농학석사학위논문

Unraveling genomic characteristics of domesticated animals and its applications using bioinformatic approaches

생물정보학을 이용한 가축유전체의 특성 규명과 그 응용에 대한 연구

2016년 2월

서울대학교 대학원

농생명공학부 동물생명공학전공

김 권 도

Unraveling genomic characteristics of domesticated animals and its applications using bioinformatic approaches

By

Kwondo Kim

Supervisor: Professor Hee-bal Kim

Feb, 2016

Department of Agricultural Biotechnology

Seoul National University

Abstract

Unraveling genomic characteristics of domesticated animals and its applications using bioinformatic approaches

Kwondo Kim

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Domesticated animals have unique genomic characteristics different to wild species due to artificial selection by human being. As their genomic properties have great effect on the production traits of them such as milk yielding, and parity, discovering and analyzing novel genomic features of domesticated animals can provide commercial as well as scientific value. Of these genomic features, single nucleotide polymorphism (SNP) has been widely utilized in a lot of researches. In domesticated animals, SNP was mainly used for discriminating breeds which have a high price in food industry. In practical, breed of individual pigs have been identified using SNP genotyping to guarantee the confidence of food products made from valuable pig breeds. Other forms of genomic variants, ranging from hundreds to millions nucleotides, structural variants cause a variation of genome sequences on a larger scale compared to SNP. One of the structural variants, transposable elements (TE), is able to transpose from one region to other

region of genome sequence, which can produce phenotypic variation along with genetic variation.

Chicken, Gallus gallus, is a valuable species both as a food source and as a model organism for scientific research. In particular, chicken genome, as a melting pot of TE, contains lots of TE sequences. In addition, there have been several instances that demonstrated the effect of TE variation on phenotypic traits. In chapter 2, the genomic DNA of Gyeongbuk Araucana (GA), a newly developed blue-egg laying chicken, was sequenced by next generation sequencing. Using generated DNA fragments, genetic variation based on the TE insertion pattern was investigated and clustering analysis was performed. From the comparative analyses with 12 chicken breeds, three TE insertions specific for GA have been identified, which could be a clue for the cause of phenotypic traits of GA, including blue egg shell formation. Furthermore, the result of clustering TE insertion pattern could provide the information for the position and origin of GA breed in chicken species.

Traceability is defined as a method that traces back the origin of animals or animal products, which is an important step to cope with contagious diseases related to animal products such as food-borne illness. It also performs a role to guarantee the food safety and enhance the consumers' confidence to animal products. Therefore, it is important to identify the origin of animal products for safety purposes. However, there have been only a few studies addressing this issue using classification. In chapter 3, 4,122 commercial pigs originating from 104 farms were genotyped using a customized SNP chip. Using these genotyping data, a model to classify individual pigs according to their origins was constructed and evaluated. In this model, several factors including genetic relationship, classifiers, and features were considered to

establish the best prediction model based on these genotyping data. This study showed that the model with LogitBoost classifier outperformed other models in terms of classification performance under most conditions. Furthermore, a greater level of accuracy was observed when a higher kinship-based cutoff was employed. These two results demonstrated the applicability of a machine learning-based approach using SNP chip data for practical traceability.

The findings in this study provided the insight for the contribution of genomic variants to phenotypic traits of domesticated animals and demonstrated the applicability of genomic variants as classification markers.

Key words: Domesticated animals, Genomic variants, Transposable elements, Traceability, Classification

Student number: 2011-21283

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 Transposable elements

1.1.1 Classification of Transposable elements

Transposable elements (TE) refer to all forms of mobile DNA segments in the genome. Recent technologies for genome sequencing have revealed that TEs of various classes constitute a large fraction of most eukaryotic genomes (Slotkin and Martienssen 2007). Also, there have been numerous studies that transposition of TEs in the genome can lead to various genetic or epigenetic effects in eukaryotic cells by diverse mechanisms.

TEs are generally classified into two classes according to their mechanisms to transpose. Type I TEs, called retrotransposons, require reverse-transcriptase to transpose, which generates RNA-intermediate status of TEs (Wicker, Sabot et al. 2007). Type I TEs are transposed in two stages: first, TE in DNA state is transcribed to RNA, and second, the RNA produced is then reverse transcribed to DNA. This reverse transcribed DNA is finally introduced at a new position in the genome (Kazazian 2004). Retrotransposons undergo duplicative transposition, which means that they increases their total number and provide the potential to expand genomes (Slotkin and Martienssen 2007).

Type I TEs are also divided into two types by the presence of repeat sequences, called long terminal repeat (LTR), at the terminal of TE sequence. LTR retrotransposon containing LTR generally contain pol and gag genes encoding proteins closely related to retroviral proteins (Slotkin and Martienssen 2007). From this point of view, LTR retrotransposons are similar to retroviruses in genomic structure (Whitelaw and Martin 2001). However, LTR retrotransposons lack or contain a remnant of an env gene, which does

not allow them to escape the cell they are in (Slotkin and Martienssen 2007). LTR retrotransposon encodes reverse transcriptase by itself, which is different to some sub-types of non-LTR retrotransposons.

Non-LTR retrotransposons contain several open reading frames (ORFs) encoding proteins that mediate transposition (Malik, Burke et al. 1999). However, all of non-LTR cannot encode reverse transcriptase. The representative sub-type of non-LTR retrotransposons are long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LINEs usually are over 5kb in length and include two ORFs (ORF1 and ORF2). ORF1 encodes an RNA binding protein and ORF2 encodes a protein having an endonuclease as well as a reverse transcriptase (Singer 1982). SINEs, less than 0.5kb in length, do not encode a functional reverse transcriptase and rely on other elements for transposition (Singer 1982).

Type II TEs, called DNA transposons, do not undergo RNA-intermediate step in transposition. Instead, a protein called transposase which is encoded by DNA transposon recognizes the terminal inverted repeats (TIRs) that flank the transposon, excises the TE out of the donor sequence, and then integrates it into a new position of the genome (Feschotte and Pritham 2007). As they are also insert into a new position without duplication, the mechanism of transposition for DNA transposons is described as cut-and-paste (Kapitonov and Jurka 2006).

TEs can be also classified into two types according to their ability for transposition. Autonomous TEs produces all the proteins that are required for transposition(Slotkin and Martienssen 2007). However non-autonomous TEs partly encode required proteins for transposition. Therefore, they need to utilize proteins encoded from the other region of genome for transposition. In

type I or type II TEs, non-autonomous TEs usually lack reverse transcriptase or transposase.

1.1.2 The effects of transposable elements on eukaryotic genome

The transposable characteristics of TEs can have numerous meanings in genome evolution and epigenetic effects. Alterations of genome sequence aroused by TEs are solely regard as genetic effect, which can also lead to epigenetic effects including a change of gene expression, moreover a modification of phenotypic traits.

It was known that TEs constitute a large fraction of eukaryotic genomes. Recent genome sequencing revealed that about 44% of human genome belong to TE sequences (Kidwell 2002). In particular, the considerable fraction of chicken genome consist of TEs (Hillier, Miller et al. 2004). Thus, TEs are closely related to genome evolution. The alterations of genome induced by transposition range from modifications in the size and arrangement of whole genomes to substitutions, deletions, and insertions of a single nucleotide (Kidwell 2002, Kazazian 2004). These modifications of whole genome can be a driver of evolution in terms of a capacity to change the fitness of host.

Epigenetic effects induced by transposition of TEs are mainly associated to regulation of gene expression. A promoter included in TEs occasionally can be active in a new site, which lead to altered expression of gene located in downstream of the new site by introducing a new transcription start site (Slotkin and Martienssen 2007). In addition, TE insertion can disrupt existing cis-regulatory elements or introduce a new cis-regulatory elements such as a transcription factor binding site. At the post-transcriptional level, the TE insertion into the 3' UTR of a gene can offer an alternative polyadenylation

site, a binding site for a microRNA or an RNA-binding protein. TE inserted into intron can also modify splicing pattern of pre-mRNA, which can result in transcription of a novel mRNA isoforms (Feschotte 2008). These TE-induced variations for regulation of gene expression are ultimately associated to the phenotypic traits of host. In practical, several dramatic changes of traits generated by TEs have been reported in animals. In mice, transposition events occurred by retroviral elements place the agouti gene under the control of a promoter present in the LTR of the retrotransposon, which results in a diverse traits of yellow fur, obesity and diabetes (Perry, Copeland et al. 1994). In contrast, TE insertion into upstream of a gene tyrosinase which is responsible for melanin synthesis, was demonstrated to suppress transcription (Wu, Rinchik et al. 1997). As mice, chicken also demonstrated impressive phenotypic changes induced by TE insertions, of which a representative case is associated with blue-egg shell (Wang, Qu et al. 2013, Wragg, Mwacharo et al. 2013). In addition, the insertion of retroviral sequence in intron 4 of the tyrosinase gene results in aberrant transcripts lacking exon 5, which causes the white coat color in the chicken (Chang, Coville et al. 2006). From all these examples, it can be inferred that there might be more cases to show the effects of transposition to phenotypes of host.

1.1.3 Transposable elements as a phylogenetic marker

In the field of phylogenetic studies, the selection of relevant phylogenetic marker should be considered to reflect the true evolutionary history. Generally, a gene or sequence fragment with clonality, neutrality, or evolutionary rate constancy have been used as a marker for phylogenetics (Galtier, Nabholz et

al. 2009). One of these, mitochondrial DNA has been by far the most popular marker over the last three decades, while the use of TE insertions as phylogenetic markers has been popular recently (Kido, Aono et al. 1991, Murata, Takasaki et al. 1993, Xing, Wang et al. 2005, Xing, Wang et al. 2007, Meyer, McLain et al. 2012). As a phylogenetic marker, TE insertion has several features as follow.

First, the probability of two identical TEs independently inserting in the exactly same position in the genome is nearly zero. Second, the insertion of TE can be assumed to be unidirectional because there is no known mechanism to completely remove TEs after their insertion in the genome and the simultaneous removal of TEs is unlikely to happen in multiple genomes (Xing, Wang et al. 2005). These two features, therefore, support the assumption that the clonal inheritance of each TE insertion could be maintained through generations.

As described in previous section, TE insertion could have an effect on phenotypic traits of host, which result in the change of fitness to environment. However, coding and regulatory regions comprise only tiny part of the genome. (in human 5%)(Lander, Linton et al. 2001). Most of the TE insertion events are likely to cause no influence on the fitness of host(Meyer, McLain et al. 2012).

Finally, the presence of a TE insertion at a same position within the genomes from two lineages could provide an evidence that the insertion occurred within a common ancestor of the two lineages(Meyer, McLain et al. 2012).

1.2 Machine-learning approach for traceability

1.2.1 Definition of traceability

In modern food industry, ordinary manufacturers and consumers are completely separated geographically. Due to this, food distribution system for transporting products from manufacturers to consumers has been developed as the scale of food industry growing. However, increasing complexity of the system induced an issue that it is hard to aware where the final products came from.

Traceability is generally defined as a method to guarantee the identification for animals or animal products within the food distribution system (Dalvit, De Marchi et al. 2007), which enable to find the origin of animal product. This is very important issue in case of the outbreak of infectious disease derived from food products.

Golan, Krissof, Kuchler, Nelson, and Price proposed that a traceability system could be characterized by its breadth, depth, and precision (Golan, Krissoff et al. 2004). The breadth describes the amount of information recorded in the traceability system. The depth is associated to how far, back or forward, the system tracks. Finally, the precision is the accuracy with which the system can pinpoint a particular product's movement, and is determined by an acceptable error rate. These characteristics of traceability system might be various depending on the objectives of the systems.

1.2.2 The history of traceability

According to the historical evidences, traceability has been started from a very early stage (3,800 years ago) (Blancou 2001). In this times, individual identification for valuable animals was conducted by means of body markings with a written records. Indelible markings was used over the following centuries. Meanwhile, traceability for disease control purposes commenced later, prompted by the major epizootics. During this period, some animal products could not be traded internationally unless attaining a certificate of origin guaranteeing safety (Blancou 2001).

In modern times, methods utilized in traceability primarily focused on breeding strategy. For example, breeds with high scarcity and values are identified by traceability methods. However, more recently, occurrences of animal-related diseases such as bovine spongiform encephalopathy (BSE) and food poisoning enhanced consumer's demand for food safety related to public health, which led to the need to trace animals and animal products along the food chain (Murphy, Pendell et al. 2008). For this reason, traceability has been interest to all those including researchers involved in livestock production and marketing. In a large number of countries, traceability is already mandatory for most animal products (Dalvit, De Marchi et al. 2007).

Means of identifying products could be diverse according to the types of products or distribution processes (Golan, Krissoff et al. 2004). Conventional method for traceability used tags or records written on papers for identification. Although, these tools present several advantages in data processing, handling and expense, it is a critical weak point that they are vulnerable to fraud or loss (Smith, Tatum et al. 2005, Dalvit, De Marchi et al. 2007).

1.2.3 Genetic traceability

After 90's, the genetic traceability has been appeared as an alternative to conventional traceability systems. Genetic traceability is same as conventional methods except using DNA information for the identification of animals or animal products (Dalvit, De Marchi et al. 2007). DNA molecule has a characteristic allowing to distinguish individuals, moreover DNA is hard to modify, and stable to various processes within food distribution, and able to be extracted from all kinds of tissues (Dalvit, De Marchi et al. 2007). These advantages has led to the growth of applications and researches on DNA marker for traceability. Despite of these advantages of DNA marker, practical problems still existed. In case of individual identification, the implementation is easy due to the need of low number of markers, but in order to conduct this technique properly, the information for DNA marker is need to be collected from every animal at birth. Nevertheless, a number of DNA markers have been developed, evaluated and applied to practical food distribution (Goffaux, China et al. 2005, Negrini, Nicoloso et al. 2008, Negrini, Nicoloso et al. 2009, Ramos, Megens et al. 2011).

Meanwhile, breed or species traceability was implemented to defend and valorize particular products. This approach is based on genetic markers specific to breed or species. Once specific genetic markers are retained, this approach can present high accuracy (Dalvit, De Marchi et al. 2007). In fact, various studies have been performed on these genetic markers, especially, related to coat color (Fernández, Fabuel et al. 2004, D'Alessandro, Fontanesi et al. 2007). In addition to this, another approach was studied related to the probability that an individual came from particular population. This method depends on the assumption that individuals will have more similar genotypes

if they come from the same population. Therefore, to produce high precision, it is necessary to apply differentiated breeds to the reference population. As mentioned above, these two approaches focus on distinguishing breeds or species, the availability is limited to particular products including indigenous variety. However, in practice, most of animal product in food market are derived from crossbred population. As a result, it is difficult to develop a method which can be applied to practical market.

1.2.4 Feature selection

In the machine-learning field, feature selection refer to a procedure that remove redundant variables from total variables so that retrieving a subset of the input set (Jain and Zongker 1997). The main objective of feature selection is improving the prediction or classification accuracy. It can also make the model faster and more cost-effective, and provide an deeper understanding into the underlying process that generated the input data (Guyon and Elisseeff 2003, Saeys, Inza et al. 2007).

In the context of classification, there are three categories of feature selection techniques depending on how they combine the feature selection search with the construction of the classification model (Saeys, Inza et al. 2007).

First, filter method extract the relevant features by using the intrinsic properties of features. Generally, low-scoring features are removed following the calculation of relevance score for each features. This method is simple and fast and independent of the classifier. However, it ignores the interaction between feature and classifier (Yu and Liu 2003).

Different to filter method, wrapper method evaluates features using a specific classification model. It searches an optimal subset from all feature subsets by assessing the classification performance of each features. Advantages of wrapper approaches include dependency to classification model which introduce the interaction between feature subset and classifier. A common drawback of wrapper method is that it has a higher risk of overfitting than filter method and are very computationally intensive (Das 2001).

The third feature selection method, embedded method, combines the step for searching an optimal subset of features and building classification model. It is similar to wrapper method in the incorporation of the interaction with the classification model, while it is less computationally intensive than wrapper method (Saeys, Inza et al. 2007).

1.2.5 Classification for traceability

In the context of classification, tracking back the origin of animal products is a multiclass classification problem because of multiple candidate origins. In general, multiclass classification is associated with more difficulties compared to binary classification (Even-Zohar and Roth 2001). The main problem is related to optimization. For example, minimization of the loss function should be performed given the training sets to build an accurate classifier. Loss function is affected by the number of classes, and minimization of this obstacle could be attained by reducing the number of classes.

Several works have attempted to develop approaches to deal with multiclass classification. One of them is the decomposition of the multiclass

problem into a set of two-class classification problems (Ex. one-against-all and error correcting output codes). The outputs from decomposed binary subproblems are then combined to obtain the multiclass prediction (Hastie and Tibshirani 1998, Lorena, De Carvalho et al. 2008).

The adaptations of the internal operations of the classifier into a multiclass have been also studied, which is, however, either impractical or not easy to perform in some cases (Hsu and Lin 2002, Lorena, De Carvalho et al. 2008).

This chapter will be published in *elsewhere*
as a partial fulfillment of Kwondo Kim's M.Sc program.

Chapter 2. Whole genome sequencing of Gyeongbuk Araucana, a newly developed blue-egg laying chicken breed, reveals its origin and genetic characteristics

2.1 Abstract

Chicken, Gallus gallus, is a valuable species both as a food source and as a model organism for scientific research. Here, the genome of Gyeongbuk Araucana, a rare chicken breed with unique phenotypic characteristics including flight ability, large body size, and laying blue-shelled eggs, was sequenced to identify its genomic features. Genomes of Gyeongbuk Araucana, Leghorn, and Korean Native Chicken were generated at a total of 33.5, 35.82, and 33.23 coverage depths, respectively. Along with the genomes of 12 Chinese breeds, genetic variation based on the transposable element insertion pattern was investigated and phylogenetic analysis was performed to elucidate the cause of phenotypic traits including blue egg shell formation. This study presents results of the first genomic analysis on data from the Gyeongbuk Araucana breed; it has potential to serve as an invaluable resource for future research on the genomic characteristics of this chicken breed as well as others.

2.2 Introduction

Chicken, Gallus gallus, is valuable not only as a food source but also as a model organism for scientific research(Darwin Charles 1859). In the last thousands of years, hundreds of chicken breeds have diverged under natural and artificial selection in a wide variety of circumstances. As a result, chickens have undergone significant phenotypic differentiation in body size, plumage, egg color, and flying ability(West and Zhou 1989).

The shell color of a hen's eggs is genetically determined; while all eggs start out white, in certain breeds pigments are deposited on the egg as it travels through the oviduct (Zhao, Xu et al. 2006). It is believed by many consumers and poultry farmers that blue tinted eggs such as those produced by Araucanas have a higher protein and lower cholesterol content, making them a favorite and preferred among health-conscious consumers. Research published by the U.S National Institute of Health has refuted these claims, revealing that blue eggs in fact contain lower total egg protein content and consistently higher cholesterol levels (SOMES, FRANCIS et al. 1977). Nevertheless, these eggs continue to be a favorite among consumers and demand for this product has encouraged selective breeding for this trait in the poultry production industry, along with the development of breeds which have been optimized for egg production rate and feed-to-egg conversion.

