농 학 박 사 학 위 논 문

Detection of Adaptive Signatures in the Livestock Genomes and Estimate of Connectedness

가축 유전체 내 적응적 진화 흔적 발굴과 혈연연결도의 추정

2021년 2월

서울대학교 대학원

농생명공학부

이 원 석

Detection of Adaptive Signatures in the Livestock Genomes and Estimate of Connectedness

By

Won Seok Lee

Supervisor: Professor Heebal Kim, Ph.D

February 2021

Department of Agricultural Biotechnology

Seoul National University

가축 유전체 내 적응적 진화흔적 발굴과 혈연연결도의 추정

지도교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함

2020 년 11 월

서울대학교 대학원

농생명공학부

이 원 석

이원석의 농학박사 학위논문을 인준함

2021 년 2 월

위 원 장 한 재 용 (인)

부위원장 김 희 발 (인)

위 원 조 서 애 (인)

위 원 유 재 웅 (인)

위 원 곽 우 리 (인)

Abstract

Detection of Adaptive Signatures in the Livestock Genomes and Estimate of Connectedness

Won Seok Lee
Department of Agricultural Biotechnology
The Graduate School
Seoul National University

For thousands of years, mutations, natural selection and artificial selection, genetic drift, inbreeding and mating have contributed to the diversification of livestock genetic backgrounds. Recent developments in bioinformatics have provided evolution history and information of livestock genetic resources. Genetic markers and molecular studies are being used to characterize livestock diversity and to reconstruct events that have now formed diversity patterns. These includes ancestry and historical movements, admixture, and genetic structures. Exploring that past information is important for understanding trends and better characterizing the current state of animal genetic resources. In 2009, the cattle became one of the first livestock species with a fully mapped genome. The results of NGS technology can be analyzed using bioinformatics and statistics. There are many techniques for determining gene profiles, including full genome sequencing. Molecular genetic studies, particularly genome-wide linkage studies and whole genome sequencing, can link adaptive traits to genomic regions, genes, or even mutations. Since certain regions of DNA contain genes that influence observable properties, they have a statistically detectable association with microevolutionary properties. Different levels of genetic diversity information can be obtained from different types of genetic markers. For example, autosomal polymorphism is used for estimation of population diversity, genetic relationships, and population genetic mixture, while mitochondrial DNA polymorphism is used to reconstruct geographic regions of domestication, reconstructing migration pathways and the number of female founders.

This doctoral dissertation, composed of five chapters, is mostly dedicated to uncovering the signatures of different natural and artificial selection forces left on the genomes of various livestock breeds selected for several economic and adaptation traits. The first chapter, Chapter 1, is the introductory chapter which highlights about genetic variations in livestock genetic resources with special emphasis on Korean native breeds and the principles behind signature of positive selection. In this chapter, the objectives, methods of detecting positive selection signatures, and reviews on previous studies of positive selection studies from genetically diverse livestock breeds of the world are presented. It is also presented the concept of connectedness and related previous researches introduced.

Chapter 2 presents the genomic signature of different natural and artificial selection on Korean native goat breeds divergently selected for various economic and adaptation traits. Together with domestication, natural and artificial selection forces have significantly modified the goat genome which resulted in morphological, production, and adaptation characteristics. Identifying genomic regions affected due to these forces in goat would give an insight into the history of selection for economically important traits and genetic adaptation to specific environments of populations under consideration. Here, I explored the genomes of Korean native goat and crossbred goat in order to decipher genomic regions affected due to selection for disease resistance and environmental adaptation traits, respectively.

The third chapter, Chapter 3, is based on the identification of signature of natural and artificial selection in the genome of Korean imported pig breeds in relation to their superior fecundity ability. To reveal the genomic regions affected these adaptation mechanisms, I compared the genomes of Korean imported pig breeds with Korean native pig breeds using cross-population statistical (Fst and heterozygosity) methods. As a result, several genes were identified under selection that are overrepresented related to reproduction function, immunity, coat color, and other traits. Several genes (e.g., *PLSCR4, AGTR1* and *CORIN*) were related to reproduction traits such as fertility, ovulation rate, and uterine function. Therefore, the genes and biological processes identified here directly and/or indirectly contribute to the superior fecundity mechanisms of Korean imported breeds.

Chapter 4 presents the genomic signature of different natural and artificial selection on thoroughbred horse breeds comparing to Korean native Jeju horses selected for various positively selected candidate genes. Especially, Thoroughbreds are known for an outstanding racing performance. I identified 98 and 200 genes that are under positive selection using XP-EHH and XP-CLR methods. Further, I performed and found 72 BR terms. These genes and BP terms are related to the ocular size, energy metabolism, immunity and function that are related to running performances.

Chapter 5 present the connectedness rating (CR) among swine herds in Korea. Using 104,380 performance and 83,200 reproduction records of three breeds (Yorkshire, Landrace and Duroc), connectedness rating (CR) was estimated for two traits: Average Daily Gain (ADG) and Number of Born Alive (NBA) of eight breeding herds in Korea. I calculated the average CR for ADG of the Yorkshire ranges from 1.32% to 28.5%. The average CR for NBA of the Yorkshire herd ranges from 0% to12.79%. This study suggested that four out of eight herds are possible to evaluate genetic values together for ADG and NBA of the Yorkshire herds since the preconditions were satisfied for the four herds. It is also possible to perform joint genetic analysis of the ADG records of all Duroc herds since the preconditions were also satisfied.

In conclusion, from these studies, a list of candidate genes were detected under positive selection from different livestock breeds that are selected for various economic and adaptation traits. These findings will increase our understanding of the adaptive events that have generated the enormous phenotypic variation observed between livestock breeds prevailing today. Molecular markers that contribute to local environmental adaptation were revealed in addition to those affecting production traits such as disease resistance, reproduction, and other associated traits. The markers identified in these studies can be used in genomic selection and breeding programs to fit different production systems. Meanwhile, to perform joint analysis for breeding value in Korean swine herds, it is essential to certain degree of connectedness. There is no study using three different imported breeds before. To develop our own swine seed, joint breeding value system is essential for standard selection procedure. Therefore, this connectedness estimation gives an essential statistical background and will help make our own swine breed. There are some limitations in these studies because of sample size and validation of genetic expression level and so on. Despite these limitations, I tried to verify as much as possible using the various statistics methods and agreement with

previous research results. Therefore, these results are provided the insight to the future livestock studies and practical advices to the related livestock industries.

Keywords: Goat, Pig, Horse, Positive selection, selective sweep, Connectedness

Student number:2014-22935

Table of Contents

List of Tables

List of Figures

Abbreviations

ADG	Average Daily Gain
BV	Breeding Value
CR	Connectedness Rating
DAVID	Database for Annotation, Visualization, and Integrated Discovery
FDR	False Discovery Rate
Fst	Fixation index based on Wright's F-statistics
GO-BP	Gene Ontology Biological Processes
Hp	Heterozygosity
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
NBA	Number of Born Alive
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PEVD	Prediction Error Variance of Difference
QTL	Quantitative Chain Reaction
SNP	Single Nucleotide Polymorphism
VED	Variance of Estimated Differences
XP-CLR	Cross-Population Composite Likelihood Ratio
XP-EHH	Cross-Population Extended Haplotype Homozygosity

General Introduction

Domestication of livestock species is one of the main causes of human evolution. Controlling food production led to a major population bulge and affected every aspect of human life. During the several thousand years, factors like mutations, natural and artificial selection, genetic drift, inbreeding and crossbreeding have contributed to the diversification of livestock genetic resources. Using systematic pedigree and performance recording and applying specific breeding objectives made genetic changes. Due to this process, the fixation of breed-specific traits and a remarkable increase in productivity. Some breeds were isolated in populations, while many breeds continued to interact with intentional cross-breeding or unintended introgression. As a result of these developments, a limited number of transboundary commercial breeds, such as the Holstein cow and Large White pig, have become very widespread and increasingly dominate livestock production globally. The wide number of livestock breeds and the genetic diversities within them mean that animal genetic resources have a substantial value to society. The different breeds provide substantial animal products and services for the benefit of human. So greater livestock diversity allows humans to be better prepared to meet future challenges. Having access to a range of diverse livestock traits may allow for greater ability to cope with environment and emerging diseases. Within breeds, the greater genetic diversity allows for continued selection for improving a given trait, such as disease resistance or tolerance to diseases and ability to superior reproduction and running performance. Especially, local breeds that were developed by a given community often have a huge cultural significance for that community. However, according to the report of the food and agriculture organization of the United Nations (FAO), 17% of the world's farm animal breeds are at risk of extinction and 58% are of unknown risk status. This means that it can be underestimated. The world's pool of animal genetic resources is also currently shrinking, with rapid and uncontrolled loss of breeds and their genes. Almost 100 livestock breeds have gone extinct between 2000 and 2014. With the loss of these breeds comes the loss of their unique adaptive traits, which are often under the control of many different genes and complex interactions between the genotype and the environment. To protect these unique traits and the diversity, especially indigenous species, collaborative global and local efforts towards the characterization and management of must be made. As a part of that effort, genomic research in livestock

has begun using bioinformatics tool and big data. Because of these research, genes that behind selection revealed their biological function. With the accumulation of these results, we can utilize and preserve the livestock more effectively.

Meanwhile, to estimate the accurate breeding value is the essential to the livestock industries. However, according to the previous research, if there is a no genetic connection between herds brings to estimate the incorrect breeding value. That is, integrated breeding value estimation needs a certain degree of genetic connection among herds. That standard is called “connectedness” and to estimate this statistic is the first step of integrative breeding for standard selection.

Chapter 1. Literature Review

1.1 Positive selection signature

1.1.1 Definition of positive selection

Selection can act in a directional manner. For example, Positive selection is the phenomenon that allele is favored and so immediately propagated. If disfavored, it is called negative selection or also called purifying selection. It is reported that random mutations are normally deleterious, therefore many novel mutations are removed (Vitti, Grossman et al. 2013).

If a specific population receives new selective pressure due to some environmental change, it will adapt to the environment through standing variation or novel mutation. Here, the standing variation means that one or more alleles are present in one locus in a population. That is, alleles that are already in the population and that are beneficial to the individual can be used immediately. This means that the rate of adaptation and the probability of fixation may be higher than the emergence of new mutations. It also means that it may have been passed in adaptation of the previous environment. When new mutation or standing variation emerges for the environment, the number increases within the population by the positive selection and this mutation increases its frequency compared to other alleles. With this increase, the neutral and nearly neutral genetic variation associated with this mutation also spreads into the population, a phenomenon known as genetic hitchhiking. Therefore, the diversity of nucleotide sequence around the gene is reduced, which is called selective sweep. The genetic diversity that is reduced by this selective sweep can be used to identify candidate selective sweep genes. This mark persists till novel mutation and recombination restore diversity to the specific population(Vitti, Grossman et al. 2013)

1.1.2 Methods to detect signature of positive selection in livestock genomes

The basic method of finding the selection signature is the neutrality test. The null hypothesis that the mutation is neutral, and the null hypothesis rejected are methods to use. Specifically, the currently used statistics largely classified the selected gene as a result of microevolution in the population in four different ways.

First, there is a method like Tajima's D that uses the difference in the pattern of the frequency spectrum (Tajima 1989). The purpose of this measure, determining whether the DNA sequence evolution is a random process. If the Tajima's D value is negative, it is regarded as a selective sweep. The other way, a positive value means balancing selection.

Second, there is a method using a linkage disequilibrium. A selective sweep causes extended haplotype homozygosity. It is a measure of linkage disequilibrium. It contains the selected allele and rises its frequency. The high peak of extended haplotype homozygosity then begins to break down to restore the diversity to the population(Vitti, Grossman et al. 2013).

These include the extended haplotype (EHH), integrated long-range haplotype test (LRH), integrated EHH (iHH), and integrated haplotype score (iHS). Recently, EHH is integrated to calculate haplotype decay, and cross-population extended haplotype homozygosity (XP-EHH), which is used for comparison among groups, is also widely used(Vitti, Grossman et al. 2013)..

Third, there is a method of using population differentiation, which includes statistics such as Fst. This Fst means a genetic change of subgroups. Fst is close to 1 and heterozygosity is close to 0 when heterogeneity of heterozygosity is lower than that of whole population(Vitti, Grossman et al. 2013).

Heterozygosity is also another good indicator for detecting selective sweep because it could be measured in sliding windows from DNA sequences from a pool of haplotypes.(Rubin, Zody et al. 2010)

Next, the XP-CLR finds a selective sweep using the allele frequency differentiation between groups as a statistic based on the Brownian motion model and the deterministic model(Chen, Patterson et al. 2010). It is a likelihood method to identify selective sweep regions using the multiclocus allele frequency differentiation between populations. It is reported that this method is much more robust to ascertainment bias in SNP sampling

According to the previous literature, every method has its own strengths and weakness (Utsunomiya et al 2015). Genomic regions identified in one way may not be identified in any other way using the same data. Because each method has different target and time scale for detecting candidate genes (Qanbari and Simianer 2014). The use of

combinations is an alternative approach to detect selection signals proposed as a means of increasing the reliability of these studies (Gouveia et al. 2014).

Table 1.1 Statistical method used for detection of positive selection signature in this thesis

Methods	**Fst**	**Heterozygosity**	**XP-EHH**	**XP-CLR**
Chapters used	3 Chapter	3 Chapter	2,4 Chapters	2,4 Chapters
Characteristics	▪ This compares allele frequencies within and between two populations ▪ Comparatively, large values of Fst at a locus indicate stark difference between population	▪ The way to determine the degree of genetic variation at a specific locus ▪ The patterns of reduced heterozygosity is used to identify positive selection signatures.	▪ Cross-population extended haplotype homozygosity ▪ This method compares haplotype lengths between populations to control for local variation in recombination rates ▪ It detects recent, fixed or nearly fixed sweep regions	▪ Cross-population composite likelihood ratio ▪ This method identifies genetic regions based on multilocus allele frequency differentiation between two populations ▪ It can detect recent and ongoing sweep regions

1.1.3 Signature of selection in the livestock genome

Since many livestock have been selected for excellent economic traits by traditional breeding, this can be a powerful selection pressure together natural selection. The selection signature for genes involved in economic traits will be very strong and the statistical methods described above can be used to select these domesticated genes. Therefore, many studies have been performed in livestock species (Biswas and Akey 2006). An exploring of the goat genome for selection signals detected regions involved in the adaptation to local conditions (Benjelloun, Alberto et al. 2015). For example, Mengistie et al revealed the positive selection for thermotolerance, beef quality in African cattle(Taye, Kim et al. 2017, Taye, Lee et al. 2017). I studied using several livestock genome(goat,pig,horse) in this thesis. Quite a few previous studies detected using these animal genomes. Goats (*Capra hircus*) were domesticated about 10,000 years ago in western Iran. They were used for many purposes. It is very important to preserve goat genome for their diversity. Especially indigenous breeds are adapted their local environment. So, studies for each indigenous one gives us some insight for behind the adaptation. Benjelloun Alberto et al studied Morroccan goats using whole genome sequencing data(Benjelloun, Alberto et al. 2015). They revealed the genes involved in the adaptation to their local conditions. And Xiaolong et al identified genes related to the production and adaptive traits in Chinese goat populations(Wang, Liu et al. 2016).

Since domestication event, both natural and artificial selection pressures have changed the genomic landscape of the pig. These resulted in hundreds of swine breeds with the dramatic changes in phenotypic traits. Several candidate genes were revealed under strong selection during pig domestication including loci controlling stature, coat color (Rubin et al. 2012). To preserve Korean native Jeju pigs, several attempts has been adapted to the Jeju pigs. Cho et al. identified the KIT gene in Jeju ‘Nanchukmacdon’ breed(Cho, Zhong et al. 2011). The Kit gene is expressed coat color.

In addition, Gu et al found positively selected genes of thoroughbred horses that associated with their athletic-performance genes (Jingjing Gu et al. 2009). These genes are mainly responsible for fatty acid oxidation, increased insulin sensitivity and muscle strength (Jingjing Gu et al. 2009).

In this thesis, the whole genomes of livestock breeds(goat,pig,horse) were scanned for adaptive signature of positive selection. As a result, the comparing the genomes of

Korean native goat with the genomes of crossbred goats discovered the disease resistance trait. Likewise, identifying candidate genes for Korean imported pigs for their superior litter size and Thoroughbred horse's racing performance contribute to understand Korean local livestock's genome characteristics.

1.2 The concept of connectedness

1.2.1 Connectedness definition

Accurate breeding value estimation in the breeding industry is extremely important. Because it can achieve not only economic benefits but also produce their golden animal seeds which is getting bigger impact on the industries. The reliability of comparisons of breeding values among farms depends on the degree of connectedness between herds. Connectedness defines to the genetic similarity between herds, but it contains statistical significance. The stronger genetic connectedness suggests more accurate breeding value estimation. So far, many different methods for calculating connectedness have been developed, among them, Connectedness Rating (CR) suggested that the most suitable and consistent indicator for connectedness (Table 1.2) (Soga 2009). The CR is defined as the correlation between estimates of fixed genetic group effects. The CR is less dependent on genetic structure and group size. It is also advantageous for relatively easy computation. In addition, it is highly correlated with prediction error variance of difference. So, many previous research have been adapted this method.

Especially, the swine breed has been improved through genomic selection in the farm units or by importing superior piglets from developed countries. However, multinational corporations have demanded royalties based on the purpose of raising their breeders. If they continue to improve pigs through imports of breeders, they will not be able to avoid huge economic loss. As an alternative way, pig breeding network business is trying to improve piglets by consolidating small domestic pig farms in Korea. However, if there is no genetic link between the herds, the evaluation of Estimated Breeding Value between different farms is not reliable and genetic evaluation is not possible. It has been reported that the accuracy of the genetic evaluation increases when the CR between the herds is increased. So, it should be estimated for future evaluation in Korean breeding value program.

Table 1.2 Comparison of connectedness statistics.

Statistic type	Computation	Requirement of genetic grouping	Correlation with Bias	Value range	Influence of data size	Influence of structure	Eliminate Bias
PEVD	Very hard	Required	High	Unlimited	Influenced	-	Eliminate all
CR	Easy	Not required	High	0~1	-	No influence	Eliminate all
GF	Very easy	Not required	Low	Unlimited	Influenced	-	Influenced
R	Hard	Not required	High	0~1	No influence	Normal	Eliminate all

PEVD: prediction error variance of difference between animals, CR: connectedness rating, GF: genetic flow, R: connectedness correlation

1.2.2 Connectedness rating between swine herds

For accurate integrated breeding program, there have been studies for CR estimation in the various countries. In Canada, Mathur et al developed and performed CR estimation(Mathur, Sullivan et al. 2002). They found that the degree of CR to most herds in national program is below the threshold.

In China, Sun et al studied CR using three different breeds(Yorkshire, Landrace and Duroc)(Sun, Wang et al. 2009). They used data on age at 100kg. The results showed that average CR was low in most herds.

In USA, Soga estimated the CR between U.S. purebred duroc herds(Soga 2009). The CR level was lower than the criteria. It means there is a risk of bias to perform integrated purebred duroc breeding program.

Likewise, there is an attempt to consolidate breeding program among Korean swine companies. Therefore, the accurate CR estimation in Korea is foundation of genetic progress of future Korean swine.

This chapter was published in *Molecules and Cells*
as a partial fulfillment of Wonseok Lee's Ph.D program

Chapter 2. Detecting positive selection of Korean native goat populations using next-generation sequencing

2.1 Abstract

Goats (*Capra hircus*) are one of the oldest domesticated animals. Among them, Korean native goats are the indigenous goats that have been raised in the Korean peninsula almost 2,000 years ago. Although their small body size, low production of milk and meat, they are known to resist lumbar paralysis. This study was performed to reveal the distinct genetic features and patterns of selection in Korean native goats by comparing the genome between Korean native goat and crossbred goat populations. I sequenced the whole genome of 15 Korean native goats and 11 crossbred goats using next-generation sequencing (Illumina platform) to perform comparative genome studies between populations. So, I found decreased nucleotide diversity in Korean native goats compared to crossbred goats. Genetic structural analysis demonstrated that the Korean native goat and crossbred goat populations shared common ancestry but were clearly distinguished. Finally, to reveal the Korean native goat's selective sweep region, positive selection signals were detected in the Korean native goat genome using cross-population extended haplotype homozygosity (XP-EHH) and cross-population composite likelihood ratio test (XP-CLR). As a result, I was able to identify candidate genes for recent selection, such as *CCR3* gene related to lumbar paralysis resistance. Combined with further studies and recent goat genome information, this study will be a cornerstone of understanding Korean native goat genome.

2.2 Introduction

Goats (*Capra hircus*) are one of the oldest domesticated animals, and their domestication started in western Iran about 10,000 years ago. With 599 breeds (Kim, Park et al. 2014) developed in different environmental conditions and selected both due to natural selection and artificial selection developing different traits, goats are used for meat, milk, wool and skin.

Korean native goats have been raised in the peninsula since 2,000 years ago (Kim, Park et al. 2014). Because Korean native goats are low producers of milk and meat, research opted to cross them with Saanen goats that are believed to produce better. As Saanens are known to be the largest goat breeds among dairy types and good milk producers, they were mostly used to form the crossbred goat group. In addition, the Boer goat, which is developed for meat production, also have been used for crossbreeding. These crossbred black goats in Korea not only have a higher milk yield, but also a higher growth rate than the native goats. In addition, when fully grown, the crossbred goats are bigger than the Korean native goat (Son 1999).

Detecting the genetic variants related to phenotypic traits is one of the important issues in livestock genomic research. Due to their large differences in body size, weight, muscle mass, milk production, and coat color (http://cemendocino.ucanr.edu/), goats are good model animals for genetic studies. However, only a few studies have analyzed the genetic resource of the goat. Moreover, studies on Korean native goat populations using the whole-genome sequence have not yet been reported and there is no selective sweep signature study. Using next-generation sequencing (NGS), it is possible to examine known and unknown SNPs in the genome. Selective sweep signatures contributing to the domestication process could also be identified. Therefore, this study was performed to discover the different genetic features of Korean native goats and elucidate the selective sweep signatures that have contributed to phenotypic appearances using whole-genome next generation sequencing data at the population level.

2.3 Materials and Methods

2.3.1 Sampling and whole-genome sequencing

11 crossbred goats and 15 Korean native goats in Korea were sampled using whole-blood samples. Blood samples from Korean native goats were collected from the Animal Genetic Resources Station, National Institute of Animal Science, Rural Development Administration in Korea. For the crossbred goat, I obtained the samples from a Korean black goat small farm. Blood (10 mL) was drawn from the carotid artery and treated with heparin to prevent clotting. DNA was isolated from whole blood using a G-DEXTMIIb Genome DNA Extraction Kit (iNtRoN Biotechnology, Seoul, Korea) according to the manufacturer's protocol. I randomly sheared 3 μg of genomic DNA using the Covaris System to generate inserts of about 300 bp. The fragments of sheared DNA were end-repaired, A-tailed, adaptor-ligated, and amplified using a TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA). Paired-end sequencing of goat genomes to about tenfold coverage were conducted at NICEM (National Instrumentation Center for Environment Management, Seoul, Korea) using the Illumina HiSeq2000 platform with TruSeq SBS Kit vs-HS (Illumina). All short-read data have been deposited at the Short Read Archive (SRA) under accession SRA160379.

2.3.2 Sequencing information and variant calling process

Almost 220,000,000~230,000,000 paired-end reads were aligned to the reference goat genome from the Goat Genome Database(http://goat.kiz.ac.cn/). Bowtie 2-2.1.0 was used with the default settings (Langmead and Salzberg 2012). And next process, to call the SNPs and INDELs, Open-source packages were used variant calling. Picard tools 1.94 (http://picard.sourceforge.net), SAMTools 0.1.19 (Li, Handsaker et al. 2009), VCFtools 4.0 (Danecek, Auton et al. 2011), and the Genome Analysis Toolkit (GATK) 2.6.4 (McKenna, Hanna et al. 2010) were used. The Read Group was added, and duplicate reads were filtered, and all mate-pair information were confirmed using the module MarkDuplicates, and FixMateInformation of Picard tools. Next, SAMtools was indexed the resulting bam format files and calculated the aligned read length with the flagstat option (Li, Handsaker et al. 2009). The GATK modules, RealignerTargetCreator,

IndelRealigner, UnifiedGenotyper, SelectVariants, and VariantFiltration were used for realignment and variant calling. The VCFtools was used for handling the VCF format files (Danecek, Auton et al. 2011). The GATK UnifiedGenotyper module used Substitution calls (McKenna, Hanna et al. 2010). The variations filtered with a Phred-scaled quality score of <30. The variants were erased based on MQ0 (mapping quality score 0) >3, (MQ0/read depth) >0.1(10%), quality depth <5, and FS (Phred-scaled *p*-value using Fisher's exact test to detect strand bias) >200. After filter process, the SNPs were filtered out again by removing those within 10 bp of INDELs. For the last SNP filtering process, only biallelic SNPs were considered the real SNPs. Haplotype information on each chromosome, BEAGLE was used (Browning and Browning 2007) to infer the haplotype phase and impute missing alleles simultaneously for the entire set of goat populations.

2.3.3 Nucleotide diversity analysis

I analyzed and compared highly variable regions between two populations using the goat reference genome(Dong, Xie et al. 2013). I used VCFtools to count the variants number and nucleotide diversity in each 1-Mb window region (Danecek, Auton et al. 2011).

2.3.4 Variant annotation in highly variable regions

I made the SNPs pool using VCFtools (Danecek, Auton et al. 2011) to detect SNPs that contribute to differences in phenotypes. SNPEff was used to annotate the variant regions (Dong, Xie et al. 2013). Then I removed variants that were not population-specific and identified to breed-specific nonsynonymous genes. To detect genes that have different allele types between populations, protein IDs of the Ensembl Genome Browser (Hubbard, Barker et al. 2002) were used.

2.3.5 Population structure analysis

STRUCTURE software (Evanno, Regnaut et al. 2005) was used to demonstrate genetic proportions of each goat individual from ancestral populations. STRUCTURE software adapted Bayesian algorithms to identify the true number of clusters, K (It is assumed the

number of ancestral populations). Beagle was used to make input files for running STRUCTURE. I used every 50 SNPs in intergenic regions to avoid bias. Then I used 100,000 iterations with 2,000 burn-in iterations in each analysis from K = 2 to K= 4.

2.3.6 Principal components analysis (PCA)

PCA (Jackson 2005) was examined population differentiation between Korean native goats and crossbred goats using genotype data from 15 Korean native goats and 11 crossbred goats. I used GCTAtool (Browning and Browning 2007), which implements PCA in EIGENSTRAT (Price, Patterson et al. 2006), to estimate eigenvectors. VCFtools (Danecek, Auton et al. 2011) and PLINK (Purcell, Neale et al. 2007) were used to prepare input data sets for GCTAtool (Yang, Lee et al. 2011).

2.3.7 Deciphering candidate genes under positive selection

To decipher signatures of selective sweeps, I used the XP-EHH method (Sabeti, Varilly et al. 2007) which calculates cross-population extended haplotype homozygosity (Sabeti, Reich et al. 2002). The calculation for XP-EHH was performed using the software xpehh (Sabeti, Varilly et al. 2007). After XP-EHH results, these numbers changed using log ratios (unstandardized XP-EHH). These were standardized to have mean 0 and variance 1, and these XP-EHH *z*-scores were assigned *p*-values assuming a normal distribution. I assumed the exception of tails that diverged from the null expectation. The top 100 P-value loci (Benjamini and Hochberg 1995) were considered positive selection signals. Next, I performed the cross-population composite likelihood ratio test (XP-CLR) using the XP-CLR software with non-overlapping windows of 50 kb (Chen, Patterson et al. 2010). XP-CLR used two models. One is Brownian motion to model genetic drift under neutrality, and the second one is a deterministic model to approximate the effect of a selective sweep on SNPs in the vicinity (Chen, Patterson et al. 2010). XP-CLR uses allele frequency differentiation between populations. Comparing to the allele frequency spectrum based methods, it is much more robust to ascertainment bias in SNP discovery (Chen, Patterson et al. 2010). I designated windows with an XP-CLR value in the top 1 % of the empirical distribution. Genes located in the regions under significant selection were annotated and as well as XP-EHH analysis, the lowest top 100 loci based on P-

value were regarded as selective sweep candidates. Candidates from both XP-EHH and XP-CLR results were annotated to the closest genes that spanned (partially or completely) the window regions using CHIR_1.0 (Dong, Xie et al. 2013). These are defined as candidate genes. DAVID (Database for Annotation, Visualization and Integrated Discovery) was used for gene ontology and pathway analyses (Dennis Jr, Sherman et al. 2003). Finally, to apprehend CCR3 sequence structure, gene association study has been performed using PLINK (Purcell, Neale et al. 2007).

2.4 Results

2.4.1 Korean native and crossbred goats' resequencing results

To obtain goat resequencing data, I generated NGS pair-end reads to about ten-fold coverage for 15 Korean native goats and 11 crossbred goats using Illumina HighSeq2000. Each individual goat, over 92% of all reads were successfully aligned to the reference goat genome (domestic goat, *Capra hircus*, 2n = 60, predicted size = 2.66 Gb; (Dong, Xie et al. 2013)). It was excluded the possible polymerase chain reaction duplicates. The mapped reads covered an average of 94.77% of the reference genome, and BEDTools was calculated read coverage of the reference genome (Quinlan and Hall 2010).

As a result, total 22,759,033 Single Nucleotide Polymorphisms (SNPs) and 2,450,921 INDELs were filtered. 26.6% of these SNPs and 26.8% of the INDELs were in genic regions, while 73.4% of the SNPs and 73.2% of the INDELs were in intergenic regions. Within the genic regions, 845,203 SNPs and 108,653 INDELs were detected in only the Korean native goat population and 1,661,071 SNPs and 108,653 INDELs were detected in only the crossbred goat population. I calculated the average nucleotide change rate, it was 1 per 110 base pairs. The SNPs distribution in the Korean native goat genome revealed that the vicinity of anode chromosome parts was highly variable. The X chromosome showed less variable than autosomes.

2.4.2 Nucleotide diversity of the Korean native goat population

I analyzed the nucleotide diversity of Korean native goats and crossbred goats using VCFtools (Danecek, Auton et al. 2011). The number of SNPs were also counted and integrated for each 1-Mb bin region of the genome using VCF format files, which contained variant information on 26 goats based on the reference goat genome (Dong, Xie et al. 2013). The overall distributions of Korean native goats and crossbred goat's nucleotide diversity were shown differently. Korean native goats generally showed lower nucleotide diversity (total average nucleotide diversity=0.0007) than crossbreed goats (total average nucleotide diversity=0.0010) and SNP density of Korean native goat (total average SNP density=6662.83) is also lower than crossbred goats SNP density (total average SNP density =7701.61) in the same genomic regions. (Figure 2.1).

