Master's Thesis of Agriculture

Linear mixed models for genome-wide association study and phenotype prediction

전장유전체연관분석과 표현형 예측 연구를 위한 선형혼합모형

2019년 2월

서울대학교 대학원

농생명공학부

원 소 영

Linear mixed models for genome-wide association study and phenotype prediction

전장유전체연관분석과 표현형 예측 연구를 위한 선형혼합모형

지도교수 김 희 발

이 논문을 농학석사 학위논문으로 제출함

2018년 12월

서울대학교 대학원

농생명공학부

원 소 영

원소영 의 농학석사 학위논문으로 인준함

2018년 12월

위 원 장 윤 철 희 (인)

부위원장 김 희 발 (인)

위 원 조 서 애 (인)

Abstract

Linear mixed models for association study and phenotype prediction

Sohyoung Won

Department of Agricultural Biotechnology

Seoul National University

With the advance of sequencing and genotyping technologies, a large amount of genomic data has been accumulated and is available for biological studies. Along with the development of statistical models and computational capabilities, sizable genomic data can be analyzed thoroughly. Processing large genomic data via statistical computation enables discerning the relationship between genotypes and phenotypes.

In this thesis, the main concern was how differences in genotypes are related to phenotypes. I conducted genome-wide association study to discover genetic variants correlated with phenotypes. Also, I constructed prediction models to precisely estimate phenotypes from genotypes. In the studies, various linear mixed models were applied to calculate the effects of genetic variants.

In chapter 2, genome-wide association study on intramuscular fat content

of pig was performed. Statistically significant single nucleotide polymorphisms were found and annotated to genes. Genes related to mitogen-activated protein kinase pathway were identified as candidate genes affecting the intramuscular fat content of pigs.

In chapter 3, genomic prediction models using haplotype alleles were constructed. The models attempt to predict carcass weight in Hanwoo. Different haplotype defining methods were implemented and the prediction accuracies of them were compared. As a result, genomic prediction accuracy was higher when haplotype alleles were used compared to when individual SNPs were used.

In chapter 4, models predicting human height from genotype were developed. I designed a genomic best linear unbiased prediction model adjusted with parental height. In addition, variables having highest effects on height were selected using bootstrap resampling. Models using only the selected variables were tested, and consequently I could obtain a model with high prediction ability.

Through these studies, I could understand how linear mixed models can be applied to explain relationships between genotypic variation and phenotypic variation. The findings of this dissertation will help to extend the use of linear mixed models for understanding the genetic architectures in animals and human.

Key words: genome-wide association study, genomic prediction, linear mixed model

Student number: 2017-22852

Table of Contents

List of Figures

List of Tables

Chapter 1. Literature Review

1.1 Linear Mixed Models

Linear mixed models are linear models with both fixed effects and random effects. Here, fixed effects represent values from the full population have specific values. On the other hand, random effects represent values of a random sample from the population and follow specific distributions such as the normal distribution. Linear mixed models are often used when there is hierarchical structure in data (Gelman and Hill 2006). In genomic data, linear mixed models can be applied to explain the additive genetic effects where population structure is present among samples. Here, the effect of each genetic variant is different but follows a common distribution. Furthermore, linear mixed models allow to fit regression models having more number of independent variables than sample size. Since genotype data may contain tens of thousands or more variants while it is very difficult to obtain a comparable sample size, linear mixed models can be useful to handle genomic data.

1.2 Genome-wide association study

With the advance of sequencing and genotyping technologies, large amount of genomic data has been accumulated and is available for biological studies. From genomic data, discovering how differences in the genome is related with phenotypes is a major subject of interest. Genome-

wide association study (GWAS) is a study using statistical models and genetic variants across the whole genome to identify which variants are associated with a certain phenotype. Since when GWAS was first introduced, many remarkable discoveries have been made by GWAS (Visscher, Wray et al. 2017). For example, genetic risk factors related to human diseases such as schizophrenia (Nature 2009), (Li, Chen et al. 2017) or type 2 diabetes (Scott, Mohlke et al. 2007), (Frayling 2007) were revealed. In animals, quantitative trait loci affecting economical traits such as milk production (Raven, Cocks et al. 2014), (Pryce, Bolormaa et al. 2010) were discovered, which could be further used to improve the rate of genetic gain from breeding.