The Gyeongbuk Araucana (GA) domestic chicken is a hybrid breed developed in Gyeongbuk, Korea by crossing the Golden Duckwing Araucana and the White Leghorn, two breeds with very distinct characteristics. The White Leghorn, a small breed with a rump, is renowned for prolific egg-laying as well as a good feed-to-egg conversion ratio. Meanwhile, the Golden

Duckwing Araucana is a similarly sized small rumpless and tufted breed that produces blue-shelled eggs at a relatively slow rate. These two breeds were crossed to produce a Korean chicken variety which would possess the qualities favorable to commercial poultry production from both parent breeds, including the blue tint of their egg shells and high egg production rate. Although both the Golden Duckwing Araucana and White Leghorn are small breeds, the GA resulting from the cross is extremely large and can fly well. Additionally, GA chickens have a combination of phenotypic traits from both parents: a rump but no ear tufts. Although GA is a relatively new breed, it has already been registered in the Domestic Animal Diversity Information System (DAD-IS) of the FAO (Food and Agriculture Organization).

Analysis of genetic information and patterns can be useful for discovering the origin of specific breed or detecting specific trait in each breed. Recent phylogenetic analyses using various genetic information revealed the origin of domesticated chickens (Eriksson, Larson et al. 2008) and the Korean native chicken (KNC) (Kwak, Song et al. 2014). Additionally, genetic variants present between several chicken breeds have been utilized to support the characterization of specific trait in chicken breeds (Dorshorst, Okimoto et al. 2010, Chang, Chen et al. 2012, Freese, Lam et al. 2014). However, although a wide range of genomic studies on domestic animals, and in chickens specifically, have been conducted to investigate the genetic architecture of these species using next generation sequencing, no studies of this nature have been performed on GA. For this reason, whole genome sequencing on GA chickens was performed. Additionally, the whole genome sequencing of Leghorn (LH) and KNC were performed and the whole genome paired-end reads for other 12 chicken breeds were obtained from the sequence

read archive (SRA) in EMBL-EBI database. Using a total of 28 chicken genomes, transposable element (TE) variants were identified based on the idea of blue egg formation induced by TE insertion. From these TE variants, the results of previous studies related to blue egg formation were confirmed. Moreover, 2 candidate genes specific to the GA breed were identified and cladograms were constructed to investigate the TE variants pattern to divide chicken breeds. This study is the first of its kind to report a comprehensive view of the GA chicken breed's TE at a genomic level.

2.3 Materials and Methods

2.3.1 DNA sequencing and sample collection

Blood samples from male GA, LH, and KNC chickens were obtained. Samples were collected from the Livestock Research Institute, Yeongju, Korea. To prevent clotting, blood drawn from the carotid artery was treated with heparin. To generate inserts of ~300 bp fragments, 3 μg of genomic DNA was randomly sheared using Covaris System. The TruSeq Dna Sample Prep. Kit (Illumina, San Diego, CA) was used to construct the library following manufacturer guidelines. Whole genome sequencing was performed using Illumina HiSeq 2000 platform. Additionally, genomic data of 14 chicken samples from the EMBL-EBI database were downloaded, which included 2 Silkies (one is from China and the other is from Taiwan), 1 Taiwanese native chicken (TNC), 1 Leghorn, 1 Tibetan chicken (TB), 1 Shouguang (SG), 1 Wenchang (WC), 1 Beijing You (BY), 1 White Plymouth Rock (WPR), 1 Dong Xiang (DX), 1 Cornish (CN), 1 Luxi Game (LG), 1 Rhode Island Red (RIR), and 1 Red Jungle Fowl (RJF). Also, data from 5 Korean Native Chickens, used in a previous study, were also included to increase the quality of variants calling and following analyses. A quality check on raw sequence data was performed using fastQC(Andrews 2010) software, and potential adapter sequences were removed prior to sequence alignment using Trimmomatic-0.32(Bolger, Lohse et al. 2014).

2.3.2 Short reads alignment

Paired-end sequence reads were mapped to the chicken reference genome (Galgal 4.75) from the Ensembl database using Bowtie2(Langmead and Salzberg 2012) with default settings. For downstream processing, several open-source software packages were used: Picard tools (http://picard.sourceforge.net), SAMtools(Li, Handsaker et al. 2009), and Genome Analysis Toolkit (GATK)(McKenna, Hanna et al. 2010). "CreateSequenceDictionary" and "MarkDuplicates" Picard command-line tools were used to read reference FASTA sequence for writing bam file with only sequence dictionary, and to filter potential PCR duplicates, respectively. Using SAMtools, index files for the reference and bam files were created. Then, local realignment of sequence reads to correct misalignment due to the presence of small insertion and deletion was performed using GATK "RealignerTargetCreator" and "IndelRealigner" arguments. Base quality score recalibration was performed to get accurate quality scores and to correct the variation in quality with machine cycle and sequence context.

2.3.3 Transposable element (TE) probes

As the majority of TEs in the chicken genome are of the retrotransposon type (Hillier, Miller et al. 2004), I focused on three types of retrotransposons and DNA transposons in this study: LTR, LINE, SINE, and DNA transposons. Then, all possible TE probe sequences were obtained from Repeatmasker Genomic datasets (http://www.repeatmasker.org/), and used these as alignment subject for TE identification. The coordinates of TEs upon Gallus

gallus reference genome (from Repeat library 4, 20140131) were also used in TE identification.

2.3.4 Transposable element variants (TEV) identification

I identified TE variants (TEV) across 28 chicken genomes using Retroseq software(Keane, Wong et al. 2013). Retroseq employs discordantly or solely mapping reads to seek candidate TE insertion sites, called breakpoints. In the alignment step, it was necessary to determine the appropriate insert size of paired reads to obtain pure discordantly mapping reads induced by TE insertion not by the mapping distance. The minimum insert size was set at 1000bp to guarantee this requisite.

Due to the difference in depth coverage across the paired-end reads data from various data production processes, it was necessary to adjust the parameters for each group from different data production processes. Additionally, for the confidence of breakpoint, I filtered the raw calls of TEVs and only recovered the calls tagged as "FL=8" which met all breakpoint criteria. Final calls of TEVs were classified into four groups (LTR, LINE, SINE, and DNA transposons) based on the nomenclature and classification used in Repeatmasker. For population analyses, TEV calls within 100 bps were clustered and regarded as a single TEV call; using these cluster positions, breed-specific breakpoint and related genes were identified.

2.3.5 Construction of cladogram based on TE presence polymorphism

Using cluster position, presence of a TE insertion was coded as “1” or “2” according to genotypes, and absence of TE insertion was coded as “0” for each individual. To improve the reliability of each TE insertion loci, loci present in less than 3 individual genomes were excluded. This data matrix was used as an input for calculation of p-distance matrix. I implemented the neighbor-joining method using MEGA software (Tamura, Stecher et al. 2013). A cladogram was constructed for each type of TE. Additionally, a cladogram using the data matrix of all TE types was also created.

2.4 Results and Discussion

2.4.1 Identification of transposable element variants (TEV)

Using Retroseq software, a total of 412,208 candidate TEVs from 28 genome sequences were obtained. Most of the TEVs (~250,000) are located in the intergenic region, while only 149 TEVs were found in the exon region (Table 2.1). For GA, 22,033 candidate TEVs different from chicken reference genome (Galgal 4.75) were identified. Most of the TEVs were annotated as intergenic sequence and only 11 TEVs as exon sequence (Table 2.1). As previous studies, the majority of TEVs belongs to the CR1 families of LINEs (CR1 : 10,524, LINE : 10,676, LTR : 8,587, DNA : 2,611, SINE : 159) (Abrusán, Krambeck et al. 2008), which was consistent in other breeds as well as GA.

The number of TE variants correlates with chromosome length across the whole genome. However, TEVs of all breeds were not evenly distributed within the chromosomes (Figure 2.2a and Figure 2.4). Particularly, when dividing the genome sequence into 50kb bins, significantly more TEVs were identified in one region (Chromosome 8: 8,900,000 ~ 8,950,000) than in other regions across almost all breeds. The genome of GA also included this region. Additionally, the GA genome contained several regions where TEVs were frequently detected. This result is equivalent to precedent studies about biased distribution of TEs for chicken genome (Abrusán, Krambeck et al. 2008).

2.4.2 Cladograms based on TE variants of 28 chicken genome

A cladogram based on the pattern of TE presence from 28 chicken genomes were constructed. Several other studies have constructed phylogenetic trees based on TE presence pattern in primates. These studies on the relationship between species showed some incongruent TE insertion sites caused by several factors, including incomplete lineage sorting and hybridization between species (Xing, Wang et al. 2007). Considering these factors within species and mating-free environment, I estimated that these factors would considerably impact TE presence patterns and consequently cause a confounding result. As predicted, I obtained results similar to those estimated for TE types like LINE, SINE, and DNA transposons (Figure 2.3). However, interestingly, for LTR the constructed cladogram was similar to the SNP based-cladogram which approximately segregates whole individuals by breed (Figure 2.2b). This result indicates that LTR polymorphism might be used as a marker for revealing the relationship of relatively close organisms. Furthermore, it suggests a possible hypothesis that there are more retroviral insertions such as LTR that occurred recently than insertions of other TE types. These results are consistent with those from a previous study on the effects of retroviral insertions on phenotypic traits of breeds (Chang, Coville et al. 2006) as well as blue egg shell formation, which indicates that there might be more cases related to the recent determination of phenotypic traits for breeds than I estimate.

A cladogram based on the LTR presence pattern is shown in Figure 2.2b. Like the SNP-based cladogram, GA and LH were clustered into one

group and close to KNC, which supports the origin of GA which is a result of hybridization between Golden Duckwing Araucana and White leghorn. In previous studies, EAV-HP insertions in the chickens from China and America were regarded as separate integration events (Wang, Qu et al. 2013). In this study, DX and GA were not close to each other in either TE presence-based or SNP-based cladogram. From these results, I can infer that independent insertions occurred in DX and Araucana chickens, which is consistent with previous studies.

2.4.3 Candidate retroviral insertions influencing phenotypic traits

The genetic determination of blue egg shell coloration has been determined in Araucana chickens; EAV-HP insertion promotes the expression of SLCO1B3 gene in the uterus of the oviduct in Araucana chickens, which causes blue egg shell formation(Wang, Qu et al. 2013, Wragg, Mwacharo et al. 2013). The results in this study identified retroviral insertions from three GA genomes in SLCO1B3 gene equally. All of these results are consistent with previous reports. Additionally, there was also EAV-HP insertion in DX genome sequence adjacent to the insertion of GA in SLCO1B3. DX has been known to lay blue eggs as that in Araucana chickens(Wang, Qu et al. 2013). In order to validate the presence of EAV-HP insertion, local de novo assembly with Velvet (Zerbino and Birney 2008) was carried out using the reads mapped within 600 upstream and downstream of the candidate breakpoint. From this process, several contigs were obtained and aligned with TE probes used in TE identification. Then I conducted multiple sequence alignment over sequences recovered from the above process of several

individuals, and retrieved partial conserved TE sequences (142bp) from 3 GA and DX genomes. No insertion was identified in this gene from breeds which don't lay blue eggs, except one sample from Korean native breed. About 7000bp away from the insertion site, LINE_CR1 TEV was identified in KNC_7 sample. Although this is located within SLCO1B3 gene, the distance between the former and the latter is far longer than the length of previously defined insertion sequence (~4,000 bp). Collectively, the commonly identified insertion of EAV-HP in three GA chickens is identical to that of previous experimental studies, which indicates that GA shares genetic characteristics with other breeds that lay blue eggs.

In addition to blue egg shell formation, retroviral insertion can influence the transcription of genes in many ways. Several studies on mice found that the effects of intronic retrovirus insertion on the transcription of the resident gene result in an alteration in the ratios of the splice variants by premature transcription, by providing a cryptic promoter or by altering splicing. In the case of blue egg shell formation, insertion upstream of resident gene can influence the expression of the gene(Isbel and Whitelaw 2012). Consequently, these transcriptional effects can modify phenotypic traits, which was shown in chicken(Chang, Coville et al. 2006) as well as mouse(Perry, Copeland et al. 1994, Vasicek, Zeng et al. 1997).

Here, I propose candidate TEVs related to phenotypic traits of GA. One LTR insertion specific to GA within SUCLG1 gene was identified. This insertion is located in the 3' UTR region of the gene and retrieved partial sequence of 188bp were conserved in all three GA genomes. SUCLG1 is a gene encoding α subunit of succinate-CoA ligase which forms a complex with nucleoside diphosphate kinase and plays an important role in the salvage of

deoxyribonucleotides for mitochondrial DNA synthesis. For this reason, mutations in the gene was known to be associated with mitochondrial DNA depletion disorder in human(Ostergaard, Schwartz et al. 2010). There is another gene including an insertion site (136 bp conserved sequence) identified only in GA. SCN5A contains a LTR insertion within intron region. SCN5A is a gene encoding cardiac-specific voltage-gated sodium channel and known to be related to many cardiovascular diseases (Schott, Alshinawi et al. 1999, Remme, Wilde et al. 2008). This result is also consistent with identification of heart and muscle development-related genes within selection signal. From this information, although the detailed mechanism is unclear, I can hypothesize that there might be a relationship between the TE insertion of SCN5A gene and the characteristic of GA related to flying.

Table 2.1 The result summary of transposable element variants annotation using Snpeff

Region Type	GA	BY	CN	DX	KNC	LH	LG	RIR
Downstream	2,067	727	704	968	14,619	4,174	866	659
Exon	11	3	2	2	48	28	2	3
Intergenic	14,059	5,715	4,803	7,724	128,283	30,101	6,364	5,024
Intron	7,890	3,693	3,067	5,034	91,577	18,947	3,951	2,842
None	17	1	1	5	36	21	1	3
Splice site acceptor	4	4	3	4	33	13	7	1
Splice site donor	7	1	0	3	29	14	2	2
Splice site region	44	11	9	24	288	81	20	10
Upstream	2,688	910	736	1,106	14,997	5,169	970	706
UTR 3’	156	38	41	73	1,298	292	66	53
UTR 5’	108	17	6	28	236	207	19	13
Total	22,033	9,371	7,825	12,671	216,315	48,797	10,251	7,887

Region Type	RJF	SG	SK	TB	TNC	WC	WPR
Downstream	908	648	2,296	829	1,984	849	1,126
Exon	4	5	16	3	15	5	7
Intergenic	7,237	5,282	10,800	6,087	7,263	6,672	8,838
Intron	4,548	3,300	6,154	3,753	3,613	4,239	5,557
None	0	3	11	2	16	3	3
Splice site acceptor	0	0	6	0	3	6	5
Splice site donor	2	0	3	2	5	3	2
Splice site region	14	10	36	13	16	18	21
Upstream	976	698	2,838	898	2,701	1,038	1,319
UTR 3’	50	49	149	49	67	58	92
UTR 5’	29	3	107	14	82	27	39
Total	11,690	8,518	17,176	9,791	10,912	10,880	14,395

Figure 2.1 Gyeongbuk Araucana chicken.

(A) blue egg shell of Gyeongbuk Araucana. (B) general appearance of Gyeongbuk Araucana chicken breed. The photographs were taken by Dr. Oh.

Figure 2.2 Results of analyses using transposable element variants (TEVs).

(A) A circular plot for distribution of TEVs on the genome of GA. Four types of TE are represented by different colors (LTR : green, LINE : red, SINE: orange, and DNA: blue). (B) A cladogram based on LTR presence patterns for 28 chicken genomes. Three groups for LH, GA, and KNC are well-defined. The neighbor-joining tree using whole genome variants and other TE types (LINE, SINE, and DNA) are shown in Figure 2.3.

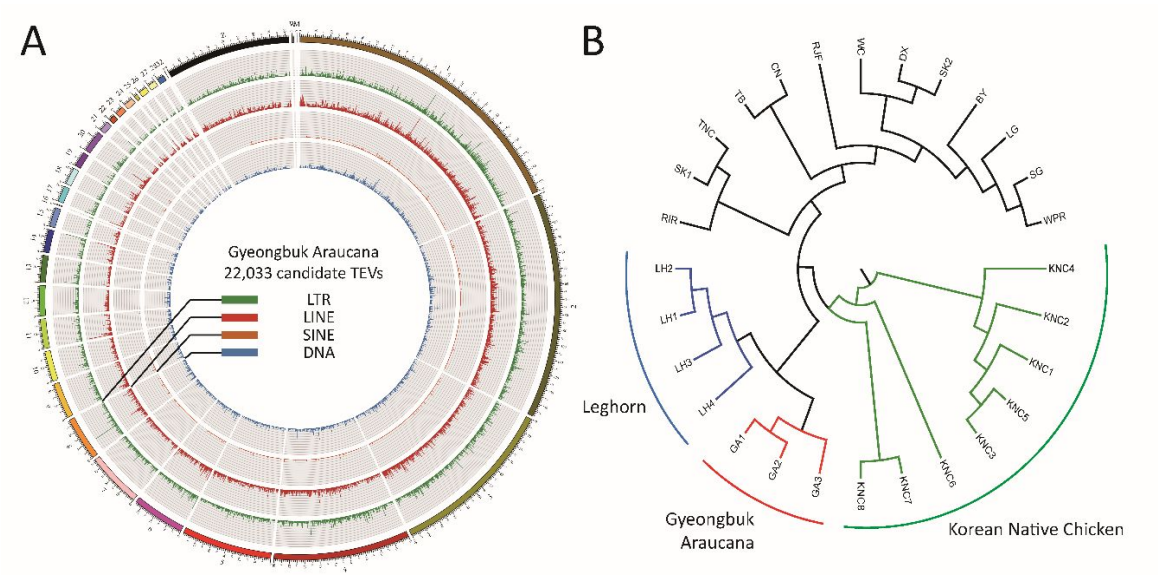


Figure 2.3 Cladograms based on (A) SNVs, (B) DNA transposons, (C) LINE, and (D) SINE transposons