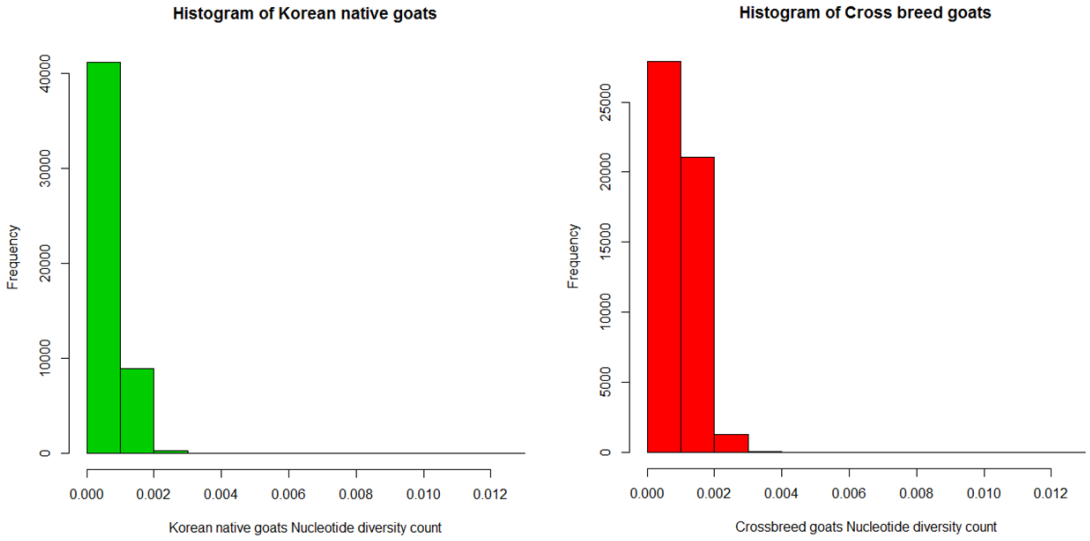


Figure 2.1. Nucleotide diversity plots of Korean native goats (Green) and crossbred goats (Red).

The numbers under horizontal axis of plot represent the nucleotide diversity of genomic region. Vertical axis indicates values of frequency. Each breed is colored differently.

2.4.3 Population structure of Korean native and crossbred goats

The result of principal components analysis (PCA) (Jackson 2005) showed that the Korean native goats were clearly distinguished from crossbred goats (Figure 2.2a). Then I analyzed the genetic structures of Korean native and crossbred goat populations admixture analysis using STRUCTURE software (Evanno, Regnaut et al. 2005). I identified that Korean native and crossbred goats shared part of ancestral proportion with crossbred and the crossbred were formed with breeding of Korean native goat and other different breeds (Figure 2.2b).

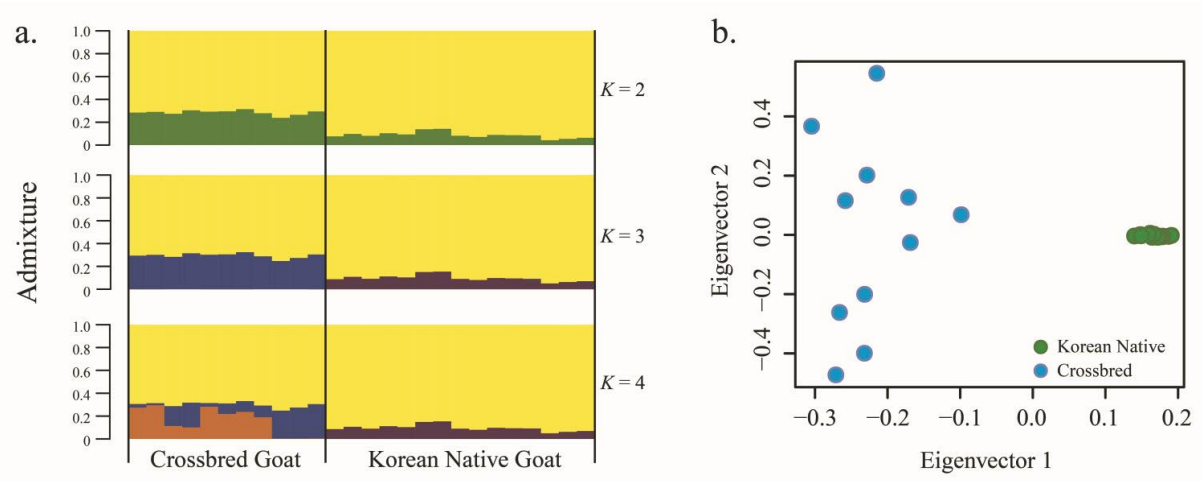


Figure 2.2 The admixture of goat populations using STRUCTURE and Principal Component Analysis. a) Each segment indicates the proportion from ancestral populations. The different individuals colored segments assume that part of the genome originated from different ancestral populations. This figure represents the genetic structure of goat populations when I assume that the number of ancestral population of goats is 2 to 4 breeds. b) The green circles are displayed as Korean native goats, and light blue circles are crossbred goats. The horizontal axis indicates eigenvector 1, and the vertical axis indicates eigenvector 2. The values of eigenvectors were estimated using GCTA tool.

2.4.4 Highly variable Korean native goat genomic regions

I selected the top 5% of the highly variable regions among all chromosomal regions in each population as significant. I used the reference from the Goat Genome Database (Dong, Xie et al. 2013) and Ensembl Genome database (Hubbard, Barker et al. 2002).I annotated to identify gene locations and annotation information. Then I process the gene ontology (GO)-term analysis of gene sets from highly variable regions. I used the DAVID analysis tool (Dennis Jr, Sherman et al. 2003) for GO-term analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Huang, Sherman et al. 2009). With the result that, significantly enriched genes were founded in sensory perception terms in Biological Process GO-term and Olfactory transduction in KEGG pathway.

2.4.5 Nonsynonymous SNP study in Korean native goat

Among 16,570,906 SNP sites, I identified significant 76 nonsynonymous SNP sites of Korean native goats using generated re-sequenced data and SNPEff Software. Here, I reported genes (Chymosin (*CYM*) and collagen, type Xi, and alpha 2 (*COL11A2*)) that may contribute to growth and body size differences between the two populations.

2.4.6 Putative selective sweep signatures in the Korean native population

I estimated the values of cross-population extended haplotype homozygosity (XP-EHH; (Sabeti, Varilly et al. 2007)) to detect selective sweep regions and performed a pairwise test of the Korean native goat and crossbred populations (Figure 2.3). The genome was split into non-overlapping segments of 50 kb and I computed the maximum XP-EHH score in each segment. To define the empirical P-value, the genomic windows were binned in increments of 500 SNPs according to the method used elsewhere (Lee, Kim et al. 2014). The regions with P-values less than 0.01 (1%) were considered strong signals of positive selection in the native goat population. Based on the P-value of XP-EHH test, I identified 82 loci under positive selection. Significant loci were annotated to the closest genes and I found 64 genes.

In addition, I used a cross-population composite likelihood ratio (XP-CLR; (Chen, Patterson et al. 2010)) test to detect selective sweep regions between the two populations

(Figure 2.3). Using the top 100 XP-CLR score regions, 161 significant genes were identified in Korean native population. The common genes from both XP-EHH and XP-CLR statistics are total 16 genes. (Table 2.1) Based on the 161 genes, then I performed GO-term analysis using DAVID analysis tool (Dennis Jr, Sherman et al. 2003). The GO-term cell adhesion (FDR < 0.05) was among the most enriched functional categories that might be related to lumbar paralysis. In addition, the term neuron development was enriched in the GO-term (FDR< 0.05). Among the genes identified in Korean native goats, I choose CCR3, involved in lumbar paralysis, and calculated the frequencies of SNPs between the native and crossbreds using PLINK program. As a result, I found 11 loci that are related to CCR3 sequence differences (Table 2.2).

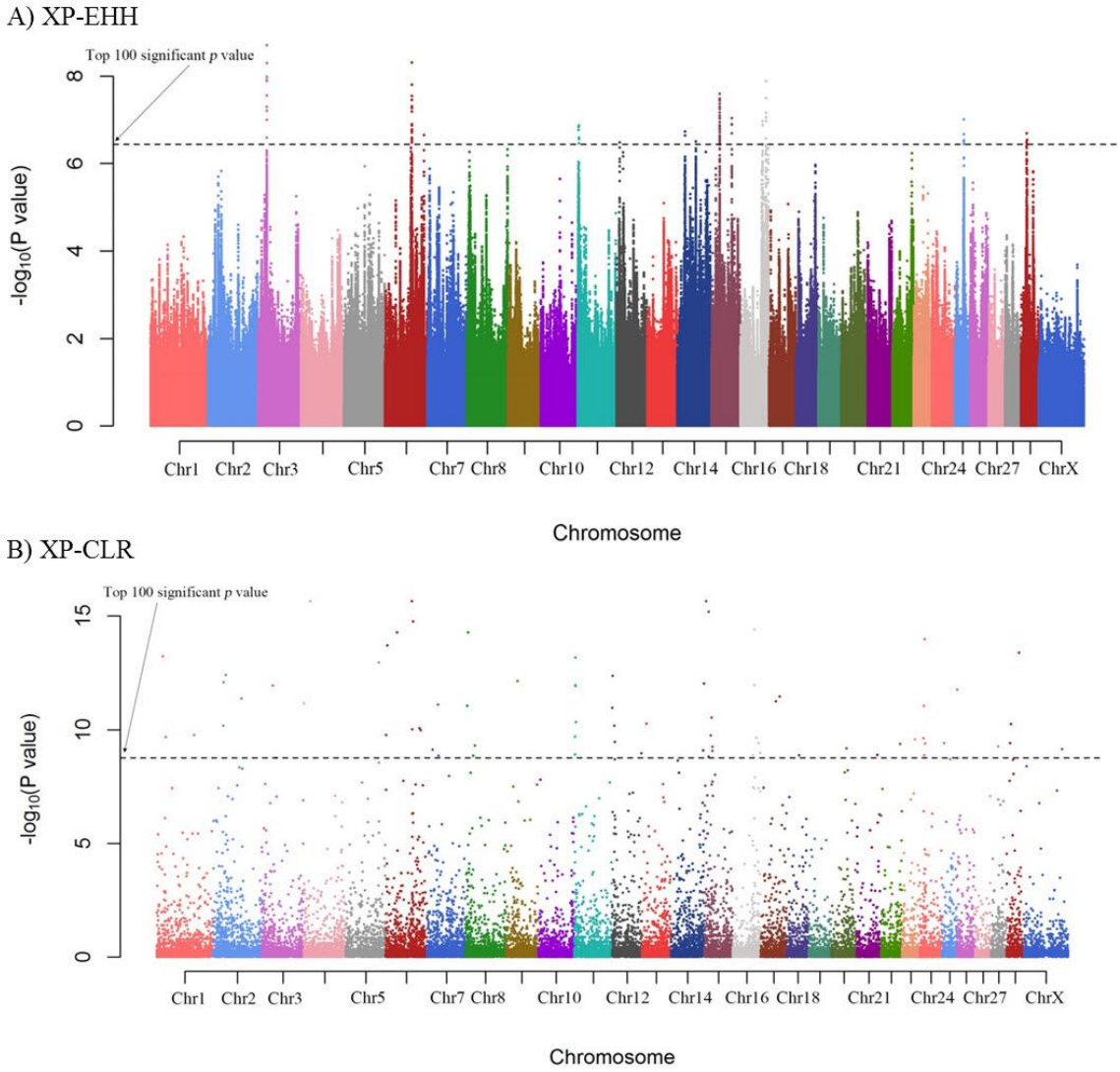


Figure 2.3 Manhattan plot shows the a) XP-EHH and b) XP-CLR between Korean native goat and crossbred goat populations. The vertical axis indicates $-\log_{10}$ (*p*-value) of XP-EHH and XP-CLR values

Table 2.1 Common genes from XP-CLR and XP-EHH analysis

Candidate genes	Chromosome	Start	End	Max XP-CLR	XP-CLR *P*-value	Max XP-EHH	XP-EHH *P*-value	Description
GSG1L	25	25387565	25553146	15.678434	0.003	3.60004	0.0079	Germ cell-specific gene 1-like protein
MAML2	15	12228464	12634778	23.027664	0.01	4.29108	0.0007	Mastermind-Like Transcriptional Coactivator 2
GRIK4	15	29483965	29884993	21.99459	0.005	3.5161	0.0099	Glutamate Receptor, Ionotropic, Kainate 4
GALNTL6	8	3811525	5285470	25.717514	0.01	3.85503	0.0067	Polypeptide N-Acetylgalactosaminyltransferase-Like 6
SNTG2	8	110216150	110309905	17.300588	0.005	4.42576	0.0004	Syntrophin, Gamma 2
ADCY8	14	20372503	20605323	16.30426	0.05	4.54814	0.0002	Adenylate Cyclase 8
KCNQ3	14	19287002	19645708	22.801613	0.05	3.56943	0.0086	Potassium Channel, Voltage Gated KQT-Like Subfamily Q, Member 3
ADAM12	26	43866756	44215139	18.464761	0.01	3.95478	0.0022	ADAM Metallopeptidase Domain 12
HHAT	16	70287240	70652353	15.983459	0.01	3.70004	0.0059	Hedgehog Acyltransferase

PRKG1	26	6008632	7296338	19.132908	0.05	4.18636	0.0010	Protein Kinase, CGMP-Dependent, Type I
ACBD6	16	59075764	59280270	19.233905	0.005	3.77667	0.0045	Acyl-CoA Binding Domain Containing 6
TMEM131	11	3263373	3430184	18.66351	0.01	4.15289	0.0012	Transmembrane Protein 131
NTM	29	33380373	34343966	28.250062	0.01	3.68597	0.0061	Neurotrimin
VWA3B	11	3541661	3726637	26.748701	0.15	3.68215	0.0062	Von Willebrand Factor A Domain Containing 3B
VAV3	3	33181438	33606999	15.790811	0.003	3.84214	0.0037	Vav 3 Guanine Nucleotide Exchange Factor
LOC102181667	6	75028430	75742734	19.582784	0.005	3.80866	0.0040	

Table 2.2 Major candidate genes obtained from XP-CLR and XP-EHH analysis

Candidate genes	Chromosome	Start	End	Max XP-CLR	XP-CLR *P*-value	Max XP-EHH	XP-EHH *P*-value	Description
CCR3	22	52849231	52882844	23.73496	0.005	-	-	Receptor for a C-C type chemokine
HM13	13	59025564	59064557	21.42064	0.01	-	-	Minor histocompatibility antigen H13
IGSF10	1	114342294	114369614	18.15857	0.01	-	-	Immunoglobulin superfamily member 10
ROBO1	1	24953553	26069449	24.10493	0.01	-	-	Roundabout homolog 1
ROBO 2	1	22148573	22808722	19.23946	0.05	-	-	Roundabout homolog 2
CLNK	6	102650000	102700000	-	-	3.61921	0.0074	Cytokine-dependent hematopoietic cell linker
NTM	29	33380373	34343966	28.250062	0.01	3.68597	0.0061	Neurotrimin
MYO5A	4	54037221	54172686	15.234912	-	-	-	Myosin, Heavy Polypeptide Kinase

Table 2.3 CCR3 region sequence information of the 11 loci

CHR[1]	BP[2]	Alt[3]	Native[4]	Cross[5]	Ref[6]	CHISQ[7]	P[8]	OR[9]
22	52854331	T	2	21	G	31.92	1.61E-08	0.01667
22	52859337	G	2	21	C	31.92	1.61E-08	0.01667
22	52860819	A	2	21	T	31.92	1.61E-08	0.01667
22	52850476	G	23	5	A	28.72	8.37E-08	38.33
22	52870519	G	21	5	A	23.01	1.61E-06	21
22	52872604	T	21	5	A	23.01	1.61E-06	21
22	52873011	A	21	5	G	23.01	1.61E-06	21
22	52880722	C	21	5	T	23.01	1.61E-06	21
22	52854052	G	20	5	C	20.46	6.08E-06	16.67
22	52851203	C	19	5	T	18.1	2.10E-05	13.57
22	52851723	T	19	5	G	18.1	2.10E-05	13.57

CHR[1] = Chromosome Number

BP[2] = SNP physical location

Ref[3] = Reference allele code

Alt[4] = Alternate allele code

Native[5] = Korean native goats alternative allele frequency

Cross[6]= Cross breed goats alternative allele frequency

CHISQ[7]= chi-square test statistic (1df)

P[8] = p-value

OR[9] = Odds ratio

2.5 Discussion

2.5.1 The genetic backgrounds of two populations

Korean native goat population and some outbred goat lines were used to form crossbred goat population based on Korean native goat to improve inferior traits of Korean native goat. From the whole-genome sequencing data, I observed a reduction in nucleotide diversity in Korean native goats compared with the crossbred goats which might be an indication of inbreeding in Korean native goats. In support of finding, scholars (Odahara, Chung et al. 2006) has shown that Korean native goats have lower genetic variability compared with other Asian goat populations.

With the result that, the admixture analysis using STRUCTURE (Evanno, Regnaut et al. 2005) is showed that the proportion of imported alleles increased during crossbreeding in crossbred, while a majority part of alleles became an indigenous allele in Korean native goat. The nucleotide diversity analysis in the crossbred genome might have increased due to the consistent inflow of new alleles, unlike the Korean native goats.

2.5.2 Genes involved in highly variable region

Recombination rate is known to influence the nucleotide diversity (Nachman 2001). The anode regions of chromosomes are famous to be variable because of recombination events. Based on the distribution of variation in the goat genome, the anode regions of chromosomes have more variable than other regions. I found that the top 5% of highly variable regions were enriched with genes involved in olfactory sensors and neurological systems. And from the result of nonsynonymous SNP analysis, I identified enrichment of amino acid substitution in genes related to olfactory sensors.

Genes involved in olfactory systems might have formed through this adaptation. Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. Odor molecules in the environment are detected by olfactory receptors. For animals, olfactory receptors are essential to finding nutritious food and avoid eating toxic substances, avoid predators, identify suitable mating partners and their offspring (Mombaerts 2004, Niimura 2009). Korean native goats have been developed under feed shortage condition that they have been forced to

graze in fields, freely or confined, that are overgrazed shrubs and bark of trees to which their digestibility is low. The positive selection of Olfactory genes might be because of this adaptation while Saanen goats were commercial breeds artificially selected for intensive production (Son 1999, Choi, Choy et al. 2006).

Korean native goat populations are also different from the crossbred goat population based on nonsynonymous SNP results of chymosin (*CYM*) and collagen, type Xi, and alpha 2 (*COL11A2*). CYM is a gene that encoded an enzyme involved in milk ingestion of young ruminant animals (Emmons and Lister 1976), and *COL11A2* is related to the osteochondrodysplasias (Vikkula, Madman et al. 1995) and Stickler syndrome (Sirko-Osadsa, Murray et al. 1998). These genes might be influenced the early cycle of the goat life and body development. It might be contributed to differences in the growth of young kids between Korean native and crossbred goats.

2.5.3 Selective sweep regions in Korean native goat populations

Korean native goats are resistant to lumbar paralysis that highly affects goats of exotic origin introduced into the peninsula (Son 1999). Lumbar paralysis is a common disease in ungulate mammals such as goats, sheep, or cattle which is transmitted through mosquito biting that carries on filarial parasites called *Setaria digitata*. This parasite invades the central nervous system (CNS) such as brain or spinal cord causing a disorder in the hind legs (Son 1999). The positive selection of *CCR3* gene may function an essential role in the resistance of the Korean native goat. This gene encodes a chemokine receptor that is expressed in eosinophils, T_H1 and T_H2 cells (Sallusto, Mackay et al. 1997) and contributes to immune response through mobilizing these immune cells when there is parasitic infection. A previous study reported that antigens of the adult *Setaria digitata* induced a type of T_H1 and T_H2 cytokine response (Dalai, Das et al. 1998). In this study, I found SNP frequency differences between the breeds compared which might be the reason for the superior resistance of lumbar paralysis via immune system response (Son 1999). Additionally, three more genes *(CLNK*, *HM13*, *IGSF10)* were identified as positively selected in relation to immune response (Leo and Schraven 2001, Dybkaer, Iqbal et al. 2007, Severino, Silva et al. 2014). In addition to immune related genes, I detected neurologically significant genes. The *ROBO1,* known as axon guidance and

neuronal cell migration, is expressed in mesenchymal stem cells (MSC) which differentiate into neuronal cells and involved in neuroprotective effect (Crigler, Robey et al. 2006, Uccelli, Benvenuto et al. 2011, Wright, Masri et al. 2011). Interestingly, some of the genes (*ROBO2, NTM*) are also involved in the development of the neural system (Hivert, Liu et al. 2002, Yu, Qian et al. 2012). These genes may play a certain role in recovering damaged nerve cells, and therefore, contribute to lumbar paralysis resistance. I expect that I could understand description of selective sweep signatures and genetic features better in Korean native goat population through this study, and this comprehension will bring about a deeper understanding of its physiology.

This chapter was published in *Genes & Genomics*
as a partial fulfillment of Wonseok Lee's Ph.D program

Chapter 3. Identifying candidate positive selection genes in Korean imported pig breeds

3.1 Abstract

Domestication and artificial selection have modified the genome landscape of the pig. The identification of selection signatures in the genome can help to elucidate the selection mechanisms and uncover the causal genes related to the phenotypic variations between domestic pig breeds. Therefore, I scanned the genomes of Korean imported pig breeds comparing to Korean native breeds using Z-transformed Fst (ZFst) and Heterozygosity (ZHp) statistics to search for the signatures of selection. As a result, I identified 411 (ZFst = 175; ZHp = 256) putatively selected genes in commercial breeds. The identified gene regions were harboring related to immunity, coat color, reproduction function and other traits. Several genes (*PLSCR4,AGTR1,CORIN,APOB,CLAUDIN1 and PON1*) were closely related to reproduction traits such as fertility, ovulation rate, and uterine function and male spermatogenesis. This study revealed genes which improve our understanding of the biological mechanisms of higher litter sizes, the phenotype of interest, in higher litter pig breeds.

3.2 Introduction

The domestic pig (*Sus scrofa domesticus*) are valuable meat producing animals originated from the Eurasian wild boar (*Sus scrofa*) 9000 YBP in the Near East in the Tigris Basin (Giuffra, Kijas et al. 2000). Since domestication, both natural and artificial selection processes have changed the genomic landscape of the pig; this resulted in hundreds of breeds in the world with dramatic changes in phenotypic traits such as behavior, body composition, reproduction, and coat color (Rubin, Megens et al. 2012). Intensive artificial selection for traits related to lean meat production including high growth rate, feed conversion efficiency, soundness and litter size achieved breeds of divergent phenotypic traits (Rauw, Kanis et al. 1998). In addition, reproductive technologies such as artificial insemination, estrus detection, and synchronization contributed a lot to the current pig production protocols and productivity in developed and developing countries (Knox 2014). It is known that selection for production traits like growth rate, compromised reproductive ability, and immunity of animals.

Efficient reproductive performance is an important trait in meat animal production. The pig is noted for its high fertility and other reproductive traits. It has short maturity age and gestation length, multiple offspring per pregnancy, and quick tendency to rebreed (Knox 2014). Among reproduction traits, prolificacy, defined as litter size at birth, has been considered the most crucial component of sow productivity. In some pig breeds, litter size has been used as selection objective and the criterion to improve reproductive performance, on economic, genetic, and ease of measurement grounds (McLaren and Bovey 1992). In addition to sow management, litter size is affected by genetics such as ovulation rate, fertility, embryonic mortality, and uterine capacity of sow. For example, the Meishan pigs from China are prolific and have higher ovulation rate and uterine capacity, which allows them to maintain their higher number of attached embryos through gestation (Haley and Lee 1992).

Pork from Korean native pigs has a preferred taste and palatability due to its higher intramuscular fat content (marbling) that contributes to higher tenderness. However, they have slower growth rate and lower litter size compared to the common imported breeds such as Duroc, Yorkshire, and landrace (Kim, Yeo et al. 2002, Choi, Chung et al. 2015). Analysis of positive selection signature of the genomes of pig breeds between high and low litter size is important to understand the biological mechanisms for higher

reproductive performances. In this study, using a whole genome sequencing data of five pig breeds, I searched for the selective sweep regions with marked allele differences for genes related to reproductive performances, using F_{ST} (ZF_{ST}) statistics and Homozygosity (ZHp) test.

3.3 Material and Methods

3.3.1 Sample preparation and whole genome re-sequencing

Whole-genome sequencing data of pigs obtained from NCBI Sequence Read Archive database under the accession numbers SRP047260 (for 48 Korean Imported Pigs, KIP) and SRP049499 (for 14 Korean Native Pigs, KNP) were used for this study. The details about the sequenced samples and sequencing methods are described in the previous researches (Kim, Yeo et al. 2002). After collecting data, the paired-end reads were mapped to the *Sus scrofa* reference genome (Sscrofa 10.2) using Bowtie2 (Langmead and Salzberg 2012). Open-source packages, including Picard tools 1.94 (http://picard.sourceforge.net), SAMtools 0.1.19 (Li, Handsaker et al. 2009), Genome Analysis Toolkit (GATK) 2.6.4 (McKenna, Hanna et al. 2010) and the VCFtools 4.0 (Danecek, Auton et al. 2011) were used for downstream processing and variant calling for SNPs. The "MarkDuplicates" Picard command-line option was used to remove potential PCR duplicates. Next, SAMtools was used to index the resulting bam format files and calculate the mapped read length. I then performed local realignment of sequence reads to correct misalignment due to the presence of small insertion and deletion, using GATK "RealignerTargetCreator" and "IndelRealigner" arguments. Also, base quality score recalibration was performed to get accurate quality scores and correct the variation in quality with machine cycle and sequence context. For candidate SNP identification, GATK "UnifiedGenotyper" and "SelectVariants" arguments were used with the following filtering criteria: all variants with 1) Phred-scaled quality score of less than 30; 2) read depth less than 5; 3) MQ0 (total count across all samples of mapping quality zero reads) > 4; and 4) Phred-scaled P-value using Fisher's exact test more than 200 were filtered out to reduce false-positive calls due to strand bias. For the haplotype information on each chromosome, I used BEAGLE (Browning and Browning 2007) to infer the haplotype phase and impute missing alleles simultaneously for the entire set in

the pig populations.

3.3.2 Construction of a phylogenetic tree, principal component, and population structure analysis

SNPhylo (Lee, Guo et al. 2014) was used to construct a phylogenetic tree with a pig data. SNPhylo is a pipeline developed to construct phylogenetic trees from large SNP data (selected SNPs based on LD blocks); the pipeline allows constructing a maximum-likelihood tree with bootstrap values. I used SNP data from a total of 72 individuals (48 KIP, 14 KNP and 10 wild boars (from Sequence Read Archive (SRA) database under the accession number: SRP047260) to construct the phylogenetic tree. In SNPhylo, the options I used are as follows: Minor allele frequency > 0.05, the number of bootstrap samples = 1,000 and wild boars were set as outgroup. FigTree (v.1.4; http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the tree.

I performed principal component analysis (PCA) to examine population differentiation among KIP and KNP pig populations using genotype data from each pig. To estimate eigenvectors, I used GCTAtool (Yang, Lee et al. 2011), which implements PCA in EIGENSTRAT. VCFtools (Danecek, Auton et al. 2011) and PLINK (Purcell, Neale et al. 2007) were used to prepare input data sets for GCTAtool.

Further, I used STRUCTURE version 2.3.4 (Hubisz, Falush et al. 2009) software to identify the genetic proportions of each pig individual from ancestral populations. STRUCTURE software uses Bayesian algorithms to detect the true number of clusters, also referred to as *K* (the number of ancestral populations). PLINK was used to generate STRUCTURE input files, using -thin option. I used the “admixture” model with 100,000 iterations and 1,000 burn-in iterations for each analysis from $K = 2–5$.

3.3.3 Selection signature statistical analysis

To detect the signature of positive selection, I compared the genomes of KIP with KNP breeds using two separate statistics, ZF_{ST} and ZHp. F_{ST} is the most commonly used metric for measuring genetic differentiation between populations (Holsinger and Weir

2009). It compares the variance in allele frequencies among populations with that of the within populations. A larger F_{ST} value means that the allele frequencies are different; therefore, populations are different. When it is small, populations are considered to be the same. In this study, I used Z-transformed F_{ST} statistics to compare the genomes of commercial pig breeds and the Korean native pigs. The VCFtools version 0.1.12 program was used to estimate F_{ST} (the method of Weir and Cockerham (Weir and Cockerham 1984)) with the arguments --fst-window-size (150 kb) and --fst-window-step (75 kb). Then, the F_{ST} values were Z-transformed.

I also performed the Z-transformed heterozygosity (ZHp) test following the protocol used by Rubin et al. (Rubin, Megens et al. 2012) With an overlapping sliding window approach, the pooled heterozygosity (H*p*) was calculated as:

$$H_p = \frac{2 \sum n_{MAJ} \sum n_{MIN}}{(\sum n_{MAJ} + \sum n_{MIN})^2},$$

Where: $\sum n_{MAJ}$ and $\sum n_{MIN}$ are the sums of major and minor allele frequency at the given 150 kb windows with a step size of 75 kb (concordant with the F_{ST} analysis).

Then, H_p values were Z-transformed: $ZH_p = \frac{(H_p - \mu H_p)}{\sigma H_p}$, where μH_p is the overall average heterozygosity and σH_p is the standard deviation for all windows within one breed group (Elferink, Megens et al. 2012). The number of major and minor allele frequencies were counted at each of the identified SNP in each KIP breed and KNP breed group using VCF tools (Danecek, Auton et al. 2011). Then I calculated the ZHp scores by an in-house python script. In two of the methods, I only used the SNP positions whose minor allele frequency was >0.05 and excluded windows that number of SNPs is below 10. The top chromosomal 1% outlier regions of the distribution were considered to be candidate region under selection in the KIP breeds. Genes that are located (partially or completely) in the window regions were presumed as candidate genes putatively under positive selection (Lee, Kim et al. 2014) and annotated based on *sus scrofa* (Sscrofa 10.2).

3.3.4 Gene ontology terms enrichment tests

The DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool was used for gene ontology and pathway analyses with the *sus scrofa* background set (Dennis, Sherman et al. 2003). I performed Functional Annotation Clustering with the highest classification stringency option. To further determine biological process, ClueGO plugin of Cytoscape (Shannon, Markiel et al. 2003, Bindea, Mlecnik et al. 2009) was performed using the combined gene list. This program integrates gene ontology (GO) categories and visualizes a functionally organized GO category networks based on the overlap between the different GO categories.

3.3.5 Candidate gene variants annotation

To identify non-synonymous coding SNPs that overlap to candidate genes that may contribute to KIP breeds phenotypes, SNPEff (version 4.3i) (Cingolani, Platts et al. 2012) was used for filtering variants annotation with a *Sus scrofa* reference genome (Sscrofa10.2).

3.4 Result

3.4.1 Sequence information

The whole genomes of sample pigs were sequenced to an approximate coverage of 12.07 fold on average, with a total of 195,908,825 bp. Sequence reads of each breed were aligned to the pig reference genome (*Sus scrofa* 10.2) from the Ensembl database using Bowtie2, and 87.79% of the sequence reads were aligned to the reference sequence (Table 3.1). After removing PCR duplicates and recalibrating base quality, I finally retained a total of 28,065,585 SNPs, and the average nucleotide change rate was 1 per ~84 base pairs (Table 3.2).