GWAS measures the probabilities of genetic variants being associated with a trait and attempts to find causal variants, which have high probabilities of association, in other words have low p-values assuming no association. The most frequently used genetic variant is single nucleotide polymorphism (SNP) which refers to a single base-pair change in the DNA sequence (Nature 2010). The trait of GWAS may be either qualitative or quantitative and different models can be used according to the type of the trait.

When the trait is quantitative, linear models are fitted to measure the significance of a SNP as an explanatory variable in the model. Quantitative traits are affected by large number of genetic variants cumulatively and

environmental effects. Therefore, a model accounting for both the additive effects of genetic variants and possible environmental effects is needed for GWAS of quantitative traits.

$$y_i = \mu + x_i\beta + g_i + \epsilon_i$$

Here, y_i is the phenotype of the ith individual, μ is the general mean, β is the vector of covariates, g_i is additive genetic effect of the ith individual, and ϵ_i is the random error. Covariates such as sex or age, or any differences other than genetic variants can be included in the model as the β term to adjust environmental effects in the phenotype. Additive genetic effect, g_i, is the total sum of the effects of the SNP alleles that an individual carry. This can be expressed as a linear combination of the SNP genotype coded in 0, 1, or 2 according to the number of a specific allele and the numeric effect of the allele, as $g_i = \Sigma\alpha_{ij}u_j$, where α_{ij} is the genotype of the jth SNP of the ith individual and u_j is the effect of the jth SNP.

The effects of SNPs, denoted as u_j in the equation above, can be treated as either fixed effects or random effects. If u_js are fixed effects, their effect and significance of each SNP can be simply calculated by linear regression. However, this is valid only when the samples are unrelated (Balding 2006). If not, false positive SNPs may be discovered due to population stratification. This is an important problem especially in GWAS with livestock, since it is almost impossible to obtain unrelated livestock

sample. To adjust population stratification avoiding spurious association, adding principal components as covariates can be considered (Price, Patterson et al. 2006). Another method to deal with population stratification is using linear mixed models (Yu, Pressoir et al. 2006). In linear mixed models, the effects of SNPs are random effects while environmental effects such as sex and age are fixed effects. Random effects follow a distribution rather than have a fixed value. In linear mixed models for GWAS, the effects of SNPs are assumed to follow normal distributions. The random errors are also random following normal distribution, and the variances of the genetic effect and environmental effect are estimated using restricted maximum likelihood. The significances of SNPs can be estimated using likelihood ratio test.

In qualitative traits occurring as cases and controls, the frequencies of an allele in case and control are compared to measure the probability of association. Generalized linear models such as the logistic model can be used. With related samples, generalized linear mixed models can be applied.

GWAS mostly concerns identifying common variants associated with a trait under the common diseases common variant hypothesis (Lander 1996). In this hypothesis, many common variants have small effects on a disease or a trait. Accordingly, it is common to leave out SNPs having lower minor allele frequencies, which seem rare, before performing GWAS. Removing rare SNPs is also due to statistical power since the statistical

power is extremely low for rare SNPs (Turner, Armstrong et al. 2011). However, common variants cannot account for all the phenotypic variation and GWAS attempting to detect rare variants are also being conducted (Cohen, Kiss et al. 2004), (Cohen, Kiss et al. 2004). Other quality controls prior to GWAS include removing SNPs severely deviating from Hardy-Weinberg equilibrium to prevent genotyping error and population stratification (Turner, Armstrong et al. 2011). Also samples or SNPs with low genotyping rates are removed to maintain data quality and statistical power (Turner, Armstrong et al. 2011).

