Attenuated Stat5 Expression and Phosphorylation. *PloS One,* 9**,** e89187.

Sasaki, S., A. Shionoya, M. Ishida, M. J. Gambello, J. Yingling, A. Wynshaw-Boris & S. Hirotsune (2000) A LIS1< i>/</i> NUDEL/Cytoplasmic Dynein Heavy Chain Complex in the Developing and Adult Nervous System. *Neuron,* 28**,** 681-696.

Schaeffer, L. (2006) Strategy for applying genome-wide selection in dairy cattle. *Journal Of Animal Breeding And Genetics,* 123**,** 218-223.

Senetar, M. A., C. L. Moncman & R. O. McCann (2007) Talin2 is induced during striated muscle differentiation and is targeted to stable adhesion complexes in mature muscle. *Cell Motility And The Cytoskeleton,* 64**,** 157-173.

Seroussi, E., G. Glick, A. Shirak, E. Yakobson, J. Weller, E. Ezra & Y. Zeron (2010) Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. *BMC Genomics,* 11**,** 673.

She, X., Z. Cheng, S. Zöllner, D. M. Church & E. E. Eichler (2008) Mouse segmental duplication and copy number variation. *Nature Genetics,* 40**,** 909-914.

Shen, T., S. X. Xu, X. H. Wang, W. H. Yu, K. Y. Zhou & G. Yang (2012) Adaptive evolution and functional constraint at TLR4 during the secondary aquatic adaptation and diversification of cetaceans. *Bmc Evolutionary Biology,* 12.

Shendure, J. & H. Ji (2008) Next-generation DNA sequencing. *Nature Biotechnology,* 26**,** 1135-1145.

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir & A. Darvasi (2003) Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics,* 12**,** 771-776.

Shin, D.-H., K.-H. Cho, K.-D. Park, H.-J. Lee & H. Kim (2013) Accurate Estimation of Effective Population Size in the Korean Dairy Cattle Based on Linkage Disequilibrium Corrected by Genomic Relationship Matrix. *Asian-Australasian Journal Of Animal Sciences (AJAS),* 26**,** 1672-1679.

Signer-Hasler, H., C. Flury, B. Haase, D. Burger, H. Simianer, T. Leeb & S. Rieder (2012) A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS One,* 7**,** e37282.

Skol, A. D., L. J. Scott, G. R. Abecasis & M. Boehnke (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics,* 38**,** 209-213.

Smedley, D., S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson & A. Kasprzyk (2009) BioMart–biological queries made easy. *BMC Genomics,* 10**,** 22.

Smit, A., R. Hubley & P. Green. 2012. RepeatMasker Open-3.0 (Institute for Systems Biology, Seattle, WA).

Snijders, A. M., N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey & K. Kimura (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics,* 29**,** 263-264.

Spielman, R. S., R. E. McGinnis & W. J. Ewens (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal Of Human Genetics,* 52**,** 506.

Stanke, M. & B. Morgenstern (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research,* 33**,** W465-W467.

Stankiewicz, P. & J. R. Lupski (2010) Structural variation in the human genome and its role in disease. *Annual Review Of Medicine,* 61**,** 437-455.

Stiening, C., J. Hoying, M. Abdallah, A. Hoying, R. Pandey, K. Greer & R. Collier (2008) The effects of endocrine and mechanical stimulation on stage I lactogenesis in bovine mammary epithelial cells. *Journal Of Dairy Science,* 91**,** 1053-1066.

Stiening, C. M., M. Ben Abdallah, J. B. Hoying & R. J. Collier (2006) Effect of lactogenic stimuli and serotonin on secretory activation of bovine mammary epithelial cells (BMEC). *The FASEB Journal,* 20**,** A616.

Stothard, P., J.-W. Choi, U. Basu, J. M. Sumner-Thomson, Y. Meng, X. Liao & S. S. Moore (2011) Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics,* 12**,** 559.

Sudmant, P. H., J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure & E. E. Eichler (2010) Diversity of human copy number variation and multicopy genes. *Science,* 330**,** 641-646.

Sun, H. & N. K. Tonks (1994) The coordinated action of protein tyrosine phosphatases and kinases in cell signaling. *Trends In Biochemical Sciences,* 19**,** 480-485.

Sved, J. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology,* 2**,** 125-141.

Svishcheva, G. R., T. I. Axenovich, N. M. Belonogova, C. M. van Duijn & Y. S. Aulchenko (2012) Rapid variance components-based method for whole-genome association analysis. *Nature Genetics,* 44**,** 1166-1170.