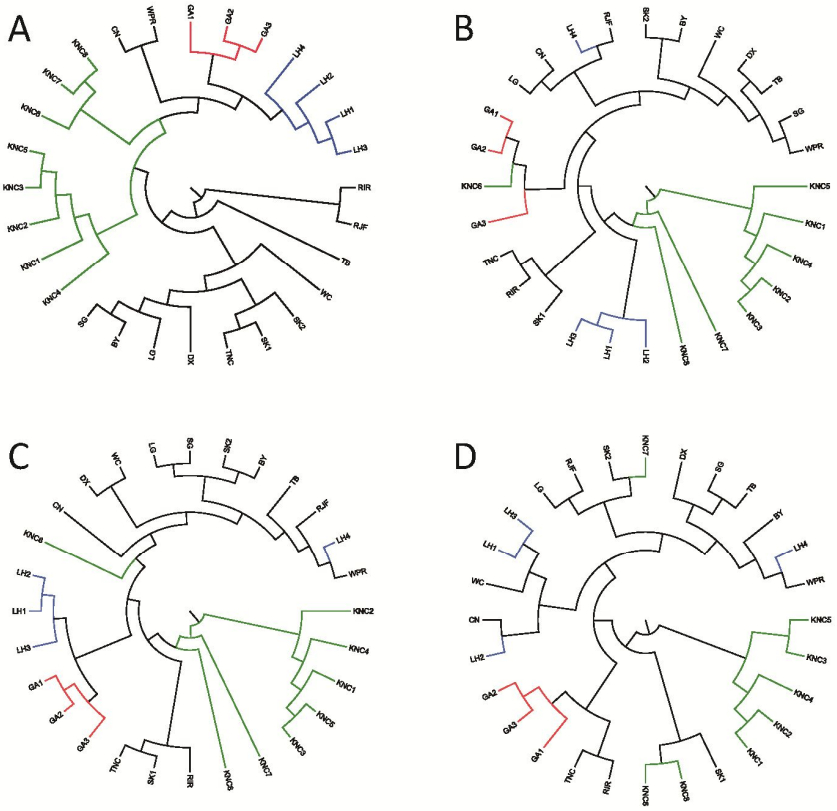
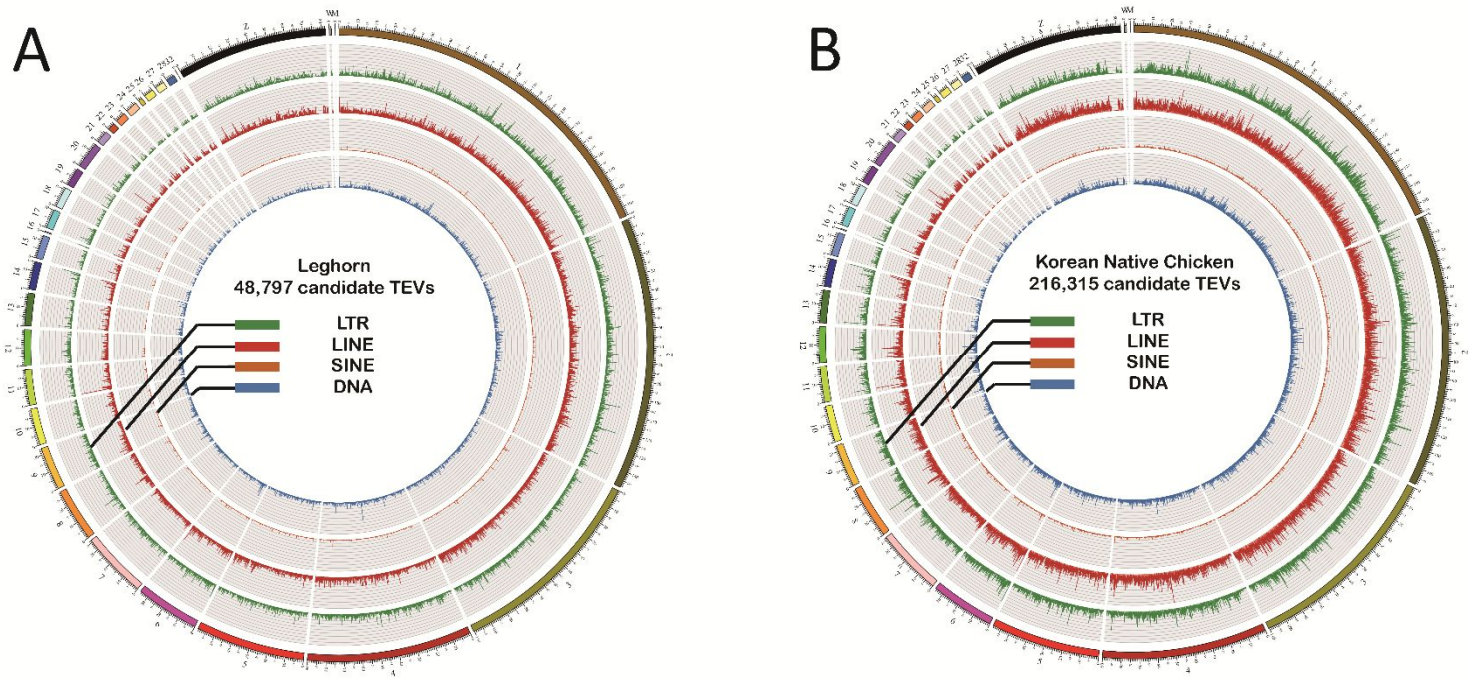


Figure 2.4 Circular plots for distributions of candidate TEVs (A) of LH, (B) of KNC



This chapter was published in *PLoS One*
as a partial fulfillment of Kwondo Kim's M.Sc program.

Chapter 3. Application of LogitBoost classifier for traceability using SNP chip data

3.1 Abstract

Consumer attention to food safety has increased rapidly due to animal-related diseases; therefore, it is important to identify their places of origin (POO) for safety purposes. However, only a few studies have addressed this issue and focused on machine learning-based approaches. In the present study, classification analyses were performed using a customized SNP chip for POO prediction. To accomplish this, 4,122 pigs originating from 104 farms were genotyped using the SNP chip. Several factors were considered to establish the best prediction model based on these data. The applicability of the suggested model was also assessed using a kinship coefficient-filtering approach. The results showed that the LogitBoost-based prediction model outperformed other classifiers in terms of classification performance under most conditions. Specifically, a greater level of accuracy was observed when a higher kinship-based cutoff was employed. These results demonstrated the applicability of a machine learning-based approach using SNP chip data for practical traceability.

3.2 Introduction

Due to the occurrence of animal-related diseases such as bovine spongiform encephalopathy (BSE) and avian influenza (AI), consumer attention to food quality has increased greatly. Accordingly, place of origin (POO) tracing systems have become important to increasing consumer confidence regarding food safety. In the food industry, these are referred to as traceability systems. Traceability is defined as a method that can guarantee the identification of animals or animal products within the food industry (Dalvit, De Marchi et al. 2007). This system is already mandatory for most animal products in a large number of countries. Product tracking has conventionally been conducted by labeling with ear tags and tattoos (Dalvit, De Marchi et al. 2007, Smith, Pendell et al. 2008). Although this technique presents several advantages, including easy application, low cost, and fast data processing, it is vulnerable to fraud or loss (Dalvit, De Marchi et al. 2007). Thus, genetic traceability has been proposed as an alternative to conventional traceability systems. Genetic traceability is the same as labeling systems in principle, except that DNA is used to identify animals or their products. It is possible to distinguish individual animals from one another based on DNA (Goffaux, China et al. 2005). Moreover, DNA molecules are difficult to falsify, can withstand various processes within the food distribution system, and can be extracted from different types of tissues (Dalvit, De Marchi et al. 2007, Negrini, Nicoloso et al. 2008). These advantages have led to increased application and research into use of DNA markers for traceability. One typical marker, the single nucleotide polymorphism (SNP), has been widely applied (Ramos, Megens et al. 2011, Dimauro, Cellesi et al. 2013, Heaton, Leymaster

et al. 2014). There are several methods for obtaining SNP information regarding a sample, including next generation sequencing (NGS), microarrays, and SNP chips. Among these, genotyping using a SNP chip is less expensive and produces SNP data for a relatively large number of samples by customizing chip design.

Numerous studies have been conducted to develop prediction models for classification using diverse biomarkers (Wang, Tetko et al. 2005, Long, Gianola et al. 2007, Iquebal, Dhanda et al. 2013); however, few of these have focused on traceability. One reason for this is that traceability involves multiclass classification. Multiclass classification is generally associated with several difficulties (Even-Zohar and Roth 2001). The main problem associated with this type of classification is optimization. For example, when training sets are given, minimization of the loss function should be performed to build an accurate classifier. Loss function is affected by the number of classes, and minimization of this obstacle could be attained by reducing the number of classes. Several classifiers such as the K-nearest neighbor (KNN) and support vector machine (SVM) have frequently been employed to overcome problems related to multiclass classification (Ding and Dubchak 2001, Yuan, Chen et al. 2008, Güney and Atasoy 2012). Some studies have used KNN and SVM to classify foods according to origin (Teye, Huang et al. 2013, Teye, Huang et al. 2014). In addition, LogitBoost can address multiclass classification problems using a parametric method (Friedman, Hastie et al. 2000, Sun, Reid et al. 2014).

In the present study, 4,122 pigs that originated from 104 farms were genotyped using a customized SNP chip. Based on these data, I attempted to develop a POO prediction model considering three variable factors: (1)

Kinship-based filtering was applied to assess the applicability of classification-based approaches for practical POO prediction; (2) the wrapper-method was used as a feature selection step to remove redundant features (Seo and Oh 2012); (3) LogitBoost, SVM, and KNN were used as classifiers. I compared classification performance using combinations of these factors to identify the optimal POO prediction model.

3.3 Materials and methods

3.3.1 Prescreening SNP markers to generate the customized SNP chip

A total of 384 pigs belonging to five major commercial breeds (19 Korean native black pigs, 17 Landrace, 168 Yorkshire, 84 Berkshire, and 96 Duroc) were genotyped using an Illumina Porcine SNP60 chip to prescreen SNP markers. SNPs were filtered according to several criteria (minor allele frequency [MAF] ≤ 0.05, missing rate ≥ 0.10, and Hardy-Weinberg equilibrium test p-value ≤ 0.001). Following this filtering step, 39,785 SNPs for Korean native black pigs, 42,156 SNPs for Landrace, 44,961 SNPs for Yorkshire, 41,408 SNPs for Berkshire, and 39,652 SNPs for Duroc were retrieved. Among these, 312 SNP markers that were identified in five breeds (MAF ≥ 0.4) were retrieved, and four to nine SNPs with lower linkage disequilibrium (LD) were selected for each chromosome. As a result, 133 SNP markers were obtained. I next performed additional genotyping for 1,045 muscle tissue samples obtained from 11 slaughterhouses (detailed information regarding slaughterhouses is provided in Table 3.3) throughout the Republic of Korea to confirm that the selected SNP markers were evenly distributed for each location. Ultimately, 96 SNP markers including known SNPs for individual animal identification were selected while taking into account the geographical distribution of SNP markers (0.3 ≤ allele frequency ≤ 0.7). These 96 SNP markers were used as features in downstream analyses, including feature selection and classification. More detailed information

regarding these markers is presented in Table 3.4. All genotyped samples were obtained from pigs slaughtered for meat production.

3.3.2 Genotyping 96 SNPs and kinship coefficient-based subset generation for development of the traceability prediction model

From April to June 2014, 4,122 slaughtered commercial pigs originating from 104 different farms were genotyped using a customized SNP chip manufactured by Illumina (provided by the S1 Dataset). Some individual animals and SNPs were filtered out (MAF < 0.01 and genotype missing rate > 0.9) using PLINK v1.07 (Purcell, Neale et al. 2007). As a result, 3,974 individual pigs and 92 SNPs remained.

Most pigs in the livestock industry are derived from a crossbred population, and sires and semen are shared with several farms. Therefore, the origins of pigs are not clearly distinguishable because of genetic similarity. However, sows are generally not shared among farms and produce piglets several times during their lives (Koketsu 2000, Petroman, Petroman et al. 2012). Therefore, I assumed that piglets produced from a single sow might have genetically close relationships. In practice, because genetic information regarding sows in a farm could be considerably dissimilar, it is necessary to screen farms consisting of unrelated individuals to distinguish pigs according to their farms. kinship coefficients (Manichaikul, Mychaleckyj et al. 2010) was employed to evaluate the genetic relationships. The King 1.4 software was used to calculate pairwise kinship coefficients within each farm (Manichaikul, Mychaleckyj et al. 2010). The relationship between two individuals is classified by a kinship coefficient > 0.353 as monozygotic twins

(0.177, 0.353), as parent-offspring or sibling pairs (0.088, 0.177), as second-degree relative pairs (such as half-siblings, avuncular pairs or grandparent-grandchild pairs; 0.044, 0.088) or as third-degree relative pairs (such as first cousins), while < 0.044 indicates unrelated pairs (Manichaikul, Mychaleckyj et al. 2010).

To infer the degree of genetic relatedness to attain reasonable classification accuracy, four subsets of data composed of farms to satisfy the following criteria were generated: mean of the kinship within a farm ≥ 0.00, 0.05, 0.10, and 0.15. The subset with a kinship mean ≥ 0.00 had 741 individuals from twenty farms, the subset with a kinship mean ≥ 0.05 included 235 individuals from eight farms, the subset with a kinship mean ≥ 0.10 included 134 individuals from five farms, and the subset with a kinship mean ≥ 0.15 contained 67 individuals from two farms. To visualize the distribution of individuals by their genetic information in the four subsets, scatter plots were generated by principal component analysis (PCA) using A Tool for Genome-wide Complex Traits Analysis (GCTA) (Yang, Lee et al. 2011).

3.3.3 Wrapper-based feature selection for removing redundant SNP markers

Feature selection is an important step for improving classification performance. Although I already performed a prescreening step to generate an

SNP marker set suitable for traceability, redundant or irrelevant features might be included in this set. Therefore, the wrapper method (Kohavi and John 1997) was utilized to extract valuable features. The wrapper method is a classifier-dependent approach designed to search for feature subsets that would produce the best accuracy. There were two approaches for extracting the best feature subset. The first was top-down selection for which a model was evaluated after eliminating one feature from the entire feature set and replacing the eliminated feature with another. This process was repeated for all features (Approach 1). The second approach was bottom-up selection in which the evaluation step was conducted using only one feature (Approach 2).

3.3.4 Classifiers for multiclass prediction

One main reason for the limited research on traceability prediction is that this type of prediction presents a representative multiclass classification problem. For multiclass data, classification is often associated with several difficulties (Hsu and Lin 2002, Wu, Lin et al. 2004). Unfortunately, most traditional classifiers were developed for binary classification, which cannot be directly employed for multiclass prediction. There are two approaches for addressing multiclass classification problems. The first is a one-against-all approach employing binary classifiers such as the support vector machine (SVM) and LogitBoost (Polat and Güneş 2009). The second is use of classifiers able to predict multiclass data, such as the k-nearest-neighbor (KNN).

LogitBoost is a recently developed boosting algorithm that can handle multiclass problems by considering multiclass logistic loss (Friedman, Hastie

et al. 2000, Sun, Reid et al. 2014). This technique has been used to predict protein structural classes known as representative multiclass problems (Cai, Feng et al. 2006). Other approaches, including SVM- and KNN-based multiclass prediction, have been implemented in many fields (Ding and Dubchak 2001, Yuan, Chen et al. 2008, Güney and Atasoy 2012). KNN, which is one of the simplest methods, classifies an instance according to a majority vote of its k nearest instances. SVM is a high-performance classifier that builds an optimal hyperplane containing the largest distances from support vectors in each given class. As a result, spaces distinguished based on the hyperplanes represent specific classes and predict unknown class data (test data).

In the present investigation, I used these classifiers with the following parameters: LogitBoost I = 20, KNN (IBk) k = 11, and SVM (SMO) kernel = Radial Basic Function (RBF) Kernel, which is implemented in the RWeka (Hornik, Zeileis et al. 2007) package of the R software. These parameter values were determined based on the results of a greedy search using various parameter values for each classifier (Figure 3.7). The default for the RWeka package was used for all other parameters.

3.3.5 Comparison of classification performance

The classification performance of three classifiers (LogiBoost, KNN, and SVM) were compared according to classification accuracy (Metz 1978), balanced accuracy (Brodersen, Ong et al. 2010), sensitivity (Metz 1978), specificity (Metz 1978), area under the curve (AUC) values (Fawcett 2006), and a receiver operating characteristic (ROC) curve (Metz 1978) with 10-fold

cross-validation to avoid overfitting. ROC curves were generated by calculating the false positive and true positive rates for continuous thresholds. I used the ROCR package (Sing, Sander et al. 2005) of the R software to calculate and visualize the ROC curves.

3.3.6 Simulation analysis for estimating the effects of biases

To investigate the effects of biases generated by the various sample sizes and number of classes in different kinship-based subsets, a simulation analysis was performed. I used the LogitBoost classifier and 92 features to estimate biases in the simulation analysis. Three types of simulations were carried out. Whole simulations were repeated 1000 times using sampling without replacement, and 10-fold cross-validations were performed to analyze classification accuracies for each repetition. The first simulation was conducted to survey the impact of the number of classes. To assess this effect, I adjusted the number of classes in the whole kinship-based subsets to two, which was the smallest value among subsets. Then, two classes were randomly selected for each repetition. Sample sizes varied according to random sampling. The second simulation was conducted to survey the effects of sample size. The sample size was fixed at 67, which was the smallest value among all of the subsets. The numbers of classes varied according to random sampling. Finally, the effects of two biases were simultaneously investigated by adjusting both sample size and number of classes.

3.4 Results and Discussion

3.4.1 Assessment of prediction model performance for traceability classification

In the present study, I applied three representative multiclass classifiers to four subsets of SNP data based on kinship-based filtering. In addition, 2 (top-down and bottom-up) × 3 (LogitBoost, SVM, and KNN) wrapper-based feature selection methods were used to generate the best prediction model for traceability. The entire pipeline for data processing including classification is presented as a schematic diagram in Figure 3.1. Specific elements (classifier, feature subset, and kinship coefficient) were expected to be directly associated with prediction accuracy. The influences of these elements were investigated by calculating the prediction accuracy from various points of view. First, I determined how distinguished the individual animals were according to the farms of origin using four subsets based on kinship coefficient-based filtering. As shown in Figure 3.2, the four subsets established based on the cutoff criteria (mean of the kinship within a farm $\geq$ 0.00, 0.05, 0.10, and 0.15, respectively) were visualized by PCA. As the cutoff criterion increased, greater segregation among farms was observed. These findings imply that traceability prediction could be performed when individuals on one farm have highly similar genetic information, which was expected. Using the PCA, I observed subsets with different numbers of samples and farms depending on the cutoff criterion. Therefore, these figures should be interpreted with caution in terms of bias due to the smaller number of classes, larger sample size, and

larger number of features, which generally improve accuracy when classification is performed.

Next, the power of explanation for traceability prediction for each SNP were calculated. I defined "feature score" as the contribution of a feature to the accuracy for classification. In Approach 1, a feature score was calculated based on the accuracy of whole features minus the accuracy associated with eliminating a feature. For Approach 2, a feature score was the accuracy associated with using that feature. As expected, only a few outliers were observed for all feature scores (Table 3.5 and Figure 3.6). Most outliers fell below the lower quantile, indicating that the majority of prescreened features were well selected (Figure 3.6). If the prescreening step had not identified meaningful features for POO prediction, outliers would be observed below the lower quantile and above the upper quantile due to randomness. Therefore, it was confirmed that the customized chip containing 96 SNPs was suitable for POO prediction. It was also demonstrated that some features should be removed from the prediction model for better accuracy.