Table 3.1 Mapping rate and the number of filtered SNPs of resequencing data of 62 pigs used in the study (Korean native, Jeju native, Duroc, Yorkshire and Landrace).

Sample	Breed	Read counts	Alignment rate	Filtered SNPs	DNA sequenced(bp)
1729	duroc	189,981,228	83.04%	5,500,208	28,928,850,841
1735	duroc	216,262,109	88.27%	5,695,891	34,772,576,313
1795	duroc	213,438,494	88.31%	5,576,732	33,287,873,267
1933	duroc	215,310,717	88.44%	2,670,494	11,042,707,831
1964	duroc	215,957,755	88.59%	3,088,844	12,118,226,691
24-64	duroc	189,954,956	88.98%	5,451,659	30,624,613,624
24-78	duroc	177,287,960	88.44%	5,597,864	30,641,023,801
25-24	duroc	212,403,850	85.90%	5,725,634	32,868,695,644
25-78	duroc	186,098,347	88.24%	5,333,114	29,973,056,298
25-80	duroc	182,106,185	87.47%	5,439,538	29,093,526,947
26-23	duroc	210,353,056	88.52%	5,609,598	33,396,732,659
26-66	duroc	200,210,010	88.92%	5,387,610	28,820,545,557
27-20	duroc	219,307,768	88.59%	5,150,757	29,810,469,903
27-78	duroc	198,151,669	87.31%	5,465,681	28,707,067,543
27-81	duroc	194,402,442	89.17%	5,438,391	31,276,913,026
DAA8330	duroc	175,110,506	89.01%	5,623,468	28,570,363,154
DAA8623	duroc	181,271,952	89.18%	5,479,224	29,816,628,792
DAA9119	duroc	175,693,544	88.90%	5,140,032	28,602,098,193
DAA9736	duroc	206,108,415	88.18%	5,403,000	31,327,740,301
DAA9738	duroc	189,818,774	88.42%	5,417,103	29,751,367,118
S_10	landrace	172,854,048	90.67%	6,796,623	28,146,194,274
S_11	landrace	177,762,997	90.79%	6,913,132	28,560,921,397
S_12	landrace	167,772,384	89.86%	7,009,900	26,830,912,166
S_13	landrace	161,965,912	89.84%	6,666,455	26,150,197,503
S_14	landrace	169,155,065	88.72%	5,743,262	21,580,605,663
S_15	landrace	164,183,890	89.60%	5,522,006	19,280,831,743
S_17	landrace	150,756,759	83.92%	4,248,996	19,800,222,780
S_18	landrace	148,854,532	83.16%	3,877,656	19,509,323,552
S_19	landrace	166,471,738	89.81%	6,454,074	25,425,629,156
S_20	landrace	161,192,359	86.73%	6,164,404	23,189,100,924

Sample	Breed	Read counts	Alignment rate	Filtered SNPs	DNA sequenced(bp)
S_7	landrace	163,981,621	90.13%	7,106,242	26,245,312,621
S_8	landrace	163,102,194	90.57%	6,830,310	26,148,565,061
S_9	landrace	166,101,951	90.72%	7,110,918	26,959,821,636
KL1	yorkshire	218,537,902	85.20%	6,780,503	24,872,602,703
KL2	yorkshire	216,835,746	84.44%	6,977,679	25,959,993,393
KL3	yorkshire	210,827,426	85.37%	6,401,205	25,322,996,354
KL4	yorkshire	209,333,064	86.03%	6,733,692	22,316,951,907
KL5	yorkshire	214,065,193	86.07%	6,511,262	22,694,241,585
KL6	yorkshire	215,146,933	87.22%	6,997,362	26,348,485,254
KL7	yorkshire	335,518,129	86.81%	6,937,685	27,379,915,490
KL8	yorkshire	205,559,812	86.69%	6,270,258	19,266,022,675
pig31	yorkshire	221,392,087	82.58%	6,006,615	19,761,767,330
pig32	yorkshire	185,643,766	80.95%	4,447,859	13,309,411,970
S_1	yorkshire	158,826,981	90.70%	6,930,786	25,889,324,075
S_2	yorkshire	169,910,292	89.78%	7,339,912	27,521,377,538
S_3	yorkshire	175,644,056	90.09%	7,363,610	28,280,174,622
S_4	yorkshire	170,772,011	89.72%	7,236,036	27,418,477,373
S_5	yorkshire	166,906,130	90.77%	6,978,822	27,077,789,551
KK1	KNP	219,383,180	86.11%	6,799,331	31,638,979,962
KK2	KNP	217,322,589	87.14%	7,229,957	36,631,567,674
KK3	KNP	217,690,622	86.76%	7,540,071	36,036,077,075
KK4	KNP	192,667,885	86.89%	6,951,274	40,565,629,709
KK5	KNP	205,207,112	87.33%	7,472,731	34,353,068,664
KK6	KNP	213,524,043	87.09%	7,294,454	34,090,842,311
KK8	JejuNative	218,986,339	87.17%	7,488,173	35,228,862,135
KK9	JejuNative	204,356,928	86.40%	6,932,005	42,763,646,742
KK10	JejuNative	242,627,234	85.34%	7,764,786	38,048,976,506
KK11	JejuNative	211,576,832	86.67%	7,907,052	35,412,469,474
10_453	JejuNative	212,859,163	89.49%	8,013,021	34,064,577,362
10_561	JejuNative	198,898,231	89.49%	7,051,058	22,971,600,902
12_98	JejuNative	214,011,967	89.63%	7,716,165	34,260,081,906
K8_I7	JejuNative	213,002,684	88.62%	8,268,713	36,144,754,624

Table 3.2 Summary of detailed variants rate of Koran Imported Pig breeds (Duroc, Yorkshire and Landrace)

Chromosome	Length	Variants	Variants rate
1	315,321,322	2,777,511	113
2	162,569,375	1,900,771	85
3	144,787,322	1,687,136	85
4	143,465,943	1,647,449	87
5	111,506,441	1,338,075	83
6	157,765,593	1,770,300	89
7	134,764,511	1,667,377	80
8	148,491,826	1,721,810	86
9	153,670,197	1,864,712	82
10	79,102,373	1,231,277	64
11	87,690,581	1,147,765	76
12	63,588,571	860,956	73
13	218,635,234	2,192,921	99
14	153,851,969	1,738,607	88
15	157,681,621	1,630,072	96
16	86,898,991	1,130,959	76
17	69,701,581	935,026	74
18	61,220,071	822,861	74

3.4.2 Construction of a phylogenetic tree, population structure, and PCA

A phylogenetic tree is a good method for inferring evolutionary relationships among various organisms. Using the SNPhylo, I constructed a non-rooted phylogenetic tree that individual pigs were able to cluster to their breed; all breeds were monophyletic with 100 % bootstrap value (Figure 3.1a). As expected, the wild boar was farthest from all breeds. The principal component analysis depicted that KIPs were clearly distinct from KNPs (Figure 3.1c). PC1 separated KIP breeds from KNPs explaining 15.2% of the total variation, and PC2 separated Duroc from Yorkshire and Landrace (11.8%). I then examined the genetic structures of KIPs and KNPs through admixture analysis using STRUCTURE software. The structure showed that the KIPs and the KNPs are clearly divided (Figure 3.1b). A previous study to assess the genetic diversity of pig breeds in Korea reported that Korean native pig breeds were distinct from other imported commercial breeds (Kim, Yeo et al. 2002).

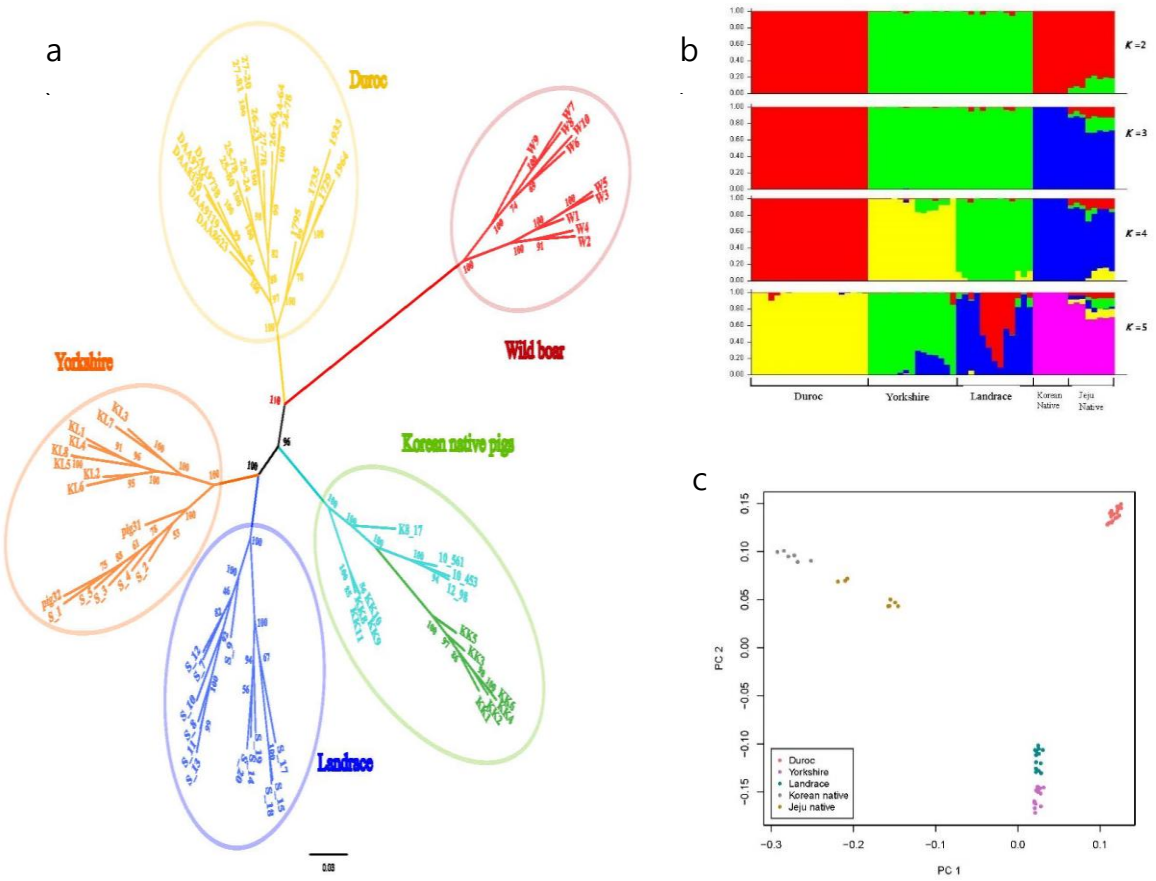


Figure 3.1 Phylogenetic tree and population structure of pig breeds used in this analysis. (a) Phylogenetic tree constructed using SNPhylo. Wild boar is set as an outgroup; (b) Structure. (c) PCA;

3.4.3 Positive selection statistical analysis

I used ZF_{ST} and ZHp statistics to detect the positive selection signature in the genomes of KIP breeds. I searched genomic regions of KIP breeds with reduced Hp and increased differentiation to the KNP breeds. Accordingly, 175 putative genes from ZF_{ST} statistics (Table 3.3) and 256 putative genes based on ZHp scores were identified to be under positive selection in the KIP breeds (Table 3.4). Twenty of the genes were common for both statistics.

Table 3.3 Summary of genes identified from ZFst statistics

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
1	148425001	148575000	602	4.121066393	SPRED1
1	148275001	148425000	528	4.115101269	FAM98B
1	148350001	148500000	594	4.110070706	SPRED1
1	148200001	148350000	435	3.76971548	RASGRP1
1	148200001	148350000	435	3.76971548	FAM98B
1	170625001	170775000	624	3.587227953	TMX3
1	170700001	170850000	890	3.525700291	TMX3
1	148125001	148275000	456	3.497860612	RASGRP1
1	148125001	148275000	456	3.497860612	FAM98B
1	148050001	148200000	252	3.462033362	RASGRP1
1	179400001	179550000	897	3.45736056	ZNF532
1	179925001	180075000	796	3.413567642	NEDD4L
1	179925001	180075000	796	3.413567642	7SK
1	143100001	143250000	953	3.329114045	CCNDBP1
1	143100001	143250000	953	3.329114045	EPB42
1	143100001	143250000	953	3.329114045	UBR1
1	169800001	169950000	733	3.298222442	DOK6
1	143175001	143325000	860	3.275968222	UBR1
1	171375001	171525000	187	3.216908383	5S_rRNA
1	169875001	170025000	1143	3.209439201	DOK6
3	125400001	125550000	1791	3.497386031	C2orf43
3	125400001	125550000	1791	3.497386031	GDF7
3	125325001	125475000	1230	3.368183051	C2orf43
3	125250001	125400000	1040	3.163448606	APOB
4	74700001	74850000	641	3.279896296	DNAJC5B
4	72825001	72975000	1190	3.157804737	PREX2
5	42750001	42900000	715	4.04228587	PHLDA1
5	42750001	42900000	715	4.04228587	NAP1L1
5	45525001	45675000	1274	3.988862599	KIAA1551
5	42825001	42975000	582	3.777411001	PHLDA1
5	42825001	42975000	582	3.777411001	NAP1L1
5	44925001	45075000	713	3.506439585	FGD4
5	44850001	4.50E+07	821	3.405908628	FGD4
5	52125001	52275000	593	3.292315728	KRAS
5	52125001	52275000	593	3.292315728	LYRM5
5	42900001	43050000	645	3.254568248	NAP1L1
5	45450001	45600000	832	3.197216903	KIAA1551
5	81750001	81900000	915	3.168961052	H1FNT

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
6	300001	450000	1023	3.740320634	SPIRE2
6	300001	450000	1023	3.740320634	DEF8
6	300001	450000	1023	3.740320634	FANCA
6	300001	450000	1023	3.740320634	ZNF276
6	300001	450000	1023	3.740320634	VPS9D1
6	225001	375000	641	3.598282051	TCF25
6	225001	375000	641	3.598282051	MC1R
6	225001	375000	641	3.598282051	SPIRE2
6	225001	375000	641	3.598282051	DEF8
6	44550001	44700000	621	3.550765494	ITPKC
6	44550001	44700000	621	3.550765494	C19orf54
6	44550001	44700000	621	3.550765494	SNRPA
6	44550001	44700000	621	3.550765494	EGLN2
6	44625001	44775000	694	3.453505498	ADCK4
6	44625001	44775000	694	3.453505498	NUMBL
6	44625001	44775000	694	3.453505498	ITPKC
6	44625001	44775000	694	3.453505498	C19orf54
6	44625001	44775000	694	3.453505498	SNRPA
6	44625001	44775000	694	3.453505498	EGLN2
6	44625001	44775000	694	3.453505498	CYP2F1
6	44700001	44850000	371	3.447934642	ADCK4
6	44700001	44850000	371	3.447934642	NUMBL
6	44700001	44850000	371	3.447934642	CYP2F1
6	44475001	44625000	339	3.440472761	BLVRB
6	44475001	44625000	339	3.440472761	PLD3
6	1425001	1575000	363	3.313146204	U6
6	48000001	48150000	631	3.188258265	SLC1A5
6	48000001	48150000	631	3.188258265	AP2S1
6	48000001	48150000	631	3.188258265	ARHGAP35
6	48075001	48225000	759	3.15286909	ARHGAP35
7	69075001	69225000	575	4.697909216	BRMS1L
7	68925001	69075000	736	4.622932645	BRMS1L
7	69000001	69150000	1031	4.607848256	BRMS1L
7	69525001	69675000	663	3.634985451	FAM177A1
7	69525001	69675000	663	3.634985451	SRP54
7	66525001	66675000	812	3.620828321	CLEC14A
7	66525001	66675000	812	3.620828321	SSTR1
7	58650001	58800000	715	3.532001273	UROC1
7	58650001	58800000	715	3.532001273	ZXDC
7	58650001	58800000	715	3.532001273	SLC41A3

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
7	58650001	58800000	715	3.532001273	CCDC37
7	69600001	69750000	875	3.498028541	FAM177A1
7	69600001	69750000	875	3.498028541	SRP54
7	69600001	69750000	875	3.498028541	U2
7	69150001	69300000	697	3.393956476	INSM2
7	69450001	69600000	897	3.318556432	PSMA6
7	58800001	58950000	1368	3.317994236	CCDC37
7	69225001	69375000	1198	3.275274603	INSM2
7	58575001	58725000	1338	3.247595551	UNC45A
7	58575001	58725000	1338	3.247595551	RCCD1
7	58575001	58725000	1338	3.247595551	UROC1
7	58575001	58725000	1338	3.247595551	ZXDC
7	58575001	58725000	1338	3.247595551	SLC41A3
7	69900001	70050000	914	3.178642514	EAPP
7	69900001	70050000	914	3.178642514	CFL2
7	69900001	70050000	914	3.178642514	U1
7	69900001	70050000	914	3.178642514	U1
7	69900001	70050000	914	3.178642514	U1
7	69900001	70050000	914	3.178642514	U1
7	69900001	70050000	914	3.178642514	U1
8	43275001	43425000	1573	4.923627463	Metazoa_SRP
8	43350001	43500000	1394	4.666134161	Metazoa_SRP
8	43950001	44100000	961	4.560178373	KDR
8	43500001	43650000	1536	4.507689662	KIT
8	43425001	43575000	1049	4.25129885	KIT
8	42900001	43050000	1155	4.162106739	GSX2
8	42900001	43050000	1155	4.162106739	PDGFRA
8	87600001	87750000	1229	4.097351923	SLC10A7
8	87600001	87750000	1229	4.097351923	LSM6
8	87375001	87525000	392	4.019206608	SLC10A7
8	87525001	87675000	1498	4.010700648	SLC10A7
8	44025001	44175000	711	4.002530545	SRD5A3
8	87675001	87825000	753	3.976304443	LSM6
8	39675001	39825000	1805	3.919668621	CORIN
8	39675001	39825000	1805	3.919668621	NFXL1
8	39675001	39825000	1805	3.919668621	U6
8	38025001	38175000	962	3.858878386	GABRG1
8	87450001	87600000	1245	3.856914348	SLC10A7
8	87225001	87375000	735	3.843144185	SLC10A7
8	87300001	87450000	96	3.784055141	SLC10A7

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
8	38100001	38250000	740	3.765728995	GABRG1
8	38100001	38250000	740	3.765728995	GABRA2
8	43875001	44025000	1298	3.613724202	KDR
8	42975001	43125000	1481	3.590046237	PDGFRA
8	42975001	43125000	1481	3.590046237	U6
8	43575001	43725000	2035	3.577123019	KIT
8	39150001	39300000	962	3.529533449	GABRB1
8	39150001	39300000	962	3.529533449	COMMD8
8	87150001	87300000	1163	3.407485698	SLC10A7
8	86850001	8.70E+07	896	3.390846142	TTC29
8	86925001	87075000	579	3.363532153	TTC29
8	39750001	39900000	1398	3.356282008	CORIN
8	39750001	39900000	1398	3.356282008	NFXL1
8	39750001	39900000	1398	3.356282008	CNGA1
8	39750001	39900000	1398	3.356282008	5S_rRNA
8	90675001	90825000	1472	3.286533135	INPP4B
8	43800001	43950000	1390	3.239447352	KDR
8	43800001	43950000	1390	3.239447352	U6
8	90600001	90750000	1497	3.232839718	INPP4B
8	37950001	38100000	1412	3.175444565	GABRG1
8	39225001	39375000	1414	3.175232829	GABRB1
8	39225001	39375000	1414	3.175232829	COMMD8
8	39225001	39375000	1414	3.175232829	ATP10D
8	39300001	39450000	1389	3.166405614	COMMD8
8	39300001	39450000	1389	3.166405614	ATP10D
8	89400001	89550000	980	3.149970492	GAB1
9	80550001	80700000	453	4.855528673	TFPI2
9	89175001	89325000	535	4.250977594	THSD7A
9	102975001	103125000	1271	4.195926144	KIAA1324L
9	103200001	103350000	924	4.140261389	KIAA1324L
9	89100001	89250000	1063	4.127929572	THSD7A
9	101775001	101925000	940	3.991308519	DBF4
9	101775001	101925000	940	3.991308519	CCDC126
9	89025001	89175000	727	3.974114067	THSD7A
9	102675001	102825000	1268	3.965045911	PGP3
9	102675001	102825000	1268	3.965045911	CROT
9	80625001	80775000	212	3.945536962	TFPI2
9	80625001	80775000	212	3.945536962	GNG11
9	82350001	82500000	316	3.917909019	ASB4
9	107025001	107175000	995	3.906533666	SEMA3E

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
9	102750001	102900000	1141	3.89397551	CROT
9	102750001	102900000	1141	3.89397551	TMEM243
9	102750001	102900000	1141	3.89397551	DMTF1
9	102750001	102900000	1141	3.89397551	U6
9	91500001	91650000	658	3.889645867	ETV1
9	88950001	89100000	313	3.878080682	THSD7A
9	89925001	90075000	1296	3.864062275	TMEM106B
9	101850001	1.02E+08	807	3.824350758	DBF4
9	101850001	1.02E+08	807	3.824350758	CCDC126
9	101850001	1.02E+08	807	3.824350758	SLC25A40
9	101850001	1.02E+08	807	3.824350758	RUNDC3B
9	82950001	83100000	350	3.819152266	DYNC1I1
9	99000001	99150000	551	3.817728522	ABCB5
9	107100001	107250000	950	3.809704444	SEMA3E
9	98850001	9.90E+07	493	3.792466185	ABCB5
9	101250001	101400000	968	3.771548094	FAM126A
9	103050001	103200000	1339	3.755463433	KIAA1324L
9	88425001	88575000	1251	3.747840925	NDUFA4
9	88425001	88575000	1251	3.747840925	SCARNA18
9	88425001	88575000	1251	3.747840925	SCARNA17
9	103125001	103275000	1144	3.746110527	KIAA1324L
9	88875001	89025000	433	3.738699755	THSD7A
9	88500001	88650000	1118	3.72130817	NDUFA4
9	88500001	88650000	1118	3.72130817	PHF14
9	88500001	88650000	1118	3.72130817	SCARNA18
9	88500001	88650000	1118	3.72130817	SCARNA17
9	88575001	88725000	730	3.720775178	PHF14
9	102900001	103050000	927	3.666994146	TMEM243
9	102900001	103050000	927	3.666994146	DMTF1
9	102900001	103050000	927	3.666994146	KIAA1324L
9	98925001	99075000	541	3.666001176	ABCB5
9	87750001	87900000	1224	3.663058771	7SK
9	87750001	87900000	1224	3.663058771	5S_rRNA
9	88650001	88800000	570	3.652588773	PHF14
9	102375001	102525000	1175	3.629283173	U6
9	82650001	82800000	1024	3.628830495	DYNC1I1
9	82650001	82800000	1024	3.628830495	U6
9	106950001	107100000	951	3.599683892	SEMA3E
9	99150001	99300000	742	3.59705544	SP8
9	87675001	87825000	592	3.577210634	7SK

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
9	87675001	87825000	592	3.577210634	5S_rRNA
9	82425001	82575000	743	3.570252539	ASB4
9	101700001	101850000	869	3.565784172	IGF2BP3
9	101700001	101850000	869	3.565784172	DBF4
9	102450001	102600000	1697	3.561724675	PGP1A
9	102450001	102600000	1697	3.561724675	U6
9	90150001	90300000	1154	3.547209784	SCIN
9	91425001	91575000	1105	3.515748683	ETV1
9	98775001	98925000	745	3.514200817	ABCB5
9	80700001	80850000	488	3.466735369	GNG11
9	80700001	80850000	488	3.466735369	BET1
9	145800001	145950000	1782	3.4628365	SYT14
9	88800001	88950000	871	3.457207234	THSD7A
9	97950001	98100000	421	3.453352172	TMEM196
9	99900001	100050000	1899	3.42508902	DNAH11
9	145725001	145875000	2154	3.419817515	SERTAD4
9	145725001	145875000	2154	3.419817515	SYT14
9	101925001	102075000	731	3.415962453	SLC25A40
9	101925001	102075000	731	3.415962453	RUNDC3B
9	90075001	90225000	894	3.396212563	VWDE
9	88725001	88875000	852	3.393328568	PHF14
9	82875001	83025000	909	3.393058422	DYNC1I1
9	82275001	82425000	129	3.370278511	PON1
9	98475001	98625000	258	3.368146545	ITGB8
9	89250001	89400000	330	3.350207366	THSD7A
9	98325001	98475000	803	3.318680554	MACC1
9	98550001	98700000	394	3.317373629	ITGB8
9	82200001	82350000	267	3.304924992	PON1
9	98250001	98400000	753	3.292994744	MACC1
9	98175001	98325000	984	3.264936028	MACC1
9	82575001	82725000	660	3.243623669	ASB4
9	82575001	82725000	660	3.243623669	PDK4
9	82575001	82725000	660	3.243623669	U6
9	98625001	98775000	357	3.242134214	ITGB8
9	98025001	98175000	798	3.235219927	TMEM196
9	99825001	99975000	1603	3.231788338	DNAH11
9	81600001	81750000	564	3.231445179	SGCE
9	81600001	81750000	564	3.231445179	U6
9	102075001	102225000	1176	3.230057941	RUNDC3B
9	102525001	102675000	2031	3.219704763	PGP1A

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
9	102525001	102675000	2031	3.219704763	PGP3
9	82500001	82650000	744	3.190068976	ASB4
9	82500001	82650000	744	3.190068976	PDK4
9	102600001	102750000	1752	3.185038412	PGP1A
9	102600001	102750000	1752	3.185038412	PGP3
9	102600001	102750000	1752	3.185038412	CROT
9	80775001	80925000	703	3.174845863	BET1
9	83025001	83175000	602	3.159009444	DYNC1I1
9	85500001	85650000	577	3.15840344	MIOS
9	82725001	82875000	1292	3.14978066	DYNC1I1
9	84300001	84450000	979	3.147312836	ACN9
9	99225001	99375000	735	3.142530515	SP8
11	18900001	19050000	1003	3.139799847	CDADC1
12	43875001	44025000	1194	3.617462443	MYO1D
12	43875001	44025000	1194	3.617462443	U6
12	39225001	39375000	1148	3.306823318	APPBP2
13	127725001	127875000	423	4.869357247	TTC14
13	127725001	127875000	423	4.869357247	CCDC39
13	127725001	127875000	423	4.869357247	U6
13	127800001	127950000	359	4.836246063	TTC14
13	127800001	127950000	359	4.836246063	CCDC39
13	127800001	127950000	359	4.836246063	U6
13	126600001	126750000	790	4.788094297	ACTL6A
13	126600001	126750000	790	4.788094297	MRPL47
13	127875001	128025000	304	4.58474709	CCDC39
13	128175001	128325000	443	4.485150693	DNAJC19
13	128175001	128325000	443	4.485150693	FXR1
13	128250001	128400000	363	4.364482879	DNAJC19
13	128100001	128250000	375	4.362810892	DNAJC19
13	128100001	128250000	375	4.362810892	FXR1
13	126675001	126825000	999	4.217654674	ACTL6A
13	126675001	126825000	999	4.217654674	MRPL47
13	126675001	126825000	999	4.217654674	NDUFB5
13	89550001	89700000	787	4.065759399	SLC25A36
13	89550001	89700000	787	4.065759399	snoU13
13	125250001	125400000	549	4.064357559	U6
13	127950001	128100000	389	3.985934796	FXR1
13	96900001	97050000	528	3.963979928	AGTR1
13	126825001	126975000	1016	3.956466938	PEX5L
13	126825001	126975000	1016	3.956466938	SNORA81

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
13	147150001	147300000	1454	3.945164598	DIRC2
13	147150001	147300000	1454	3.945164598	SEMA5B
13	147225001	147375000	1744	3.924560461	DIRC2
13	147225001	147375000	1744	3.924560461	SEMA5B
13	147225001	147375000	1744	3.924560461	HSPBAP1
13	86475001	86625000	851	3.867033886	ARMC8
13	86475001	86625000	851	3.867033886	DBR1
13	86475001	86625000	851	3.867033886	A4GNT
13	86475001	86625000	851	3.867033886	DZIP1L
13	96975001	97125000	896	3.852796441	AGTR1
13	96975001	97125000	896	3.852796441	CPB1
13	86400001	86550000	854	3.850328618	CLDN18
13	86400001	86550000	854	3.850328618	ARMC8
13	86400001	86550000	854	3.850328618	DBR1
13	86400001	86550000	854	3.850328618	A4GNT
13	91500001	91650000	1015	3.781003217	U2SURP
13	91500001	91650000	1015	3.781003217	CHST2
13	91500001	91650000	1015	3.781003217	SNORD112
13	126450001	126600000	278	3.756324981	PIK3CA
13	91425001	91575000	898	3.709370621	PAQR9
13	91425001	91575000	898	3.709370621	U2SURP
13	126750001	126900000	1134	3.702909011	MRPL47
13	126750001	126900000	1134	3.702909011	NDUFB5
13	126750001	126900000	1134	3.702909011	SNORA81
13	93900001	94050000	403	3.697790833	U6
13	86550001	86700000	825	3.694811921	DZIP1L
13	137925001	138075000	978	3.60130477	OSTN
13	137925001	138075000	978	3.60130477	CCDC50
13	147075001	147225000	718	3.580671428	DIRC2
13	93975001	94125000	512	3.576130048	U6
13	128025001	128175000	274	3.558322291	FXR1
13	97200001	97350000	633	3.54491719	GYG1
13	97200001	97350000	633	3.54491719	HLTF
13	137850001	1.38E+08	989	3.532986942	OSTN
13	86100001	86250000	841	3.440677196	SOX14
13	90150001	90300000	1175	3.40301003	RASA2
13	147300001	147450000	1776	3.389407795	SEMA5B
13	147300001	147450000	1776	3.389407795	HSPBAP1
13	147300001	147450000	1776	3.389407795	PARP14
13	97125001	97275000	965	3.371804473	GYG1