Many tools are developed for the practical application of GWAS, for example PLINK (Purcell, Neale et al. 2007), GCTA (Yang, Lee et al. 2011) and GEMMA (Zhou and Stephens 2012). After preparing standard input files according to the manual of a certain tool, GWAS can be performed by entering some simple command lines. The results of GWAS can be presented as a Manhattan plot, which shows the physical positions and significances of SNPs. Gene annotations of significant SNPs followed by functional classifications or pathway analyses of the annotated genes are generally performed as the next step of GWAS. Since statistical significance does not always mean biological significance, these are crucial in further interpreting the results of GWAS.

1.3 Genomic prediction

Genomic prediction is an effective way of predicting breeding values or phenotypes from a large number of genetic markers distributed across the entire genome (Hayes and Goddard 2001). Even though GWAS has discovered numbers of quantitative trait loci (QTL) associated with various complex traits, the effects of the QTLs could only account for a limited part of the genetic effects since very large number of genetic variants contribute to the genetic variances of complex traits (Hayes and Goddard 2010). To better account for genetic variances, an approach to use all available genetic markers simultaneously was developed, which is genomic prediction. In genomic prediction, a genome-wide panel of dense SNPs are used as explanatory variables of the prediction model, and their effects are calculated using statistical models. Then, the predicted values are obtained as the sum of the effects of all SNPs used. In this way, all QTL is assumed to be in linkage disequilibrium (LD) with at least one SNP (Meuwissen and Goddard 2001). Here, LD is the nonrandom association between different loci (Slatkin 2008). As all QTLs are in LD with some SNP used for prediction, all the possible genetic effects of a complex trait can be explained from the effects of SNPs.

For the statistical prediction model of genomic prediction, linear mixed models as follows is used,

$$\mathrm{y} = \mathrm{Xb} + \mathrm{Zu} + \epsilon, \qquad \mathrm{u} \sim \mathrm{N}\left(0, \sigma_g^2 G\right), \epsilon \sim N(0, \sigma_e^2 I)$$

,where y is a vector of phenotypes, b is the vector of fixed effects such as sex and age, u is the vector of additive genetic effects, ϵ is the vector of random errors, X and Z are design matrices. Here, additive genetic effect u and random error ϵ follow normal distributions with mean 0 and variance $\sigma_g^2 G$ and $\sigma_e^2 I$ respectively. σ_g^2 and σ_e^2 are variance components, where σ_g^2 is the additive genetic variance and σ_ϵ^2 is the environmental variance, and there sum equals phenotypic variance. I is an $n \times n$ identity matrix and G is the genomic relationship matrix also $n \times n$, when n is the number of individuals in the model.

The genomic relationship matrix, G, expresses the degree of genetic relatedness among individuals, in other words the proportion of genome shared by individuals. The scale of G is 0 to 1, where 0 means being perfectly unrelated and 1 means being genetically identical. The diagonals of G is the relatedness of an individual and itself, which is 1, and the $(i \times j)$th element is the relatedness of the ith individual and the jth. The genomic relationship matrix can be calculated from SNP genotypes by measuring how many alleles two individuals have in common adjusted with allele frequencies. This is realized using the following equation (VanRaden 2008).

$$ZZ' = (M - P)(M - P)',$$

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)} .$$

Here, M is the matrix of genotype coded in 1, 0, -1 so that the dialogs of MM' is the number of homozygous alleles of individuals and off-diagonals of MM' is the number of alleles shared by two individuals. P is the matrix of allele frequencies and p_is are minor allele frequencies. The numerator expresses degrees of relatedness and denominator is for scaling. With the genomic relationship matrix, the additive genetic effects can be predicted as the best unbiased linear prediction (BLUP) solutions from the following equation. Alternatively, non-linear methods are applied for genomic prediction such as Bayesian methods (Hayes and Goddard 2001) (Habier, Fernando et al. 2011).