Next, feature selection step was performed before carrying out classification analysis. As shown in Figure 3.3, the classification accuracies were calculated for the different classifiers and the number of features (features were added to the feature set for the prediction model in order of the feature score generated in Approaches 1 and 2). Four subsets were used to compare classification performance depending on the classifiers and feature sets. Accuracy was determined using 10-fold cross-validation to avoid overfitting. As expected, a subset with more features and higher kinship had better classification accuracy. Overall, I observed a pattern in which accuracy gradually increased with the number of features. These findings indicated that

the customized chip was appropriately designed for traceability because only a few irrelevant features might be included for the 96 SNPs. Generally, including a large number of irrelevant features in a whole feature-set does not increase accuracy, although the features are included in the prediction model. Thus, I again concluded that the 96 pre-selected SNPs were suitable for traceability.

Interestingly, the LogitBoost classifier showed better performance in terms of accuracy than the other classifiers in most situations. This remarkable result indicated that the LogitBoost classifier was more suitable for predicting animal or food origin. It is difficult to constantly obtain a better performance with a specific classifier in diverse situations, as shown by comparison of the SVM and KNN classifiers. Nevertheless, with the exception of one situation (kinship $\geq$ 0.15 and Approach 1), the LogitBoost classifier consistently performed better than the others. In addition, the classification accuracy achieved with LogitBoost had a smaller variance than that of the other classifiers in most situations (Table 3.1). LogitBoost also outperformed the other classifiers in terms of efficiency, with greater levels of accuracy observed when using a relatively small number of features. Overall, the results of this study demonstrated that LogitBoost appears to be the best method for POO prediction in terms of performance assessment when using accuracy as a measurement.

Although classification accuracy is good for evaluating classifiers, unintended bias is occasionally generated. For example, intact accuracy may produce misleading information about general performance when a classifier is evaluated using an imbalanced dataset. In such cases, classification accuracy is not a reliable measure for assessing a prediction model. To avoid

an inflated performance estimation for imbalanced data, another measure, balanced accuracy was employed. The balanced accuracies were calculated as the average accuracies for each class. To further compare classification accuracy and balanced accuracy, balanced accuracies were calculated for the same combinations of factors as shown in Table 3.1. As indicated in Table 3.2, LogitBoost still generally produced better results than the other methods in terms of balanced accuracy. The D89 class that contained only four individual animals always showed poor performance, as indicated by a balanced accuracy of zero. Although the D89 class had a high kinship coefficients mean, PCA demonstrated that individuals in this class overlapped entirely with individuals in other classes (Figure 3.2). Collectively, the characteristics of the D89 class including small sample size and an overlap of individuals with animals from other classes caused poor performance in terms of balanced accuracy. It was also found that the D62 class had particularly high balanced accuracy for the LogitBoost classifier throughout all kinship-based subsets. For example, the LogitBoost classifier had a balanced accuracy of 0.923 for the kinship-based subset (kinship $\geq$ 0.00), while the KNN- and SVM-based approaches had balanced accuracy values of 0.469 and 0.475, respectively. This phenomenon was consistently observed for the other kinship-based subsets.

LogitBoost-based approaches constantly showed better balanced accuracy than other techniques, except for the subset with a kinship mean $\geq$ 0.15, for which the KNN had a more balanced accuracy. However, the overall balanced accuracies were relatively low compared to analysis of the classification accuracy. These findings indicated that there were biases caused by imbalanced classes, which led to overestimation during analysis of the

classification accuracy. Nevertheless, the findings from the accuracy and balanced accuracy analyses demonstrated that LogitBoost had better performance than the other methods, with a few exceptions. Overall, LogitBoost appears to be a more suitable model for POO prediction in terms of consistency. It was also found that balanced accuracy increased with a higher mean kinship coefficient for the subsets.

By assessing the prediction model based on accuracy and balanced accuracy, it was found that the LogitBoost classifier outperformed previously known classifiers for POO prediction. When balanced accuracy was used as a measurement, a strong class-specific accuracy pattern was also observed. To further investigate this pattern, ROC curves were produced as another technique for predicting performance (Figure 3.4). Strong farm-specific curves were observed. It was again found that the D89 class had the lowest performance. The distinct difference between curves for the D89 class and those for the other classes can be interpreted as differences in suitability for the prediction model. Thus, ROC curves can be used to screen out a class that is unsuitable for the prediction model. Additionally, ROC curves showed better performance when the mean kinship coefficient increased, as indicated by the AUC values shown in Table 3.6.

3.4.2 Effects of biases of the kinship-based filtering approach on assessment of the prediction model

Although several performance measures including accuracy, balanced accuracy, ROC curves, and AUC values showed better performance for POO prediction when the kinship cut-off criterion was greater, some bias-

associated problems that prevented accurate model assessment remained. There are two types of bias, difference in sample size and difference in number of classes. In general, reducing the number of classes and/or a large training sample size leads to greater classification accuracy. In the current study, kinship-based filtering subsets had diverse sample sizes and numbers of classes. For this reason, suggested kinship-based filtering approach was affected by the two types of bias, which represented a limitation of study design this research. Therefore, the effects of the biases were investigated. To accomplish this, I performed three simulation analyses by adjusting the number of classes, the sample size, or both. The results of the first simulation analysis are shown in the top of Figure 3.5. The data in this figure confirmed that the previous assessment results were underestimated owing to the effects of the number of classes. The previous accuracies fell below median levels in all kinship subsets. In addition, four kinship-based subsets had similar median levels of accuracy when the number of classes was adjusted. Contrary to the first simulation, the second simulation showed that accuracies were overestimated because of the effects of sample size. As shown in Figure 3.5, the previous accuracies represented by red points were located above the median levels for all kinship subsets. In addition, the median accuracies for the four kinship-based subsets differed significantly. The two simulations described above confirmed that the number of classes has a significant influence on classification accuracy because accuracies in the second simulation varied more drastically according to differences in the number of classes, contrary to those in the first simulation. Finally, the effects of the two biases were simultaneously evaluated (as shown at the bottom of Figure 3.5), which revealed that the results were generally underestimated. However, the

standard deviation of the accuracies decreased as the kinship coefficient cutoff increased.

Taken together, the results of the simulation studies indicated that the number of classes has a greater effect on classification accuracy than sample size. In addition, a higher kinship coefficient cutoff produced a lower standard deviation for the accuracies when both sample size and number of classes were constant. These findings indicated that we can expect to gain greater classification accuracy for populations with a higher kinship coefficient if the effects of sample size or number of classes are controlled. Although I controlled these biased factors in the simulation analysis, there was no practical method for fixing these two factors at equal values. This is because I did not collect samples while considering kinship coefficient values because the primary study design focused on identifying SNPs for individual identification. I actually screened the samples according to kinship coefficient after sample collection, which was a major limitation of this study. Nevertheless, the overall relationship between kinship coefficient and classification accuracy was consistent. Consequently, it was determined that greater classification accuracy accompanied an increased kinship coefficient mean. I also obtained a reasonable accuracy distribution for the subset with a kinship coefficient greater than 0.10. These results imply that we can utilize a kinship coefficient of 0.10 as a criterion for pig traceability.

3.4.3 Application of the prediction model for a practical traceability system

In this study, I concluded that the LogitBoost method was most suitable for POO prediction. LogitBoost has been utilized for various areas of data analysis such as protein structure prediction (Cai, Feng et al. 2006). This method outperformed the SVM classifier for predicting protein structural classes. In addition, LogitBoost was employed for tumor classification using gene expression data (Dettling and Bühlmann 2003). Other types of data analysis such as text classification were also included in Logitboost applications (Kotsiantis, Athanasopoulou et al. 2006). Furthermore, the classifier has been employed in various fields that deal with multiclass prediction. Since POO prediction was also a representative type of multiclass classification, it was anticipated that LogitBoost would be applicable. Not surprisingly, LogitBoost was successfully used for POO prediction. To the best of our knowledge, this is the first time the LogitBoost classifier has been implemented for traceability classification with genotyping data. Consequently, a few improvements should be made to enable the practical use of suggested approaches.

It is clear that when individual organisms originate from the same population they will have similar genotypes (Cornuet, Piry et al. 1999). In the current study, kinship coefficients was used to measure the degree of the relationship between individuals based on this assumption. The results showed that subsets with a higher kinship coefficient had better performance. In particular, individuals within groups with a kinship coefficient higher than 0.1 were identified with reasonable accuracy using all of the evaluated statistics. If an original population was bound with an adequate relationship (pairwise kinship coefficient mean $\geq$ 0.10), it was possible to identify the original population of a given individual with reasonable accuracy. The

findings revealed that the suggested prediction model would be helpful for improving current traceability systems.

Table 3.1 Best classification accuracies for diverse situations (two different feature selection approaches, four different kinship filtered sets, and three classifiers).

Levels of accuracy were calculated by 10-fold cross-validation and expressed as the means ± 10-fold variance. Bold represents greater accuracy than other classifiers for each kinship-based filtered subset.

Subset	Algorithm	Approach 1		Approach 2	
		# of Features	Mean ± Variance	# of Features	Mean ± Variance
Kinship ≥ 0.00	LogitBoost	**83**	**0.652 ± 0.002**	**81**	**0.661 ± 0.004**
	KNN (IBk)	80	0.557 ± 0.006	90	0.549 ± 0.001
	SVM (SMO)	86	0.588 ± 0.004	87	0.578 ± 0.001
Kinship ≥ 0.05	LogitBoost	**72**	**0.878 ± 0.002**	**85**	**0.868 ± 0.005**
	KNN (IBk)	88	0.720 ± 0.015	81	0.726 ± 0.010
	SVM (SMO)	88	0.784 ± 0.004	90	0.747 ± 0.009
Kinship ≥ 0.10	LogitBoost	**47**	**0.950 ± 0.002**	**64**	**0.942 ± 0.003**
	KNN (IBk)	24	0.833 ± 0.013	28	0.850 ± 0.005
	SVM (SMO)	73	0.792 ± 0.009	50	0.790 ± 0.008
Kinship ≥ 0.15	LogitBoost	**53**	**0.992 ± 0.001**	**4**	**0.992 ± 0.001**
	KNN (IBk)	82	0.983 ± 0.003	20	0.992 ± 0.001
	SVM (SMO)	73	0.967 ± 0.005	72	0.909 ± 0.007

Table 3.2 Evaluation of predicted performance according to balanced accuracy.

The balanced accuracies were calculated by 10-fold cross-validation. Values represent the mean ± 10-fold variance. Figures written in bold represent a higher level of balanced accuracy than those of the other classifiers in each class. Figures given in parentheses represent the number of features used in classifiers.

Kinship ≥ 0.00						
Class	Approach 1			Approach 2		
	LogitBoostt (83)	KNN (IBk) (80)	SVM (SMO) (86)	LogitBoostt (81)	KNN (IBk) (90)	SVM (SMO) (87)
D1	**0.387 ± 0.124**	0.050 ± 0.025	0.000 ± 0.000	**0.446 ± 0.146**	0.133 ± 0.104	0.017 ± 0.003
D2	0.683 ± 0.057	0.729 ± 0.039	**0.810 ± 0.035**	**0.829 ± 0.026**	0.808 ± 0.044	0.742 ± 0.030
D10	0.422 ± 0.068	0.421 ± 0.088	**0.701 ± 0.039**	0.513 ± 0.061	0.389 ± 0.114	**0.612 ± 0.032**
D11	**0.713 ± 0.033**	0.676 ± 0.107	0.695 ± 0.042	**0.735 ± 0.036**	0.624 ± 0.100	0.625 ± 0.108
D13	0.751 ± 0.039	0.823 ± 0.024	**0.905 ± 0.011**	0.703 ± 0.025	0.810 ± 0.027	**0.913 ± 0.018**
D18	**0.418 ± 0.041**	0.245 ± 0.071	0.277 ± 0.117	**0.547 ± 0.048**	0.254 ± 0.032	0.291 ± 0.109
D27	**0.540 ± 0.067**	0.532 ± 0.065	0.513 ± 0.076	**0.585 ± 0.083**	0.416 ± 0.051	0.521 ± 0.062
D38	**0.774 ± 0.045**	0.712 ± 0.074	0.682 ± 0.064	**0.857 ± 0.057**	0.685 ± 0.051	0.697 ± 0.074
D59	**0.797 ± 0.060**	0.642 ± 0.069	0.648 ± 0.098	**0.768 ± 0.050**	0.698 ± 0.047	0.677 ± 0.063
D60	**0.755 ± 0.091**	0.067 ± 0.044	0.145 ± 0.038	**0.611 ± 0.114**	0.108 ± 0.034	0.361 ± 0.118
D61	**0.605 ± 0.150**	0.462 ± 0.160	0.150 ± 0.114	0.185 ± 0.082	**0.273 ± 0.044**	0.000 ± 0.000
D62	**0.923 ± 0.017**	0.469 ± 0.064	0.475 ± 0.131	**0.840 ± 0.040**	0.608 ± 0.062	0.486 ± 0.066
D66	**0.655 ± 0.082**	0.448 ± 0.170	0.340 ± 0.062	**0.463 ± 0.115**	0.407 ± 0.038	0.411 ± 0.145
D89	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

D90	0.693 ± 0.080	**0.827 ± 0.038**	0.527 ± 0.132	0.708 ± 0.104	**0.717 ± 0.068**	0.689 ± 0.133
D100	0.638 ± 0.082	**0.889 ± 0.035**	0.663 ± 0.094	0.705 ± 0.110	**0.833 ± 0.125**	0.746 ± 0.048
D102	**0.770 ± 0.037**	0.355 ± 0.055	0.683 ± 0.120	**0.862 ± 0.030**	0.364 ± 0.038	0.702 ± 0.049
D103	0.452 ± 0.031	**0.607 ± 0.095**	0.492 ± 0.045	0.418 ± 0.100	**0.517 ± 0.081**	0.468 ± 0.100
D107	0.733 ± 0.063	**0.787 ± 0.045**	0.757 ± 0.040	**0.818 ± 0.076**	0.718 ± 0.080	0.795 ± 0.024
D114	**0.766 ± 0.056**	0.630 ± 0.077	0.678 ± 0.050	0.683 ± 0.102	0.628 ± 0.055	**0.747 ± 0.040**
Balanced Accuracy	**0.624 ± 0.061**	0.518 ± 0.067	0.507 ± 0.065	**0.614 ± 0.070**	0.500 ± 0.060	0.525 ± 0.061

Kinship ≥ 0.05

Class	Approach 1			Approach 2		
	LogitBoostt (72)	KNN (IBk) (88)	SVM (SMO) (88)	LogitBoostt (85)	KNN (IBk) (81)	SVM (SMO) (90)
D11	**0.975 ± 0.006**	0.940 ± 0.018	0.933 ± 0.028	0.975 ± 0.006	**0.980 ± 0.004**	0.933 ± 0.012
D59	0.793 ± 0.103	**0.904 ± 0.029**	0.832 ± 0.032	**0.938 ± 0.010**	0.848 ± 0.032	0.751 ± 0.057
D60	**0.843 ± 0.029**	0.150 ± 0.065	0.597 ± 0.142	**0.717 ± 0.073**	0.204 ± 0.045	0.575 ± 0.132
D62	**0.955 ± 0.009**	0.767 ± 0.063	0.727 ± 0.052	**0.963 ± 0.006**	0.693 ± 0.066	0.718 ± 0.100
D66	**0.900 ± 0.024**	0.661 ± 0.061	0.696 ± 0.080	**0.795 ± 0.116**	0.591 ± 0.081	0.623 ± 0.090
D89	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
D90	0.838 ± 0.035	0.802 ± 0.076	**0.947 ± 0.014**	0.875 ± 0.045	0.806 ± 0.106	**0.944 ± 0.014**
D100	0.900 ± 0.044	**0.941 ± 0.015**	0.867 ± 0.048	**0.967 ± 0.011**	0.889 ± 0.111	0.817 ± 0.114
Balanced Accuracy	**0.776 ± 0.031**	0.646 ± 0.041	0.700 ± 0.049	**0.779 ± 0.033**	0.626 ± 0.056	0.670 ± 0.065

Kinship ≥ 0.10

Class	Approach 1			Approach 2		
	LogitBoostt (47)	KNN (IBk) (24)	SVM (SMO) (73)	LogitBoostt (64)	KNN (IBk) (28)	SVM (SMO) (50)
D59	**1.000 ± 0.000**	0.896 ± 0.022	0.947 ± 0.007	**1.000 ± 0.000**	0.925 ± 0.016	0.950 ± 0.013
D62	**1.000 ± 0.000**	0.917 ± 0.031	0.745 ± 0.051	**1.000 ± 0.000**	0.942 ± 0.016	0.922 ± 0.017

D66	**0.942 ± 0.016**	0.785 ± 0.060	0.847 ± 0.046	**0.930 ± 0.027**	0.848 ± 0.028	0.677 ± 0.108
D89	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
D100	**0.950 ± 0.025**	0.933 ± 0.020	0.811 ± 0.050	**0.963 ± 0.012**	0.900 ± 0.026	0.622 ± 0.129
Balanced Accuracy	**0.778 ± 0.008**	0.706 ± 0.026	0.670 ± 0.031	**0.779 ± 0.008**	0.723 ± 0.017	0.634 ± 0.053

Kinship ≥ 0.15

Class	Approach 1			Approach 2		
	LogitBoost (72)	KNN (IBk) (88)	SVM (SMO) (88)	LogitBoost (85)	KNN (IBk) (81)	SVM (SMO) (90)
D59	0.950 ± 0.025	**0.989 ± 0.001**	0.988 ± 0.002	**1.000 ± 0.000**	**1.000 ± 0.000**	0.975 ± 0.006
D100	0.852 ± 0.031	**0.933 ± 0.020**	0.858 ± 0.037	**0.950 ± 0.025**	0.875 ± 0.051	0.775 ± 0.068
Balanced Accuracy	0.901 ± 0.028	**0.961 ± 0.010**	0.923 ± 0.019	**0.975 ± 0.013**	0.938 ± 0.026	0.875 ± 0.037

Table 3.3 Slaughterhouses

No.	Location	Slaughterhouse	Address	Phone
1	Gangwon-do	Gangwon LPC Inc.	438-3, Gahyeon-dong, Wonju-si, Gangwon-do, Korea	82-33-732-1300
2	Gyeonggi-do	Bucheon Livestock Joint Market	12-4, Samjeong-dong, Ojeong-gu, Bucheon-si, Gyeonggi-do, Korea	82-32-620-5000
3	Gyeonggi-do	Hyupsin Food Inc.	298, Bakdal-ro, Manan-gu, Anyang-si, Gyeonggi-do, Korea	82-31-447-9001
4	Gyeongsangnam-do	Jinju SK Industry Inc.	1369, Namgang-ro, Jinju-si, Gyeongsangnam-do, Korea	82-55-755-5508
5	Gyeongsangnam-do	Bukyung Livestock Joint Market	6-9, Eobang-dong, Gimhae-si, Gyeongsangnam-do, Korea	82-55-325-1331
6	Jeollanam-do	Manna Inc.	6, Seotae-ri, Hwasun-eup, Hwasun-gun, Jeollanam-do, Korea	82-61-373-6144
7	Jeollabuk-do	Gimje Meat Processing Factory	630, Guseong-gil, Geumsan-myeon, Gimje-si, Jeollabuk-do, Korea	82-63-540-6700
8	Chungcheongnam-do	Hongju Meat Inc.	539, Sangjeong-ri, Gwangcheon-eup, Hongseong-gun, Chungcheongnam-do, Korea	82-41-630-7000
9	Chungcheonbuk-do	Farm Story LPC Inc.	421-3, Seongjae-ri, Ohchang-eup, Choengwon-gun, Chungcheonbuk-do, Korea	82-43-210-4269
10	Gyeongsangbuk-do	Lotte Food Inc.	94, Gongdan 3-gil, Gimcheon-si, Gyeongsangbuk-do, Korea	82-54-420-2533
11	Jeju-do	Jeju Livestock Joint Market	2533, Eoeum-ri, Aewol-eup, Jeju-si, Jeju-do, Korea	82-64-799-5135