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZFst	Gene symbol
13	97125001	97275000	965	3.371804473	HLTF
13	97050001	97200000	1187	3.361736045	CPB1
13	96825001	96975000	463	3.359735501	AGTR1
13	126900001	127050000	733	3.303150788	PEX5L
13	141975001	142125000	492	3.293732171	APOD
13	125175001	125325000	704	3.289351419	U6
13	90450001	90600000	852	3.273522302	GRK7
13	82575001	82725000	456	3.230642041	SLCO2A1
13	82575001	82725000	456	3.230642041	Y_RNA
13	89175001	89325000	1515	3.229911915	CLSTN2
13	136950001	137100000	922	3.192186339	CLAUDIN1
13	136950001	137100000	922	3.192186339	CLDN16
13	136950001	137100000	922	3.192186339	5S_rRNA
13	137475001	137625000	970	3.178825046	GMNC
13	137025001	137175000	499	3.174400486	CLDN16
13	137025001	137175000	499	3.174400486	5S_rRNA
13	94350001	94500000	454	3.17086668	PLSCR4
13	216450001	216600000	1764	3.167442392	WDR4
13	216450001	216600000	1764	3.167442392	NDUFV3
13	136350001	136500000	573	3.167237957	TP63
13	136350001	136500000	573	3.167237957	U3
13	136275001	136425000	393	3.146290661	TP63
13	136275001	136425000	393	3.146290661	U3
18	6225001	6375000	2139	3.291892255	NUB1
18	6225001	6375000	2139	3.291892255	SMARCD3
18	6225001	6375000	2139	3.291892255	ssc-mir-671
18	6225001	6375000	2139	3.291892255	WDR86
18	6225001	6375000	2139	3.291892255	ABCF2
18	6225001	6375000	2139	3.291892255	ssc-mir-671
18	5325001	5475000	866	3.187477031	CCT8L2
18	6150001	6300000	1230	3.174634126	NUB1
18	6150001	6300000	1230	3.174634126	SMARCD3
18	6150001	6300000	1230	3.174634126	ssc-mir-671
18	6150001	6300000	1230	3.174634126	ssc-mir-671

Table 3.4 Summary of genes identified from ZHp statistics

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
1	275325000	275475000	604	-2.509997346	ABCA1
1	142875000	143025000	674	-2.622020323	ADAL
1	117675000	117825000	61	-2.604127984	ANKDD1A
1	125850000	126000000	151	-2.424550796	AQP9
1	208200000	208350000	201	-2.633745286	ARID4A
1	249375000	249525000	350	-2.650284927	C9orf135
1	249300000	249450000	347	-2.510218989	C9orf135
1	143100000	143250000	642	-3.016905921	CCNDBP1
1	143400000	143550000	162	-2.626513361	CDAN1
1	257700000	257850000	693	-2.889708382	CEP78
1	224775000	224925000	100	-3.571964882	CH242-112J16.10
1	224775000	224925000	100	-3.571964882	CH242-142L3.2
1	180750000	180900000	385	-2.447576575	CILP
1	56550000	56700000	82	-2.918305502	COL19A1
1	56550000	56700000	82	-2.918305502	COL9A1
1	273150000	273300000	356	-2.416273331	CYLC2
1	182025000	182175000	667	-2.514258937	DIS3L
1	182100000	182250000	511	-2.417769988	DIS3L
1	93300000	93450000	576	-2.648522786	DOPEY1
1	110325000	110475000	97	-2.57637249	ELAC1
1	143100000	143250000	642	-3.016905921	EPB42
1	16725000	16875000	310	-2.700269755	ER
1	270000000	270150000	105	-3.186360505	ERP44
1	153300000	153450000	327	-2.438226294	FAM169B
1	146700000	146850000	382	-2.474937309	FSIP1
1	298950000	299100000	612	-2.644732078	GPR144
1	224775000	224925000	100	-3.571964882	IFN-ALPHA-10
1	224775000	224925000	100	-3.571964882	IFN-ALPHA-8
1	224775000	224925000	100	-3.571964882	IFN-DELTA-3
1	224775000	224925000	100	-3.571964882	IFN-OMEGA-2
1	176700000	176850000	447	-2.441615113	KIAA1468
1	142875000	143025000	674	-2.622020323	LCMT2
1	92625000	92775000	732	-2.476281767	LGSN
1	182100000	182250000	511	-2.417769988	MAP2K1
1	182100000	182250000	511	-2.417769988	MAP2K1
1	110325000	110475000	97	-2.57637249	ME2
1	298950000	299100000	612	-2.644732078	NR6A1
1	213075000	213225000	373	-2.669041438	NTRK3

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
1	275325000	275475000	604	-2.509997346	OR13C8
1	180750000	180900000	385	-2.447576575	PARP16
1	176700000	176850000	447	-2.441615113	PIGN
1	257700000	257850000	693	-2.889708382	PSAT1
1	208200000	208350000	201	-2.633745286	PSMA3
1	298950000	299100000	612	-2.644732078	PSMB7
1	249375000	249525000	350	-2.650284927	PTAR1
1	249300000	249450000	347	-2.510218989	PTAR1
1	249450000	249600000	378	-2.473913965	PTAR1
1	92625000	92775000	732	-2.476281767	RIPPLY2
1	182100000	182250000	511	-2.417769988	RPL4
1	182025000	182175000	667	-2.514258937	SCARNA14
1	182100000	182250000	511	-2.417769988	SCARNA14
1	298950000	299100000	612	-2.644732078	SF-1
1	92700000	92850000	655	-2.518844757	SNAP91
1	182100000	182250000	511	-2.417769988	SNAPC5
1	182025000	182175000	667	-2.514258937	SNORA31
1	182100000	182250000	511	-2.417769988	SNORA31
1	182100000	182250000	511	-2.417769988	SNORD16
1	182100000	182250000	511	-2.417769988	SNORD16
1	182100000	182250000	511	-2.417769988	SNORD18
1	182100000	182250000	511	-2.417769988	SNORD18
1	182100000	182250000	511	-2.417769988	SNORD18
1	142800000	142950000	425	-2.54331281	ssc-mir-2366-2
1	270000000	270150000	105	-3.186360505	STX17
1	182025000	182175000	667	-2.514258937	TIPIN
1	182100000	182250000	511	-2.417769988	TIPIN
1	170700000	170850000	141	-2.587161899	TMX3
1	159600000	159750000	196	-2.765014413	TRPM1
1	143250000	143400000	305	-2.718022889	TTBK2
1	143400000	143550000	162	-2.626513361	TTBK2
1	142875000	143025000	674	-2.622020323	TUBGCP4
1	142800000	142950000	425	-2.54331281	TUBGCP4
1	257700000	257850000	693	-2.889708382	U6
1	19650000	19800000	219	-2.733267306	U6
1	208200000	208350000	201	-2.633745286	U6
1	143100000	143250000	642	-3.016905921	UBR1
1	143175000	143325000	567	-2.891655398	UBR1
1	143250000	143400000	305	-2.718022889	UBR1
1	129900000	130050000	116	-2.832360144	UNC13C

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
1	142875000	143025000	674	-2.622020323	ZSCAN29
1	142800000	142950000	425	-2.54331281	ZSCAN29
1	182100000	182250000	511	-2.417769988	ZWILCH
2	73350000	73500000	383	-2.538334632	ACSBG2
2	75000000	75150000	298	-2.419460379	ANKRD24
2	148125000	148275000	623	-2.862156626	APBB3
2	148200000	148350000	703	-2.46633961	APBB3
2	20550000	20700000	491	-2.536896022	API5
2	75000000	75150000	298	-2.419460379	CCDC94
2	148200000	148350000	703	-2.46633961	CD14
2	73050000	73200000	283	-2.664028058	CD70
2	72975000	73125000	221	-2.7403812	CRB3
2	73050000	73200000	283	-2.664028058	CRB3
2	72900000	73050000	125	-2.439501807	CRB3
2	75000000	75150000	298	-2.419460379	CREB3L3
2	72975000	73125000	221	-2.7403812	DENND1C
2	72900000	73050000	125	-2.439501807	DENND1C
2	148200000	148350000	703	-2.46633961	DND1
2	75000000	75150000	298	-2.419460379	EBI3
2	148125000	148275000	623	-2.862156626	EIF4EBP3
2	148200000	148350000	703	-2.46633961	EIF4EBP3
2	148200000	148350000	703	-2.46633961	HARS
2	148200000	148350000	703	-2.46633961	HARS2
2	82650000	82800000	251	-2.432318687	HK3
2	73050000	73200000	283	-2.664028058	KHSRP
2	73350000	73500000	383	-2.538334632	MLLT1
2	73500000	73650000	274	-2.502802959	NRTN
2	73500000	73650000	274	-2.502802959	RFX2
2	72900000	73050000	125	-2.439501807	SCAMC-3
2	75000000	75150000	298	-2.419460379	SIRT6
2	73050000	73200000	283	-2.664028058	SLC25A41
2	148125000	148275000	623	-2.862156626	SLC35A4
2	148200000	148350000	703	-2.46633961	SLC35A4
2	148125000	148275000	623	-2.862156626	SRA1
2	148200000	148350000	703	-2.46633961	SRA1
2	148200000	148350000	703	-2.46633961	TMCO6
2	72975000	73125000	221	-2.7403812	TNFSF14
2	72900000	73050000	125	-2.439501807	TNFSF14
2	72975000	73125000	221	-2.7403812	TNFSF9
2	73050000	73200000	283	-2.664028058	TNFSF9

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
2	20550000	20700000	491	-2.536896022	TTC17
2	72975000	73125000	221	-2.7403812	TUBB4A
2	73050000	73200000	283	-2.664028058	TUBB4A
2	148125000	148275000	623	-2.862156626	U6
2	73950000	74100000	164	-2.548051054	U6
2	73350000	73500000	383	-2.538334632	U6
2	148200000	148350000	703	-2.46633961	U6
2	82650000	82800000	251	-2.432318687	UNC5A
2	148200000	148350000	703	-2.46633961	WDR55
3	80475000	80625000	211	-2.766691313	ACTR2
3	125400000	125550000	260	-2.766080623	C2orf43
3	74625000	74775000	743	-2.805889963	DYSF
3	74550000	74700000	572	-2.754711604	DYSF
3	17775000	17925000	253	-2.536304701	FBXL19
3	125400000	125550000	260	-2.766080623	GDF7
3	17775000	17925000	253	-2.536304701	HSD3B7
3	54300000	54450000	422	-2.565019051	IL1RL1
3	121275000	121425000	600	-2.454228355	NCOA1
3	17775000	17925000	253	-2.536304701	ORAI3
3	134250000	134400000	270	-2.850231872	ROCK2
3	17775000	17925000	253	-2.536304701	SETD1A
3	134250000	134400000	270	-2.850231872	SNORA31
3	17775000	17925000	253	-2.536304701	STX1B
4	134700000	134850000	1280	-2.487077038	ABCA4
4	90750000	90900000	491	-2.767616188	ADCY10
4	90675000	90825000	205	-2.600548451	ADCY10
4	119550000	119700000	538	-2.515269021	CD53
4	37500000	37650000	425	-2.558306192	CU459197.3
4	37500000	37650000	425	-2.558306192	CU459197.4
4	90675000	90825000	205	-2.600548451	DCAF6
4	90600000	90750000	143	-2.425401394	DCAF6
4	74775000	74925000	432	-2.711234518	DNAJC5B
4	74700000	74850000	414	-2.560536406	DNAJC5B
4	119550000	119700000	538	-2.515269021	DRAM2
4	46350000	46500000	179	-3.085747694	FAM92A1
4	46275000	46425000	459	-2.825935515	FAM92A1
4	119550000	119700000	538	-2.515269021	LRIF1
4	90750000	90900000	491	-2.767616188	MPC2
4	90675000	90825000	205	-2.600548451	MPC2
4	90600000	90750000	143	-2.425401394	MPC2

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
4	74925000	75075000	402	-2.47921895	MTFR1
4	74925000	75075000	402	-2.47921895	PDE7A
4	91350000	91500000	168	-2.469159485	POU2F1
4	46350000	46500000	179	-3.085747694	RBM12B
4	46275000	46425000	459	-2.825935515	RBM12B
4	90750000	90900000	491	-2.767616188	SNORD70
4	83850000	84000000	1032	-2.525149858	SOX17
5	24750000	24900000	185	-3.02492031	GLI1
5	81750000	81900000	478	-2.46911949	H1FNT
5	24750000	24900000	185	-3.02492031	INHBE
5	24750000	24900000	185	-3.02492031	KIF5A
5	24750000	24900000	185	-3.02492031	MBD6
6	99675000	99825000	799	-2.440515233	GATA6
6	99000000	99150000	231	-2.417366087	GREB1L
6	102000000	102150000	471	-2.834205626	HRH4
6	102075000	102225000	488	-2.817881938	HRH4
6	101925000	102075000	328	-2.622488424	HRH4
6	102000000	102150000	471	-2.834205626	IMPACT
6	101925000	102075000	328	-2.622488424	IMPACT
6	32475000	32625000	536	-2.708045281	ITFG1
6	99525000	99675000	259	-3.043080222	MIB1
6	32325000	32475000	399	-2.903191337	PHKB
6	32400000	32550000	649	-2.825176459	PHKB
6	32250000	32400000	170	-2.73967565	PHKB
6	32250000	32400000	170	-2.73967565	U6
6	32250000	32400000	170	-2.73967565	U6
6	32475000	32625000	536	-2.708045281	U6
6	101925000	102075000	328	-2.622488424	U6
7	59475000	59625000	179	-2.678565894	ACAN
7	28875000	29025000	171	-3.569028964	BTNL2
7	28875000	29025000	171	-3.569028964	BTNL2
7	28875000	29025000	171	-3.569028964	BTNL3
7	28875000	29025000	171	-3.569028964	BTNL4
7	62925000	63075000	595	-2.522903784	C15orf39
7	80250000	80400000	447	-2.455511303	CIDE-B
7	66525000	66675000	448	-2.917870794	CLEC14A
7	62850000	63000000	506	-2.790986882	COMMD4
7	62925000	63075000	595	-2.522903784	COMMD4
7	80325000	80475000	427	-2.516155888	DCAF11
7	80250000	80400000	447	-2.455511303	DHRS1

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
7	80325000	80475000	427	-2.516155888	EMC9
7	69600000	69750000	185	-2.5634363	FAM177A1
7	80325000	80475000	427	-2.516155888	FITM1
7	80325000	80475000	427	-2.516155888	GMPR2
7	80250000	80400000	447	-2.455511303	GMPR2
7	59475000	59625000	179	-2.678565894	HAPLN3
7	80325000	80475000	427	-2.516155888	IPO4
7	80250000	80400000	447	-2.455511303	IPO4
7	80325000	80475000	427	-2.516155888	IRF9
7	62850000	63000000	506	-2.790986882	MAN2C1
7	62925000	63075000	595	-2.522903784	MAN2C1
7	59475000	59625000	179	-2.678565894	MFGE8
7	80325000	80475000	427	-2.516155888	NEDD8
7	80250000	80400000	447	-2.455511303	NEDD8
7	62850000	63000000	506	-2.790986882	NEIL1
7	62925000	63075000	595	-2.522903784	NEIL1
7	80250000	80400000	447	-2.455511303	NOP9
7	80325000	80475000	427	-2.516155888	PCK2
7	122325000	122475000	79	-4.507189051	PPP4R4
7	80325000	80475000	427	-2.516155888	PSME1
7	80325000	80475000	427	-2.516155888	PSME2
7	61725000	61875000	618	-2.798167046	PSTPIP1
7	62850000	63000000	506	-2.790986882	PTPN9
7	62775000	62925000	375	-2.45801889	PTPN9
7	80250000	80400000	447	-2.455511303	RABGGTA
7	10200000	10350000	431	-2.490892302	RANBP9
7	61725000	61875000	618	-2.798167046	RCN2
7	61650000	61800000	420	-2.677180731	RCN2
7	80325000	80475000	427	-2.516155888	REC8
7	80250000	80400000	447	-2.455511303	REC8
7	61725000	61875000	618	-2.798167046	SCAPER
7	61650000	61800000	420	-2.677180731	SCAPER
7	62850000	63000000	506	-2.790986882	SIN3A
7	62925000	63075000	595	-2.522903784	SIN3A
7	62775000	62925000	375	-2.45801889	SIN3A
7	10200000	10350000	431	-2.490892302	SIRT5
7	69600000	69750000	185	-2.5634363	SRP54
7	66525000	66675000	448	-2.917870794	SSTR1
7	80250000	80400000	447	-2.455511303	TGM1
7	80325000	80475000	427	-2.516155888	TINF2

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
7	80250000	80400000	447	-2.455511303	TINF2
7	80325000	80475000	427	-2.516155888	TSSK4
7	80250000	80400000	447	-2.455511303	TSSK4
7	66975000	67125000	376	-2.441244057	TTC6
7	69600000	69750000	185	-2.5634363	U2
7	62850000	63000000	506	-2.790986882	U6
7	62850000	63000000	506	-2.790986882	U6
7	62925000	63075000	595	-2.522903784	U6
7	62775000	62925000	375	-2.45801889	U6
8	49200000	49350000	312	-2.734691918	FAM198B
8	49275000	49425000	509	-2.511442634	FAM198B
8	59025000	59175000	600	-2.467661903	IGFBP7
8	12675000	12825000	148	-2.919543725	LCORL
8	59025000	59175000	600	-2.467661903	POLR2B
8	49275000	49425000	509	-2.511442634	TMEM144
9	80625000	80775000	131	-2.768880176	GNG11
9	96825000	96975000	40	-3.146185677	HDAC9
9	101925000	102075000	364	-2.4616533	RUNDC3B
9	101925000	102075000	364	-2.4616533	SLC25A40
9	80625000	80775000	131	-2.768880176	TFPI2
9	104400000	104550000	118	-2.806679231	U6
10	72000000	72150000	271	-3.521459456	AKR
10	72000000	72150000	271	-3.521459456	AKR1C3
10	72000000	72150000	271	-3.521459456	AKR1E2
11	53775000	53925000	605	-2.567688099	FBXL3
11	5325000	5475000	1237	-2.591870306	FLT1
11	5250000	5400000	904	-2.469895661	FLT1
11	53775000	53925000	605	-2.567688099	MYCBP2
11	5250000	5400000	904	-2.469895661	PAN3
12	55650000	55800000	344	-2.595260577	ALOX12B
12	55725000	55875000	643	-2.501584854	ALOX12B
12	55650000	55800000	344	-2.595260577	ALOX15B
12	55725000	55875000	643	-2.501584854	ALOX15B
12	20925000	21075000	1158	-2.486893962	CNP
12	20925000	21075000	1158	-2.486893962	DHX58
12	20925000	21075000	1158	-2.486893962	DNAJC7
12	55725000	55875000	643	-2.501584854	HES7
12	20925000	21075000	1158	-2.486893962	HSPB9
12	20925000	21075000	1158	-2.486893962	KAT2A
12	20925000	21075000	1158	-2.486893962	NKIRAS2

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
12	55725000	55875000	643	-2.501584854	PER1
12	20925000	21075000	1158	-2.486893962	RAB5C
12	55725000	55875000	643	-2.501584854	TMEM107
12	55725000	55875000	643	-2.501584854	U8
12	20925000	21075000	1158	-2.486893962	ZNF385C
13	142950000	143100000	646	-2.842563807	CEP19
13	143025000	143175000	616	-2.825142077	CEP19
13	168600000	168750000	56	-3.470693768	CMSS1
13	133950000	134100000	373	-2.504193372	EIF4A2
13	133950000	134100000	373	-2.504193372	KNG1
13	143025000	143175000	616	-2.825142077	NRROS
13	137850000	138000000	211	-2.497064466	OSTN
13	142950000	143100000	646	-2.842563807	PIGX
13	143025000	143175000	616	-2.825142077	PIGX
13	133950000	134100000	373	-2.504193372	RFC4
13	133950000	134100000	373	-2.504193372	SNORA27
13	133950000	134100000	373	-2.504193372	SNORA4
13	133950000	134100000	373	-2.504193372	SNORA63
13	133950000	134100000	373	-2.504193372	SNORA63
13	133950000	134100000	373	-2.504193372	SNORA81
13	133950000	134100000	373	-2.504193372	SNORD2
14	125175000	125325000	536	-2.836641943	CCDC147
14	125175000	125325000	536	-2.836641943	GSTO1
14	125100000	125250000	510	-2.643932549	GSTO1
14	125175000	125325000	536	-2.836641943	ITPRIP
14	125100000	125250000	510	-2.643932549	ITPRIP
14	21525000	21675000	485	-2.649003133	NEK1
14	8250000	8400000	217	-3.886512777	RHOBTB2
14	124950000	125100000	483	-2.525979407	SFR1
14	125100000	125250000	510	-2.643932549	WDR96
14	125025000	125175000	413	-2.541185523	WDR96
14	124950000	125100000	483	-2.525979407	WDR96
15	102450000	102600000	439	-3.033991467	CALCRL
15	102375000	102525000	407	-2.488501478	CALCRL
15	116025000	116175000	118	-2.601632964	CASP10
15	116850000	117000000	139	-3.397647527	CDK15
15	115950000	116100000	116	-2.621655672	FAM126B
15	116025000	116175000	118	-2.601632964	FAM126B
15	116025000	116175000	118	-2.601632964	FLIP-L
15	66375000	66525000	319	-3.27084113	GPR17

Chromosome	Window Start (bp)	Window end (bp)	Number of SNPs	ZHp	Gene symbol
15	66300000	66450000	221	-2.429047775	GPR17
15	69450000	69600000	324	-2.83600745	KCNJ3
15	66375000	66525000	319	-3.27084113	LIMS2
15	66300000	66450000	221	-2.429047775	LIMS2
15	106500000	106650000	323	-2.488298286	NAB1
15	115950000	116100000	116	-2.621655672	NDUFB3
15	116025000	116175000	118	-2.601632964	NDUFB3
15	115950000	116100000	116	-2.621655672	ORC2
15	66375000	66525000	319	-3.27084113	SFT2D3
15	66300000	66450000	221	-2.429047775	SFT2D3
15	81075000	81225000	147	-2.813044402	SNORA70
15	102450000	102600000	439	-3.033991467	TFPI
15	106500000	106650000	323	-2.488298286	TMEM194B
15	102450000	102600000	439	-3.033991467	U5
15	102375000	102525000	407	-2.488501478	U5
15	102225000	102375000	438	-2.427414421	U5
15	113700000	113850000	94	-2.925491381	U6
15	113775000	113925000	199	-2.776552048	U6
15	65850000	66000000	276	-2.455340783	U6
15	157650000	157681499	261	-2.44866488	U6
15	124200000	124350000	841	-2.425539097	UNC80
15	66375000	66525000	319	-3.27084113	WDR33
15	66300000	66450000	221	-2.429047775	WDR33
16	58500000	58650000	359	-2.958143711	FAM196B
16	58425000	58575000	425	-2.714676226	FAM196B
17	15750000	15900000	541	-2.454379957	CHGB
17	15750000	15900000	541	-2.454379957	TRMT6
18	31275000	31425000	128	-3.080725858	CAPZA2
18	21000000	21150000	895	-2.49595663	FAM71F2
18	21000000	21150000	895	-2.49595663	IMPDH1
18	24300000	24450000	409	-2.800200503	POT1
18	49725000	49875000	40	-3.104739473	ssc-mir-196b-2
18	600000	750000	772	-2.431383076	U6
18	600000	750000	772	-2.431383076	VIPR2

Next, I performed gene enrichment analysis using DAVID gene enrichment test and obtained two significant biological process terms ($p<0.05$; Table 3.5) and five significant ($p<0.05$) KEEG pathways (Table 3.6). Among the significantly enriched KEGG pathways is central carbon metabolism in cancer (ssc05230, $p=1.09E-04$). In this pathway, genes relevant to the determination of coat color, meat quality, and feed intake phenotypes are annotated.

Table 3.5 DAVID biological process terms, functional annotation clustering of genes obtained. (FDR<0.05).

Category	Term	Count	%	Genes
GOTERM_BP_DIRECT	GO:0007601~visual perception	6	1.754386	DRAM2, SOX14, GRK7, ABCA4, DNAJC19, TRPM1
GOTERM_BP_DIRECT	GO:0007493~endodermal cell fate determination	2	0.584795	GATA6, SOX17

Table 3.6 KEGG pathways enriched from genes identified as positively selected in Korean Imported Pig breeds from ZF_{ST} and ZHp statistics.

Term	P-Value	Genes	Fold Enrichment
ssc05230:Central carbon metabolism in cancer	1.09E-04	NTRK3, SLC1A5, KRAS, MAP2K1, HK3, PIK3CA, SIRT6, KIT	7.142494
ssc03050:Proteasome	0.008312	PSMB7, PSMA6, PSME1, PSME2, PSMA3	6.150481
ssc05211:Renal cell carcinoma	0.027365	KRAS, MAP2K1, GAB1, EGLN2, PIK3CA	4.324557
ssc04014:Ras signaling pathway	0.043533	KRAS, FLT1, MAP2K1, RAB5C, GAB1, PIK3CA, GNG11, KIT, KDR	2.264495
ssc04725:Cholinergic synapse	0.048672	KRAS, MAP2K1, PIK3CA, GNG11, CREB3L3, KCNJ3	2.992126

The ClueGO Cytoscape plugin analysis, using "use GO Term Fusion option", clustered genes involved in different biological functions. Fibroblast growth factor receptor signaling pathway, regulation of Notch signaling pathway, and regulation of fibroblast apoptotic process are clusters related to reproduction traits at different stages of development. The Notch signaling pathway plays a key role in cell-cell communication and further, regulates embryonic development. The cluster term "long-chain fatty acid metabolic process" is related to meat quality (Zhang, Yang et al. 2016).

The sweep regions identified in this analysis harbor genes involved in several biological functions including immune function (e.g., *EPB42, ABCB5, IL1RL1, TGM1, IGFBP7, CD70, NRROS, TNFSF14, TNFSF9*), coat color (e.g., *MC1R, KIT, KRAS, DNAJC5B*), meat production and quality (e.g., *CAPZA2, COL9A1, COL19A1, FITM1, ROCK2, EPB42, TMX3*), body size (*LCORL*), reproduction function (e.g., *ER, PLSCR4, AGTR1, PDGFRA*), and residual feed intake (*EPB42*). Mutations in *MC1R* and *KIT* genes have been reported to affect coat color in pigs (Klomtong, Chaweewan et al. 2015). *EPB42* is an erythrocyte membrane protein which is involved in the regulation of erythrocyte shape and mechanical properties. This gene have been previously reported upregulated in the low residual feed intake pigs (Vincent, Louveau et al. 2015) and affect meat pH (Zambonelli, Davoli et al. 2013). *TMX3* is found selected in pigs with an effect on eye development and body growth (Zhou, Li et al. 2016); visual perception is found associated with the growth of pigs.

In analysis, several genes associated with reproduction traits were detected (Table 3.7). *PLSCR4* is a membrane protein linked to uterine function and ovulation. It was previously found expressed in the endometrial and myometrial layers of pregnant rat uterus (Phillippe, Bradley et al. 2006). In a whole-genome association study of reproductive traits in pig genome, *PLSCR4* is found to be associated with a total number of piglets born, and the number of those born alive (Onteru, Fan et al. 2012). Figure 3.2 shows the plot of ZF_{ST} and ZHp values of *PLSCR4* gene region. In this gene region, I identified two missense variants (rs320433969: p.Val264Ile, and rs336494357: p.Ser332Thr). *AGTR1* plays a fundamental role in follicular development, deviation, atresia, and ovulation in a species-specific manner (Gonçalves, Ferreira et al. 2012). It is a physiological co-factor necessary for the expression of genes in granulosa cells that are critical for ovulation. It also stimulates the meiotic maturation of ovulated ova and follicular oocytes in the absence of gonadotropin in rabbit ova (Yoshimura, Karube et al.

1996). The ovarian follicular development involves cell proliferation and angiogenesis to which PDGFs and their receptors play a crucial role. *PDGFRA* is a receptor found localized in the oocyte, theca and pre-granulosa/granulosa of the rat ovary with a key function for the development of the ovary and the follicle (Sleer and Taylor 2007). It is found expressed in corpus luteum of a rat localized in the luteal parenchymal and vascular cells (Sleer and Taylor 2007). This gene is found linked with Fec^B gene in the Booroola Merino ewe known for its high fecundity rate (Baird and Campbell 1998). *TFPI2*, called placental protein 5, was significantly overexpressed during pig follicular development (Bonnet, Le Cao et al. 2008).

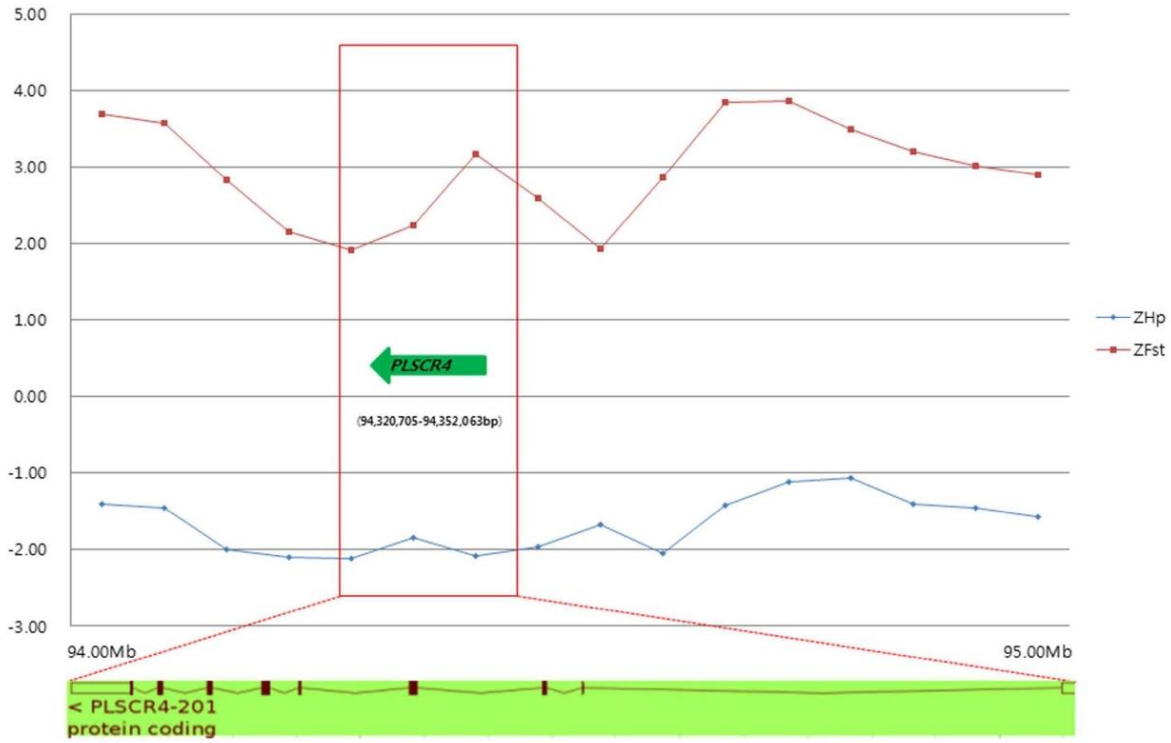


Figure 3.2 Plot of the ZFst and ZHp value of *PLSCR4* gene region. The box in the plot indicates the gene region and the points are the ZFst and ZHp values overlapped within 50 kb region.

AGTR1 regulate placental development and generation of extravillous trophoblasts (Tower, Lui et al. 2010). Placenta, an organ that connects the developing fetus to the mother, is required for the development of an embryo. It allows nutritional and gas exchanges between the fetus and the maternal organism. It has endocrinological and immunological functions that are essential in pregnancy and for fetal growth (Tarrade, Kuen et al. 2001). *CORIN* plays a role in female pregnancy by promoting trophoblast invasion and spiral artery remodeling in the uterus (Cui, Wang et al. 2012, Soares, Chakraborty et al. 2014). It is expressed in the pregnant mouse and human uterus to which its impaired expression is associated with preeclampsia, a major risk factor for placental abruption (Cui, Wang et al. 2012, Nagashima, Li et al. 2013). I scanned the gene region for non-synonymous mutations and identified 16 missense variants in this gene region (Table 3.8). *IGF2BP3* is expressed in placenta, testis, and various cancers, and plays a role in trophoblast differentiation during placental development which is a prerequisite for successful pregnancy (Li, Liu et al. 2014). Placental function influences the health of the fetus and contributes to uterine capacity (Vallet and Freking 2007).