Genomic prediction has provided a powerful tool in animal selection by accurately estimating the genetic merits of animals. The breeding value estimated using genomic prediction is called genomic estimated breeding value (GEBV) and the selecting animals according to their GEBV is called genomic selection. Genomic selection is a form of marker assisted selection (MAS), which is an indirect selection process based on markers rather than the trait itself. The difference between MAS using DNA markers and genomic selection is that MAS uses only some significant markers while genomic selection uses a large number of markers genome-wide. As the effects of individual SNPs are very small in complex

traits, using all available SNPs can improve the prediction accuracy making genomic selection outperform MAS (Hayes 2007).

Another selection method that can be compared with genomic selection is pedigree based selection. Similar linear mixed models are used in pedigree based selection and genomic selection. However, different methods are used to estimate the variance matrix of additive genetic effects, in other words the relationship matrix. The average relatedness is estimated in pedigree based method while the actual degree of DNA shared is estimated in genomic selection as described above. Because genomic selection can account for random recombination during meiosis, it can more accurately estimate the relationship matrix resulting in better prediction (Villanueva, Pong-Wong et al. 2005).

Nonetheless, there are cases when genotype data are available for only a small part of the population. In this case, a model jointly incorporating both pedigree and genotype to use all phenotypes can be a solution. This is achieved by constructing a relatedness matrix H that combines pedigree and genomic relationships as follows (Legarra, Aguilar et al. 2009) (Christensen and Lund 2010).

$$\mathrm{H} = \mathrm{A} + \begin{vmatrix} A_{11}A_{22}^{-1} & 0 \\ 0 & I \end{vmatrix} \begin{bmatrix} I \\ I \end{bmatrix} [G - A_{22}][I \quad I] \begin{vmatrix} A_{22}^{-1}A_{21} & 0 \\ 0 & I \end{vmatrix}$$

$$\mathrm{H}^{-1} = \mathrm{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

Subscripts 1 and 2 denotes ungenotyped and genotyped animals respectively while A is the pedigree relationship matrix and G is the genomic relationship matrix. The H matrix is used as the variance-covariance matrix of additive genetic values in the linear mixed model. This method is referred to as single-step GBLUP.

In practical, genomic selection is consisted of two stages. First, a model is trained for genomic prediction from genotypes and phenotypes. The effects of markers and covariates are estimated at this stage. Secondly, GEBVs are calculated from genotypes using the model from the first stage. Once a model is constructed, breeding values can be estimated without measuring phenotypes. Thus, genomic selection is especially effective when phenotypes are difficult to measure, for example when phenotypes are expressed in only one sex or phenotypes should be measured after death. Also, genomic selection enables selecting young animals before they can produce phenotypes, consequently reducing the time for selection and accelerating genetic gain.

Genomic prediction is widely used for animals (Hayes, Bowman et al. 2009) and plants (Jannink, Lorenz et al. 2010), yet carefully being implemented for estimating phenotypes of human. Still, there are studies estimating human complex traits such as height, high-density lipoproteins, and body mass index by genomic prediction (Lello, Avery et al. 2018) (Rudan, Campbell et al.). Genomic prediction may be a powerful tool to

predict health indexes of risks for diseases, should be correctly used with ethical concerns.

This chapter was published in *Asian-Australasian Journal of Animal Sciences* as a partial fulfillment of Sohyoung Won's Master program.