Table 3.4 Selected SNP markers

SNP	Assay ID	SNP name	Allele	rs number	Chromosome	Position
SNP1	GTA0029134	ALGA0002500	TC	rs81353459	1	36691125
SNP2	GTA0029124	ALGA0003632	AG	rs80818014	1	62989866
SNP3	GTA0029127	ALGA0005188	TC	rs81001361	1	108032798
SNP4	GTA0027167	ALGA0010607	AG	rs81001439	1	302880686
SNP5	GTA0027175	ALGA0012333	AG	rs81368483	2	19693385
SNP6	GTA0027197	ALGA0017166	AG	rs81373103	3	3059882
SNP7	GTA0027150	ALGA0017261	AG	rs81379588	3	5840882
SNP8	GTA0029128	ALGA0020170	TG	rs81373795	3	100347076
SNP9	GTA0027187	ALGA0020295	AG	rs81374145	3	103487445
SNP10	GTA0027148	ALGA0023180	AC	rs80976115	4	11956752
SNP11	GTA0027256	ALGA0026994	AG	rs81382424	4	104284028
SNP12	GTA0029118	ALGA0028052	AG	rs81380202	4	120752641
SNP13	GTA0027206	ALGA0030335	AG	rs81385823	5	8872938
SNP14	GTA0029115	ALGA0033986	AG	rs80922042	5	105182709
SNP15	GTA0027188	ALGA0034886	AG	rs81394644	6	22138877
SNP16	GTA0029136	ALGA0037105	TC	rs81392460	6	134307926
SNP17	GTA0027174	ALGA0038431	AG	rs80814806	7	8844382
SNP18	GTA0027215	ALGA0038635	AC	rs80903447	7	11867962
SNP19	GTA0029121	ALGA0043483	TC	rs80875831	7	95283610
SNP20	GTA0027229	ALGA0052166	AG	rs81408300	9	30555946
SNP21	GTA0027236	ALGA0056803	AG	rs81428674	10	9683745
SNP22	GTA0029112	ALGA0056924	AG	rs81428973	10	11062697
SNP23	GTA0027153	ALGA0059061	AG	rs81425082	10	52086866
SNP24	GTA0029116	ALGA0064392	TG	rs81433418	12	4992763
SNP25	GTA0029138	ALGA0065426	TC	rs81440978	12	17788471
SNP26	GTA0029119	ALGA0067483	AG	rs80931112	13	3445254
SNP27	GTA0027183	ALGA0071504	AG	rs81447525	13	99980492
SNP28	GTA0029129	ALGA0072858	AG	rs80939920	13	183883486
SNP29	GTA0027151	ALGA0073188	AC	rs81441710	13	192229132
SNP30	GTA0027205	ALGA0075911	AG	rs80803891	14	20105162
SNP31	GTA0029131	ALGA0079359	TC	rs80973431	14	90459793
SNP32	GTA0027224	ALGA0083823	AG	rs80810051	15	1047733
SNP33	GTA0027178	ALGA0084361	AC	rs81451849	15	24623255
SNP34	GTA0029117	ALGA0085130	AG	rs80966936	15	50746568
SNP35	GTA0027204	ALGA0088449	AG	rs81244935	15	157102798
SNP36	GTA0029120	ALGA0089251	TG	rs81464737	16	15711247

SNP37	GTA0029126	ALGA0092844	TC	rs80962528	17	5506024
SNP38	GTA0027225	ALGA0093942	AG	rs81465558	17	28475777
SNP39	GTA0027254	ALGA0095059	AG	rs80831567	17	45969331
SNP40	GTA0029137	ALGA0097474	TC	rs81467738	18	24947919
SNP41	GTA0029130	ALGA0097857	AG	rs81468642	18	35731229
SNP42	GTA0027237	ALGA0109641	AG	rs81477834	11	29405328
SNP43	GTA0027257	ALGA0110410	AG	rs81338661	2	26730378
SNP44	GTA0027180	ALGA0115847	AC	rs81345194	8	138063274
SNP45	GTA0027250	ALGA0119982	AG	rs81327268	13	13928440
SNP46	GTA0029135	ALGA0124374	AG	rs81305532	3	11691606
SNP47	GTA0027217	ASGA0001168	AC	rs81348505	1	15731768
SNP48	GTA0027246	ASGA0003689	AG	rs81354990	1	93179162
SNP49	GTA0027209	ASGA0006871	AG	rs81351913	1	285219286
SNP50	GTA0027239	ASGA0009403	AG	rs81368238	2	18805832
SNP51	GTA0027231	ASGA0011793	AG	rs81364493	2	133593694
SNP52	GTA0027149	ASGA0017082	AG	rs80886731	4	2363714
SNP53	GTA0027171	ASGA0018449	AG	rs80897680	4	13602527
SNP54	GTA0027198	ASGA0029755	AG	rs80983079	-	-
SNP55	GTA0027211	ASGA0031089	AG	rs80801891	7	9721700
SNP56	GTA0027159	ASGA0035039	AG	rs80791412	7	92270821
SNP57	GTA0027169	ASGA0035601	AG	rs81396105	7	105616538
SNP58	GTA0029114	ASGA0040082	TC	rs81404763	8	139708737
SNP59	GTA0027258	ASGA0041336	AG	rs81413894	9	9224376
SNP60	GTA0027249	ASGA0042099	CG	rs81407644	9	26191661
SNP61	GTA0027199	ASGA0048625	AC	rs81426512	10	64825425
SNP62	GTA0027192	ASGA0060257	AG	rs81443163	13	215340754
SNP63	GTA0027170	ASGA0060872	AG	rs81450975	14	7471763
SNP64	GTA0027160	ASGA0094977	AG	rs81314288	3	16913373
SNP65	GTA0029113	ASGA0096881	TC	rs81316705	-	-
SNP66	GTA0029125	H3GA0000077	AC	rs81355602	1	1614750
SNP67	GTA0027185	H3GA0000926	AG	rs80963451	1	17622619
SNP68	GTA0027163	H3GA0006218	AG	rs81368467	6	147237891
SNP69	GTA0027244	H3GA0009291	AG	rs81369179	3	36270941
SNP70	GTA0027193	H3GA0012015	AG	rs80918208	4	13008743
SNP71	GTA0027165	H3GA0026867	AC	rs81408241	9	30203501
SNP72	GTA0027219	H3GA0027004	AG	rs81409265	9	43779851
SNP73	GTA0029133	H3GA0028278	AG	rs81416948	9	136602162
SNP74	GTA0027194	H3GA0030549	AC	rs81426586	10	64940895
SNP75	GTA0029123	H3GA0031292	AG	rs80962437	11	10198249

SNP76	GTA0027156	H3GA0031439	AG	rs81430022	11	16289120
SNP77	GTA0027243	H3GA0038523	AG	rs80883745	14	5972386
SNP78	GTA0027161	H3GA0046698	AG	rs81460044	16	57921341
SNP79	GTA0027235	H3GA0048952	AC	rs80976267	17	45602515
SNP80	GTA0027208	H3GA0054041	AG	rs81325463	1	287683176
SNP81	GTA0027162	H3GA0056419	AG	rs81323954	12	7759970
SNP82	GTA0029132	M1GA0001903	AC	rs80820154	1	305774749
SNP83	GTA0027173	M1GA0008026	AG	rs81385563	5	79039700
SNP84	GTA0027201	M1GA0011894	AG	rs81399380	8	33090664
SNP85	GTA0027222	M1GA0018145	AG	rs80914882	14	7363912
SNP86	GTA0027184	M1GA0022894	AG	rs80960907	17	69302804
SNP87	GTA0027196	M1GA0024249	AG	rs81331703	12	22259328
SNP88	GTA0027200	MARC0001787	AG	-	-	-
SNP89	GTA0029122	MARC0004720	TC	-	-	-
SNP90	GTA0027176	MARC0008528	AG	-	-	-
SNP91	GTA0027212	MARC0034477	AC	-	-	-
SNP92	GTA0027207	MARC0055696	AG	-	-	-
SNP93	GTA0027238	MARC0056053	AG	-	-	-
SNP94	GTA0027203	MARC0065987	AG	-	-	-
SNP95	GTA0027168	MARC0073259	AC	-	-	-
SNP96	GTA0027251	MARC0076283	AG	-	-	-

Dashes indicate missing information.

Table 3.5 Feature scores for each SNP calculated with three classifiers and two approaches

(A) Kinship ≥ 0.00

	LogitBoost		KNN(IBk)		SVM(SMO)	
SNP	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
ALGA0003632	44.804	65.047	52.227	54.386	29.015	38.057
ALGA0005188	40.756	51.417	53.981	43.725	55.870	40.081
ALGA0010607	34.143	46.154	24.291	25.506	32.659	41.296
ALGA0012333	57.625	56.680	31.579	46.154	53.576	42.240
ALGA0028052	57.895	34.143	25.506	25.776	45.884	43.860
ALGA0033986	61.943	53.171	11.876	36.032	41.430	46.019
ALGA0034886	49.663	51.822	54.386	31.984	35.628	45.209
ALGA0038635	48.313	41.296	41.700	34.413	12.686	45.614
ALGA0043483	40.891	56.005	53.441	46.694	56.545	46.559
ALGA0056803	58.165	63.428	40.081	53.711	37.247	9.717
ALGA0059061	50.607	42.915	36.707	37.787	58.974	60.999
ALGA0064392	44.534	55.466	12.551	46.424	10.661	51.687
ALGA0067483	27.126	21.457	42.105	47.368	21.997	59.109
ALGA0072858	47.368	36.842	42.375	44.669	38.596	52.632
ALGA0075911	58.974	16.329	46.559	15.789	8.907	8.772
ALGA0085130	14.305	55.466	33.468	52.632	50.067	52.227
ALGA0089251	64.642	52.632	27.126	47.638	49.798	53.036
ALGA0092844	61.673	45.074	20.378	29.825	22.402	52.767
ALGA0093942	25.506	55.061	27.530	48.853	13.630	60.594
ALGA0095059	39.676	41.296	40.621	36.032	59.649	56.410
ALGA0097857	41.970	46.964	29.825	31.309	38.327	56.815
ALGA0110410	40.756	41.700	54.386	26.451	32.794	40.486
ALGA0115847	46.829	55.331	51.147	50.202	53.441	47.638
ALGA0119982	58.435	28.880	14.170	20.243	10.121	59.514
ASGA0006871	45.344	51.012	19.568	36.707	59.379	41.026
ASGA0009403	45.344	50.607	31.714	42.375	14.575	59.244
ASGA0011793	36.707	48.853	50.067	38.057	44.534	43.725
ASGA0017082	38.192	49.933	31.309	46.019	18.758	42.240
ASGA0018449	50.877	55.466	53.441	47.503	12.281	44.804
ASGA0031089	39.541	54.926	41.296	49.258	51.957	46.694
ASGA0035601	64.103	51.822	50.877	40.081	12.281	47.773
ASGA0040082	44.939	54.386	42.915	50.472	54.386	48.853

ASGA0041336	63.833	59.109	50.067	49.798	39.001	48.313
ASGA0042099	60.594	62.078	40.081	54.386	56.140	49.393
ASGA0060257	50.067	64.777	45.074	55.331	39.001	53.576
ASGA0060872	47.503	51.417	53.846	47.233	46.424	59.649
ASGA0094977	44.130	49.393	54.116	42.375	9.312	44.265
ASGA0096881	27.395	48.853	29.555	47.503	9.717	9.042
GTA0027154	30.769	33.468	36.032	19.298	49.798	26.586
GTA0027157	34.548	27.800	41.430	13.900	56.545	32.389
GTA0027158	18.084	62.753	36.302	54.386	9.312	31.174
GTA0027179	17.274	53.036	54.926	47.773	24.022	15.115
GTA0027181	11.606	65.317	51.687	54.116	40.081	15.520
GTA0027182	31.849	39.946	33.738	44.265	58.030	24.561
GTA0027186	64.777	55.601	37.652	48.718	52.092	60.864
GTA0027189	36.167	48.583	21.862	37.517	27.530	34.143
GTA0027213	27.260	53.846	54.521	44.130	35.897	17.139
GTA0027214	39.946	59.379	29.555	50.877	55.061	32.389
GTA0027218	51.687	46.289	40.891	42.240	12.281	21.592
GTA0027230	62.078	42.240	42.240	18.489	51.012	13.765
GTA0027233	64.777	62.753	49.123	53.171	48.043	20.918
GTA0027234	64.103	28.205	29.420	35.358	48.313	8.907
GTA0027240	64.372	30.769	41.430	18.084	51.417	58.300
GTA0027245	41.700	63.293	19.433	52.901	50.337	59.109
GTA0027247	33.738	46.289	53.306	26.586	40.891	24.831
GTA0058588	38.866	51.687	33.738	45.074	47.638	23.887
GTA0058589	63.563	56.275	52.497	49.258	51.012	34.143
GTA0058592	45.749	48.448	33.198	39.271	32.254	58.974
GTA0058594	59.379	39.136	51.417	35.088	24.966	24.022
GTA0058596	47.638	48.448	48.313	43.860	39.676	15.924
GTA0058597	48.043	52.632	36.842	48.313	53.846	30.904
GTA0058598	44.130	56.005	54.386	52.767	41.161	28.205
GTA0058600	26.586	48.448	54.926	45.614	17.409	29.285
GTA0058601	7.962	22.942	23.347	45.884	8.907	34.548
GTA0058602	41.565	49.258	37.112	45.479	27.530	25.236
GTA0059294	42.105	20.918	10.121	14.035	8.907	8.772
GTA0059295	24.561	58.974	54.656	51.822	46.289	55.061
GTA0059296	42.915	55.196	24.291	50.877	56.545	42.915
GTA0059297	38.731	27.260	46.289	16.194	50.067	9.852
GTA0059299	63.563	63.968	20.378	55.061	11.606	52.092
GTA0059301	32.389	47.638	23.617	48.178	59.109	43.320

GTA0059302	47.233	42.645	53.441	45.884	58.300	43.455
GTA0059303	28.880	50.337	50.607	45.074	9.042	50.067
GTA0059304	47.773	64.777	13.090	47.773	36.437	46.289
GTA0059305	38.596	51.282	36.437	49.798	53.981	42.915
GTA0059306	39.271	22.267	31.309	16.194	45.074	41.565
GTA0059307	64.103	60.459	15.655	52.767	17.409	55.870
GTA0059822	49.663	54.791	53.711	52.901	43.455	37.787
H3GA0000077	46.964	63.023	48.718	50.607	52.497	37.922
H3GA0009291	45.209	42.375	24.561	32.659	49.798	9.852
H3GA0012015	58.300	54.791	55.331	43.860	9.447	58.300
H3GA0027004	63.833	56.275	39.001	54.116	48.448	48.853
H3GA0031439	27.935	48.043	54.251	34.278	54.251	50.067
H3GA0046698	42.780	21.457	42.915	16.734	27.800	8.772
H3GA0048952	35.223	33.063	41.970	35.628	29.960	55.601
M1GA0008026	7.962	7.287	42.780	6.208	21.188	8.907
M1GA0011894	32.794	51.822	26.721	53.171	29.825	47.099
MARC0004720	31.309	48.178	27.935	43.995	55.601	10.931
MARC0008528	44.669	49.798	36.437	38.731	29.960	11.606
MARC0055696	43.185	45.074	49.528	40.081	57.760	13.090
MARC0065987	60.594	40.081	47.773	31.174	59.109	58.570
MARC0076283	49.528	53.441	50.067	40.891	49.393	12.955
Mean	43.951	47.775	38.944	40.889	37.652	38.205
Variance	193.173	150.206	160.087	141.233	283.874	275.354

(B) Kinship ≥ 0.05

	LogitBoost		KNN(IBk)		SVM(SMO)	
SNP	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
ALGA0003632	78.723	47.234	46.383	39.149	69.362	52.340
ALGA0005188	82.128	65.532	54.894	40.851	65.106	71.915
ALGA0010607	71.489	82.553	56.596	67.660	73.617	71.489
ALGA0012333	81.702	78.723	58.298	68.511	71.915	25.532
ALGA0028052	82.979	83.830	66.383	69.787	75.745	76.596
ALGA0033986	72.340	65.106	55.745	57.021	67.234	71.064
ALGA0034886	80.000	47.660	57.447	42.979	42.553	74.894
ALGA0038635	79.574	85.106	54.894	68.085	68.511	71.064
ALGA0043483	82.128	82.553	22.979	66.809	72.340	72.340
ALGA0056803	77.872	74.468	38.723	62.128	72.766	75.745
ALGA0059061	81.702	51.489	57.872	43.830	67.660	77.447

ALGA0064392	77.872	82.128	67.660	71.064	65.532	53.617
ALGA0067483	21.702	31.064	44.681	65.957	32.766	18.298
ALGA0072858	83.830	73.617	54.894	67.234	33.617	77.021
ALGA0075911	83.404	63.830	67.660	61.702	62.979	78.298
ALGA0085130	30.638	81.277	60.851	66.383	17.872	80.000
ALGA0089251	75.319	82.128	51.489	67.660	74.894	77.021
ALGA0092844	65.957	72.340	57.447	60.851	72.340	79.574
ALGA0093942	82.979	73.617	48.085	62.979	73.191	77.872
ALGA0095059	82.553	82.979	41.277	63.404	35.319	77.872
ALGA0097857	83.830	85.532	57.872	64.681	71.064	79.574
ALGA0110410	68.085	82.128	35.745	64.681	18.723	54.043
ALGA0115847	75.319	67.660	65.532	66.809	61.702	74.043
ALGA0119982	79.149	82.128	67.660	68.511	22.128	77.021
ASGA0006871	64.255	75.319	60.851	64.681	51.915	27.660
ASGA0009403	82.553	82.553	37.447	65.957	42.128	76.596
ASGA0011793	83.404	64.681	51.915	63.830	74.468	35.319
ASGA0017082	74.043	77.021	63.830	64.255	52.766	77.021
ASGA0018449	72.766	78.298	66.383	67.660	54.043	56.596
ASGA0031089	69.787	74.043	69.362	43.830	70.638	71.915
ASGA0035601	69.362	70.638	66.383	54.894	68.085	57.447
ASGA0040082	82.128	83.404	45.957	63.404	74.468	28.085
ASGA0041336	80.851	77.021	45.957	62.979	71.915	36.170
ASGA0042099	47.234	80.000	59.149	58.298	65.106	76.596
ASGA0060257	72.340	82.128	68.085	67.234	42.553	17.872
ASGA0060872	71.489	74.043	57.021	66.383	74.043	78.723
ASGA0094977	69.787	84.681	68.085	63.830	61.277	73.617
ASGA0096881	72.340	69.787	60.426	68.511	74.043	78.298
GTA0027154	64.681	82.128	61.277	66.809	73.191	70.213
GTA0027157	81.277	69.787	60.851	64.255	67.234	70.638
GTA0027158	80.000	83.830	43.404	68.085	25.532	73.191
GTA0027179	80.426	70.213	57.447	55.745	21.702	61.702
GTA0027181	65.957	54.043	52.766	35.319	74.043	17.447
GTA0027182	71.064	74.468	57.447	63.404	60.000	24.681
GTA0027186	71.064	63.404	33.617	51.915	60.000	55.319
GTA0027189	80.851	66.809	53.191	65.106	62.128	75.319
GTA0027213	77.021	82.553	44.681	67.234	37.021	48.936
GTA0027214	81.702	84.255	41.277	63.404	54.043	77.872
GTA0027218	65.957	47.660	40.426	39.149	24.681	19.149
GTA0027230	77.447	82.553	17.872	65.532	50.213	78.298