APOB encodes two versions of apolipoprotein B to which both are components of lipoproteins that carry fats and fat-like substances (such as cholesterol) in the blood. They are constituents of lipid that are important for male gamete membrane and liquid content (Peterlin, Zorn et al. 2006). *APOB* is expressed in the testis and epididymis in mouse, and polymorphisms in *APOB* is associated with male infertility in humans (Peterlin, Zorn et al. 2006) and mouse (Huang, Voyiaziakis et al. 1996). Claudins function as major constituents of the tight junction complexes that regulate the permeability of epithelia. A gene called *CLAUDIN1* is expressed in the epididymis. The epididymis is responsible for post-testicular sperm maturation, transport, protection and storage to which the sperm gets its motility and the ability to fertilize. That said, *CLAUDIN1* is involved in the formation of functional tight junctions to which its malfunction results in epididymal dysfunction leading to male infertility in humans (Dubé, Dufresne et al. 2010). Previous reports suggest it is under selection in Landrace and Yorkshire breeds (Choi, Chung et al. 2015). *PON1* is a high-density lipoprotein-associated enzyme that prevents low-density lipoprotein oxidation and its polymorphism is associated with both male and female infertility (Marsillach, Checa et al. 2010, Lazaros, Xita et al. 2011). It is localized in the seminiferous tubules and spermatozoa and have been implicated in the pathogenesis of male infertility. *DNAJC5B,* a testicular

tissue protein Li 55, is a heat shock protein to which its polymorphism is associated with improved fertilization rate and/or improved embryo survival rate in cattle (Zhang, unpublished), and white coat color in Yorkshire pigs (Moon, Kim et al. 2015). ER, estrogen receptor, is a nuclear receptor family of transcription factors with key functions in reproduction and fertility. Estrogen controls many cellular processes including growth, differentiation, and function of the reproductive system (Lazari, Lucas et al. 2009). *MFGE8* is a sperm surface protein involved in fertilization. It is expressed in the epididymis, oviduct, and uterus of the pig. Polymorphism in this gene has been previously found associated with fertility traits in Holstein cattle (Fontanesi, Calò et al. 2014). A mutation in *MFGE8* gene causes a protein change (rs327367193: p.Ser272Asn).

SSTR1 is a receptor for somatostatin (SRIH) which functions as an endocrine signaling for growth, immune resistance and reproduction (Geris, De Groef et al. 2003). *Sstr1* is found upregulated in the brain of postpartum mice compared to virgin females which is associated with postpartum process, maternal behavior (Zhao, Saul et al. 2012). One missense SNP was identified to change the amino acid in the SSTR1 gene (rs345286477: p.Thr33Ala).

Taken all together, KIP breeds are productive with better growth, fecundity, and other reproductive traits. They have been reported to have larger litter size compared to those of KNP breeds (Kim, Yeo et al. 2002, Choi, Chung et al. 2015). The number of ova shed during ovulation, fertility, uterine capacity, and embryo mortality are determinant factors on the number of piglets born alive per sow (Haley and Lee 1992). The genes identified as positively selected in KIP breeds in relation to reproductive traits might contribute to their superior reproductive performances and overall productivity. In addition to genes that contribute to larger litter size, postnatal maternal care is important for proper piglet growth when KIP are used as a maternal line.

Table 3.7 Candidate genes affecting reproduction traits in Korean Imported Pig breeds detected as positively selected based on ZF_{ST} and ZHp.

Chr.	Start	End	Gene	ZF_{ST}	ZHp
1	16725000	16875000	ER	-	-2.70
3	125250000	125400000	APOB	3.16	-
4	74700000	74850000	DNAJC5B	3.28	-2.56
7	59475000	59625000	MFGE8	-	-2.68
7	66525000	66675000	SSTR1	3.62	-2.92
8	39675000	39825000	CORIN	3.92	-
8	42900000	43050000	PDGFRA	4.16	-
9	80550000	80700000	TFPI2	4.86	-2.77
9	82275000	82425000	PON1	3.37	-
9	101700000	101850000	IGF2BP3	3.57	-
13	94350000	94500000	PLSCR4	3.17	-
13	96900000	97050000	AGTR1	3.96	-
13	136950000	137100000	CLAUDIN1	3.19	-

ZF st: The Z transformed Fst (weighted Fst) value; ZHp: The Z transformed heterozygosity

Table 3.8 Causative variants of candidate genes in Korean imported pig breeds.

Gene	Chr.	Position	SNP_id	Ref	Alt	Protein change
MFGE8	7	59510271	rs327367193	C	T	p.Ser272Asn
SSTR1	7	66597098	rs345286477	T	C	p.Thr33Ala
CORIN	8	39506034	.	C	A	p.Gly1097Val
		39506048	.	C	A	p.Lys1092Asn
		39506063	rs327585581	C	G	p.Arg1088Pro
		39506090	rs340997149	C	G	p.Cys1079Ser
		39506118	rs318444517	C	T	p.Ala1070Thr
		39506121	rs329582727	C	A	p.Gly1069Trp
		39506158	rs321296418	C	A	p.Arg1056Ser
		39516314	rs329597137	C	T	p.Arg1019His
		39518279	rs324332716	C	T	p.Arg957His
		39518328	rs335001283	T	G	p.Met941Leu
		39572176	.	G	A	p.Pro653Leu
		39666192	rs343167087	T	C	p.Ile323Val
		39732066	rs339634644	T	C	p.Glu200Gly
		39732130	rs322282133	T	C	p.Arg179Gly
		39732168	rs324743354	C	T	p.Gly166Asp
		39781485	rs330849429	G	A	p.Pro37Ser
PLSCR4	13	94322827	rs336494357	A	T	p.Ser332Thr
		94323823	rs320433969	C	T	p.Val264Ile

Chr. – chromosome; Position – causal snp position; Snp_id – causal snp id; Ref – reference nucleotide; alt – alternative nucleotide

3.5 Discussion

I identified several genes under positive selection in imported pig breeds that are widely used in Korea. These genes are involved in several biological functions such as immune response, growth, reproduction, and coat color determination. This result will help in improving our understanding of biological mechanisms and pathways that are related to the phenotypes of these breeds.

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

Chapter 4. Analysis of Cross-Population Differentiation Between Thoroughbred and Jeju Horses

4.1 Abstract

This study was intended to identify genes positively selected in Thoroughbred horses (THBs) that potentially contribute to their running performances. THB is one of the fastest horse breeds and has the larger eyes in horse breed. However, the relation between their speed and eye size has not identified yet. Therefore, I studied to reveal this relationships comparing the genomes of THB and Jeju horses (JH, Korean native horse). I performed cross-population extended haplotype homozygosity (XP-EHH) and cross-population composite likelihood ratio test (XP-CLR) statistical methods for analysis using whole genome resequencing data of 14 THB and 6 JH. As a result, I identified 98 (XP-EHH) and 200 (XP-CLR) genes that are under positive selection in THB. Gene enrichment analysis identified 72 BP terms. The genes and BP terms explained some of THB's characteristics such as immunity, energy metabolism and eye size and function related to running performances. GO terms that play key roles in several cell signaling mechanisms, which affected ocular size and visual functions were identified. GO term Eye photoreceptor cell differentiation is among the terms annotated presumed to affect eye size. This analysis revealed some positively selected candidate genes in THB related to their racing performances. The genes detected are related to the immunity, ocular size and function, and energy metabolism.

4.2 Introduction

Horses were domesticated 6,000 years ago in the Eurasian steppe (Gu, Orr et al. 2009). Domestication and artificial selection strongly affected differentiation of horse breeds to increase horse capacity related to racing or packing type. Especially, Thoroughbreds (THB) became an outstanding horse breed for racing preferable to any other horse breed. The athletic performance of THB has come from the intense selection that resulted in different anatomical and physiological characteristics (Hinchcliff, Kaneps et al. 2008). Among the physiological characteristics, typical of THB are large muscle mass to body weight ratio, high skeletal muscle mitochondrial density and oxidative enzyme activity, and considerable intramuscular stores of energy substrates (Hinchcliff, Kaneps et al. 2008). The anatomical characteristics of THB are their long legs and a lean body (Montgomery 1971). In addition to these characteristics, THB have larger eyes compared to their relatives (Howland, Merola et al. 2004) which might contribute to their running performances. Ocular size is hypothesized to have an effect on running speed in animals. According to Leuckart's Law (Heard-Booth and Kirk 2012), animals capable of achieving fast speeds require large eyes to enhance visual acuity and avoid collisions with obstacles in their environment. This law is an empirical law in zoology and applied to vertebrate animals (Heard-Booth and Kirk 2012).

Selective sweep is among the major factors which can increase genetic differentiation between two populations and causes allele frequency spectra to depart from expectation under neutrality (Chen, Patterson et al. 2010). Most methods of identifying evidence of positive selection are based on the decay of linkage disequilibrium (LD) and distortion in the variation of allele frequency spectra (Ma, Zhang et al. 2014). Using heterozygosity statistics, Gu et al. (Gu, Orr et al. 2009) reported the positive selection of candidate athletic-performance gene regions that are responsible for fatty acid oxidation, increased insulin sensitivity and muscle strength in thoroughbred horses. In another study, Park et al. (Park, Kim et al. 2014) identified positively selected genes related to exercise response in horses using cross-population extended haplotype homozygosity (XP-EHH) method. In this study, I used XP-EHH (Sabeti, Varilly et al. 2007) and cross-population composite likelihood ratio method (XP-CLR) (Chen, Patterson et al. 2010) methods to test for signatures of selective sweeps in THB. XP-EHH calculates haplotype decay separately for each group using the EHH

(Sabeti, Varilly et al. 2007) and XP-CLR is a likelihood method for detecting selective sweeps using jointly modeling the multilocus allele frequency differentiation between the two groups (Ma, Zhang et al. 2014). XP-CLR provides higher power than other approaches to detect selective sweeps and good localization of the selected allele. Additionally, it has been reported that XP-CLR is much more robust to ascertainment bias in SNP discovery than methods based on the allele frequency spectrum (Chen, Patterson et al. 2010).

Here, using XP-EHH and XP-CLR population statistics, I compared THB and Jeju horse (JH) populations to identify positive selection sweep regions in THB. JH is a Korean native breed in Jeju Island located far south of the Korean peninsula. They are hardy with a small to medium body size (Kim, Yang et al. 1999). JH are general breeds that have been raised for several purposes as riding, racing, and meat, and not intensively selected for a special purpose (Chang-Yeon, Sung-Heum et al. 2008).

4.3 Materials and Methods

4.3.1 Samples and ethics statement

Blood samples were collected from THB and JH horses by trained veterinarians according to relevant international as well as national guidelines and under permission from the Guide for the Care and Use of Laboratory Animals of Pusan National University. All experimental procedures used in this study were approved by the Institutional Animal Care and Use Committee of the Pusan National University (PNU-2013-0417).

4.3.2 Pre-processing of DNA resequencing data

Whole-blood samples (10 ml) were collected from 14 THB and 6 JH. Sequence data of these 20 samples were generated using the Illumina HiSeq2000 platform. The DNA sequencing data has been submitted to the NCBI Sequence Read Archive (SRA) database with accession numbers (SRS345323 to SRS345338 and SRS346577 to SRS346580) (Park, Kim et al. 2014).

Then, I carried out a base sequence quality check using the fastQC (ver 0.10) software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). I removed the potential adapter sequence using Trimmomatic-0.32. Paired-end sequence reads were mapped to the reference Equus caballus (ver 2.66) genome using Bowtie2 (Langmead and Salzberg 2012) with the default setting. The overall alignment rate of reads to the reference sequence was 94.58 % (91.24 % to 98.76 %) with an average read depth of 15.87 x (12.13 x to 22.26 x). On average across the whole samples, the reads covered 97.66 % (97.53 % to 97.77 %) of the genome. For downstream processing and variant calling, I used open-source software packages. Using Picard (ver 1.56) potential PCR duplicates were filtered. Then, I used SAMtools (ver 0.1.18) (Li, Handsaker et al. 2009) to make index files for reference and bam files. After preparation of these files, Genome Analysis ToolKit 1.4 (GATK) was used to perform local realignment of reads to correct misalignments due to the presence of indels (Realigner Target Creator and In del Realigner arguments) (McKenna, Hanna et al. 2010). The Unified Genotyper and Select

Variants arguments of GATK were used for calling candidate single nucleotide polymorphisms (SNPs). To filter variants and remove possible false positives, option "VariantFiltration" was adopted with the following command options: 1) All SNPs with a phred-scaled quality score of less than 30 and with MQ0 (mapping quality zero); 2) Total count across all samples of mapping quality zero reads) >4 were filtered; 3) Quality depth (unfiltered depth of non-reference samples; low scores are indicative of false positives and artifacts) less than 5 were filtered; and 4) SNPs with FS (phred-scaled P value using Fisher's exact test) >200 were filtered as FS represents variation on either the forward or the reverse strand, which are implied of false-positive calls. After this, it remained with ~12.9 million autosomal SNPs. These SNPs Ire phased and imputed using BEAGLE Version 3.3.2 (Browning and Browning 2007).

4.3.3 Population structure analysis

For principal component analysis (PCA), I used the genome-wide complex trait analysis (GCTA) (Yang, Lee et al. 2011) to estimate the eigenvectors incorporating genotype data from THB and JH. Structure admixture analysis between the two breeds was performed. I limited the genotype data to a random subset of approximately 0.1% of total SNPs using PLINK (-thin option) (Price, Patterson et al. 2006, Purcell, Neale et al. 2007) and conducted the STRUCTURE (ver 2.3.4) with 2 options: the "admixture model" and K=2. Then I used ancestry graphs implemented Treemix 1.12 (Pickrell and Pritchard 2012) to show the historical relationship between these two populations, using –m flag option in this study to infer migration events with 1,000 replicated bootstraps.

4.3.4 Selective sweep analysis and gene annotation

I performed two analyses to detect positive selection signatures in THB population. Whole SNP sets were used from both THB and JH for the analysis. Initially, the XP-EHH that measures cross-population extended haplotype homozygosity was used to identify positive selection regions. The calculation for XP-EHH was performed using the software xpehh [(Sabeti, Varilly et al. 2007); http://hgdp.uchicago.edu/Software/]. I

assumed that genetic distance was equal to physical distance. These log ratios (unstandardized XP-EHH) were standardized to have a mean of zero and a variance of one. Then, I split the genome into non-overlapping segments of 50 kb and computed the maximum XP-EHH score in each segment. Top 1% regions with high XP-EHH values were considered strong signals in the THB population.

Next, the cross-population composite likelihood ratio (XP-CLR) (ver 1.0) test for detecting selective sweep regions that involve jointly modeling the multilocus allele frequency between two populations (Chen, Patterson et al. 2010). Whole SNP sets were used from both THB and JH for the analysis. The parameters used were as follows: Non-overlapping sliding windows of 50kb, the maximum number of SNPs within each window as 400, and correlation level from which the SNPs contribution to XP-CLR result was down-weighted 0.95 following Lee *et al*. (Lee, Kim et al. 2014). The regions with the XP-CLR values in the top 1% using XP-CLR score were designated candidate sweeps. Significant genomic regions identified from XP-EHH and XP-CLR were annotated to nearby genes (Equ cab 2). Genes that are located (partially or completely) in the window regions were presumed as candidate genes (Lee, Kim et al. 2014).

DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool was used for annotation and pathway analyses. In addition, using these positively selected genes, ClueGO plugin of Cytoscape was used to cluster by gene ontology and visualized them (Bindea, Mlecnik et al. 2009).

4.4 Results

4.4.1 DNA re-sequencing

From the re-sequencing of DNA from 14 THB and 6 JH whole genome, I obtained sequencing approximately 15.87x coverages on average, with a total of approximately 39 billion bp in 40 million reads per sample. Sequence reads of each sample were aligned with an overall alignment rate of 94.58 % of the whole genome area.(Table 4.1) I finally obtained a total of ~12.9 million autosomal SNPs used for sweep analysis.

Table 4.1. Summary of sequencing data

Sample ID	DNA Sequenced (bp)	Total Reads	Alignment Rate (%)	Read Depth	Genome Coverage (%)
Thoroughbred 1	37,246,295,613	382,315,321	93.73	15.05	97.73
Thoroughbred 2	30,888,431,883	325,731,309	91.24	12.48	97.64
Thoroughbred 3	31,450,675,227	328,079,041	92.32	12.72	97.62
Thoroughbred 4	35,963,836,413	372,542,460	92.89	14.54	97.72
Thoroughbred 5	33,694,868,912	347,933,273	93.24	13.63	97.69
Thoroughbred 6	35,377,479,583	366,035,305	93	14.3	97.69
Thoroughbred 7	34,425,811,410	359,701,273	92.11	12.13	97.68
Thoroughbred 8	31,609,589,758	328,493,910	92.63	12.78	97.6
Thoroughbred 9	30,000,054,254	309,864,749	93.15	12.13	97.59
Thoroughbred 10	37,775,043,083	375,306,333	96.94	15.29	97.72
Thoroughbred 11	33,509,411,544	350,176,922	92.08	13.54	97.72
Thoroughbred 12	41,399,035,283	420,765,583	94.66	16.73	97.72
Thoroughbred 13	44,528,923,038	428,687,902	96.24	18.01	97.73
Thoroughbred 14	46,311,892,576	458,855,646	97.1	18.72	97.77
Jeju Horse1	35,538,209,275	369,702,396	92.5	14.36	97.72
Jeju Horse2	33,926,243,752	349,622,049	93.38	13.71	97.66
Jeju Horse3	52,145,917,996	535,283,964	98.7	21.13	97.57
Jeju Horse4	54,694,350,831	562,662,095	98.49	22.16	97.53
Jeju Horse5	54,931,054,131	564,955,386	98.51	22.26	97.55
Jeju Horse6	53,495,854,448	548,708,850	98.76	21.68	97.57

4.4.2 Population structure

I performed Structure analysis in a randomly sampled subset of 12,855 SNPs (~ 0.1% of the total autosomal SNPs in this study) to understand the admixture level between the breeds considered (Figure 4.1a) that showed clear differences. This was supported by principal component analysis (PCA, Figure 4.1b) which infers global patterns of genetic structure without breed membership as unsupervised analysis. The largest principal component (PC1) positioned THB apart from JH explaining 17.8% of the variation. In addition, I performed the Treemix 1.12 analysis to infer the migration events of THB and JH. However, I did not find any potential migration events between the two breeds (Figure 4.1c). Given this information, I suggest that they are clearly divided into two groups for downstream analysis. This suggests that THB and JH have evolved separately in different places. These results are consistent with a study using microsatellite markers (Cho 2007).

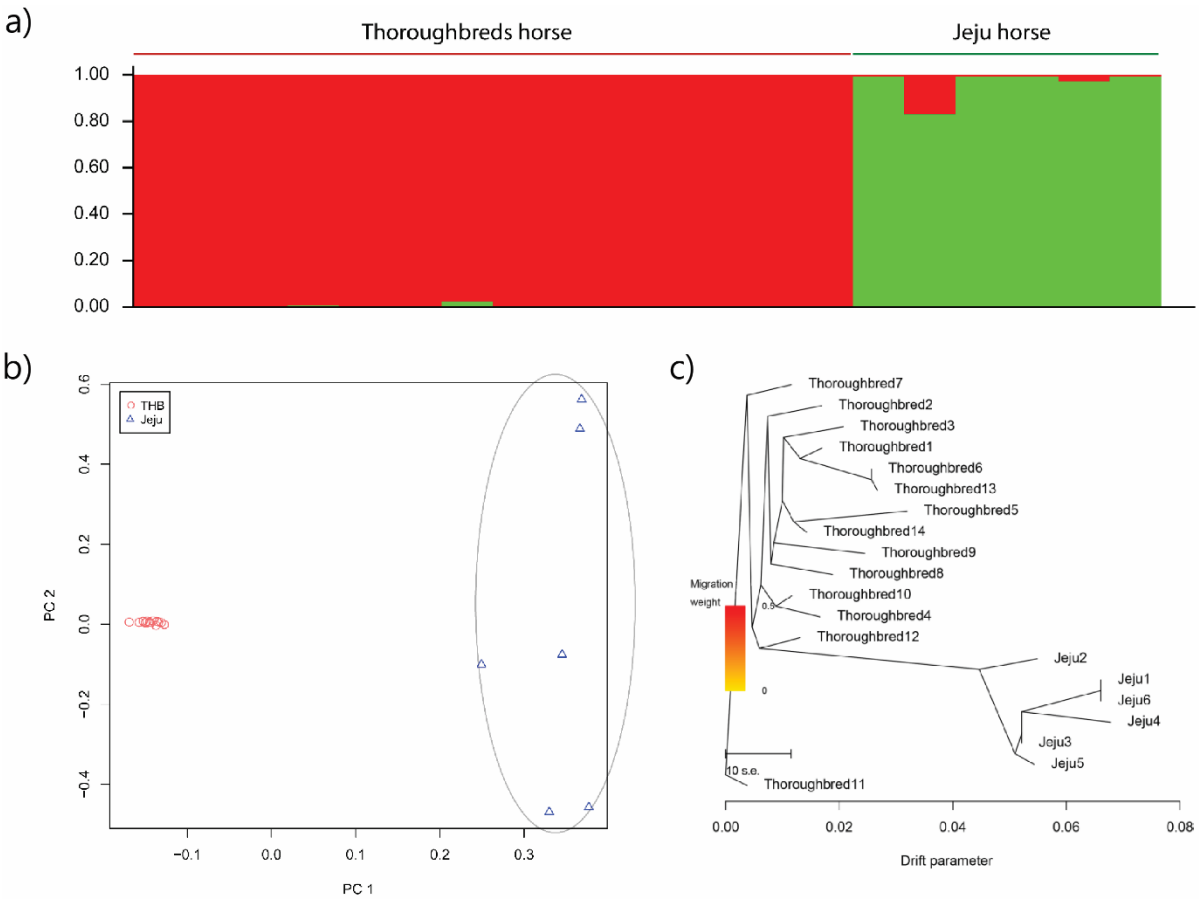


Figure 4.1 Population stratification of Thoroughbred and Jeju Horses a) Population structure, b) PCA plot, and c) Treemix analysis. a) Each segment represents the proportion of a horse individual genome from ancestral populations. Different colored segments in individuals assume that part of the genome originated from different ancestral populations. This figure shows the genetic structure of horse breeds when I assume that the number of ancestral populations of horse is 2. b) Red circles are individuals in Thoroughbreds horses, and blue triangles are individuals in Jeju horses. The horizontal axis indicates eigenvector 1, and the vertical axis indicates eigenvector 2. Values of eigenvectors were estimated using GCTA tool. c) The result of TreeMix shows pattern of population splits and mixture between the two horse breeds. The drift parameter means proportional to Ne generations, where Ne is the effective population size. The scale bar shows ten times the average standard error of the estimated entries in the sample covariance matrix.

4.4.3 Putative positive selection signals in THB horses

I used the XP-EHH method (Sabeti, Varilly et al. 2007) to find genes under positive selection in THB, which calculates haplotype decay separately for each group using the EHH (Ricard, Bruns et al. 2000). In addition, I calculated the XP-CLR statistics between THB and JH breeds. This statistic searches for the selective sweep on single nucleotide polymorphisms (SNPs) in the vicinity of the selected allele, using Brownian motion to model genetic drift under neutrality through allele frequency differentiation between populations (Chen, Patterson et al. 2010). The Manhattan plot of the $-\log_{10}$ transformed XP-EHH and XP-CLR score p-values is presented in Figure 2a and 2b, respectively. Using the top 1% outlier regions, a total of 288 genes were detected using XP-EHH (98 genes)(Table 4.2) and XP-CLR (200 genes)(Table 4.3) population statistics.

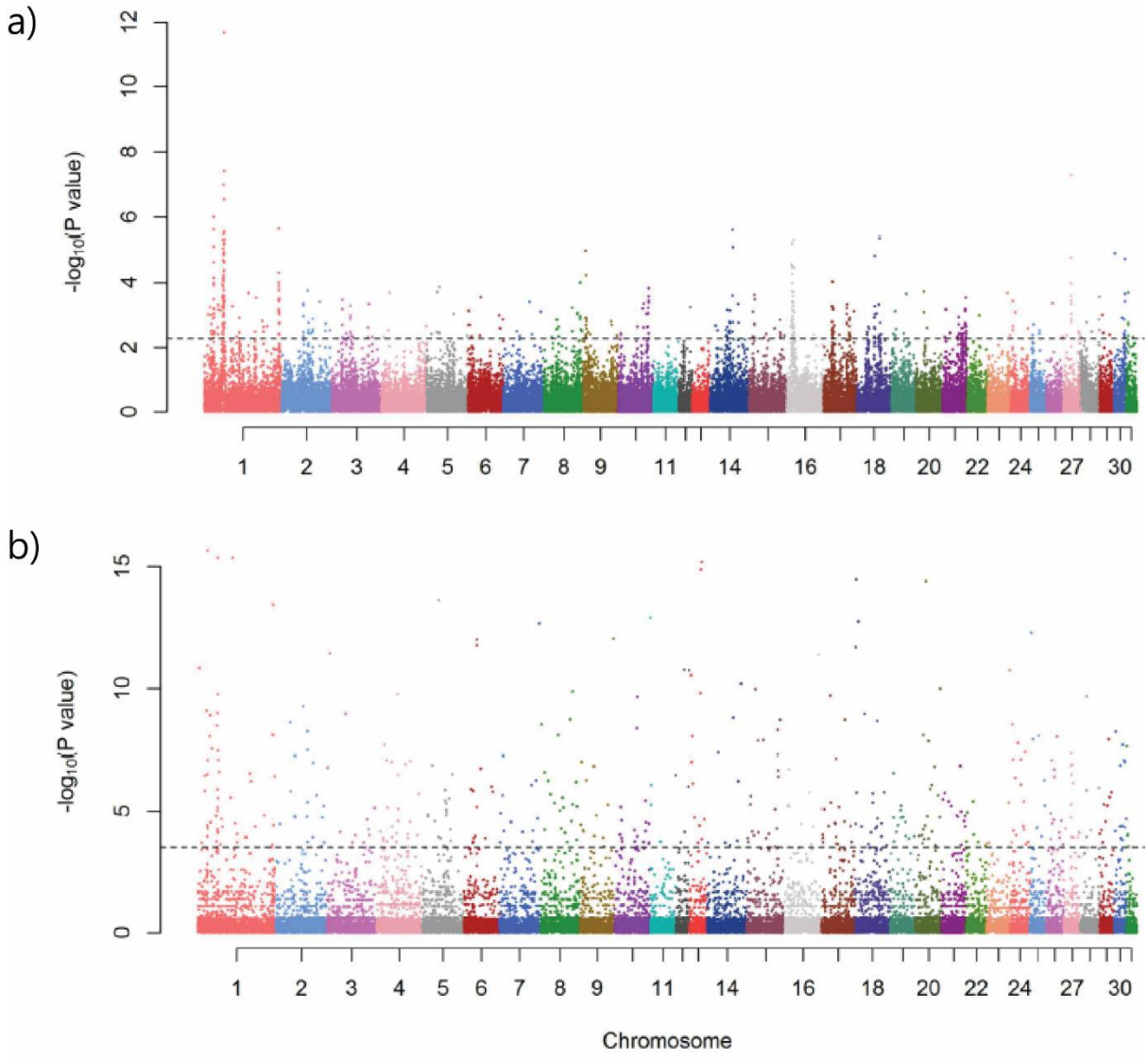


Figure 4.2 Manhattan Plot of $-\log_{10}$ transformed a) XP-EHH values, and b) XP-CLR score P-values of Thoroughbred Horses as compared to Jeju horses. The y- axis indicates $-\log_{10}$ (p-value) of XP-EHH and XP-CLR values and the x-axis is the chromosomal position.