Chapter 2. Identification of genes related to intramuscular fat content of pig using genome-wide association study

2.1 Abstract

The aim of this study is to identify SNPs and genes related to pig IMF and estimate the heritability of IMF. Genome-wide association study (GWAS) on 704 inbred Berkshires was performed for intramuscular fat content (IMF). To consider the inbreeding among samples, associations of the SNPs with IMF were tested as random effects in a mixed linear model using the genetic relationship matrix by GEMMA. Significant genes were compared with reported pig IMF QTL regions and functional classification of the identified genes were also performed. Heritability of IMF was estimated by GCTA tool. Total 365 SNPs were found to be significant from a cutoff of p-value <0.01 and the 365 significant SNPs were annotated across 120 genes. 25 genes were on pig IMF QTL regions. BMPER, FOXO1, EDAR, RNF149, CD40, PTPN1, SOX9, MYC and MIF were related to mitogen-activated protein kinase (MAPK) pathway which regulates the differentiation to adipocytes. These genes and the genes mapped on QTLs could be the candidate genes affecting IMF. Heritability of IMF was estimated as 0.52, which was relatively high, suggesting that a considerable portion of the total variance of IMF is explained by the SNP information. Our results can contribute to breeding pig with better IMF and therefore, producing pork with better sensory qualities.

2.2 Introduction

Intramuscular fat content (IMF), which stands for the amount of fat located throughout skeletal muscles, is a major quality trait of meat affecting sensory attributes such as flavor and texture. IMF is decided by the number and size of intramuscular adipocytes, and is directly related to the juiciness and tenderness of meat (Hocquette, Gondret et al. 2010). Pork with higher IMF tends to have better flavor, juiciness and tenderness, resulting in higher overall acceptability (Fernandez, Monin et al. 1999). Therefore, by breeding pigs to have higher IMF, more palatable pork can be produced.

GWAS enabled to find out the impact of genetic variants on various traits of animals affecting productivity. By using GWAS and genotyped SNP data, genes associated to a certain economic trait of animals can be discovered. Previous GWAS studies about IMF of pigs have found out that H-FABP and ACSL4 polymorphisms to have association with IMF of different pig populations (Chen, Jiang et al. 2014). Also, SFRS18 gene is reported to be related to the regulation of intramuscular fat deposition in pigs (Wang, Xue et al. 2009). Polymorphic microsatellite loci CSSM34 and ETH10 were associated with marbling scores, which show the IMF in the Angus, Shorthorn and Wagyu cattle (Hocquette, Gondret et al. 2010).

Many previous studies using GWAS to find out the association of genomic data with meat quality traits such as IMF focused on finding quantitative trait locus (QTL) (de Koning, Janss et al. 1999), (Paszek,

Wilkie et al. 2001), (Ovilo, Pérez-Enciso et al. 2000). Out bred line-cross model analysis suggested QTLs on chromosomes 2, 4, and 6, and the half-sib model analysis suggested linkage for chromosomes 4 and 7 (de Koning, Janss et al. 1999). The data of QTLs discovered from previous studies is accumulated as a QTL database. The QTL database shows where QTL regions are located throughout chromosomes for each economic trait and animal. Using the QTL database, we can check whether a gene associated with a specific trait is within the known QTL region of the trait or not.

In this study, we analyzed the SNP data and IMF of pigs using genome wide association study (GWAS) to identify SNPs associated with IMF. To adjust the effect of inbreeding, a genetic relationship matrix was constructed and used during GWAS. Significant SNPs were matched to the nearest genes within 100kb. We compared the identified genes with the QTL database of pig IMF and classified the function of the identified genes. We also estimated the heritability of IMF using the data. This study aims to search genes associated to IMF of pig and furthermore, to provide knowledge for breeding pigs having better IMF consequently, better sensory qualities.

2.3 Materials and Methods

Ethics statement

The study protocol and the standard operating procedures (No. 2009-077, C-grade) of Berkshire pigs were reviewed and approved by National Institute of Animal Science's Institutional Animal Care and Use Committee.

Animals and phenotype records

Inbred Berkshire population was used for analysis, and IMF of the Berkshire sample were measured. A total of 704 samples were examined. Among them, 367 samples were male, 204 samples were female and the sex of 133 samples was unknown. Chemical fat extraction procedures were performed to measure IMF of each pig.