GTA0027233	82.128	71.915	37.447	52.766	53.191	62.553
GTA0027234	84.681	70.638	57.872	65.532	69.362	30.638
GTA0027240	70.638	68.085	51.064	64.255	72.766	63.404
GTA0027245	37.021	56.170	57.021	43.830	41.702	76.170
GTA0027247	78.298	71.064	50.638	45.957	53.617	29.362
GTA0058588	82.979	75.319	41.702	67.660	53.617	65.106
GTA0058589	75.745	78.298	48.511	62.979	28.085	76.596
GTA0058592	80.851	76.596	59.149	66.809	75.319	22.553
GTA0058594	80.000	73.191	51.915	64.255	68.085	62.553
GTA0058596	83.404	52.340	59.574	40.426	51.064	65.106
GTA0058597	78.298	86.383	56.170	68.511	42.553	70.213
GTA0058598	69.787	56.596	51.489	55.745	66.809	76.596
GTA0058600	67.660	67.234	56.170	56.596	73.191	70.213
GTA0058601	28.936	17.872	56.596	39.574	20.851	71.489
GTA0058602	80.426	63.830	64.255	51.064	17.872	56.170
GTA0059294	70.213	80.000	51.489	65.106	68.085	77.872
GTA0059295	60.000	80.000	68.936	63.404	46.809	60.000
GTA0059296	77.021	80.000	57.447	63.830	64.255	35.319
GTA0059297	78.298	77.021	66.809	64.255	73.617	74.468
GTA0059299	80.000	70.213	60.426	59.149	72.766	39.574
GTA0059301	65.532	84.681	64.681	67.234	76.596	32.340
GTA0059302	82.979	66.809	54.043	52.340	67.234	72.340
GTA0059303	82.128	63.830	47.234	56.596	41.702	20.426
GTA0059304	72.340	72.766	65.532	56.170	57.447	16.170
GTA0059305	71.064	68.511	59.149	55.745	62.553	75.745
GTA0059306	81.277	74.468	65.957	60.000	66.383	73.191
GTA0059307	72.340	78.298	65.106	66.809	41.277	41.277
GTA0059822	67.234	68.511	29.362	55.319	72.766	55.319
H3GA0000077	84.681	33.617	57.872	25.106	30.213	74.468
H3GA0009291	82.553	77.021	58.298	66.809	69.787	73.617
H3GA0012015	72.766	72.766	54.894	62.979	74.894	73.191
H3GA0027004	73.191	71.064	66.809	57.447	54.894	74.894
H3GA0031439	80.000	84.681	54.043	67.660	41.702	77.872
H3GA0046698	77.021	72.766	51.915	63.404	69.787	78.298
H3GA0048952	77.872	75.319	66.809	58.298	61.277	42.979
M1GA0008026	26.383	20.851	53.617	19.574	68.085	35.745
M1GA0011894	77.447	43.830	35.745	28.936	52.766	74.468
MARC0004720	77.021	74.468	68.936	63.404	66.383	63.404
MARC0008528	79.574	80.000	40.851	65.532	44.681	77.021

MARC0055696	82.979	68.085	58.723	64.681	68.511	65.957
MARC0065987	85.106	39.149	24.681	33.191	51.915	45.532
MARC0076283	80.426	49.787	67.234	42.128	74.468	45.957
Mean	73.797	70.370	54.265	58.821	57.664	60.712
Variance	161.363	203.136	124.054	122.254	283.227	389.687

(C) Kinship ≥ 0.10

	LogitBoost		KNN(IBk)		SVM(SMO)	
SNP	**Approach 1**	**Approach 2**	**Approach 1**	**Approach 2**	**Approach 1**	**Approach 2**
ALGA0003632	44.776	91.045	54.478	75.373	73.881	70.149
ALGA0005188	82.090	64.925	32.836	41.791	54.478	29.851
ALGA0010607	82.836	90.299	64.925	76.119	76.866	73.881
ALGA0012333	65.672	93.284	74.627	74.627	69.403	73.134
ALGA0028052	81.343	88.060	73.134	72.388	71.642	76.866
ALGA0033986	83.582	88.806	57.463	74.627	67.910	74.627
ALGA0034886	93.284	91.791	75.373	78.358	73.134	76.866
ALGA0038635	50.000	92.537	76.119	79.104	71.642	75.373
ALGA0043483	79.851	88.806	76.119	70.149	70.896	74.627
ALGA0056803	92.537	90.299	70.896	74.627	68.657	70.896
ALGA0059061	77.612	91.791	76.119	72.388	74.627	73.881
ALGA0064392	91.045	90.299	56.716	74.627	39.552	29.851
ALGA0067483	56.716	38.806	38.806	42.537	76.866	41.791
ALGA0072858	60.448	92.537	73.881	77.612	76.119	71.642
ALGA0075911	84.328	88.806	74.627	74.627	52.239	75.373
ALGA0085130	93.284	92.537	78.358	77.612	29.851	73.881
ALGA0089251	91.045	92.537	54.478	74.627	72.388	73.881
ALGA0092844	82.836	91.791	55.970	76.119	70.896	76.119
ALGA0093942	80.597	88.806	73.134	74.627	55.970	40.299
ALGA0095059	63.433	89.552	68.657	77.612	72.388	73.134
ALGA0097857	88.060	91.045	53.731	78.358	74.627	71.642
ALGA0110410	91.791	91.045	75.373	72.388	73.881	68.657
ALGA0115847	91.791	65.672	70.896	58.955	50.746	71.642
ALGA0119982	63.433	92.537	78.358	77.612	76.119	72.388
ASGA0006871	61.194	90.299	72.388	77.612	70.149	76.866
ASGA0009403	81.343	93.284	73.881	75.373	70.896	73.134
ASGA0011793	81.343	87.313	61.194	77.612	66.418	72.388
ASGA0017082	61.940	91.791	64.179	79.851	35.075	75.373
ASGA0018449	80.597	90.299	51.493	53.731	72.388	71.642

ASGA0031089	64.179	91.791	78.358	79.104	68.657	73.134
ASGA0035601	63.433	91.791	76.866	78.358	70.149	75.373
ASGA0040082	64.179	91.045	74.627	77.612	34.328	77.612
ASGA0041336	81.343	92.537	71.642	80.597	75.373	70.149
ASGA0042099	94.776	88.060	74.627	73.881	69.403	70.896
ASGA0060257	92.537	92.537	70.896	73.881	68.657	72.388
ASGA0060872	79.104	90.299	74.627	80.597	73.134	73.134
ASGA0094977	79.104	91.791	61.940	78.358	58.209	76.119
ASGA0096881	82.090	86.567	70.896	76.119	75.373	47.015
GTA0027154	83.582	93.284	73.134	76.119	72.388	76.866
GTA0027157	93.284	91.791	76.866	77.612	74.627	67.164
GTA0027158	92.537	91.045	72.388	61.194	50.746	75.373
GTA0027179	93.284	44.030	67.910	43.284	61.194	58.209
GTA0027181	82.090	89.552	67.164	73.881	29.851	61.940
GTA0027182	81.343	90.299	77.612	76.866	75.373	65.672
GTA0027186	59.701	87.313	52.239	66.418	70.896	76.119
GTA0027189	82.090	88.806	70.149	73.881	73.881	41.045
GTA0027213	83.582	91.045	73.881	77.612	76.119	61.194
GTA0027214	64.925	90.299	74.627	73.881	42.537	66.418
GTA0027218	84.328	41.791	70.149	47.761	52.985	63.433
GTA0027230	73.881	91.045	53.731	79.104	66.418	59.701
GTA0027233	70.149	91.791	67.910	78.358	77.612	64.179
GTA0027234	82.836	85.821	70.149	73.134	73.134	41.791
GTA0027240	79.851	93.284	67.910	79.851	61.194	71.642
GTA0027245	92.537	67.910	55.970	66.418	42.537	63.433
GTA0027247	91.791	90.299	76.866	74.627	42.537	63.433
GTA0058588	82.090	93.284	76.119	77.612	29.851	76.119
GTA0058589	75.373	91.791	71.642	73.881	72.388	76.119
GTA0058592	93.284	93.284	67.910	76.119	42.537	73.881
GTA0058594	89.552	91.045	78.358	74.627	72.388	73.881
GTA0058596	76.866	86.567	71.642	73.881	59.701	29.851
GTA0058597	82.836	88.806	73.134	65.672	67.910	70.149
GTA0058598	76.119	87.313	72.388	75.373	49.254	71.642
GTA0058600	66.418	94.776	73.134	77.612	75.373	78.358
GTA0058601	38.060	43.284	58.209	65.672	29.851	29.851
GTA0058602	83.582	88.806	71.642	73.881	32.090	69.403
GTA0059294	81.343	86.567	79.851	73.881	49.254	72.388
GTA0059295	70.149	55.970	61.194	49.254	77.612	47.015
GTA0059296	43.284	87.313	71.642	78.358	52.985	70.896

GTA0059297	94.030	88.060	76.866	60.448	72.388	72.388
GTA0059299	79.104	91.791	73.134	74.627	76.866	41.791
GTA0059301	45.522	91.045	74.627	79.851	80.597	38.806
GTA0059302	93.284	91.791	71.642	75.373	47.015	70.149
GTA0059303	93.284	93.284	72.388	75.373	52.239	72.388
GTA0059304	92.537	92.537	73.881	77.612	73.134	74.627
GTA0059305	85.075	93.284	71.642	79.851	73.881	76.866
GTA0059306	76.866	92.537	74.627	75.373	69.403	67.910
GTA0059307	92.537	92.537	78.358	75.373	72.388	76.119
GTA0059822	94.030	90.299	71.642	56.716	63.433	70.149
H3GA0000077	85.075	52.239	67.910	43.284	62.687	71.642
H3GA0009291	61.194	92.537	73.134	73.134	71.642	73.134
H3GA0012015	94.776	91.791	63.433	64.179	58.955	73.134
H3GA0027004	92.537	89.552	64.179	76.866	70.896	41.045
H3GA0031439	65.672	91.045	72.388	80.597	49.254	72.388
H3GA0046698	81.343	89.552	74.627	74.627	64.925	73.134
H3GA0048952	85.075	89.552	77.612	76.119	72.388	76.119
M1GA0008026	82.836	90.299	64.925	76.119	56.716	29.851
M1GA0011894	81.343	92.537	71.642	75.373	58.209	72.388
MARC0004720	77.612	91.045	71.642	79.104	73.881	49.254
MARC0008528	65.672	90.299	49.254	76.119	59.701	47.015
MARC0055696	38.806	91.791	65.672	59.701	73.881	56.716
MARC0065987	74.627	89.552	73.134	75.373	29.851	58.955
MARC0076283	76.866	88.806	70.149	76.866	61.940	76.866
Mean	78.261	86.989	69.095	72.429	63.092	65.931
Variance	188.387	143.395	77.211	83.236	195.367	181.144

(D) Kinship ≥ 0.15

	LogitBoost		KNN(IBk)		SVM(SMO)	
SNP	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
ALGA0003632	89.552	98.507	67.164	77.612	59.701	73.134
ALGA0005188	86.567	98.507	82.090	83.582	59.701	74.627
ALGA0010607	82.090	97.015	70.149	88.060	85.075	77.612
ALGA0012333	85.075	98.507	83.582	83.582	91.045	77.612
ALGA0028052	94.030	98.507	91.045	89.552	88.060	85.075
ALGA0033986	89.552	98.507	91.045	94.030	89.552	89.552
ALGA0034886	88.060	97.015	91.045	92.537	85.075	89.552
ALGA0038635	89.552	95.522	89.552	89.552	92.537	89.552

ALGA0043483	89.552	97.015	91.045	91.045	59.701	89.552
ALGA0056803	98.507	98.507	94.030	95.522	88.060	91.045
ALGA0059061	98.507	98.507	89.552	88.060	61.194	91.045
ALGA0064392	97.015	97.015	89.552	89.552	91.045	91.045
ALGA0067483	98.507	50.746	73.134	53.731	59.701	59.701
ALGA0072858	98.507	98.507	91.045	95.522	71.642	91.045
ALGA0075911	98.507	98.507	95.522	91.045	86.567	59.701
ALGA0085130	98.507	98.507	89.552	92.537	88.060	91.045
ALGA0089251	98.507	97.015	88.060	95.522	88.060	88.060
ALGA0092844	98.507	97.015	91.045	82.090	80.597	59.701
ALGA0093942	98.507	97.015	89.552	89.552	85.075	92.537
ALGA0095059	98.507	95.522	86.567	92.537	88.060	94.030
ALGA0097857	98.507	98.507	88.060	94.030	59.701	92.537
ALGA0110410	86.567	98.507	68.657	94.030	85.075	76.119
ALGA0115847	94.030	97.015	89.552	85.075	62.687	92.537
ALGA0119982	98.507	98.507	91.045	95.522	85.075	92.537
ASGA0006871	83.582	97.015	83.582	95.522	68.657	76.119
ASGA0009403	85.075	98.507	92.537	94.030	92.537	73.134
ASGA0011793	91.045	98.507	91.045	91.045	86.567	91.045
ASGA0017082	89.552	98.507	92.537	91.045	83.582	91.045
ASGA0018449	91.045	98.507	94.030	94.030	71.642	89.552
ASGA0031089	89.552	98.507	91.045	92.537	89.552	89.552
ASGA0035601	94.030	97.015	89.552	88.060	74.627	89.552
ASGA0040082	94.030	98.507	91.045	94.030	88.060	89.552
ASGA0041336	95.522	98.507	92.537	91.045	88.060	89.552
ASGA0042099	98.507	98.507	95.522	79.104	88.060	92.537
ASGA0060257	98.507	98.507	91.045	83.582	76.119	92.537
ASGA0060872	98.507	97.015	89.552	85.075	89.552	91.045
ASGA0094977	89.552	98.507	91.045	95.522	59.701	88.060
ASGA0096881	64.179	98.507	64.179	83.582	82.090	59.701
GTA0027154	85.075	98.507	92.537	94.030	86.567	70.149
GTA0027157	88.060	98.507	91.045	80.597	59.701	70.149
GTA0027158	83.582	98.507	92.537	83.582	85.075	73.134
GTA0027179	83.582	98.507	79.104	88.060	82.090	59.701
GTA0027181	86.567	98.507	74.627	85.075	59.701	59.701
GTA0027182	76.119	98.507	88.060	83.582	82.090	61.194
GTA0027186	86.567	62.687	91.045	52.239	62.687	59.701
GTA0027189	95.522	98.507	92.537	85.075	82.090	68.657
GTA0027213	77.612	98.507	70.149	83.582	88.060	61.194

GTA0027214	77.612	98.507	92.537	85.075	80.597	70.149
GTA0027218	80.597	98.507	77.612	83.582	59.701	62.687
GTA0027230	88.060	95.522	91.045	82.090	61.194	59.701
GTA0027233	83.582	98.507	77.612	86.567	59.701	62.687
GTA0027234	77.612	97.015	91.045	82.090	59.701	73.134
GTA0027240	83.582	98.507	94.030	86.567	82.090	68.657
GTA0027245	76.119	98.507	89.552	85.075	95.522	65.672
GTA0027247	80.597	98.507	89.552	83.582	85.075	62.687
GTA0058588	85.075	97.015	88.060	83.582	59.701	65.672
GTA0058589	89.552	98.507	91.045	83.582	79.104	67.164
GTA0058592	82.090	98.507	79.104	79.104	85.075	59.701
GTA0058594	74.627	98.507	76.119	83.582	59.701	61.194
GTA0058596	73.134	98.507	74.627	85.075	83.582	59.701
GTA0058597	83.582	98.507	59.701	86.567	82.090	68.657
GTA0058598	88.060	98.507	88.060	80.597	89.552	70.149
GTA0058600	82.090	97.015	79.104	82.090	59.701	64.179
GTA0058601	85.075	98.507	70.149	86.567	79.104	68.657
GTA0058602	82.090	98.507	71.642	83.582	82.090	62.687
GTA0059294	98.507	98.507	65.672	83.582	83.582	91.045
GTA0059295	98.507	46.269	91.045	46.269	76.119	94.030
GTA0059296	85.075	97.015	92.537	82.090	89.552	88.060
GTA0059297	98.507	98.507	91.045	98.507	61.194	91.045
GTA0059299	98.507	98.507	92.537	91.045	89.552	89.552
GTA0059301	94.030	98.507	92.537	94.030	83.582	85.075
GTA0059302	89.552	98.507	58.209	91.045	68.657	89.552
GTA0059303	98.507	98.507	94.030	76.119	86.567	89.552
GTA0059304	92.537	98.507	91.045	95.522	64.179	85.075
GTA0059305	92.537	98.507	92.537	95.522	59.701	89.552
GTA0059306	88.060	98.507	92.537	94.030	64.179	77.612
GTA0059307	98.507	98.507	92.537	94.030	82.090	91.045
GTA0059822	89.552	98.507	89.552	79.104	59.701	67.164
H3GA0000077	89.552	98.507	89.552	95.522	59.701	68.657
H3GA0009291	94.030	98.507	92.537	88.060	89.552	91.045
H3GA0012015	92.537	98.507	92.537	86.567	89.552	88.060
H3GA0027004	98.507	97.015	92.537	89.552	73.134	89.552
H3GA0031439	98.507	98.507	88.060	83.582	86.567	89.552
H3GA0046698	98.507	98.507	73.134	91.045	76.119	91.045
H3GA0048952	98.507	98.507	94.030	94.030	91.045	91.045
M1GA0008026	94.030	98.507	68.657	89.552	88.060	89.552

M1GA0011894	88.060	98.507	89.552	80.597	88.060	88.060
MARC0004720	70.149	98.507	86.567	83.582	80.597	59.701
MARC0008528	67.164	98.507	89.552	86.567	62.687	59.701
MARC0055696	61.194	98.507	79.104	77.612	88.060	59.701
MARC0065987	82.090	98.507	82.090	91.045	94.030	59.701
MARC0076283	88.060	98.507	82.090	79.104	59.701	59.701
Mean	89.082	96.674	85.967	86.567	77.790	78.066
Variance	73.779	66.443	79.611	72.558	143.787	168.017

Table 3.6 AUC values for each class calculated with LogitBoost and two approaches

Class	Approach 1	Approach 2
Kinship ≥ 0.00		
D1	0.902	0.848
D2	0.970	0.952
D10	0.866	0.811
D11	0.930	0.907
D13	0.934	0.944
D18	0.945	0.941
D27	0.868	0.848
D38	0.942	0.929
D59	0.984	0.744
D60	0.989	0.857
D61	0.798	0.738
D62	0.987	0.988
D66	0.872	0.836
D89	0.635	0.906
D90	0.973	0.914
D100	0.976	0.976
D102	0.976	0.930
D103	0.846	0.821
D107	0.968	0.927
D114	0.943	0.927
Mean ± Variance	0.915 ± 0.007	0.887 ± 0.005
Kinship ≥0.05		
D11	0.997	0.993
D59	0.943	0.933
D60	0.955	0.952
D62	1.000	1.000
D66	0.909	0.964
D89	0.662	0.748
D90	0.949	0.974
D100	0.991	0.991
Mean ± Variance	0.926 ± 0.012	0.944 ± 0.007
Kinship ≥ 0.10		
D59	0.895	1.000
D62	0.982	1.000
D66	0.915	0.982
D89	0.833	1.000
D100	0.950	0.972
Mean ± Variance	0.915 ± 0.003	0.991 ± 0.000

Kinship ≥ 0.15		
D59	0.993	1.000
D100	0.993	1.000
Mean ± Variance	0.993 ± 0.000	1.000 ± 0.000

Figure 3.1 A diagram representing the processes of building the prediction model for traceability.