Table 4.2 Genes overlapped with selective regions in Thoroughbreds compared to Jeju horses(XP-EHH)

Chr	Position	XP-EHH score	Gene Symbol	Gene Start	Gene End
1	13700000-13750000	5.86158	PRLHR	13735786	13736973
1	15550000-15600000	5.69065	PNLIP	15534773	15551621
1	15600000-15650000	5.31381	PNLIPRP3	15603161	15636794
1	16450000-16500000	5.4833	ATRNL1	16113986	16849098
1	16500000-16550000	5.2465	ATRNL1	16113986	16849098
1	25350000-25400000	5.27234	SORCS3	25204207	25768630
1	32900000-32950000	5.39809	BLNK	32920430	32978785
1	33200000-33250000	5.22926	C10orf131	33190582	33200629
1	33250000-33300000	5.90838	ENTPD1	33272430	33319727
1	33400000-33450000	5.80411	ALDH18A1	33436203	33476492
1	33400000-33450000	5.80411	TCTN3	33396225	33418740
1	42700000-42750000	5.06026	PRKG1	42400353	42839102
1	43200000-43250000	5.87416	MBL2	43245495	43249033
1	46400000-46450000	6.49695	ZWINT	46418493	46421253
1	64400000-64450000	5.05381	KCNMA1	64178413	64684808
1	64450000-64500000	5.07086	KCNMA1	64178413	64684808
1	77100000-77150000	5.18054	SLC35F3	76905154	77273031
1	84300000-84350000	5.27587	GRID1	84021400	84686687
1	84550000-84600000	5.62312	GRID1	84021400	84686687
1	84600000-84650000	5.13165	GRID1	84021400	84686687
1	121550000-121600000	5.11129	THSD4	121514028	121697263
1	139900000-139950000	5.55544	SLC27A2	139944322	139990189
1	139900000-139950000	5.55544	HDC	139917440	139938572
1	180500000-180550000	5.7126	U6	180536708	180536810
2	62650000-62700000	6.20426	GALNTL6	62489262	62681183
3	25650000-25700000	5.45974	SYCE1L	25675996	25681658
3	25650000-25700000	5.45974	MON1B	25666657	25671458
3	39600000-39650000	5.15622	LAMTOR3	39617734	39628537
3	39600000-39650000	5.15622	DNAJB14	39578196	39607247
3	39600000-39650000	5.15622	DAPP1	39644056	39690466
3	44900000-44950000	5.30024	GRID2	44794591	45895007
3	44950000-45000000	5.59438	GRID2	44794591	45895007
3	45050000-45100000	5.68747	GRID2	44794591	45895007
3	45100000-45150000	5.63662	GRID2	44794591	45895007
3	45150000-45200000	5.26046	GRID2	44794591	45895007
3	45200000-45250000	5.89623	GRID2	44794591	45895007
3	45850000-45900000	5.09328	GRID2	44794591	45895007
3	50100000-50150000	5.17407	DMP1	50146395	50152805
3	89350000-89400000	5.92846	C4orf19	89397506	89402420
3	89350000-89400000	5.92846	RELL1	89324004	89367130
4	6150000-6200000	5.17238	CDHR3	6186918	6234732
5	30350000-30400000	6.26976	eca-mir-29c-2	30369007	30369094
5	44150000-44200000	5.04912	S100A12	44157872	44159412
5	98200000-98250000	5.20441	INADL	97978035	98261790
6	3850000-3900000	5.46432	U6	3883149	3883255
6	3850000-3900000	5.46432	SPAG16	3796222	4460726
6	74000000-74050000	5.67948	TIMELESS	74016170	74027809
6	74000000-74050000	5.67948	MIP	74039621	74043018
6	84150000-84200000	5.59724	CPSF6	84181612	84194476
7	69100000-69150000	5.2344	XRRA1	69105591	69159184

7	69100000-69150000	5.2344	SPCS2	69080473	69100266
7	89150000--89200000	5.75848	NELL1	88433704	89174004
8	34550000-34600000	5.41974	TXNDC2	34571990	34691966
8	34550000-34600000	5.41974	RAB31	34550671	34667158
8	34600000-34650000	5.17978	TXNDC2	34571990	34691966
8	34600000-34650000	5.17978	RAB31	34550671	34667158
8	80300000-80350000	5.73443	SERPINB8	80320350	80329754
8	90500000-90550000	5.30671	ZNF516	90542842	90611367
9	6150000-6200000	6.4953	SNX16	6171006	6205303
9	6300000-6350000	5.46072	IMPA1	6292748	6306961
9	6800000-6850000	5.37018	PAG1	6824382	6837911
9	16200000-16250000	5.2304	SULF1	16074405	16224037
9	66400000-66450000	5.54417	FER1L6	66308735	66423491
9	79800000-79850000	5.20337	PTK2	79818506	80002758
10	7250000-7300000	5.25597	COX7A1	7269471	7271181
10	7250000-7300000	5.25597	CAPNS1	7262426	7268651
10	7300000-7350000	5.2253	ZNF565	7307788	7339451
10	61400000-61450000	5.40652	LAMA4	61402132	61536414
10	61450000-61500000	5.1339	LAMA4	61402132	61536414
10	69850000-69900000	5.71679	SERINC1	69859908	69897651
10	69850000-69900000	5.71679	HSF2	69834195	69853354
10	73200000-73250000	5.99878	CENPW	73213328	73221110
14	39100000-39150000	5.34111	7SK	39120973	39121303
14	39100000-39150000	5.34111	SPOCK1	39058444	39358820
14	39850000-39900000	5.21437	TRPC7	39766718	39890007
14	41700000-41750000	5.07947	C5orf15	41739245	41752621
14	41700000-41750000	5.07947	VDAC1	41721444	41735701
14	47500000-47550000	5.26567	C5orf63	47506713	47511322
14	47650000-47700000	5.53935	LMNB1	47692226	47739561
14	47650000-47700000	5.53935	MARCH3	47635117	47670086
14	47700000-47750000	5.33336	LMNB1	47692226	47739561
14	84600000-84650000	5.09522	SSBP2	84400526	84608573
15	36450000-36500000	5.15836	SPRED2	36492830	36523750
15	75100000-75150000	5.2719	LAPTM4A	75114020	75130250
15	85900000-85950000	5.19836	RNF144A	85914791	85956240
16	12700000-12750000	5.96111	CRBN	12725357	12749878
16	12700000-12750000	5.96111	TRNT1	12724876	12768252
16	12750000-12800000	5.18303	TRNT1	12724876	12768252
16	12750000-12800000	5.18303	IL5RA	12787710	12817071
16	12800000-12850000	5.43263	CNTN4	12829317	13266006
16	12800000-12850000	5.43263	IL5RA	12787710	12817071
16	12850000-12900000	6.51972	CNTN4	12829317	13266006
16	13000000-13050000	5.2077	CNTN4	12829317	13266006
16	13050000-13100000	5.18385	CNTN4	12829317	13266006
16	13100000-13150000	6.66298	CNTN4	12829317	13266006
16	13150000-13200000	5.65654	CNTN4	12829317	13266006
16	14500000-14550000	6.09174	CNTN6	14278424	14573446
16	15150000-15200000	5.48224	CHL1	15165246	15247062
16	15200000-15250000	5.36809	CHL1	15165246	15247062
16	18900000-18950000	5.08297	FOXP1	18770909	19256184
17	21550000-21600000	5.2091	PHF11	21587654	21608810
17	21550000-21600000	5.2091	RCBTB1	21554166	21582916
17	21700000-21750000	5.25187	CAB39L	21719935	21817311
17	21800000-21850000	6.38355	CAB39L	21719935	21817311
17	21800000-21850000	6.38355	CDADC1	21840188	21887601
17	21850000-21900000	5.142	FNDC3A	21894149	21987908
17	21850000-21900000	5.142	CDADC1	21840188	21887601

17	22900000-22950000	5.81028	NUDT15	22928322	22937604
17	22900000-22950000	5.81028	MED4	22898988	22917936
17	24800000-24850000	5.29218	SIAH3	24802461	24871435
17	63950000-64000000	5.6257	U3	63966017	63966106
18	42500000-42550000	5.17347	GCG	42514001	42519890
18	42500000-42550000	5.17347	FAP	42541300	42613886
18	42600000-42650000	5.41773	FAP	42541300	42613886
18	42600000-42650000	5.41773	IFIH1	42641015	42693091
18	42700000-42750000	5.63498	GCA	42711787	42724685
18	42700000-42750000	5.63498	KCNH7	42736751	42903512
18	42750000-42800000	5.70166	KCNH7	42736751	42903512
18	42800000-42850000	5.88412	KCNH7	42736751	42903512
18	42850000-42900000	6.84833	KCNH7	42736751	42903512
18	55950000-56000000	5.07136	PDE11A	55879303	56265586
18	56650000-56700000	5.10757	PLEKHA3	56647504	56665871
18	56650000-56700000	5.10757	TTN	56684352	56950669
18	68200000-68250000	5.07252	TMEFF2	68182723	68408125
19	9600000-9650000	5.20187	MECOM	9611359	9672717
19	30200000-30250000	5.25779	HRASLS	30202826	30213052
19	30200000-30250000	5.25779	ATP13A5	30217310	30320365
19	30250000-30300000	5.06828	ATP13A5	30217310	30320365
19	36800000-36850000	6.14349	PARP15	36820276	36860670
21	4800000-4850000	5.15386	CCDC125	4800230	4835583
21	4800000-4850000	5.15386	CDK7	4823527	4862725
21	46200000-46250000	5.24347	DNAH5	45963018	46358179
21	46300000-46350000	5.38774	DNAH5	45963018	46358179
21	53900000-53950000	5.78185	DNAJA1	53922754	53924220
22	16600000-16650000	5.07233	BMP2	16421811	16700617
24	41900000-41950000	5.38119	EML1	41841755	41914666
26	39300000-39350000	5.15975	TRAPPC10	39295705	39369686
29	5700000-5750000	5.69076	EPC1	5698985	5730185
30	2850000-2900000	6.89461	EXO1	2830467	2869501
30	24950000-25000000	5.59626	ZBTB41	24919687	24956608
30	24950000-25000000	5.59626	CRB1	24999325	25195424
30	25000000-25050000	5.24008	CRB1	24999325	25195424
30	25100000-25150000	5.19162	CRB1	24999325	25195424
30	25150000-25200000	5.18014	CRB1	24999325	25195424
30	25550000-25600000	5.15655	LHX9	25548666	25565340
30	26250000-26300000	5.16313	PTPRC	26241146	26299185

Table 4.3 Genes overlapped with selective regions in Thoroughbreds compared to Jeju horses(XP-CLR)

Chr	Position	XP-CLR score	Gene Symbol	Gene Start	Gene End
1	126725612-126775612	20.676	DENND4A	126,684,369	126,758,683
			SLC24A1	126,764,822	126,790,389
1	136425612-136475612	17.440	UNC13C	136,245,628	136,766,038
1	15425612-15475612	20.051	HSPA12A	15,427,489	15,460,863
			C10orf82	15,467,311	15,472,418
1	15675612-15725612	21.056	CCDC172	15,685,508	15,747,853
1	157825612-157875612	18.287	METTL17	157,867,776	157,874,131
			SLC39A2	157,875,487	157,877,348
1	23525612-23575612	32.388	7SK	23,561,099	23,561,429
1	28325612-28375612	23.433	FBXW4	28,317,811	28,394,997
1	29725612-29775612	24.623	DNMBP	29,761,898	29,828,213
1	30725612-30775612	17.742	HPSE2	30,713,590	30,967,094
1	33225612-33275612	22.710	ENTPD1	33,272,430	33,319,727
1	42425612-42475612	18.250	PRKG1	42,400,353	42,839,102
1	46375612-46425612	17.195	ZWINT	46,418,493	46,421,253
1	50275612-50325612	21.304	RHOBTB1	50,306,440	50,343,205
1	58725612-58775612	17.403	ADAMTS14	58,710,202	58,792,012
1	76875612-76925612	19.575	SLC35F3	76,905,154	77,273,031
1	8525612-8575612	16.082	OAT	8,548,828	8,560,937
1	93425612-93475612	16.845	AP3S2	93,469,167	93,515,923
1	93675612-93725612	16.115	WDR93	93,656,285	93,697,469
			PEX11A	93,708,555	93,713,201
			PLIN1	93,717,373	93,726,793
2	34475100-34525100	24.222	MINOS1	34,444,953	34,475,724
2	46625100-46675100	18.185	PRDM16	46,652,454	46,701,985
2	75125100-75175100	22.628	RAPGEF2	75,147,690	75,312,625
2	75175100-75225100	19.246	RAPGEF2	75,147,690	75,312,625
2	75475100-75525100	23.743	C4orf45	75,480,525	75,527,216
3	114978116-115028116	18.841	SORCS2	114,992,496	115,473,178
3	2778116-2828116	21.565	ZNF423	2,577,934	2,838,642
3	69028116-69078116	16.749	EPHA5	68,911,940	69,237,154
4	100225018-100275018	19.843	CNTNAP2	99,376,000	100,675,021
4	16075018-16125018	16.251	ADCY1	16,102,379	16,243,343
4	16275018-16325018	16.490	IGFBP3	16,295,112	16,300,689
			IGFBP1	16,280,516	16,284,828
4	19725018-19775018	22.948	VWC2	19,717,223	19,825,574
4	24375018-24425018	22.024	SEC61G	24,372,218	24,380,086
4	26725018-26775018	18.858	CACNA2D1	26,623,297	27,110,715
			U6	26,758,240	26,758,346
4	36275018-36325018	21.925	FAM133B	36,291,522	36,312,750
4	37025018-37075018	17.461	CALCR	37,068,659	37,131,493
4	37375018-37425018	16.560	GNGT1	37,417,410	37,515,636
4	48075018-48125018	16.894	AGMO	47,940,706	48,126,316
4	69225018-69275018	21.775	LRRN3	69,069,584	69,264,494
4	74025018-74075018	16.820	ST7	74,038,076	74,295,387
5	14975139-15025139	18.237	IQCJ-SCHIP1	14,688,233	15,215,414
5	15175139-15225139	18.360	IQCJ-SCHIP1	14,688,233	15,215,414
5	51725139-51775139	15.823	MAN1A2	51,589,486	51,773,010
5	54975139-55025139	19.384	MAGI3	55,015,010	55,232,268
			PHTF1	54,958,358	55,004,256
5	56875139-56925139	15.720	ATP5F1	56,905,428	56,918,714
			C1orf162	56,885,158	56,885,977

			ADORA3	56,865,300	56,880,853
			WDR77	56,918,930	56,926,645
5	56925139-56975139	18.277	OVGP1	56,934,870	56,943,943
			PIFO	56,973,316	56,978,156
			WDR77	56,918,930	56,926,645
5	64325139-64375139	19.519	OLFM3	64,324,886	64,506,868
6	20925133-20975133	16.086	SNORD112	20,943,206	20,943,277
6	22075133-22125133	16.149	AGAP1	22,070,624	22,434,439
6	22325133-22375133	20.050	AGAP1	22,070,624	22,434,439
6	23575133-23625133	16.497	MLPH	23,607,273	23,652,480
			RAB17	23,613,902	23,685,934
6	31075133-31125133	28.473	TSPAN9	31,002,891	31,197,638
6	31375133-31425133	18.916	PRMT8	31,381,202	31,469,652
6	40525133-40575133	21.520	APOLD1	40,567,019	40,567,849
6	66425133-66475133	20.350	LALBA	66,475,022	66,477,097
6	69675133-69725133	19.995	KRT72	69,680,384	69,692,566
			KRT73	69,698,559	69,708,600
7	77575058-77625058	20.456	RIC3	77,606,947	77,665,883
			TUB	77,570,067	77,631,674
7	80325058-80375058	16.181	GALNT18	80,256,649	80,597,024
7	88525058-88575058	20.734	NELL1	88,433,704	89,174,004
8	10378197-10428197	21.297	MYO18B	10,181,087	10,411,809
8	18728197-18778197	20.721	RBM19	18,745,890	18,873,916
8	53428197-53478197	19.598	KLHL14	53,418,094	53,510,049
8	55078197-55128197	18.861	DTNA	55,074,654	55,330,726
8	70928197-70978197	24.395	DCC	70,765,480	71,472,757
8	85128197-85178197	20.633	DOK6	84,877,188	85,288,608
9	15975414-16025414	20.736	SLCO5A1	15,938,273	16,062,069
9	2025414-2075414	17.598	CNBD1	1,831,922	2,142,743
9	65725414-65775414	19.060	ZHX1	65,748,399	65,762,668
9	7075414-7125414	17.128	ZNF704	7,080,227	7,114,212
10	12225448-12275448	16.784	C19orf69	11,969,029	12,305,319
10	13825448-13875448	18.901	CD79A	13,823,300	13,826,489
			LYPD4	13,850,603	13,854,810
			DMRTC2	13,844,388	13,848,201
			RPS19	13,830,074	13,837,793
10	15875448-15925448	16.494	MARK4	15,852,069	15,876,746
			KLC3	15,900,600	15,905,546
			DACT3	15,688,161	16,899,653
			CKM	15,881,172	15,890,008
10	16375448-16425448	17.935	NOVA2	16,384,190	16,393,862
			DACT3	15,688,161	16,899,653
10	16775448-16825448	16.186	DACT3	15,688,161	16,899,653
10	39775448-39825448	33.871	HTR1E	39,801,452	39,870,896
10	68925448-68975448	16.134	TBC1D32	68,764,771	68,950,109
			U2	68,967,924	68,968,118
10	72925448-72975448	17.050	TRMT11	72,921,154	72,971,638
10	74425448-74475448	17.740	THEMIS	74,465,373	74,641,267
10	80775448-80825448	15.987	HBS1L	80,764,958	80,842,492
10	81525448-81575448	16.061	PDE7B	81,476,052	81,765,672
10	81625448-81675448	17.651	PDE7B	81,476,052	81,765,672
11	24775090-24825090	16.196	TTLL6	24,774,459	24,812,864
			CALCOCO2	24,821,372	24,844,044
11	275090-325090	29.455	TBCD	297,524	482,406
			B3GNTL1	165,925	288,330
11	2775090-2825090	19.073	EIF4A3	2,814,986	2,828,951
			CARD14	2,772,943	2,798,639

11	3025090-3075090	20.446	CBX4	3,063,802	3,064,803
12	12675087-12725087	16.106	OR8K5	12,681,826	12,682,767
12	29575087-29625087	20.333	SHANK2	29,316,173	29,799,539
12	32125087-32175087	26.965	DUSP8	31,534,226	32,674,266
			MUC6	32,129,549	32,141,574
13	20826251-20876251	16.075	NSMCE1	20,871,434	20,899,516
13	23476251-23526251	17.501	PRKCB	23,368,966	23,672,898
13	25876251-25926251	25.804	UQCRC2	25,858,966	25,881,655
13	27026251-27076251	31.604	GP2	27,028,368	27,042,804
14	2275407-2325407	16.063	GFPT2	2,252,113	2,280,346
14	2275407-2325407	16.063	MAPK9	2,303,318	2,346,926
14	42475407-42525407	16.087	FSTL4	42,281,900	42,481,829
14	71825407-71875407	16.276	LNPEP	71,808,888	71,895,463
15	10425122-10475122	19.674	KIAA1211L	10,460,246	10,516,120
15	17625122-17675122	16.689	FABP1	17,645,656	17,655,539
			SMYD1	17,668,603	17,708,719
15	28075122-28125122	16.953	TACR1	27,972,112	28,122,537
15	4025122-4075122	16.142	NCK2	4,066,301	4,112,284
15	68275122-68325122	15.731	BRE	68,075,291	68,495,845
16	11025589-11075589	18.938	ARL8B	11,046,061	11,087,247
			EDEM1	11,011,208	11,035,979
16	39325589-39375589	17.635	ELP6	39,370,138	39,380,115
16	6675589-6725589	18.943	SEC13	6,667,915	6,693,504
			GHRL	6,704,413	6,709,418
16	67675589-67725589	17.425	CPNE4	67,434,752	67,979,351
16	87275589-87325589	16.377	GMPS	87,269,128	87,305,934
17	21475591-21525591	25.682	EBPL	21,469,751	21,476,860
			ARL11	21,504,968	21,505,495
17	35075591-35125591	16.605	DIAPH3	35,062,159	35,497,994
17	35325591-35375591	22.145	DIAPH3	35,062,159	35,497,994
17	69425591-69475591	17.470	PCCA	69,097,619	69,492,947
18	31275498-31325498	15.964	EPC2	31,175,313	31,305,229
18	3175498-3225498	19.919	MYO7B	3,154,614	3,221,832
18	41675498-41725498	19.863	TANK	41,680,821	41,756,857
18	42275498-42325498	19.221	SLC4A10	42,072,060	42,346,581
18	48325498-48375498	17.284	CERS6	48,176,453	48,465,399
18	51175498-51225498	17.457	METAP1D	51,182,790	51,253,663
18	53025498-53075498	18.063	OLA1	52,895,533	53,062,128
18	68375498-68425498	19.964	TMEFF2	68,182,723	68,408,125
18	74625498-74675498	15.843	FTCDNL1	74,664,313	74,697,642
19	10275538-10325538	16.171	LRRC31	10,281,706	10,310,009
			LRRIQ4	10,269,951	10,279,498
19	26425538-26475538	18.994	LPP	26,156,935	26,575,534
19	27175538-27225538	18.109	TP63	27,207,319	27,428,546
19	36775538-36825538	17.779	PARP14	36,754,148	36,795,593
			PARP15	36,820,276	36,860,670
19	38025538-38075538	16.656	STXBP5L	37,827,844	38,176,937
19	45275538-45325538	16.449	EEF1A1	45,277,340	45,278,728
20	16425065-16475065	16.720	CAP2	16,381,199	16,511,216
20	23725065-23775065	17.512	SLC17A1	23,713,124	23,741,030
			SLC17A3	23,767,516	23,814,310
20	24675065-24725065	38.572	ABT1	24,705,444	24,707,109
20	28875065-28925065	18.226	TRIM10	28,904,045	28,910,889
			TRIM40	28,886,521	28,897,617
			TRIM15	28,913,189	28,919,167
20	32575065-32625065	16.463	BTNL2	32,624,134	32,635,273
20	36225065-36275065	20.169	PNPLA1	36,247,031	36,284,176

20	40675065-40725065	17.008	NCR2	40,702,448	40,718,695
20	55275065-55325065	19.748	KHDRBS2	55,166,761	55,674,261
20	60525065-60575065	26.040	BAI3	60,514,643	61,156,448
21	2325625-2375625	19.455	MVB12A	2,374,359	2,377,643
			BST2	2,362,120	2,362,368
			PLVAP	2,339,197	2,346,592
21	44325625-44375625	21.672	MARCH11	44,250,022	44,356,382
21	46125625-46175625	18.588	DNAH5	45,963,018	46,358,179
22	17875778-17925778	19.311	PROKR2	17,891,663	17,897,674
22	19275778-19325778	16.789	ATRN	19,279,380	19,328,716
22	20325778-20375778	49.233	TGM3	20,308,216	20,347,901
22	29225778-29275778	16.142	DHX35	29,084,472	29,249,586
23	5178476-5228476	16.068	NTRK2	5,103,223	5,441,937
23	54328476-54378476	26.961	BARX1	54,329,182	54,329,843
23	678476-728476	15.982	ZNF510	678,829	683,407
24	11575650-11625650	20.921	SPTB	11,524,261	11,583,792
24	13675650-13725650	16.752	EIF2S1	13,678,234	13,692,660
			PLEK2	13,695,923	13,714,379
24	32525650-32575650	17.201	7SK	32,564,588	32,564,827
24	35575650-35625650	22.517	RIN3	35,536,972	35,801,389
25	21425066-21475066	23.462	ASTN2	20,887,845	21,589,955
25	28825066-28875066	16.623	GOLGA1	28,825,836	28,868,824
25	34575066-34625066	20.731	NTNG2	34,590,566	34,664,747
26	30886161-30936161	16.939	KCNE2	30,892,092	30,898,292
			C21orf140	30,924,353	30,925,084
			SMIM11	30,910,646	30,910,822
27	38475354-38525354	16.883	ARHGEF10	38,491,515	38,576,907
28	14275118-14325118	20.069	C12orf50	14,271,518	14,306,876
			C12orf29	14,310,367	14,321,768
			CEP290	14,322,732	14,409,061
28	24625118-24675118	17.134	GAS2L3	24,616,666	24,642,118
29	10575088-10625088	19.078	MYO3A	10,537,654	10,735,234
29	14525088-14575088	17.817	MLLT10	14,171,706	14,672,770
29	18125088-18175088	19.009	ST8SIA6	18,027,487	18,145,292
29	21825088-21875088	23.268	7SK	21,845,014	21,845,301
29	21825088-21875088	23.268	CCDC3	21,769,233	21,847,581
29	28325088-28375088	19.942	CALML3	28,371,632	28,372,081
30	16277987-16327987	21.691	USH2A	15,772,313	16,515,816
30	20977987-21027987	22.942	RGS21	21,002,828	21,020,497
30	27527987-27577987	33.585	SCARNA4	27,527,953	27,528,082
30	4027987-4077987	17.060	FMN2	4,027,214	4,369,602
30	4627987-4677987	23.725	KIF26B	4,577,169	5,068,285
31	175501-225501	22.875	ZDHHC14	93,019	371,433
31	4375501-4425501	15.743	PACRG	4,229,872	4,689,096

A comparison between THB and JH was appropriate because these populations have been bred under different environments for a long time. I calculated the XP-EHH values and XP-CLR scores as the window statistic of a total 44,826 and 44,844 genetic regions, respectively. By dividing the genome into a non-overlapping segment of 50 kb, I compared the genomic regions across populations and defined those genetic regions on whole horse genome. Empirical distributions using total regions can be constructed due to whole genome sequencing data. XP-EHH scores of 44,826 genetic regions and XP-CLR scores of 44,844 genetic regions showed normal distribution as expected. In this analysis, I used the outlier approach in distribution to detect a significant selective region (Kelley, Madeoy et al. 2006). I defined the top 1 percent of the XP-EHH and the XP-CLR score as a significant selective region and identified 448 significant genetic regions each which were a selective region in THB compared to JH. I identified 98 genes (XP-EHH) and 200 genes (XP-CLR) in 116 and 164 (with annotation) of total 448 significant regions.

I thought that regions with outlier XP-EHH and XP-CLR score provided several important pieces of evidence of THB domestication and selection. I constructed a biological network using Gene Ontology analysis which resulted in 72 GO BP terms (Figure 4.3). Then, the BP terms were grouped into 20 categories based on genes involved in which I focused on those supporting THB's characteristics. I hypothesize that the BP terms enriched that are related to immune function, ocular size and visual function, and energy metabolism might contribute to the THB's superior racing performances (Table 4.5).

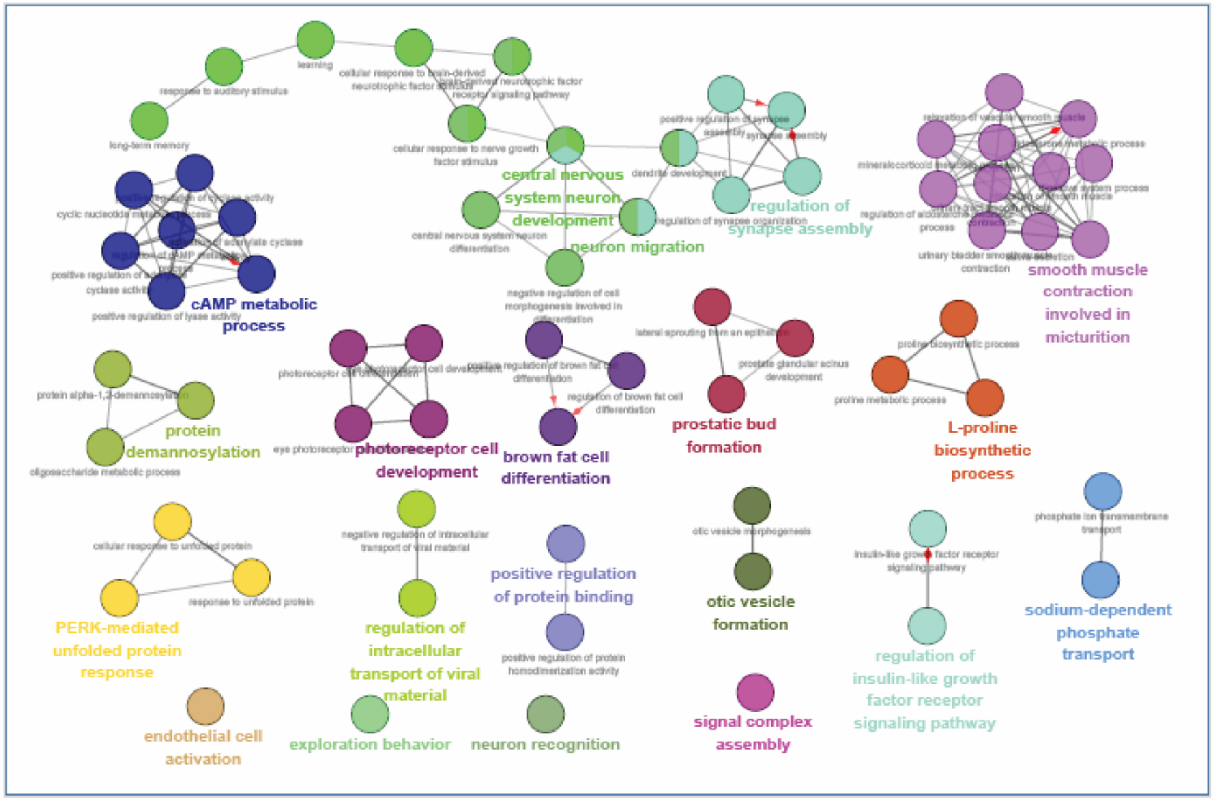


Figure 4.3 Biological network using genes related to selective regions in Thoroughbreds. GO network analysis of biological processes in Thoroughbreds and Jeju horses. GO terms visualized by ClueGo plugin of Cytoscape. Nodes are represented by a circle and imply that two GO terms share genes from the considered gene lists. The size of the circle corresponds to the number of genes related to the GO term. Edges are connections between GO groups defined by 50% genes in common.

Table 4.4 Genes in Gene Ontology terms related to eye in selective regions in Thoroughbred horse (FDR<0.05).

Gene ontology biological process	Genes in selective region	Chr.	XP-EHH value	XP-CLR scores
Dendrite development	*PRKG1*	1	5.060	18.250
	RAPGEF2	2	-	22.628
	RAB17	6	-	16.497
	NELL1	7	5.758	20.734
	DCC	8	-	24.395
	NCK2	15	-	16.142
	GHRL	16	-	18.943
	ADGRB3	20	-	26.040
	NTRK2	23	-	16.068
Photoreceptor cell development	*GNGT1*	4	-	16.560
	OLFM3	5	-	19.519
	NTRK2	23	-	16.068
	CEP290	28	-	20.069
	CRB1	30	5.596	-
Regulation of synapse assembly	*PRKG1*	1	-	18.250
	RAPGEF2	2	-	22.628
	EPHA5	3	-	16.749
	LRRN3	4	-	21.775
	RAB17	6	-	16.497
	NELL1	7	-	20.734
	PTK2	9	5.203	-
	SHANK2	12	-	20.333
	SPOCK1	14	5.341	-
	NCK2	15	-	16.142
	GHRL	16	-	18.943
	CHL1	16	5.482	-
	SLC4A10	18	-	19.221
	NTRK2	23	-	16.068

The BP term categories "Negative regulation of intracellular transport of viral material" is related to immunity. The genes involved in this term (*BST2* and *TRIM5)* are associated with negative regulation of intracellular transport of viral material term, referring to any process that stops, prevents or reduces the frequency, rate or extent of intracellular transport of viral material. *BST2* is associated with growth and development of B-cells. It is an interferon inducible transmembrane protein that provides innate immune response activity by inhibiting members of the retrovirus, filovirus, arenavirus, and herpesvirus families (Evans, Serra-Moreno et al. 2010). Equine tetherin orthologues without dual tyrosine motif could potently activate the NF-κB signaling. NF-κB plays a key role in regulating the immune response to infection (Yin, Guo et al. 2014). *TRIM5* gene encodes a member of the tripartite motif (*TRIM*) family that include three zinc-binding domains, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region. The protein forms homo-oligomers via the coiled-coil region and localizes to cytoplasmic bodies. It appears to function as an E3 ubiquitin-ligase and ubiquitinates itself to regulate its subcellular localization. It may play a role in retroviral restriction. Multiple alternatively spliced transcript variants encoding different isoforms have been described for this gene (O'Leary, Wright et al. 2016). Another immune-related gene (RIN3) which plays a role in the maturation of phagosomes that engulf pathogens have been previously found under positive selection in THB associated with racing performance (Moon, Lee et al. 2015)

THB are the fastest runners among the horse breeds used in the horse racing industry. Quite a few researchers studied why they run faster (Gu, Orr et al. 2009, Heard-Booth and Kirk 2012, Moon, Lee et al. 2015). Here, I report genes and BP terms that potentially contribute for the superior racing performances of THB. THB has been selected for structural and functional adaptions that contribute to its fast running performance (Moon, Lee et al. 2015). I identified that photoreceptor cell development and otic vesicle morphogenesis BP terms were enriched in relation to eye and ear development, respectively. Genes related to eye photoreceptor cell differentiation include *CEP290, GNGT1, CRB1, OLFM3, and NTRK2*. I inferred that strong selection of eye photoreceptor cell differentiation can directly affect increment of ocular size which leads to increased horse eyesight in the view of biological evolution at intra-species level. In vertebrate animals, ocular characteristic is influenced by many factors including body or head size, diet, and activity pattern. Heard-Booth and his colleague stressed that

maximum locomotive speed plays a key role in determining ocular shape in mammals (Heard-Booth and Kirk 2012). Leuckart's Law describes the relationship between a measure of axial eye diameter and maximum speed (Heard-Booth and Kirk 2012). It has been reported that absolute ocular diameter is significantly correlated to maximum running speed in mammals (Hinchcliff, Kaneps et al. 2008). This law also proposed that animals capable of achieving fast running speed require large eyes to enhance visual acuity and avoid collisions with environmental obstacles. The relationship between maximum running speed and eye size in a diverse sample of mammals proved this law (Heard-Booth and Kirk 2012). Additionally, there were two more GO terms which supported directly or indirectly positive selection of ocular size and function in THB in this study; dendrite development, and regulation of synapse assembly. Several genes (*PRKG1, RAPGEF2, RAB17, NELL1, DCC, NCK2, GHRL, BAI3* and *NTRK2)* were identified that trigger dendrite development (Jan and Jan 2003, Quach, Wilson et al. 2013, Ohshima 2014). When light reaches retina after traveling through cornea and lens, ganglion cells take electronic signal through dendrite and send this signal down to the optic nerve. *EPHA5, RAB17, SHANK2,* and *GHRL* are related to regulation of synapse assembly (Dalva, Takasu et al. 2000, Zerial and McBride 2001, Waites, Craig et al. 2005). Based on this knowledge related to optic nerve, I reasoned that eyesight is closely related to synapse because the retina has several neuron layers and communication among these several neuron layers is very important in eye function. *ADCY1*, involved in the regulatory processes in the central nervous system that play a role in memory and learning, have been found to be under selection in racehorse populations (Moon, Lee et al. 2015).