Szustakowski, J. D., J.-H. Lee, C. A. Marrese, P. A. Kosinski, N. Nirmala & D. M. Kemp (2006) Identification of novel pathway regulation during myogenic differentiation. *Genomics,* 87**,** 129-138.

Takada, Y., X. Ye & S. Simon (2007) The integrins. *Genome Biolology,* 8**,** 215.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard & P. M. Visscher (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research,* 17**,** 520-526.

Thevenon, S., G.-K. Dayo, S. Sylla, I. Sidibe, D. Berthier, H. Legros, D. Boichard, A. Eggen & M. Gautier (2007) The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. *Animal Genetics,* 38**,** 277-286.

Thewissen, J., L. N. Cooper, J. C. George & S. Bajpai (2009) From land to water: the origin of whales, dolphins, and porpoises. *Evolution: Education And Outreach,* 2**,** 272-288.

Thomas, D., R. Xie & M. Gebregziabher (2004) Two-Stage sampling designs for gene association studies. *Genetic Epidemiology,* 27**,** 401-414.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society. Series B (Methodological)***,** 267-288.

Tomizawa, K., N. Iga, Y.-F. Lu, A. Moriwaki, M. Matsushita, S.-T. Li, O. Miyamoto, T. Itano & H. Matsui (2003) Oxytocin improves long-lasting spatial memory during motherhood through MAP kinase cascade. *Nature Neuroscience,* 6**,** 384-390.

Toosi, A., R. Fernando & J. Dekkers (2010) Genomic selection in admixed and crossbred populations. *Journal Of Animal Science,* 88**,** 32-46.

Tozaki, T., E. Hill, K. Hirota, H. Kakoi, H. Gawahara, T. Miyake, S. Sugita, T. Hasegawa, N. Ishida & Y. Nakano (2011) A cohort study of racing performance in Japanese Thoroughbred racehorses using genome information on ECA18. *Animal Genetics,* 43**,** 42-52.

Tozaki, T., T. Miyake, H. Kakoi, H. Gawahara, S. Sugita, T. Hasegawa, N. Ishida, K. Hirota & Y. Nakano (2010) A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. *Animal Genetics,* 41**,** 28-35.

Uhen, M. D. (2010) The Origin(s) of Whales. *Annual Review of Earth and Planetary Sciences, Vol 38,* 38**,** 189-219.

Våge, D. I., M. Husdal, M. P. Kent, G. Klemetsdal & I. A. Boman (2013) A missense mutation in growth differentiation factor 9 (GDF9) is strongly associated with litter size in sheep. *BMC Genetics,* 14**,** 1.

Veit, G., B. Kobbe, D. R. Keene, M. Paulsson, M. Koch & R. Wagener (2006) Collagen XXVIII, a novel von Willebrand factor A domain-containing protein with many imperfections in the collagenous domain. *Journal Of Biological Chemistry,* 281**,** 3494-3504.

Velleman, S. G., J. Shin, X. Li & Y. Song (2012) Review: The skeletal muscle extracellular matrix: Possible roles in the regulation of muscle development and growth. *Canadian Journal Of Animal Science,* 92**,** 1-10.

Vidal, F., C. Baudoin, C. Miquel, M.-F. Galliano, A. M. Christiano, J. Uitto, J.-P. Ortonne & G. Meneguzzi (1995) Cloning of the laminin α3 chain gene (LAMA3) and identification of a homozygous deletion in a patient with Herlitz junctional epidermolysis bullosa. *Genomics,* 30**,** 273-280.

Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell & J. J. Grefenstette (2009) High-resolution haplotype block structure in the cattle genome. *BMC Genetics,* 10**,** 19.

Watanabe, N., Y. Satoh, T. Fujita, T. Ohta, H. Kose, Y. Muramatsu, T. Yamamoto & T. Yamada (2011) Distribution of allele frequencies at TTN g. 231054C> T, RPL27A g. 3109537C> T and AKIRIN2 c.* 188G> A between Japanese Black and four other cattle breeds with differing historical selection for marbling. *BMC Research Notes,* 4**,** 10.

Wellings, S. & J. Philp (1964) The function of the Golgi apparatus in lactating cells of the BALB/cCrgl Mouse. *Zeitschrift für Zellforschung und Mikroskopische Anatomie,* 61**,** 871-882.

Winchester, L., C. Yau & J. Ragoussis (2009) Comparing CNV detection methods for SNP arrays. *Briefings In Functional Genomics & Proteomics,* 8**,** 353-366.