The prescreening process for selecting the SNP markers consists of two major steps: retrieval of common SNPs for five pig breeds and selection of SNP markers based on geographical distribution (farm location). Farms were filtered by the kinship coefficient mean and four subsets were generated. The feature selection process for removing redundant features was performed using two approaches (detailed descriptions of these techniques are provided in the manuscript) and three classifiers. Using the selected features, classification performance was evaluated based on three factors (classification accuracy, balanced accuracy, and ROC curves).

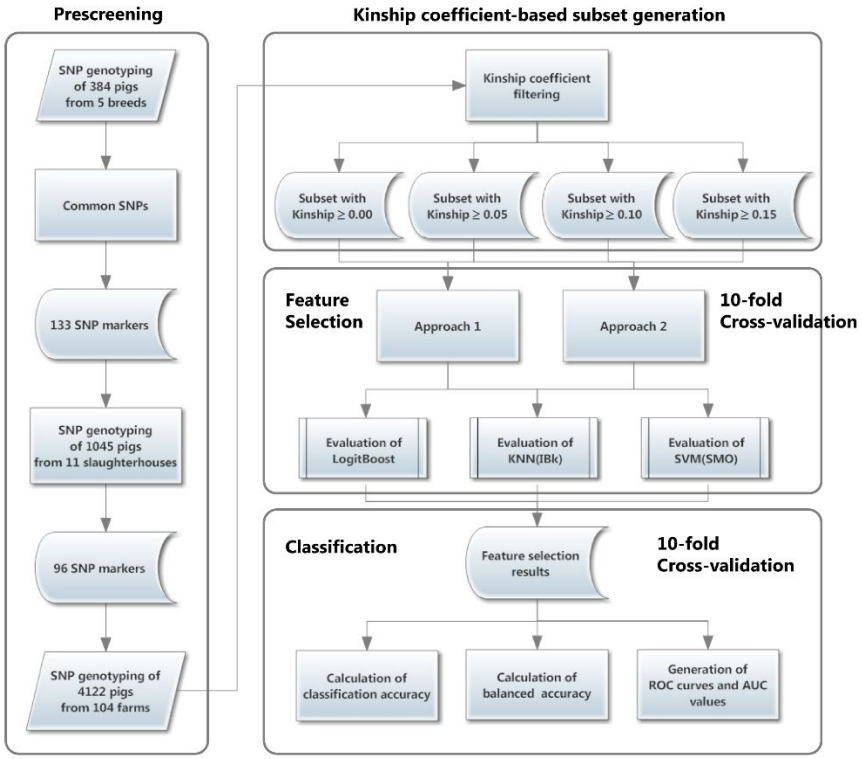


Figure 3.2 Scatter plots for four subsets with different kinship coefficient criteria (X-axis: Eigen vector 1 and Y-axis: Eigen vector 2).

Scatter plots were generated by PCA using GCTA (Yang, Lee et al. 2011). Each point represents an individual animal and is colored based on the farm information. When the kinship cutoff increased, each farm was more clearly distinguishable.

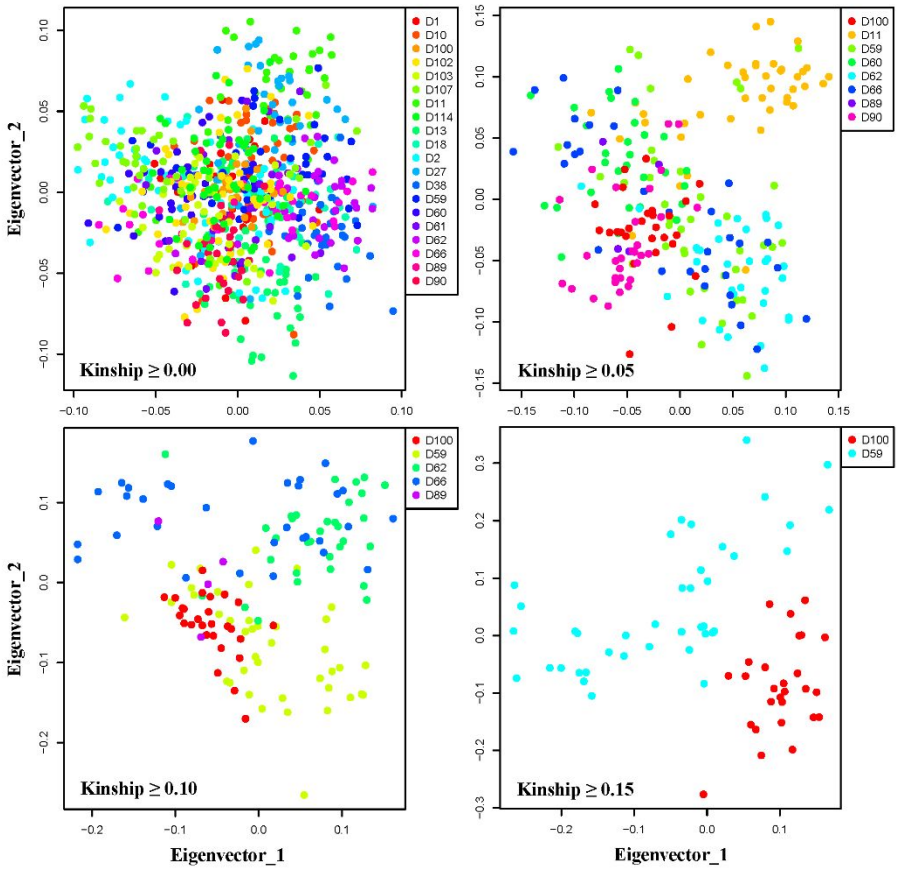


Figure 3.3 Line plots for comparing classification accuracy according to several factors, including classifiers, feature subsets, and kinship-based filtered subsets.

The X-axis contains the number of features (1 to 92 SNPs), while the Y-axis shows classification accuracy. Approach 1 is the top-down feature selection method while Approach 2 is the bottom-up feature selection technique. LogitBoost-based classification accuracy is represented by the red line. Lines corresponding to the KNN and SVM classification methods are green and blue, respectively.

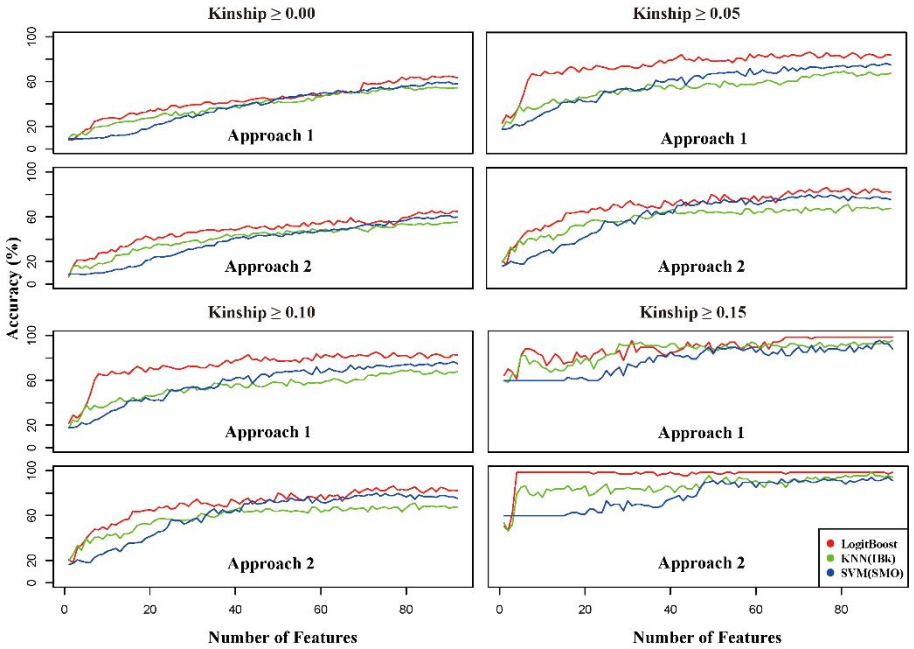


Figure 3.4 ROC curves for different kinship-based subsets to evaluate the suitability of specific farm groups with the LogitBoost classifier.

To calculate sensitivity and specificity, data were divided in half and used as a training and test set. Threshold-specific performance could then be monitored using continuous cutoffs based on the ROC curves. All processes were conducted for the four subsets with two approaches. The D89 class showed the lowest performance in most cases.

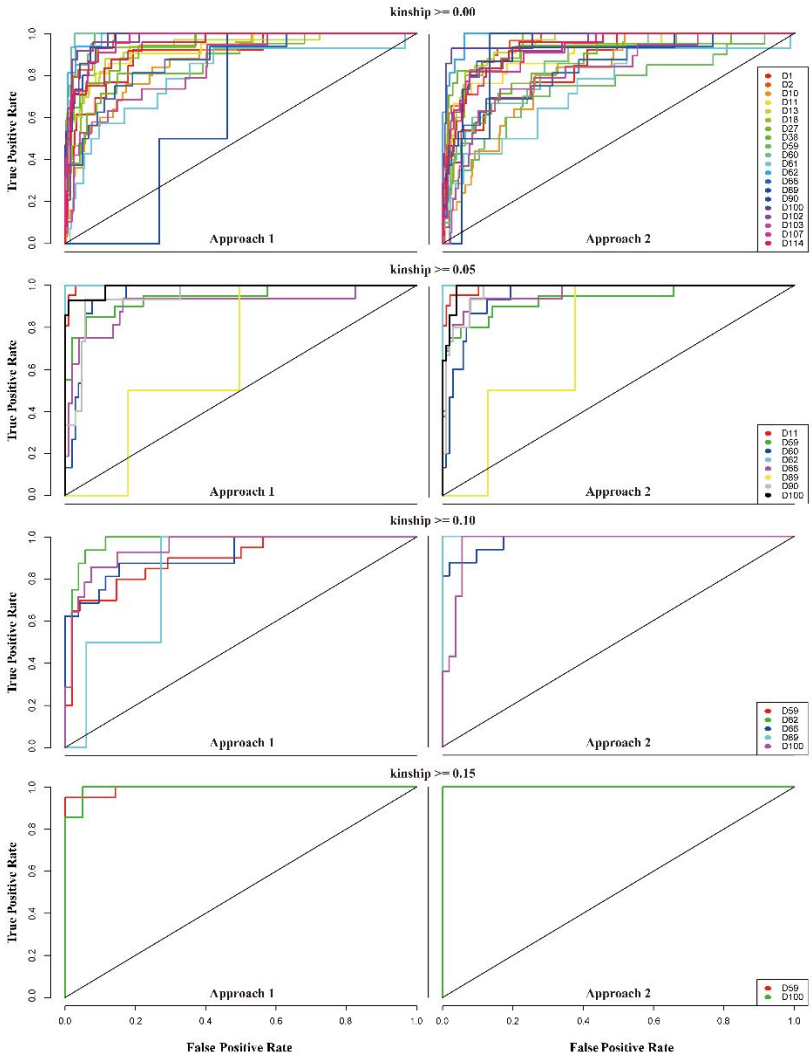


Figure 3.5 Results of sample size and number of classes correction.

Data for the three simulation analyses were generated by adjusting three factors (sample size, number of classes, or both). For the top box-plot, sample size was set at 67, which was the smallest of the four subsets. For the middle box-plot, the number of classes was set at two, which was also the smallest for the four subsets. Finally, the bottom box-plot was generated using 26 samples (the smallest sample size among all classes) for each class (binary class). To determine the classification accuracies, 10-fold cross-validations were performed. All of these processes were conducted 1000 times using 92 features. Red dots represent the previously calculated observed accuracies.

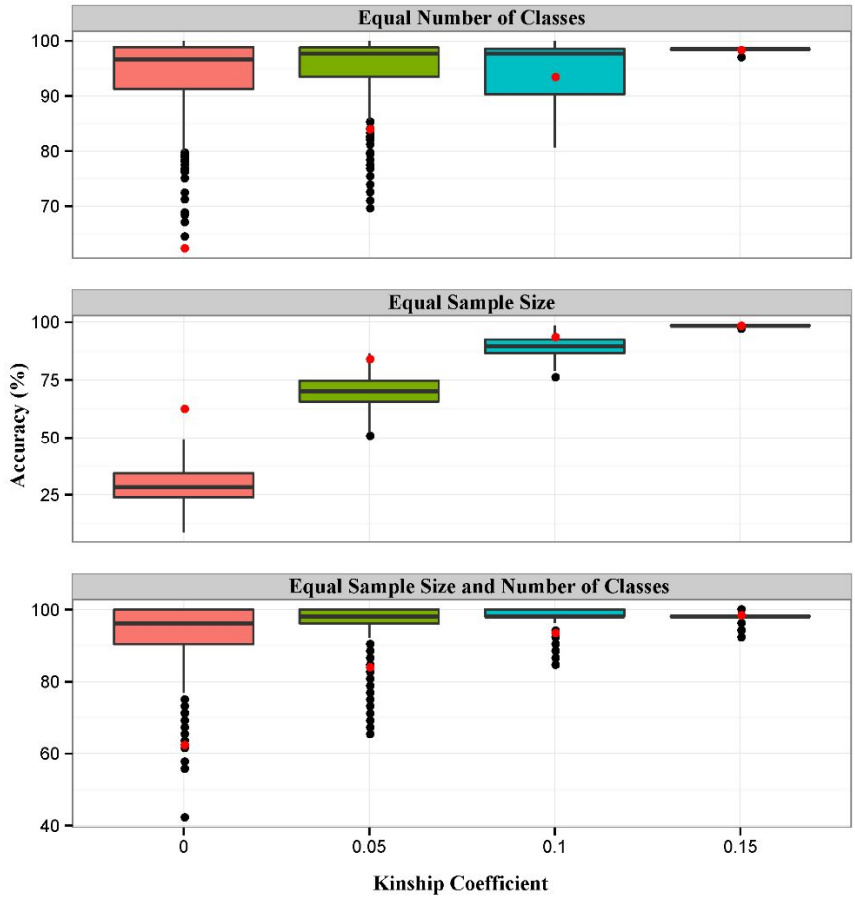


Figure 3.6 Box-plots of feature scores calculated with three classifiers and two approaches.

L, K, and S indicate LogitBoost, KNN, and SVM, respectively. 1 and 2 indicate Approach 1 and Approach 2, respectively.

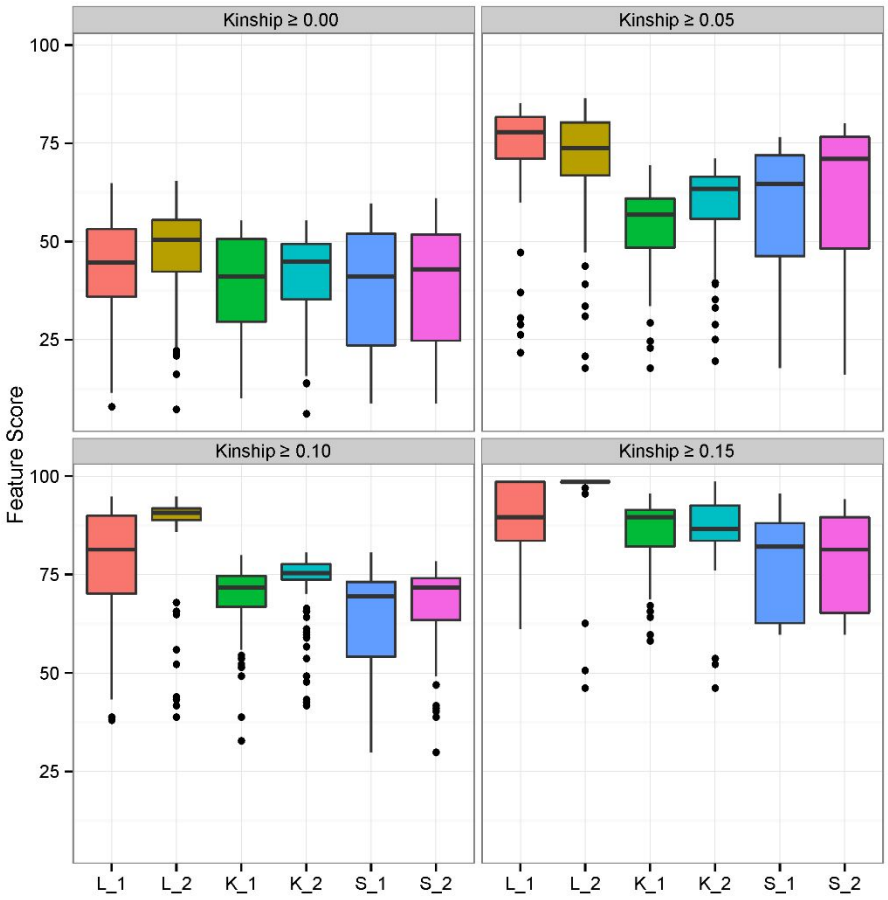
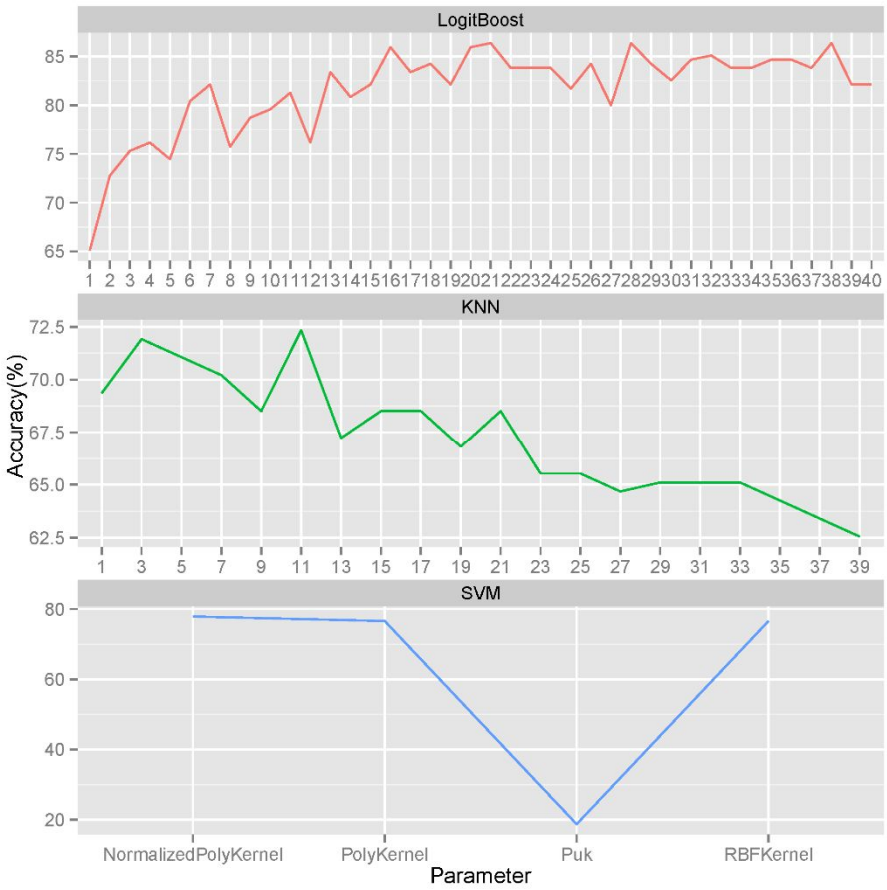


Figure 3.7 Line plots for the results of parameter optimization.