The BP term brown fat cell differentiation, defined as the process in which a relatively unspecialized cells acquire specialized features of a brown adipocyte, is an animal connective tissue cell involved in adaptive thermogenesis (Puigserver and Spiegelman 2003). Brown adipose tissue differs from white adipose tissue in the way they expend energy (Gu, Orr et al. 2009). The type, intensity, and duration of exercise determine the amount of form of fuel used (carbohydrate vs free fatty acid) that, aerobic activities (long duration, low intensity) use more free fatty acids as fuel than anaerobic activities (short duration, high intensity), which use more glucose. However, the horse is almost always using both types to some degree, at the same time. As activity level (e.g. running speed) increases, oxygen consumption rises to meet increased demand for ATP

production. Brown fat has more mitochondria than other cells. When the body needs to use energy, it uses ATP. ATP is mainly produced in the mitochondria of cells. When brown fat is activated, it creates a protein called UCP-1, which prevents ATP production from mitochondria. Instead of generating ATP, heat energy is generated to increase body temperature. The effect of fat supplementation of horse diet on horse performances has been reported. Genes including *PEX11A, LAMA4, ZNF516,* and *PRDM16* are related to brown fat cell differentiation. The positive selection of genes involved in brown adipose tissue differentiation has been previously identified in THB (Gu, Orr et al. 2009).

Insulin receptor signaling pathway is another pathway enriched which control critical energy functions such as glucose and lipid metabolism. It has been found previously to be enriched in THB horses (Gu, Orr et al. 2009) in relation to racing performance. It has also a role in the differentiation of brown adipocytes (Sharma, Huard et al. 2014).

Through QTL analysis, I identified six QTL regions that overlapped to genes in selective regions of THB (Table 4.2). *CERS6* (QTL chr18:48212639_48319679) is well-known racing distance associated gene and *BAI3* (QTL chr20:60009473_60987311) is closely related to recurrent uveitis disease of the eye. Recurrent uveitis is an acute, non-granulomatous inflammation of the uveal tract of the eye, occurring commonly in horses of all types of breeds universally (Laurie, Olsakovsky et al. 2008).

Table 4.5 QTL overlapped with selective regions in Thoroughbreds compared to Jeju horses.

Gene Name	Chr	Gene Begin	Gene End	QTL ID [a]	Related traits
SEC61G	4	24,372,218	24,380,086	qtl_4_24010915_24868953	Insect bite hypersensitivity (29305)
CERS6	18	48,176,453	48,465,399	qtl_18_48212639_48319679	Racing distance (32133)
BAI3	20	60,514,643	61,156,448	qtl_20_60009473_60987311	Recurrent uveitis (29387)
SLC17A1	20	23,713,124	23,741,030	qtl_20_23723503_23816767	Equine sarcoids (28919)
SLC17A3	20	23,767,516	23,814,310	qtl_20_23723503_23816767	Equine sarcoids (28919)
GOLGA1	25	28,825,836	28,868,824	qtl_25_24227654_30109054	Equine sarcoids (28921)

[1)]QTL(Quantitative trait locus) ID was made in this study as followed: qtl+"chromosome"+"qtl begin"+"qtl end"

THB are the epitomes of variation under domestication, yet much of the evolutionary processes underlying the genetics of this diversity are poorly understood. So, I tried to detect novel selective regions which were not reported, previously. I attained novel selective regions using XP-CLR analysis which helped us to observe the relationship between THB and JH in a different angle. These results can be used to characterize functional variants and explore the specificity of the Thoroughbred breed.

4.4.4 Limitations of the study

The possibility of obtaining false positive results is common in such kind of study. Therefore, gene expression analysis, and/or candidate gene approach experimental procedures are required to validate the candidate genes.

4.5 Discussion

I explored the whole genome and detected several positively selected genes involved in different biological and cellular functions affecting THB horses' characteristics. The genes identified in relation to THB characteristics are involved in immunity and eye size, and function that might contribute for THB's superior racing performances. These results provide a basis for further research on the genomic characteristics of THB.

This chapter will be published elsewhere as a partial fulfillment of Wonseok Lee's Ph.D program

Chapter 5. Estimation of connectedness among Korean swine breeding herds

5.1 Abstract

The aim of this study is to estimate the connectedness rate among Korean swine breeding herds. In order to calculate the connectedness, I use 104,380 performance and 83,200 reproduction records of three breeds (Yorkshire, Landrace and Duroc), connectedness rating (CR) was estimated for two traits: average daily gain (ADG) and number born alive (NBA) of eight breeding herds in Korea. The average CR for ADG of the Yorkshire ranges from 1.32% to 28.5% depending on the farm. The average CR for NBA of the Yorkshire herd ranges from 0% to12.79%. A total of 60% of Yorkshire and Duroc herds satisfied for the preconditions suggested to evaluate genetic analysis among herds. The precondition for the genetic evaluation of CR for ADG, as performance trait, was higher than 3% and that of NBA, as reproductive trait, was higher than 1.5%. The highest average CR was for the trait of ADG of the Yorkshire herds. However, the average CR of the Landrace herds for ADG was lower than the criteria. The prediction error variance of the difference (PEVD) was used to test the validation of the CR. Most of the PEVD were fluctuate together with the CR of among herds. A certain degree of connectedness is essential to estimate breeding value comparisons between herds. This study suggested that four out of eight herds are possible to evaluate genetic values together for ADG and NBA of the Yorkshire herds since the preconditions were satisfied for the four herds. It is also possible to perform joint genetic analysis of the ADG records of all Duroc herds since the preconditions were also satisfied. This study provides new insight for understanding the genetic connectedness among Korean swine herds. CR validated by PEVD could be utilized to accelerate the genetic progress of Korean swine breeding herds.

5.2 Introduction

Accurate estimation of the breeding value (BV) for important economic traits is crucial in animal breeding programs(Soga 2009). The accuracy of estimation relies on Connectedness to evaluate genetic analysis between herds. Connectedness refers to the genetic similarity among herds(Soga 2009). These genetic links are important because they can affect the prediction error variance of Difference(PEVD) of estimated breeding values(Škorput, Gorjanc et al. 2012). Many methods have been proposed to estimate Connectedness (Foulley, Hanocq et al. 1992, Kennedy and Trus 1993, Laloë, Phocas et al. 1996, Soga 2009). Among them, Mathur et al.(Mathur, Sullivan et al. 1998) suggested the Connectedness Rating(CR) as a good indicator for Connectedness. Later studies showed that the CR is consistent in the results from Connectedness analysis(Sun, Wang et al. 2009, Soga, Spangler et al. 2010, Škorput, Gorjanc et al. 2012).

In Korea, the genetic progress on swine has been achieved mainly by importing breeding pigs from oversea countries. However, as consolidation among breeding companies and breeding farms has progressed, fewer pig genetic resources are imported each year. In addition, multiple breeding farms are planning work together to evaluate pig genetic analysis to maximize genetic progress and to mitigate the need to import breeding pigs into Korea. However, if there is no genetic link between the herds, the evaluation of estimated breeding value (EBV) between different farms is not reliable and is less accurate. It has been reported that the accuracy of the genetic evaluation increases when the CR between the herds is increased (Mathur, Sullivan et al. 2002, Sun, Wang et al. 2009, Soga, Spangler et al. 2010, Škorput, Gorjanc et al. 2012).

Therefore, the aim of this research is to estimate the connectedness among swine herds using three different breeds (Yorkshire, Landrace, and Duroc) in Korea for the traits of average daily gain (ADG) and number of born alive (NBA).

5.3 Materials and Methods

5.3.1 Data preparation

Performance and reproduction data were collected from fifteen Korean swine herds (8

Yorkshire herds, 5 Landrace herds, and 4 Duroc herds), born between 1997 and 2016. To calculate the connectedness between pairs of herds, two traits were considered: Average daily gain (ADG) and Number of born alive (NBA). The numbers of records per breed and farm are presented in Tables 5.1 and 5.2.

Table 5.1 Number of records for average daily gain (ADG)

Farm	Yorkshire	Landrace	Duroc
A	20,460	327	759
B	8,620	205	580
C	9,710	3,812	3,492
D	1,296	-	-
E	17,888	357	-
F	2,971	1,094	-
G	5,476	-	2,261
H	14,138	10,484	-
TOTAL	**80,559**	**16,279**	**7,092**

Table 5.2 Number of records for number of born alive (NBA)

Farm	Yorkshire	Landrace	Duroc
A	5,127	327	759
B	2,773	205	580
C	9,710	3,812	3,492
D	1,296	-	-
E	17,888	357	-
F	2,971	1,094	-
G	5,476	-	2,261
H	14,138	10,484	-
TOTAL	**59,379**	**16,279**	**7,092**

5.3.2 Statistical model for Breeding Value

For estimating the breeding value, both ADG and NBA data sets were analyzed for each breed using the following statistical model(1).

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Hd} + \mathbf{e} \qquad \textbf{(1)}$$

where **y** is the phenotype observations for ADG and NBA, **b** is a vector of fixed effects (herd effects), **a** is the vectors of random effects (additive animal genetic effects), **d** is the vector of common litter effects, and **e** is a vector for environmental residuals ($\boldsymbol{e} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}\boldsymbol{\sigma}^2{}_{\boldsymbol{e}})$). **X**, **Z** and **H** were used as incidence matrices corresponding to vectors **b, a** and **d** related to the random additive genetic effects ($\boldsymbol{a} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{A}\boldsymbol{\sigma}^2{}_{\boldsymbol{a}}), \boldsymbol{d} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}\boldsymbol{\sigma}^2{}_{\boldsymbol{d}})$). $\boldsymbol{\sigma}^2{}_{\boldsymbol{a}}, \boldsymbol{\sigma}^2{}_{\boldsymbol{d}}$ and $\boldsymbol{\sigma}^2{}_{\boldsymbol{e}}$ represent the additive genetic variance, litter variance, and environmental residual variance, respectively.

5.3.3 Mixed model equation construction

Using this model, we constructed the mixed model equation (MME) for the corresponding equation (1).

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} & \boldsymbol{X'H} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{A}^{-1}\boldsymbol{\alpha}\mathbf{1} & \boldsymbol{Z'H} \\ \boldsymbol{H'X} & \boldsymbol{H'Z} & \boldsymbol{H'H} + \boldsymbol{I}\boldsymbol{\alpha}\mathbf{2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{h}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{d}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \\ \boldsymbol{H'y} \end{bmatrix}$$

(2)

Here, **A** is the numerator genetic relationship matrix for animals, $\boldsymbol{\alpha}\mathbf{1}$ refers to $\boldsymbol{\sigma}^2{}_{\boldsymbol{e}} / \boldsymbol{\sigma}^2{}_{\boldsymbol{a}}$ and $\boldsymbol{\alpha}\mathbf{2}$ refers to $\boldsymbol{\sigma}^2{}_{\boldsymbol{e}} / \boldsymbol{\sigma}^2{}_{\boldsymbol{d}}$.

The ASReml package(Gilmour, Gogel et al. 2015) was used for solving the above equation (2).

5.3.4 Estimation of connectedness rating

Connectedness rating was defined by the following equation (3):

$$CR_{ij} = \frac{cov(\hat{h}_i, \hat{h}_j)}{\sqrt{var(\hat{h}_i)var(\hat{h}_j)}} \quad \textbf{(3)}$$

The covariance for herd and variance of estimation of each herd effect i and j were obtained by solving equation (2).

5.3.5 Evaluation for CR

According to *Mathur et al.*(Mathur 2005), the accuracy of an individual EBV is estimated using the prediction error variance corresponding to animals. The prediction error variance of difference (PEVD) can be used to test validation of accuracy of EBVs of two individuals. PEVD is formularized as:

$$\mathbf{PEVD} = \mathbf{Var}(\hat{a}_i - \hat{a}_j) = Var(\hat{a}_i) + Var(\hat{a}_j) - 2cov(\hat{a}_{ij})$$

(4)

Then PEVD can be substituted as the Variance of Estimated Differences between Herd effects (VED)(Mathur 2005):

$$VED_{ij} = Average[PEV(\hat{a}_{ik} - \hat{a}_{ik'})] \approx Var(\hat{h}_i - \hat{h}_j) \cong Var(\hat{h}_i) + Var(\hat{h}_j) - 2cov(\hat{h}_{ij}) \quad \textbf{(5)}$$

5.4 Result

5.4.1 Connectedness Ratings for ADG trait

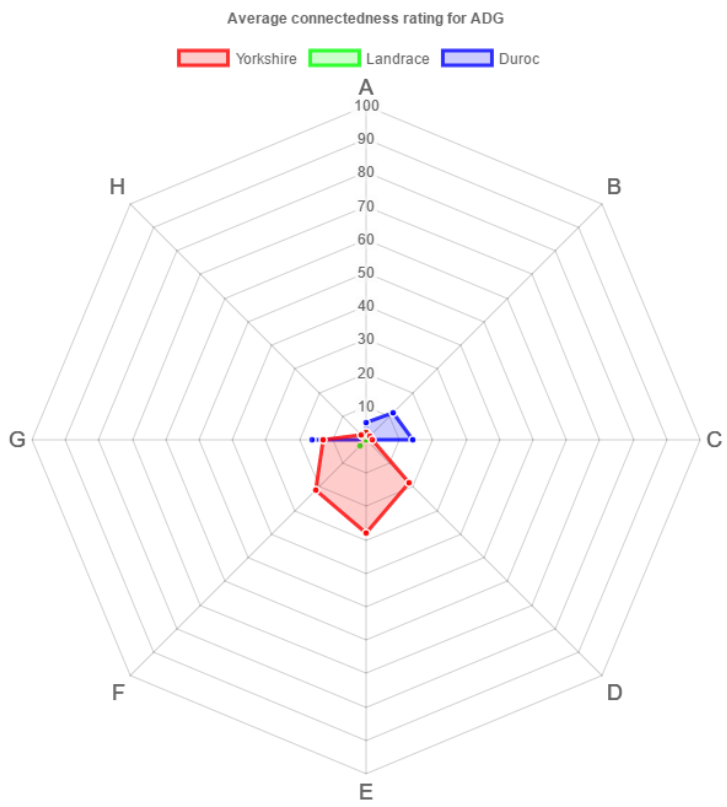
The average connectedness ratings (CR) for ADG with each breed are listed in Table 5.3 and Figure 5.1. In the Korean Yorkshire breed for the ADG trait, a total of 8 herds were analyzed. The average CR between two herds ranges from 1.32 (B herd) to 28.05 (E herd). The maximum CR was 93.44 between herd E and G and the lowest CR value was 4.4 between herds B and G. In the Korean Landrace herd for the ADG trait, a total of 5 herds were used in this analysis. The highest average CR was 2.50 between herds A and F and the lowest average CR was 0.55 between herds C and H. All 4 herds are used in the Korean Duroc analysis. The highest average CR was 16.14 between herds C and G

and the lowest average CR was 5.03 between herds A and C.

Table 5.3. Connectedness rating (CR) for ADG among herds

Herds	Yorkshire connectedness rating (%)			Landrace connectedness rating (%)			Duroc connectedness rating (%)		
	Mean	Max	Most connected herd	Mean	Max	Most connected herd	Mean	Max	Most connected herd
A	2.18	9.56	F	2.08	12.49	F	5.03	10.77	C
B	1.32	4.41	G	0.55	2.44	H	11.21	27.81	G
C	1.68	7.42	H	0.88	0.32	H	13.8	30.92	G
D	18.29	86.81	F	-	-	-	-	-	-
E	28.05	93.44	G	-	-	-	-	-	-
F	21.18	86.81	E	2.50	12.49	A	-	-	-
G	12.92	93.44	E	-	-	-	16.14	30.92	C
H	2.01	7.42	C	1.31	4.1	C	-	-	-

Figure 5.1 Average connectedness rating (CR) for ADG with three breeds



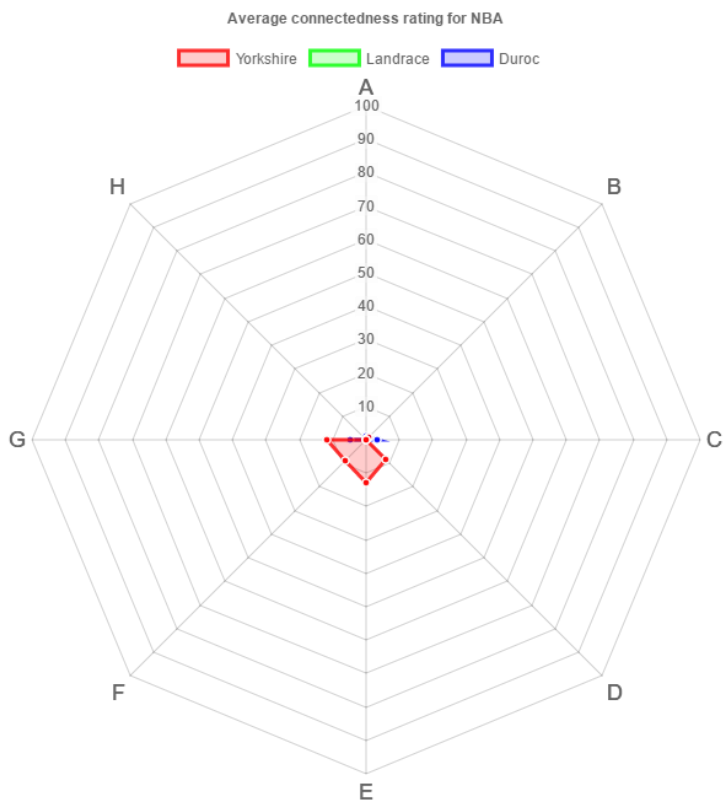
5.4.2 Connectedness Ratings for NBA trait

Similar to the ADG analysis, the average connectedness ratings (CRs) for NBA with each breed are presented in Table 5.4 and Figure 5.2. The same number of herds was analyzed as that for the ADG analysis. In the Korean Yorkshire breed for NBA, the minimum average CR ranged from ~0 (herd A) to 12.79 (herd E). The maximum CR was 89.38 between herds E and G and the lowest CR value was identified between herds A and F. In the Korean Landrace herd, for NBA, the highest average CR was 0.09(herd H) and the lowest average CR was ~0 (herds A and F). In the analysis of the Korean Duroc herd, the highest average CR of NBA was ranges of 1.17 ~ 4.70. According to Mathur et al.(Mathur, Sullivan et al. 1998), the recommended that minimum average CR for ADG and NBA is 3% and 1.5%, respectively. When these criteria for both performance and reproductive traits are met respectively, the estimated breeding value (EBV) comparison between herds can be performed accurately. The average CR values for the Landrace herd are below this criterion, so the values for the Landrace herd were excluded in the next evaluation step.

Table 5.4. Connectedness rating (CR) for NBA among herds

Herds	Yorkshire connectedness rating (%)			Landrace connectedness rating (%)			Duroc connectedness rating (%)		
	Mean	Max	Most connected herd	Mean	Max	Most connected herd	Mean	Max	Most connected herd
A	~0	~0	F	~0	~0	F	1.17	3.40	G
B	0.82	3.65	G	0.02	0.1	H	1.27	4.01	G
C	0.17	0.86	H	0.08	0.37	H	3.23	11.4	G
D	8.40	59.60	F	-	-	-	-	-	-
E	12.79	89.38	G	-	-	-	-	-	-
F	8.71	59.55	E	~0	~0	A	-	-	-
G	11.7	89.38	E	-	-	-	4.70	11.4	C
H	0.20	0.86	C	0.09	0.37	C	-	-	-

Figure 5.2 Average connectedness rating (CR) for NBA with three breeds



5.4.3 Evaluation of connectedness ratings (CR) using the Variance of Estimated Differences between Herd effects (VED)

If two herds are highly connected, the Prediction Error Variance of the Difference (PEVD) decreases. The accuracy of the estimated breeding value (EBV) is therefore greater when a pair of herds is evaluated jointly. According to Kennedy and Trus (Kennedy and Trus 1993), the variance of estimated differences between herd effects (VED) is highly correlated with the average PEV of pairwise comparisons of EBVs. Therefore, VED can be used as an evaluation for CR. Except for the Korean Landrace CR result, the VED for the Korean Yorkshire and Duroc herds was calculated. The VED for ADG and NBA traits in Korean Yorkshire and Duroc breeds are shown in Table 5.5~5.8. The VED decreases as the CR increases, of which refers that the VED can be used as validation indicator of accuracy of the CR.

Table 5.5 The Variance of Estimated Differences between Herd effects (VED) for ADG among Korean Yorkshire herds

Herds	CR (%) > 3	VED
A,D	5.86	0.0024
A,F	9.56	0.0028
B,E	4.04	0.0022
B,G	4.41	0.0069
C,E	3.86	0.0017
C,H	7.42	0.0067
D,E	47.97	0.0010
D,F	86.81	0.0004
D,G	5.53	0.0059
E,F	65.86	0.0006
E,G	93.44	0.0001
E,H	7.22	0.0027

Table 5.6 VED for ADG among Korean Duroc herds

Herds	CR (%) > 3	VED
A,B	3.53	0.00413
A,C	10.77	0.00400
A,G	5.82	0.00402
B,C	13.50	0.00462
B,G	27.81	0.00342

Table 5.7 VED for NBA among Korean Yorkshire herds

Herds	CR (%) > 1.5	VED
B,E	2.40	0.0456
B,G	3.65	0.0762
D,E	7.39	0.0456
D,F	59.55	0.0203
E,F	9.86	0.0337
E,G	89.38	0.0048

Table 5.8. VED for NBA among Korean Duroc herds

Herds	CR (%) > 1.5	VED
A,G	3.40	0.0809
B,G	4.01	0.0847
C,G	11.40	0.0469

5.5 Discussion

A certain level of connectedness is needed for accurate estimated breeding value comparisons between herds. In this study, 104,380 performance data items and 83,200 reproduction data items from three different swine breeds across a total of eight farms were used to analyze connectedness using the CR method. The range of values of CR for ADG in Korean swine herds was ranged from 0.55 to 28.05. From the results, while two breeds, the Yorkshire and Duroc were deemed satisfactory with an average CR greater than 3%, however, those of the Landrace was lower than 3%. It is concluded that it is possible to compare genetic evaluation results among the Yorkshire and Duroc herds for the ADG trait. The range of values of CR for NBA in Korean swine herds was between ~0 and 12.79 of the Yorkshire herd. However, almost 65% of the average CR for NBA was not greater than 1.5% and the average CR in the Landrace herd does not meet either the ADG or NBA criterion. The validation for CR was performed by the variance of estimated difference between herd effects (VED). PEVD decreased 80% of the pair validations when the joint evaluation was performed.

Connectedness analysis in Korea swine breeding herds has not yet been carried out for the combined evaluation of the genetic process for the three different breeds. Therefore, this study provides new insight for understanding the connectedness among Korean swine breeding herds. It was found that special effort is needed to enhance connectedness in the Landrace breed. This result may help to improve future joint Korean swine breeding evaluation programs.

General Discussion

General Discussion

Various livestock genomes now have been created through efforts to adapt to environment. This kind of struggle like a natural and artificial selection leaves unique marks in the livestock genomes. Those traces can be detected through the development of bioinformatics and the accumulation of livestock genome data. Positive selection regions can be detected using the pattern of the frequency spectrum, linkage disequilibrium and population differentiation. Each detection methods have different target and time scale for its use. So, the composite methods approach can be a increasing the reliability of these analyzes.

In this dissertation, I tried to reveal the selective sweep signals in the whole genome of Korean indigenous goat breeds using XP-EHH and XP-CLR statistics. By comparing the genome of Korean crossbred goat breeds, distinct genetic features can be identified. It is studied that Native Korean goats are resistant to lumbar paralysis which has a severe effect on goats. Among the putative candidate selective sweep genes, *CCR3* gene may play an essential role in the native Korean goat's resistance to the lumbar paralysis. Because this gene encodes a chemokine receptor that is important to cytokine response. I found SNP frequency differences between the breeds in this gene region.

Korean native pigs have lower litter size and slower growth rate compared to the Korean imported pigs. I searched for the selective sweep regions related to the reproductive performances using Fst and Homozygosity test. Several genes were associated with reproduction traits. *PLSCR4* is a membrane protein linked to uterine function and ovulation that is reported with a total number of piglets. I identified two missense variants in *PLSCR4* regions. Several putative genes were also identified the reproductive ability of sows and boars under selection in imported pig breeds that are widely used in Korea. Residual feed efficiency is a one of the major challenges for the pig industry. This is because it is one of the largest consumptions of the pig industry directly related to the economic performance. In this study, I found the *EPB42* that is related to the residual feed intake. It is reported that this gene is upregulated in swines with the low residual feed intake.

Thoroughbred is known for speed and the epitomes of variation under domestication.

But it is not yet fully analyzed using whole genome data. So, I tried to reveal the novel selective regions comparing to Jeju horses which have been bred under different environments for a long time. According to Leuckart's law, the size of the eye of an animal is related to its maximum speed of movement. Thoroughbred is the fastest among the horse breeds, so I aimed to study why they run faster using XP-EHH and XP-CLR statistics. I identified genes related to the eye photoreceptor cell development. The strong selection of the eye photoreceptor cell development maybe directly affected the increment of ocular size. *CRB1* is a protein that plays an essential role in normal vision. This protein is found in the brain and the retina, which is the specialized tissue at the back of the eye that detects light and color. In the retina, the *CRB1* protein appears to be critical for the normal development of light-sensing cells called photoreceptors and reported to the structure and orientation of photoreceptor.

However, limitations of these detecting positive selection signatures studies might have a chance of false positive results. Therefore, gene expression analysis and candidate gene approach experimental procedures are required to validate the candidate genes.

Finally, I tried to estimate the connectedness which means genetic similarity between herds. It can affect the prediction error variance of difference in estimated breeding values. The Korean swine genomic selection has been improved individually. But confronting large foreign companies, our swine industries need to consolidate to improve our swine breed. This is the first step of the integrated analysis. I analyzed the connectedness using Connectedness Rating(CR). CR is numerously reported as a good indicator for connectedness. This study suggested that four out of eight herds are possible to evaluate genetic values together for Yorkshire breed. In Duroc, it is enabled to perform joint analysis, however Landrace need to more genetic exchange.

References

Baird, D. T. and B. K. Campbell (1998). "Follicle selection in sheep with breed differences in ovulation rate." Molecular and cellular endocrinology **145**(1): 89-95.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological): 289-300.

Benjelloun, B., F. J. Alberto, I. Streeter, F. Boyer, E. Coissac, S. Stucki, M. BenBati, M. Ibnelbachyr, M. Chentouf and A. Bechchari (2015). "Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data." Frontiers in Genetics **6**: 107.

Bindea, G., B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski and J. Galon (2009). "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks." Bioinformatics **25**(8): 1091-1093.

Bonnet, A., K.-A. Le Cao, M. SanCristobal, F. Benne, C. Robert-Granié, G. Law-So, S. Fabre, P. Besse, E. De Billy, H. Quesnel, F. Hatey and G. Tosser-Klopp (2008). "In vivo gene expression in granulosa cells during pig terminal follicular development." Reproduction **136**(2): 211-224.

Browning, S. R. and B. L. Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering." The American Journal of Human Genetics **81**(5): 1084-1097.

Chang-Yeon, C., Y. Sung-Heum, C. Byung-Wook and C. Gil-Jae (2008). "Genetic characterization and polymorphisms for parentage testing of the Jeju horse using 20 microsatellite loci." Journal of Veterinary Medical Science **70**(10): 1111-1115.

Chen, H., N. Patterson and D. Reich (2010). "Population differentiation as a test for

selective sweeps." Genome research **20**(3): 393-402.

Cho, G.-J. (2007). "Genetic Relationship and Characteristics Using Microsatellite DNA Loci in Horse Breeds." Journal of Life Science **17**(5): 699-705.

Cho, I.-C., T. Zhong, B.-Y. Seo, E.-J. Jung, C.-K. Yoo, J.-H. Kim, J.-B. Lee, H.-T. Lim, B.-W. Kim and J.-H. Lee (2011). "Whole-genome association study for the roan coat color in an intercrossed pig population between Landrace and Korean native pig." Genes & Genomics **33**(1): 17-23.

Choi, J.-W., W.-H. Chung, K.-T. Lee, E.-S. Cho, S.-W. Lee, B.-H. Choi, S.-H. Lee, W. Lim, D. Lim, Y.-G. Lee, J.-K. Hong, D.-W. Kim, H.-J. Jeon, J. Kim, N. Kim and T.-H. Kim (2015). "Whole-genome resequencing analyses of five pig breeds, including Korean wild and native, and three European origin breeds." DNA Research **22**(4): 259-267.

Choi, S., Y. Choy, Y. Kim and S. Hur (2006). "Effects of feeding browses on growth and meat quality of Korean Black Goats." Small Ruminant Research **65**(3): 193-199.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly **6**(2): 80-92.

Crigler, L., R. C. Robey, A. Asawachaicharn, D. Gaupp and D. G. Phinney (2006). "Human mesenchymal stem cell subpopulations express a variety of neuro-regulatory molecules and promote neuronal cell survival and neuritogenesis." Experimental neurology **198**(1): 54-64.

Cui, Y., W. Wang, N. Dong, J. Lou, D. K. Srinivasan, W. Cheng, X. Huang, M. Liu, C. Fang and J. Peng (2012). "Role of corin in trophoblast invasion and uterine spiral artery remodelling in pregnancy." Nature **484**(7393): 246-250.

Dalai, S. K., D. Das and S. K. Kar (1998). "Setaria DigitataAdult 14-to 20-kDa Antigens

Induce Differential Thl/Th2 Cytokine Responses in the Lymphocytes of Endemic Normals and Asymptomatic Microfilariae Carriers in Bancroftian Filariasis." Journal of clinical immunology **18**(2): 114-123.