Winters, S. J. & J. P. Moore (2011) PACAP, an autocrine/paracrine regulator of gonadotrophs. *Biology Of Reproduction,* 84**,** 844-850.

Winther, M., V. Berezin & P. S. Walmod (2012) NCAM2/OCAM/RNCAM: cell adhesion molecule with a role in neuronal compartmentalization. *The International Journal Of Biochemistry & Cell Biology,* 44**,** 441-446.

Wittkowski, K. M., V. Sonakya, B. Bigio, M. K. Tonn, F. Shic, M. Ascano, C. Nasca & G. Gold-Von Simson (2014) A novel computational biostatistics approach implies impaired dephosphorylation of growth factor receptors as associated with severity of autism. *Translational Psychiatry,* 4**,** e354.

Wright, S. (1949) The genetical structure of populations. *Annals Of Eugenics,* 15**,** 323-354.

Xu, X.-Z. S., P. D. Wes, H. Chen, H.-S. Li, M. Yu, S. Morgan, Y. Liu & C. Montell (1998) Retinal targets for calmodulin include proteins implicated in synaptic transmission. *Journal Of Biological Chemistry,* 273**,** 31297-31307.

Yagi, T. & M. Takeichi (2000) Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes & Development,* 14**,** 1169-1180.

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane, X. Gan, C. Nellåker, L. Goodstadt, J. Nicod & A. Bhomra (2011) Sequence-based characterization of structural variation in the mouse genome. *Nature,* 477**,** 326-329.

Yamada, T., S. Sasaki, S. Sukegawa, S. Yoshioka, Y. Takahagi, M. Morita, H. Murakami, F. Morimatsu, T. Fujita & T. Miyake (2009) Association of a single nucleotide polymorphism in titin gene with marbling in Japanese Black beef cattle. *BMC Research Notes,* 2**,** 78.

Yang, J., S. H. Lee, M. E. Goddard & P. M. Visscher (2011a) GCTA: a tool for genome-wide complex trait analysis. *The American Journal Of Human Genetics,* 88**,** 76-82.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold & M. G. Hayes (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics,* 43**,** 519-525.

Yang, Y., E. K. Chung, Y. L. Wu, S. L. Savelli, H. N. Nagaraja, B. Zhou, M. Hebert, K. N. Jones, Y. Shu & K. Kitzmiller (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *The American Journal Of Human Genetics,* 80**,** 1037-1054.

Yim, H.-S., Y. S. Cho, X. Guang, S. G. Kang, J.-Y. Jeong, S.-S. Cha, H.-M. Oh, J.-H. Lee, E. C. Yang & K. K. Kwon (2014) Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics,* 46**,** 88-92.

Yin, H.-Q., M. Kim, J.-H. Kim, G. Kong, K.-S. Kang, H.-L. Kim, B.-I. Yoon, M.-O. Lee & B.-H. Lee (2007) Differential gene expression and lipid metabolism in fatty liver induced by acute ethanol treatment in mice. *Toxicology And Applied Pharmacology,* 223**,** 225-233.

Yoneyama, M., K. Kawada, T. Shiba & K. Ogita (2011) Endogenous nitric oxide generation linked to ryanodine receptors activates cyclic GMP/protein kinase G pathway for cell proliferation of neural stem/progenitor cells derived from embryonic hippocampus. *Journal Of Pharmacological Sciences,* 115**,** 182-195.

Youn, H.-D., C. M. Grozinger & J. O. Liu (2000) Calcium regulates transcriptional repression of myocyte enhancer factor 2 by histone deacetylase 4. *Journal Of Biological Chemistry,* 275**,** 22563-22567.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen & J. B. Holland (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics,* 38**,** 203-208.

Yucesoy, B., L. E. Charles, B. Baker & C. M. Burchfiel (2013) Occupational and genetic risk factors for osteoarthritis: a review. *Work: A Journal of Prevention, Assessment And Rehabilitation*.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren & G. Andersen (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics,* 40**,** 638-645.

Zhang, F., W. Gu, M. E. Hurles & J. R. Lupski (2009) Copy number variation in human health, disease, and evolution. *Annual Review Of Genomics And Human Genetics,* 10**,** 451-481.

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

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell & T. S. Sonstegard (2009) A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology,* 10**,** R42.

Zou, H. & T. Hastie (2005) Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology),* 67**,** 301-320.