The X-axis is the range of parameters used for each classifier (LogitBoost: iteration, KNN: K-nearest neighbors, and SVM: Kernel). The Y-axis represents classification accuracy calculated by 10-fold cross-validation.



General discussion

For a few decades, the advancement of sequencing technologies has accomplished outstanding achievement in science, particularly in genomics. In the field of genomics, genomic variants or features are useful tools for studying the underlying mechanism within genetic inheritances of a specific organism.

The representative genomic variants related to single nucleotide substitutions including SNP have been widely used to discriminate breeds with unique phenotypic traits such as coat colors in domesticated animals. In addition, although TEs took little attention compared to the other types of genomic features, there have been many studies that demonstrated the considerable effects of TE insertions on gene expression levels or phenotypic traits.

In these respects, this study could aid to understand the genetic mechanism underlying phenotypic traits of domesticated animals by revealing their genomic characteristics. Furthermore, this study presented the availability of application which used genomic characteristics as a tool for commercial purpose.

References

Abrusán, G., et al. (2008). "Biased distributions and decay of long interspersed nuclear elements in the chicken genome." Genetics **178**(1): 573-581.

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data." Reference Source.

Blancou, J. (2001). "A history of the traceability of animals and animal products." Revue scientifique et technique (International Office of Epizootics) **20**(2): 413-425.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Brodersen, K. H., et al. (2010). The balanced accuracy and its posterior distribution. Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE.

Cai, Y.-D., et al. (2006). "Using LogitBoost classifier to predict protein structural classes." Journal of theoretical biology **238**(1): 172-176.

Chang, C.-M., et al. (2006). "Complete association between a retroviral insertion in the tyrosinase gene and the recessive white mutation in chickens." BMC genomics **7**(1): 19.

Chang, C., et al. (2012). "A global analysis of molecular markers and phenotypic traits in local chicken breeds in Taiwan." Animal genetics **43**(2): 172-182.

Cornuet, J.-M., et al. (1999). "New methods employing multilocus genotypes to select or exclude populations as origins of individuals." Genetics **153**(4): 1989-2000.

D'Alessandro, E., et al. (2007). "Analysis of the MC1R gene in the Nero Siciliano pig breed and usefulness of this locus for breed traceability." Veterinary research communications **31**: 389-392.

Dalvit, C., et al. (2007). "Genetic traceability of livestock products: A review." Meat Science **77**(4): 437-449.

Darwin Charles, R. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, Murray, London.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. ICML, Citeseer.

Dettling, M. and P. Bühlmann (2003). "Boosting for tumor classification with gene expression data." Bioinformatics **19**(9): 1061-1069.

Dimauro, C., et al. (2013). "Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes." Animal genetics **44**(4): 377-382.

Ding, C. H. and I. Dubchak (2001). "Multi-class protein fold recognition using support vector machines and neural networks." Bioinformatics **17**(4): 349-358.

Dorshorst, B., et al. (2010). "Genomic regions associated with dermal hyperpigmentation, polydactyly and other morphological traits in the Silkie chicken." Journal of Heredity **101**(3): 339-350.

Eriksson, J., et al. (2008). "Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken." PLoS genetics **4**(2): e1000010.

Even-Zohar, Y. and D. Roth (2001). "A sequential model for multi-class classification." arXiv preprint cs/0106044.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern recognition letters **27**(8): 861-874.

Fernández, A., et al. (2004). "DNA tests based on coat colour genes for authentication of the raw material of meat products from Iberian pigs." Journal of the Science of Food and Agriculture **84**(14): 1855-1860.

Feschotte, C. (2008). "Transposable elements and the evolution of regulatory networks." Nature Reviews Genetics **9**(5): 397-405.

Feschotte, C. and E. J. Pritham (2007). "DNA transposons and the evolution of eukaryotic genomes." Annual review of genetics **41**: 331.

Freese, N. H., et al. (2014). "A Novel Gain-Of-Function Mutation of the Proneural IRX1 and IRX2 Genes Disrupts Axis Elongation in the Araucana Rumpless Chicken." PloS one **9**(11): e112364.

Friedman, J., et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics **28**(2): 337-407.

Güney, S. and A. Atasoy (2012). "Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronic nose." Sensors and Actuators B: Chemical **166**: 721-725.

Galtier, N., et al. (2009). "Mitochondrial DNA as a marker of molecular diversity: a reappraisal." Molecular ecology **18**(22): 4541-4550.

Goffaux, F., et al. (2005). "Development of a genetic traceability test in pig based on single nucleotide polymorphism detection." Forensic science international **151**(2): 239-247.

Golan, E. H., et al. (2004). Traceability in the US food supply: economic theory and industry studies, US Department of Agriculture, Economic Research Service Washington, DC.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." The Journal of Machine Learning Research **3**: 1157-1182.

Hastie, T. and R. Tibshirani (1998). "Classification by pairwise coupling." The annals of statistics **26**(2): 451-471.

Heaton, M. P., et al. (2014). "SNPs for parentage testing and traceability in globally diverse breeds of sheep." PloS one **9**(4): e94851.

Hillier, L. W., et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Hornik, K., et al. (2007). "RWeka: an R interface to Weka." R package version 0.3-4., URL http://CRAN. R-project. org/package= RWeka.

Hsu, C.-W. and C.-J. Lin (2002). "A comparison of methods for multiclass support vector machines." Neural Networks, IEEE Transactions on **13**(2): 415-425.

Iquebal, M. A., et al. (2013). "Development of a model webserver for breed identification using microsatellite DNA marker." BMC genetics **14**(1): 118.

Isbel, L. and E. Whitelaw (2012). "Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes." BioEssays **34**(9): 734-738.

Jain, A. and D. Zongker (1997). "Feature selection: Evaluation, application, and small sample performance." Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(2): 153-158.

Kapitonov, V. V. and J. Jurka (2006). "Self-synthesizing DNA transposons in eukaryotes." Proceedings of the National Academy of Sciences of the United States of America **103**(12): 4540-4545.

Kazazian, H. H. (2004). "Mobile elements: drivers of genome evolution." science **303**(5664): 1626-1632.

Keane, T. M., et al. (2013). "RetroSeq: transposable element discovery from next-generation sequencing data." Bioinformatics **29**(3): 389-390.

Kido, Y., et al. (1991). "Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution." Proceedings of the National Academy of Sciences **88**(6): 2326-2330.

Kidwell, M. G. (2002). "Transposable elements and the evolution of genome size in eukaryotes." Genetica **115**(1): 49-63.

Kohavi, R. and G. H. John (1997). "Wrappers for feature subset selection." Artificial intelligence **97**(1): 273-324.

Koketsu, Y. (2000). "Productivity characteristics of high-performing commercial swine breeding farms." Journal Of The American Veterinary Medical Association **216**(3): 376-379.

Kotsiantis, S., et al. (2006). "Logitboost of multinomial Bayesian classifier for text classification." International Review on Computers and Software (IRECOS) **1**(3): 243-250.

Kwak, W., et al. (2014). "Uncovering Genomic Features and Maternal Origin of Korean Native Chicken by Whole Genome Sequencing." PloS one **9**(12): e114763.

Lander, E. S., et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

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

Li, H., et al. (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Long, N., et al. (2007). "Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers." Journal of animal breeding and genetics **124**(6): 377-389.

Lorena, A. C., et al. (2008). "A review on the combination of binary classifiers in multiclass problems." Artificial Intelligence Review **30**(1-4): 19-37.

Malik, H. S., et al. (1999). "The age and evolution of non-LTR retrotransposable elements." Molecular biology and evolution **16**(6): 793-805.

Manichaikul, A., et al. (2010). "Robust relationship inference in genome-wide association studies." Bioinformatics **26**(22): 2867-2873.

McKenna, A., et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in nuclear medicine, Elsevier.

Meyer, T. J., et al. (2012). "An Alu-based phylogeny of gibbons (Hylobatidae)." Molecular biology and evolution **29**(11): 3441-3450.

Murata, S., et al. (1993). "Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution." Proceedings of the National Academy of Sciences **90**(15): 6995-6999.

Murphy, R., et al. (2008). "Review: animal identification systems in North America." The Professional Animal Scientist **24**(4): 277-286.

Negrini, R., et al. (2009). "Assessing SNP markers for assigning individuals to cattle populations." Animal genetics **40**(1): 18-26.

Negrini, R., et al. (2008). "Traceability of four European protected geographic indication (PGI) beef products using single nucleotide polymorphisms (SNP) and Bayesian statistics." Meat Science **80**(4): 1212-1217.

Ostergaard, E., et al. (2010). "A novel missense mutation in SUCLG1 associated with mitochondrial DNA depletion, encephalomyopathic form, with methylmalonic aciduria." European journal of pediatrics **169**(2): 201-205.

Perry, W. L., et al. (1994). "The molecular basis for dominant yellow agouti coat color mutations." Bioessays **16**(10): 705-707.

Petroman, C., et al. (2012). "Management of sow replacement rate." Porcine Research **2**(1): 16-18.

Polat, K. and S. Güneş (2009). "A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems." Expert Systems with Applications **36**(2): 1587-1592.

Purcell, S., et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics **81**(3): 559-575.

Ramos, A., et al. (2011). "Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing." Animal genetics **42**(6): 613-620.

Remme, C. A., et al. (2008). "Cardiac Sodium Channel Overlap Syndromes: Different Faces of< i> SCN5A</i> Mutations." Trends in cardiovascular medicine **18**(3): 78-87.

Saeys, Y., et al. (2007). "A review of feature selection techniques in bioinformatics." Bioinformatics **23**(19): 2507-2517.

Schott, J.-J., et al. (1999). "Cardiac conduction defects associate with mutations in SCN5A." Nature genetics **23**(1): 20-21.

Seo, M. and S. Oh (2012). "CBFS: High performance feature selection algorithm based on feature clearness."

Sing, T., et al. (2005). "ROCR: visualizing classifier performance in R." Bioinformatics **21**(20): 3940-3941.

Singer, M. F. (1982). "SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes." Cell **28**(3): 433-434.

Slotkin, R. K. and R. Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome." Nature Reviews Genetics **8**(4): 272-285.

Smith, G., et al. (2008). "Post-slaughter traceability." Meat Science **80**(1): 66-74.

Smith, G., et al. (2005). "Traceability from a US perspective." Meat Science **71**(1): 174-193.

SOMES, R. G., et al. (1977). "Protein and cholesterol content of Araucana chicken eggs." Poultry science **56**(5): 1636-1640.

Sun, P., et al. (2014). "An improved multiclass LogitBoost using adaptive-one-vs-one." Machine Learning **97**(3): 295-326.

Tamura, K., et al. (2013). "MEGA6: molecular evolutionary genetics analysis version 6.0." Molecular biology and evolution **30**(12): 2725-2729.

Teye, E., et al. (2013). "Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **114**: 183-189.

Teye, E., et al. (2014). "Discrimination of cocoa beans according to geographical origin by electronic tongue and multivariate algorithms." Food Analytical Methods **7**(2): 360-365.

Vasicek, T. J., et al. (1997). "Two dominant mutations in the mouse fused gene are the result of transposon insertions." Genetics **147**(2): 777-786.

Wang, Y., et al. (2005). "Gene selection from microarray data for cancer classification—a machine learning approach." Computational biology and chemistry **29**(1): 37-46.

Wang, Z., et al. (2013). "An EAV-HP insertion in 5′ Flanking region of SLCO1B3 causes blue eggshell in the chicken."

Wang, Z., et al. (2013). "An EAV-HP Insertion in 5′ Flanking Region of SLCO1B3 Causes Blue Eggshell in the Chicken." PLoS genetics **9**(1): e1003183.

West, B. and B.-x. Zhou (1989). "Did chickens go north? New evidence for domestication." World's Poultry Science Journal **45**(03): 205-218.

Whitelaw, E. and D. I. Martin (2001). "Retrotransposons as epigenetic mediators of phenotypic variation in mammals." Nature genetics **27**(4): 361-365.

Wicker, T., et al. (2007). "A unified classification system for eukaryotic transposable elements." Nature Reviews Genetics **8**(12): 973-982.

Wragg, D., et al. (2013). "Endogenous retrovirus EAV-HP linked to blue egg phenotype in Mapuche fowl." PloS one **8**(8): e71393.

Wu, M., et al. (1997). "Inherited somatic mosaicism caused by an intracisternal A particle insertion in the mouse tyrosinase gene." Proceedings of the National Academy of Sciences **94**(3): 890-894.

Wu, T.-F., et al. (2004). "Probability estimates for multi-class classification by pairwise coupling." The Journal of Machine Learning Research **5**: 975-1005.

Xing, J., et al. (2005). "A mobile element based phylogeny of Old World monkeys." Molecular phylogenetics and evolution **37**(3): 872-880.

Xing, J., et al. (2007). "A mobile element-based evolutionary history of guenons (tribe Cercopithecini)." BMC biology **5**(1): 5.

Yang, J., et al. (2011). "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics **88**(1): 76-82.

Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. ICML.

Yuan, P., et al. (2008). MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification. Semantic Computing and Systems, 2008. WSCS'08. IEEE International Workshop on, IEEE.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

Zhao, R., et al. (2006). "A study on eggshell pigmentation: biliverdin in blue-shelled chickens." Poultry science **85**(3): 546-549.

요약(국문초록)

생물정보학적 접근방법을 이용한 가축유전체의 특성 규명과 그 응용에 대한 연구

김권도

농생명공학부 동물생명공학전공

서울대학교 대학원 농업생명과학대학

가축화된 동물들은 인간에 의한 인위적 선택으로 인해 자연상태의 동물과는 다른 유전체적 특성을 가지고 있다. 또한, 그들의 유전체적 특성은 유생산, 산자수와 같은 생산형질에 큰 영향을 미칠 수 있기 때문에 가축화된 동물의 새로운 유전적 특성을 규명하고 분석하는 것은 산업적으로나 학문적으로 큰 가치를 지니고 있다고 할 수 있다. 유전적 특성 중, 단일 염기 다형성은 많은 연구에서 활용되어왔는데, 특히 가축에서는 산업적으로 큰 가치를 지닌 품종을 구분하기 위하여 연구되었다. 예를 들어, 단일 염기 다형성에 대한 유전형 분석을 통하여 희소가치가 높은 돼지를 구분하는 방법이 실제로 활용되고 있다. 또 다른 형태의 유전체 특성인 구조 변이는 단일염기 다형성 보다

대규모인 수백에서 수만개에 이르는 염기서열 변화를 일으킨다. 이러한 구조 변이 중 하나인 전이 인자는 특징적으로 유전체내에서의 이동이 가능하며 이러한 이동은 유전적 변이와 더불어 개체의 형질을 변화시킬 수 있다.

한편, 닭의 유전체는 다른 동물과 다르게 전이 인자를 다량보유하고 있으며 이러한 전이 인자로 인한 형질 변화에 대한 여러 연구가 보고 되었다. 제 2 장에서는 파란 달걀을 생산하는 닭인 경북 아라우카나의 유전체 단편 서열정보를 차세대 염기서열 분석방법을 이용하여 얻어내었다. 이를 이용하여 특이적 전이 인자의 탐색과 군집 분석을 수행하였고, 경북 아라우카나의 특성과 관련 된 3 개의 후보 전이 인자를 발굴하였다. 또한 군집 분석의 결과를 통해 경북 아라우카나의 기원과 종 내에서의 위치에 대한 정보를 얻을 수 있었다.

생산이력제는 동물 또는 동물성 식품의 생산지를 추적하는 방법을 말한다. 이는 식중독과 같은 식품과 관련된 전염성 질병을 예방하거나 대처하는 데 매우 중요한 방법이다. 생산이력제는 또한 동물성 식품에 대한 소비자의 신뢰도를 향상 시키는 역할을 할 수 있다. 그러나 기계학습을 통한 생산이력제에 대한 연구는 현재까지 거의 진행되지 않았다. 제 3 장에서는 104 개의 농장에서 생산된 4,122 마리의 돼지를 이용하여 유전형을 분석하고 이를 이용하여 각각의 돼지를 농장에 따라 분류할 수 있는 모형을 구축하였다. 거의 모든 경우에서 LogitBoost 분류기를 이용한 모형이 분류 정확도 측면에서 다른 모형을 능가하였으며, 유전적 관계가 높은

집단에서 더 높은 정확도를 나타내었다. 이 두 결과는 단일염기 다형성을 이용한 기계학습 접근 방법의 생산이력제에 대한 응용가능성을 보여준다.

주요어: 가축화 동물, 유전체 변이, 전이 인자, 생산이력제, 분류 분석

학번: 2011-21283