Dalva, M. B., M. A. Takasu, M. Z. Lin, S. M. Shamah, L. Hu, N. W. Gale and M. E. Greenberg (2000). "EphB receptors interact with NMDA receptors and regulate excitatory synapse formation." Cell **103**(6): 945-956.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth and S. T. Sherry (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). "DAVID: database for annotation, visualization, and integrated discovery." Genome biology **4**(9): R60.

Dong, Y., M. Xie, Y. Jiang, N. Xiao, X. Du, W. Zhang, G. Tosser-Klopp, J. Wang, S. Yang and J. Liang (2013). "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus)." Nature biotechnology **31**(2): 135-141.

Dubé, É., J. Dufresne, P. T. Chan, L. Hermo and D. G. Cyr (2010). "Assessing the role of claudins in maintaining the integrity of epididymal tight junctions using novel human epididymal cell lines." Biology of reproduction **82**(6): 1119-1128.

Dybkaer, K., J. Iqbal, G. Zhou, H. Geng, L. Xiao, A. Schmitz, F. d'Amore and W. C. Chan (2007). "Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways." BMC genomics **8**(1): 1.

Elferink, M. G., H.-J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans and M. A. Groenen (2012). "Signatures of selection in the genomes of commercial and non-commercial chicken breeds." PLoS One **7**(2): e32720.

Emmons, D. and E. Lister (1976). "Quality of protein in milk replacers for young calves. I. Factors affecting in vitro curd formation by rennet (chymosin, rennin) from reconstituted skim milk powder." Canadian Journal of Animal Science **56**(2): 317-325.

Evanno, G., S. Regnaut and J. Goudet (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." Molecular ecology **14**(8): 2611-2620.

Evans, D. T., R. Serra-Moreno, R. K. Singh and J. C. Guatelli (2010). "BST-2/tetherin: a new component of the innate immune response to enveloped viruses." Trends in microbiology **18**(9): 388-396.

Fontanesi, L., D. Calò, G. Galimberti, R. Negrini, R. Marino, A. Nardone, P. Ajmone-Marsan and V. Russo (2014). "A candidate gene association study for nine economically important traits in Italian Holstein cattle." Animal genetics **45**(4): 576-580.

Foulley, J. L., E. Hanocq and D. Boichard (1992). "A criterion for measuring the degree of connectedness in linear models of genetic evaluation." Genetics Selection Evolution **24**(4): 315-330.

Geris, K., B. De Groef, S. Rohrer, S. Geelissen, E. Kuhn and V. Darras (2003). "Identification of somatostatin receptors controlling growth hormone and thyrotropin secretion in the chicken using receptor subtype-specific agonists." Journal of endocrinology **177**(2): 279-286.

Gilmour, A., B. Gogel, B. Cullis, S. Welham and R. Thompson (2015). "ASReml user guide release 4.1 structural specification." Hemel hempstead: VSN international ltd.

Giuffra, E., J. Kijas, V. Amarger, Ö. Carlborg, J.-T. Jeon and L. Andersson (2000). "The origin of the domestic pig: independent domestication and subsequent introgression." Genetics **154**(4): 1785-1791.

Gonçalves, P. B., R. Ferreira, B. Gasperin and J. F. Oliveira (2012). "Role of angiotensin

in ovarian follicular development and ovulation in mammals: a review of recent advances." Reproduction **143**(1): 11-20.

Gu, J., N. Orr, S. D. Park, L. M. Katz, G. Sulimova, D. E. MacHugh and E. W. Hill (2009). "A genome scan for positive selection in thoroughbred horses." PloS one **4**(6): e5767.

Haley, C. and G. Lee (1992). "Genetic basis of prolificacy in Meishan pigs." Journal of reproduction and fertility. Supplement **48**: 247-259.

Heard-Booth, A. N. and E. C. Kirk (2012). "The influence of maximum running speed on eye size: a test of Leuckart's Law in mammals." The Anatomical Record **295**(6): 1053-1062.

Hinchcliff, K. W., A. J. Kaneps and R. J. Geor (2008). Equine exercise physiology: The science of exercise in the athletic horse, Elsevier Health Sciences.

Hivert, B., Z. Liu, C.-Y. Chuang, P. Doherty and V. Sundaresan (2002). "Robo1 and Robo2 are homophilic binding molecules that promote axonal growth." Molecular and Cellular Neuroscience **21**(4): 534-545.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting FST." Nature Reviews Genetics **10**(9): 639-650.

Howland, H. C., S. Merola and J. R. Basarab (2004). "The allometry and scaling of the size of vertebrate eyes." Vision research **44**(17): 2043-2065.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic acids research **37**(1): 1-13.

Huang, L.-S., E. Voyiaziakis, H. L. Chen, E. M. Rubin and J. W. Gordon (1996). "A

novel functional role for apolipoprotein B in male infertility in heterozygous apolipoprotein B knockout mice." Proceedings of the National Academy of Sciences **93**(20): 10903-10907.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen and T. Down (2002). "The Ensembl genome database project." Nucleic acids research **30**(1): 38-41.

Hubisz, M. J., D. Falush, M. Stephens and J. K. Pritchard (2009). "Inferring weak population structure with the assistance of sample group information." Molecular ecology resources **9**(5): 1322-1332.

Jackson, J. E. (2005). A user's guide to principal components, Wiley. com.

Jan, Y.-N. and L. Y. Jan (2003). "The control of dendrite development." Neuron **40**(2): 229-242.

Kelley, J. L., J. Madeoy, J. C. Calhoun, W. Swanson and J. M. Akey (2006). "Genomic signatures of positive selection in humans and the limits of outlier approaches." Genome research **16**(8): 980-989.

Kennedy, B. and D. Trus (1993). "Considerations on genetic connectedness between management units under an animal model." Journal of animal science **71**(9): 2341-2352.

Kim, K., Y. H. Yang, S. S. Lee, C. Park, R. Ma, J. Bouzat and H. Lewin (1999). "Phylogenetic relationships of Cheju horses to other horse breeds as determined by mtDNA D-loop sequence polymorphism." Animal Genetics **30**(2): 102-108.

Kim, K. S., J. S. Yeo and J. W. Kim (2002). "Assessment of genetic diversity of Korean native pig (Sus scrofa) using AFLP markers." Genes & genetic systems **77**(5): 361-368.

Kim, S. W., S. B. Park, M. J. Kim, D. H. Kim and D.-G. Yim (2014). "Effects of Different Levels of Concentrate in the Diet on Physicochemical Traits of Korean Native Black Goat Meats." Korean Journal for Food Science of Animal Resources **34**(4): 457.

Klomtong, P., K. Chaweewan, Y. Phasuk and M. Duangjinda (2015). "MC1R, KIT, IGF2, and NR6A1 as markers for genetic differentiation in Thai native, wild boars, and Duroc and Chinese Meishan pigs." Genetics and Molecular Research **14**(4): 12723-12732.

Knox, R. V. (2014). Impact of swine reproductive technologies on pig and global food production. Current and Future Reproductive Technologies and World Food Production. Nicolas D. G. Cli Lamb. New York, Springer. **752:** 131-160.

Laloë, D., F. Phocas and F. Menissier (1996). "Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation." Genetics selection evolution **28**(4): 359-378.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Laurie, G. W., L. A. Olsakovsky, B. P. Conway, R. L. McKown, K. Kitagawa and J. J. Nichols (2008). "Dry eye and designer ophthalmics." Optometry and vision science: official publication of the American Academy of Optometry **85**(8): 643.

Lazari, M. F. M., T. F. G. Lucas, F. Yasuhara, G. R. O. Gomes, E. R. Siu, C. Royer, S. A. F. Fernandes and C. S. Porto (2009). "Estrogen receptors and function in the male reproductive system." Arquivos Brasileiros de Endocrinologia & Metabologia **53**(8): 923-933.

Lazaros, L. A., N. V. Xita, E. G. Hatzi, A. I. Kaponis, T. J. Stefos, N. I. Plachouras, G. V. Makrydimas, N. V. Sofikitis, K. A. Zikopoulos and I. A. Georgiou (2011). "Association of paraoxonase gene polymorphisms with sperm parameters." Journal of andrology **32**(4): 394-401.

Lee, H.-J., J. Kim, T. Lee, J. K. Son, H.-B. Yoon, K.-S. Baek, J. Y. Jeong, Y.-M. Cho, K.-T. Lee and B.-C. Yang (2014). "Deciphering the genetic blueprint behind Holstein milk proteins and production." Genome biology and evolution **6**(6): 1366-1374.

Lee, T.-H., H. Guo, X. Wang, C. Kim and A. H. Paterson (2014). "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data." BMC genomics **15**(1): 162.

Leo, A. and B. Schraven (2001). "Adapters in lymphocyte signalling." Current opinion in immunology **13**(3): 307-316.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, W., D. Liu, W. Chang, X. Lu, Y. Wang, H. Wang, C. Zhu, H. Lin, Y. Zhang and J. Zhou (2014). "Role of IGF2BP3 in trophoblast cell invasion and migration." Cell death & disease **5**(1): e1025.

Ma, Y., H. Zhang, Q. Zhang and X. Ding (2014). "Identification of selection footprints on the X chromosome in pig." PLoS One **9**(4): e94911.

Marsillach, J., M. A. Checa, J. Pedro-Botet, R. Carreras, J. Joven and J. Camps (2010). "Paraoxonase-1 in female infertility: a possible role against oxidative stress–induced inflammation." Fertility and sterility **94**(3): 1132-1134.

Mathur, P. (2005). "Importance of 'connectedness' between herds for effective across herd genetic evaluation." Journal of South China Agricultural University **26**: 61-68.

Mathur, P., B. Sullivan and J. Chesnais (1998). "A practical method for estimating connectedness in large livestock populations with an application to Canadian swine herds." Journal of Animal Science **76**.

Mathur, P., B. Sullivan and J. Chesnais (2002). Estimation of the degree connectedness between herds or management groups in the canadian swine population.

Mathur, P., B. Sullivan and J. Chesnais (2002). Measuring connectedness: concept and

application to a large industry breeding program. Proc. 7th World Congr. Genet. Appl. to Livest. Prod.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel and M. Daly (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

McLaren, D. G. and M. Bovey (1992). "Genetic influences on reproductive performance." Veterinary Clinics of North America: Food Animal Practice **8**(3): 435-459.

Mombaerts, P. (2004). "Genes and ligands for odorant, vomeronasal and taste receptors." Nature Reviews Neuroscience **5**(4): 263-278.

Montgomery, E. S. (1971). The Thoroughbred, London, UK, Thomas Yoseloff Ltd.

Moon, S., T.-H. Kim, K.-T. Lee, W. Kwak, T. Lee, S.-W. Lee, M.-J. Kim, K. Cho, N. Kim and W.-H. Chung (2015). "A genome-wide scan for signatures of directional selection in domesticated pigs." BMC genomics **16**(1): 1.

Moon, S., J. W. Lee, D. Shin, K.-Y. Shin, J. Kim, I.-Y. Choi, J. Kim and H. Kim (2015). "A genome-wide scan for selective sweeps in racing horses." Asian-Australasian journal of animal sciences **28**(11): 1525-1531.

Nachman, M. W. (2001). "Single nucleotide polymorphisms and recombination rate in humans." TRENDS in Genetics **17**(9): 481-485.

Nagashima, T., Q. Li, C. Clementi, J. P. Lydon, F. J. DeMayo and M. M. Matzuk (2013). "BMPR2 is required for postimplantation uterine function and pregnancy maintenance." The Journal of clinical investigation **123**(6): 2539-2550.

Niimura, Y. (2009). "Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents." Human genomics **4**(2): 1.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White and D. Ako-Adjei (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic acids research **44**(D1): D733-D745.

Odahara, S., H. Chung, S. Choi, S. Yu, S. Sasazaki, H. Mannen, C. Park and J. Lee (2006). "Mitochondrial DNA diversity of Korean native goats." ASIAN AUSTRALASIAN JOURNAL OF ANIMAL SCIENCES **19**(4): 482.

Ohshima, T. (2014). "Neuronal migration and protein kinases." Frontiers in neuroscience **8**.

Onteru, S., B. Fan, Z. Q. Du, D. Garrick, K. Stalder and M. Rothschild (2012). "A whole-genome association study for pig reproductive traits." Animal Genetics **43**(1): 18-26.

Park, W., J. Kim, H. Kim, J. Choi, J. Park and M. Robinson-Rechavi (2014). "Investigation of De Novo Unique Differentially Expressed Genes Related to Evolution in Exercise." PLoS One **9**(3): e91418.

Peterlin, B., B. Zorn, M. Volk and T. Kunej (2006). "Association between the apolipoprotein B signal peptide gene insertion/deletion polymorphism and male infertility." Molecular human reproduction **12**(12): 777-779.

Phillippe, M., D. F. Bradley, H. Ji, K. H. Oppenheimer and E. K. Chien (2006). "Phospholipid scramblase isoform expression in pregnant rat uterus." Journal of the Society for Gynecologic Investigation **13**(7): 497-501.

Pickrell, J. K. and J. K. Pritchard (2012). "Inference of population splits and mixtures from genome-wide allele frequency data." PLoS Genet **8**(11): e1002967.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick and D. Reich

(2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics **38**(8): 904-909.

Puigserver, P. and B. M. Spiegelman (2003). "Peroxisome proliferator-activated receptor-γ coactivator 1α (PGC-1α): transcriptional coactivator and metabolic regulator." Endocrine reviews **24**(1): 78-90.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker and M. J. Daly (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics **81**(3): 559-575.

Quach, T. T., S. M. Wilson, V. Rogemond, N. Chounlamountri, P. E. Kolattukudy, S. Martinez, M. Khanna, M.-F. Belin, R. Khanna and J. Honnorat (2013). "Mapping CRMP3 domains involved in dendrite morphogenesis and voltage-gated calcium channel regulation." J Cell Sci **126**(18): 4262-4273.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Rauw, W., E. Kanis, E. Noordhuizen-Stassen and F. Grommers (1998). "Undesirable side effects of selection for high production efficiency in farm animals: a review." Livestock Production Science **56**(1): 15-33.

Ricard, A., E. Bruns and E. Cunningham (2000). "Genetics of performance traits." The genetics of the horse: 411-538.

Rubin, C.-J., H.-J. Megens, A. M. Barrio, K. Maqbool, S. Sayyab, D. Schwochow, C. Wang, Ö. Carlborg, P. Jern and C. B. Jørgensen (2012). "Strong signatures of selection in the domestic pig genome." Proceedings of the National Academy of Sciences **109**(48): 19529-19536.

Rubin, C.-J., M. C. Zody, J. Eriksson, J. R. Meadows, E. Sherwood, M. T. Webster, L.

Jiang, M. Ingman, T. Sharpe and S. Ka (2010). "Whole-genome resequencing reveals loci under selection during chicken domestication." Nature **464**(7288): 587-591.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson and G. J. McDonald (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll and R. Gaudet (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.

Sallusto, F., C. R. Mackay and A. Lanzavecchia (1997). "Selective expression of the eotaxin receptor CCR3 by human T helper 2 cells." Science **277**(5334): 2005-2007.

Severino, P., E. Silva, G. L. Baggio-Zappia, M. K. C. Brunialti, L. A. Nucci, O. Rigato Jr, I. D. C. G. da Silva, F. R. Machado and R. Salomao (2014). "Patterns of gene expression in peripheral blood mononuclear cells and outcomes from patients with sepsis secondary to community acquired pneumonia." PloS one **9**(3): e91886.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research **13**(11): 2498-2504.

Sharma, A., C. Huard, C. Vernochet, D. Ziemek, K. M. Knowlton, E. Tyminski, T. Paradis, Y. Zhang, J. E. Jones and D. von Schack (2014). "Brown fat determination and development from muscle precursor cells by novel action of bone morphogenetic protein 6." PloS one **9**(3): e92608.

Sirko-Osadsa, D. A., M. A. Murray, J. A. Scott, M. A. Lavery, M. L. Warman and N. H. Robin (1998). "Stickler syndrome without eye involvement is caused by mutations in COL11A2, the gene encoding the α2(XI) chain of type XI collagen." The Journal of pediatrics **132**(2): 368-371.

Škorput, D., G. Gorjanc and Z. Luković (2012). "Evaluation of connectedness between the management units of Landrace breed of pigs in Croatia." Acta agriculturae Slovenica **3**: 182.

Sleer, L. S. and C. C. Taylor (2007). "Cell-type localization of platelet-derived growth factors and receptors in the postnatal rat ovary and follicle." Biology of reproduction **76**(3): 379-390.

Sleer, L. S. and C. C. Taylor (2007). "Platelet-derived growth factors and receptors in the rat corpus luteum: localization and identification of an effect on luteogenesis." Biology of reproduction **76**(3): 391-400.

Soares, M. J., D. Chakraborty, K. Kubota, S. J. Renaud and M. K. Rumi (2014). "Adaptive mechanisms controlling uterine spiral artery remodeling during the establishment of pregnancy." The International journal of developmental biology **58**: 247.

Soga, N. (2009). "The effect of connectedness on the bias and accuracy for prediction of breeding value in swine herds."

Soga, N., M. L. Spangler, C. Schwab, P. Berger and T. Baas (2010). "Comparison of connectedness measures and changes in connectedness of the US Duroc population."

Son, Y. (1999). "Production and uses of Korean native Black goat." Small Ruminant Research **34**(3): 303-308.

Sun, C., C. Wang, Y. Wang, Y. Zhang and Q. Zhang (2009). "Evaluation of connectedness between herds for three pig breeds in China." animal **3**(4): 482-485.

Tarrade, A., R. L. Kuen, A. Malassiné, V. Tricottet, P. Blain, M. Vidaud and D. Evain-Brion (2001). "Characterization of human villous and extravillous trophoblasts isolated from first trimester placenta." Laboratory investigation **81**(9): 1199-1211.

Taye, M., J. Kim, S. H. Yoon, W. Lee, O. Hanotte, T. Dessie, S. Kemp, O. A. Mwai, K. Caetano-Anolles and S. Cho (2017). "Whole genome scan reveals the genetic signature of African Ankole cattle breed and potential for higher quality beef." BMC genetics **18**(1): 11.

Taye, M., W. Lee, K. Caetano-Anolles, T. Dessie, O. Hanotte, O. A. Mwai, S. Kemp, S. Cho, S. J. Oh, H. K. Lee and H. Kim (2017). "Whole genome detection of signature of positive selection in African cattle reveals selection for thermotolerance." Anim Sci J **88**(12): 1889-1901.

Tower, C. L., S. Lui, N. R. Charlesworth, S. D. Smith, J. D. Aplin and R. L. Jones (2010). "Differential expression of angiotensin II type 1 and type 2 receptors at the maternal–fetal interface: potential roles in early placental development." Reproduction **140**(6): 931-942.

Uccelli, A., F. Benvenuto, A. Laroni and D. Giunti (2011). "Neuroprotective features of mesenchymal stem cells." Best practice & research Clinical haematology **24**(1): 59-64.

Vallet, J. and B. Freking (2007). "Differences in placental structure during gestation associated with large and small pig fetuses." Journal of animal science **85**(12): 3267-3275.

Vikkula, M., E. Madman, V. C. Lui, N. I. Zhidkova, G. E. Tiller, M. B. Goldring, S. E. van Beersum, M. C. de Waal Malefijt, F. H. van den Hoogen and H.-H. Ropers (1995). "Autosomal dominant and recessive osteochondrodysplasias associated with the< i> COL11A2</i> locus." Cell **80**(3): 431-437.

Vincent, A., I. Louveau, F. Gondret, C. Tréfeu, H. Gilbert and L. Lefaucheur (2015). "Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle." Journal of animal science **93**(6): 2745-2758.

Vitti, J. J., S. R. Grossman and P. C. Sabeti (2013). "Detecting natural selection in genomic data." Annual review of genetics **47**: 97-120.

Waites, C. L., A. M. Craig and C. C. Garner (2005). "Mechanisms of vertebrate synaptogenesis." Annu. Rev. Neurosci. **28**: 251-274.

Wang, X., J. Liu, G. Zhou, J. Guo, H. Yan, Y. Niu, Y. Li, C. Yuan, R. Geng and X. Lan (2016). "Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits." Scientific reports **6**: 38932.

Weir, B. S. and C. C. Cockerham (1984). "Estimating F-statistics for the analysis of population structure." Evolution: 1358-1370.

Wright, K. T., W. E. Masri, A. Osman, J. Chowdhury and W. E. Johnson (2011). "Concise review: bone marrow for the treatment of spinal cord injury: mechanisms and clinical applications." Stem cells **29**(2): 169-178.

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher (2011). "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics **88**(1): 76-82.

Yin, X., M. Guo, Q. Gu, X. Wu, P. Wei and X. Wang (2014). "Antiviral potency and functional analysis of tetherin orthologues encoded by horse and donkey." Virology journal **11**(1): 151.

Yoshimura, Y., M. Karube, H. Aoki, T. Oda, N. Koyama, A. Nagai, Y. Akimoto, H. Hirano and Y. Nakamura (1996). "Angiotensin II induces ovulation and oocyte maturation in rabbit ovaries via the AT2 receptor subtype." Endocrinology **137**(4): 1204-1211.

Yu, B., T. Qian, Y. Wang, S. Zhou, G. Ding, F. Ding and X. Gu (2012). "miR-182 inhibits Schwann cell proliferation and migration by targeting FGF9 and NTM, respectively at an early stage following sciatic nerve injury." Nucleic acids research **40**(20): 10356-10365.

Zambonelli, P., R. Davoli, M. Bigi, S. Braglia, L. F. De Paolis, L. Buttazzoni, M. Gallo and V. Russo (2013). "SNPs detection in DHPS-WDR83 overlapping genes mapping on porcine chromosome 2 in a QTL region for meat pH." BMC genetics **14**(1): 1.

Zerial, M. and H. McBride (2001). "Rab proteins as membrane organizers." Nature reviews Molecular cell biology **2**(2): 107-117.

Zhang, W., B. Yang, J. Zhang, L. Cui, J. Ma, C. Chen, H. Ai, S. Xiao, J. Ren and L. Huang (2016). "Genome-wide association studies for fatty acid metabolic traits in five divergent pig populations." Scientific reports **6**.

Zhao, C., M. C. Saul, T. Driessen and S. C. Gammie (2012). "Gene expression changes in the septum: possible implications for microRNAs in sculpting the maternal brain." PloS one **7**(6): e38602.

Zhou, L., J. Li, J. Yang, C. Liu, X. Xie, Y. He, X. Liu, W. Xin, W. Zhang and J. Ren (2016). "Genome-wide mapping of copy number variations in commercial hybrid pigs using a high-density SNP genotyping array." Russian Journal of Genetics **52**(1): 85-92.

국문초록

가축 유전체 내 적응적 진화 흔적 발굴과 혈연연결도의 추정

이원석
농생명공학부 동물생명공학전공
서울대학교 대학원 농업생명과학대학

수천년 동안, 돌연변이, 자연 또는 인위선택, 유전적 부동, 근교교배와 선발 등으로 가축의 유전적 다양성이 다양화될 수 있었다. 최근 생물정보학의 발전으로 그 가축의 역사와 최신 유전자원 정보를 제공해 주고 있다. 이를 이용한 유전 마커와 분자연구는 이와 같은 가축 다양성이 과거에 어떻게 진행되어 현재에 다양한 모습으로 이르게 되었는지 단서를 제공하고 있다. 여기에는 가축의 조상 정보 구성 및 그들의 이동경로와 유전구조가 포함된다. 이와 같은 과거의 정보를 이해하는 것은 두말할 것없이 현재의 가축유전자원을 이해하는데 도움이 된다. 2009년에 소의 전장유전체 정보가 가축 중 처음으로 밝혀졌다. NGS기술의 결과인 유전정보는 생물정보학과 통계학을 이용하여 분석이 가능해졌다. 유전체를 이용하여 이러한 가축의 유전적 배경을 설명할 수 있는 많은 기술들이 출현했다. 분자유전연구학분야는 적응 진화 흔적을 발견하는데 큰 역할을 할 수 있게 되었다. 왜냐하면, 표현형에 영향을 줄 수 있는 DNA의 특정영역을 통계적으로 분석할 수 있게 되었기 때문이다. 가축의 유전적 다양성 정보는 다른 유형의 유전자마커에서 얻을 수 있다. 예를 들어, 상염색체다양성은 집단의 다양성, 유전적 관계, 집단의 유전적 혼합등을 추정하는데 사용할 수 있으며, 반면에, 미토콘드리아 DNA의 다양성을 이용해서는 가축화 당시의 지리적 지역의 추정이나 이주경로를 재구성할 수 있고, 모계창시자의 수등을 추정할 수 있다.

5개의 챕터로 구성된 이 박사 학위 논문은 주로 다양한 가축의 유전체에 남겨진 진화적 흔적을 찾아 이와 관련된 주요한 경제형질과 적응형질을 밝히고 마지막 챕터에서는 유전적 선발 통합분석의 기본이 되는 혈연연결도를 추정하는데 주안점을 두었다.

1장은 기존 연구들에 관한 리뷰로 여러 가축들 특히, 우리나라의 토종품종을 포함한 가축들의 유전자원으로서의 성격을 제시하였고, 양성선택의 흔적에 남겨진 원리에 대해서도 설명하였다. 이 장에서는, 양성선택의 흔적을 발굴하는 목적과 방법과 이전에 전세계에서 유전적으로 다양한 가축들의 양성선택 연구 결과들을 제시하였다. 또한, 혈연연결도의 정의와 이를 측정하기 위해 사용된 여러 통계치 중 CR을 소개하였으며, 이를 이용한 기존의 연구에 대해 리뷰하였다.

2장에서는 한국토종염소에서 자연 또는 인위선택으로 남겨진 유전적 흔적과 관련된 찾아낸 경제형질과 적응형질을 제시하였다. 가축화, 자연 혹은 인위선택 이 모두는 염소의 적응형질에 영향을 미치는 염소유전체 조성을 크게 변화시켰다. 발굴한 염소에서 이러한 영향을 받은 유전 지역과 유전자를 통해 고려할 수 있는 진화역사, 경제적형질과 환경에 적응할 수 있었던 적응형질에 대한 통찰력을 얻을 수 있다. 결론적으로 이 장에서, 한국토종염소에서 교잡종에 비해 상대적으로 강한 뇌척수사상충증 질병저항성 형질과 관련된 유전자를 발굴할 수 있었다. 또한, 초기 생장과 발육에 중요한 유전자 또한 발굴하였다.

3장에서는 한국토종돼지보다 월등한 번식능력을 가진 수입종들의 번식 성적을 설명할 수 있는 유전자를 발굴하기 위해 연구를 진행하였다. 이 기전에 영향을 미치는 유전영역을 찾기 위해, 수입종과 한국토종돼지의 유전체를 이용하여 이들의 상호집단분석을 Fst와 이형접합성 방법을 이용하여 비교하였다. 그 결과 번식능력, 면역, 털색깔 등의 형질과 관련된 유전자를 찾을 수 있었다. 예를 들어, 생식력, 배란율 및 자궁능력과 관련된 *PLSCR4, AGTR1*과 *CORIN* 유전자를 발굴하였다. 그러므로 이 장에서 밝혀낸 이러한 유전자들이 수입종들의 우수한 번식능력에 직간접적으로 영향을 미칠 것이라 사료된다.

4장에서는 서러브레드(경주마)와 한국 토종 제주마의 유전체를 비교하여 다양한 양성선택 흔적을 남긴 유전자를 찾아 보고하였다. 특히, 서러브레드는 경주마로 명성이 높다. XP-EHH와 XP-CLR 방법을 이용하여 각각 98개와

200개의 경주마의 양성선택과 관련된 유전자를 발굴하였다. 더 나아가, 이 유전자들을 이용하여 72개의 BP terms를 분석할 수 있었다. 이 유전자들과 BP terms는 면역, 에너지 대사, 눈크기와 기능에 관련된 형질로 이러한 형질들은 경주능력과 관련이 있다고 추정된다.

5장에서는 우리나라의 양돈농장간 Connectedness rating (CR)을 구하여 제시하였다. 여기에서는 여덟 개의 양돈농장의 세가지 수입종(Yorkshire, Landrace and Duroc)의 104,380개의 산육형질과 83,200개의 번식형질 데이터를 이용하여 일당증체량(ADG)과 산자수(NBA)에 대한 농장간 CR을 구하였다. 그 결과, 요크셔 일당증체량에 대한 평균 CR은 1.32%에서 28.5%였다. 요크셔의 산자수에 대한 평균 CR은 0%에서 12.79%였다. 이는 여덟개의 농장 중 네개의 농장간에 통합유전평가가 가능함을 의미한다. 또한, 두록의 일당증체량에 대한 통합유전평가도 가능하다.

결론적으로, 상기된 연구를 통하여 다양한 가축에서 여러 경제형질과 적응형질에 관련된 기존에 발표되지 않은 유전자를 발굴할 수 있었다. 이러한 발굴이 현재 많이 사육되고 있는 이러한 가축 품종 간에 관찰되는 다양한 표현형 변화에 원인이 되는 적응적 기전에 대한 우리의 이해를 높일 수 있을 것이라 생각된다. 이와 더불어 현지환경적응에 기여하는 분자마커 또한 발굴할 수 있었다. 향후 이러한 마커들이 유전선발과 육종프로그램에 사용될 수 있을 것이라 생각된다. 또한, 종돈장간의 연결도 분석을 통하여 종돈장간의 통합유전분석의 기초를 다져 앞으로 더 정확한 분석이 가능해질 수 있으며 한국형 종돈의 개발에 기초가 될 것이라 사료된다.

핵심어: 염소, 돼지, 말, 양성선택, selective sweep, Connectedness

학번: 2014-22935

감사의 글

우선 학위를 마칠 수 있게 2014년 흔쾌히 연구실에 받아주신 김희발 교수님 감사합니다. 배우는 기간 내내 배려해 주시고 좋은 가르침을 주셨습니다. 또한, 조앤김지노믹스 조서애대표님도 진심으로 감사합니다.

같이 도와가며 여러해 동안 공부하고 연구한 연구실 선배들과 후배들에게 감사합니다. 특히, 입학시기가 비슷해 오래 같이 공부할 수 있었던 멩기스티 학우에게 진심으로 감사합니다. 항상 옆에서 많은 대화를 해주고 연구에서는 날카로운 조언을 주는 진정한 친구였습니다. 그리고 일일이 거론하지는 않지만 몇 년동안 동거동락해준 연구실 여러 친구들이 저에게 큰 힘이 되었습니다. 다시 한번 감사합니다. 그리고 프로그래밍에 어려움을 겪던 제게 자상하게 가르쳐준 조앤김지노믹스 성삼선 누나에게 감사합니다. 여러분들이 없었으면 논문을 마칠 수 없었을 것입니다.

마지막으로 이 자리까지 있게해 준 가족들에게 감사합니다.

아버지 하늘에서 편안히 쉬시길 바랍니다.

어머니 항상 감사합니다.

시안,지안 사랑합니다.