국문초록

차세대 염기서열 및 단일염기다형성 데이터를 이용한 포유류의 유전체 변이와 유효집단크기 해독

신동현

농생명공학부 동물생명공학전공

서울대학교 대학원 농업생명과학대학

이 학위논문은 차세대 염기서열 및 단일염기다형성 데이터를 이용하여 포유류의 유전체 변이와 유효집단 크기를 추정하는 연구들로 구성되어 있다. 유효집단 크기 추정은 동물 집단의 다양성을 측정할 수 있을 뿐만 아니라 유전체 연구를 수행할 때에도 데이터의 크기와 질을 파악하는데 이용할 수 있는 수치이다. 유효집단 크기 추정 이후, 유전체 변이에 관한 연구를 수행하게 되는데 동물의 표현형과 유전체 데이터와의 관계를 파악하여 동물의 경제형질과 관련된 변이를 추정하였고, 차세대 염기서열 데이터를 이용하여 집단 수준에서 유전체 구조 변이를 추정하고, 진화분석 도구를 이용하여 가축화 과정에 영향을 미치는 유전체 구조 변이를 추정하는 연구를 수행하였다.

제 1 장에서는 동물 유효집단크기와 유전체 변이 추정에 관한 배경지식을 요약하였다. 먼저 집단의 유전적 다양성을 측정하기 위한 유효집단크기의 정의가 무엇이고, 현재 사용되고 있는 각 연구에서 유효집단크기 추정 기법들이 어떻게 유래되었는지에 대하여 집단

유전학적 관점에서 기술하였다. 또한 유전체 변이와 표현형과의 관계를 설명하는데 가장 널리 사용되는 유전체 연관 분석에 대하여 정의와 일반적인 방법 그리고 가장 최신 기법들에 관하여 서술하였다. 또한 유전체 구조 변이에 대하여 설명을 하고 이 유전체 변이의 생물학적 유래와 유전체 데이터 상에서 어떻게 추정되고 동물의 유전체를 이해하기 위하여 어떻게 연구를 하고 있는지에 대하여 서술하였다.

최근에는 유전체 데이터를 이용하여 동물의 경제형질과의 연관성을 파악하고 다양한 목적으로 가축의 유전체 데이터를 이용하려는 연구와 산업적 시도가 이루어져 있는데 이를 위해서는 집단의 유전적 다양성을 파악하는 것이 중요하다. 2 장에서는 한국의 젖소 혈통자료를 참고하여, 최대한 혈통이 겹치지 않게 샘플링한 다음, 단일염기다형성 데이터 분석을 한 후에, 이를 바탕으로 한국의 젖소의 유전적 다양성을 추정하였다. 3 장에서는 밍크 고래의 차세대 염기서열 데이터를 이용하여 최근 현대화를 거치면서 일어난 밍크 고래 집단의 급격한 감소에 대한 시기와 비율 추정하였다. 이는 보존 생물학적 관점에서 밍크 고래의 집단의 다양성이 언제, 얼마나 줄어들었는지를 알 수 있다는 점에서 의미가 있다.

유전체 데이터를 이용하여 표현형을 확보한 샘플에 대해서는 두 데이터 간의 관계를 파악하여 경제형질과 밀접한 관련을 가지는 유전체 변이를 추정하는 연구가 시도되고 있다. 4 장에서는 말의 경주 기록에 관한 추정 육종가를 구하고 이를 이용하여 말의 경주 형질에 영향을 미치는 유전체 변이를 찾았다. 5 장에서는 최신 방법을 이용하여 육종가를 구하고 유전체 연관분석을 진행하였다. 육종가를 구하는 최신 방법을 이용하면 직접적으로 개별 육종가가 나오는 대신 각 유전체 변이

효과를 추정할 수 있고 이를 바탕으로 개별 육종가를 추정한다. 이와 같은 방법으로 젖소의 우유 생성에 관한 유전체 변이를 찾았다.

많은 유전체 변이 연구가 단일염기다형성 변이에 집중이 되었지만 아직도 많은 부분이 설명되고 있지 않다. 이러한 점에서 복제수 변이와 같은 유전체 구조 변이는 매우 중요한 분야이다. 6 장에서는 한우와 젖소의 차세대 염기서열 데이터를 가지고 집단 수준에서 유전체 구조 변이를 추정하고 가축화 과정에 영향을 끼치는 구조 변이와 각 종특이성에 영향을 끼치는 유전체 구조 변이를 추정하였다.

주요어: 유효집단크기, 육종가, 유전체 연관분석, 유전체 구조 변이

학번: 2009-21260