Genotyping and quality control

The genomic DNAs of pig were genotyped on the Illumina Porcine 60 K SNP Beadchip. 62,163 SNPs were genotyped. We discarded the markers with low minor allele frequency (<0.05), significant deviation from Hardy-Weinberg equilibrium ($p < 10^{-3}$), and low genotype call rate (<95%). Among 62,163 SNPs, 40,191 SNPs passed quality control. 3,651 SNPs failed the Hardy-Weinberg test, 3,304 SNPs failed the genotype missingness test, and 19,829 SNPs failed the minor allele frequency test.

Genome-wide association analysis

The phenotype (IMF) was standardized to z-scores by subtracting the mean and then dividing by the standard deviation, in each sex group (male, female, unknown) separately. Single trait, univariate linear mixed model was used for the analysis assuming additive effect of SNPs. SNP effects were treated as random effects and sex was added as a covariate. Software GEMMA was used to calculate the genetic relationship matrix of individuals and to test the effects of SNPs by likelihood ratio test (Zhou and Stephens 2012). The cutoff for statistical significance of genes was $p < 0.01$.

Gene annotation and functional classification

Gene annotation of significant SNPs was based on the Ensembl Genes 89 database of Sus scrofa genes (Sscrofa 10.2). Significant SNPs were annotated to the nearest genes within a distance of 100kb. Functional classification of genes was performed on DAVID, an online functional annotation database to see where the functions of the identified genes were mainly categorized into. Sus scrofa was selected as both species and background option. The cutoff of gene ontology was $p < 0.05$.

Heritability estimation

The GCTA tool (Yang, Lee et al. 2011) was used to calculate

heritability for IMF. We calculated the genetic relationship matrix (GRM) between all pairs of samples using all the autosomal SNPs. We then estimated the variance of genetic component by restricted maximum likelihood analysis, and heritability by dividing the estimated genetic variance by the total variance measured.

2.4 Results

Identification of significant SNPs

Totally 365 SNPs from all 19 chromosomes were identified as significant SNPs as the result of GWAS in this study. Chromosome 14 contained 53 significant SNPs which was the largest number among all chromosomes. There were 40 and 35 significant SNPs on chromosome 7 and 11 respectively, which contained second and third many significant SNPs. The statistical significance values of the association between each SNP and IMF calculated as -log10(p-value) across 18 autosomal chromosomes and chromosome X was plotted in the form of a Manhattan plot (Figure 2-1).

The 365 significant SNPs found from GWAS were annotated to the nearest genes within 100kb. 153 SNPs among the 365 significantly identified SNPs were annotated across 120 genes. There were some SNPs annotated to same genes and none of the significant SNPs on chromosome 8 and 15 had genes within 100kb distance. Full information of significant SNPs, their chromosome number, position, closest gene, distance from the closest gene, raw p-value is listed on Supplementary Table S2-1.

Mapping on QTL database

Identified genes were compared with IMF QTL regions base on the Pig QTLdb. Total 25 genes from the 120 significant genes, which is 20.8%,

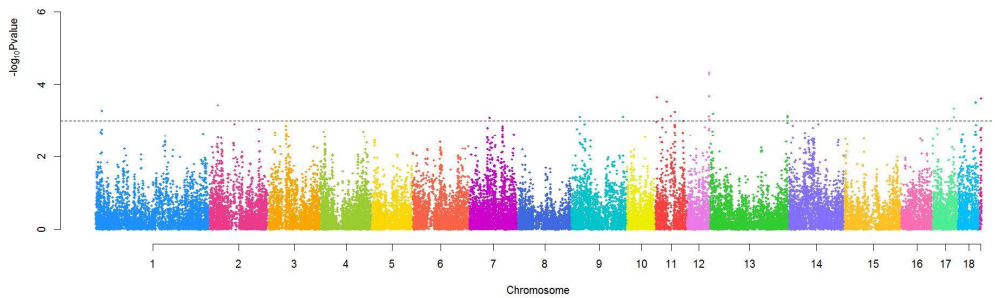

Figure 2-1. Statistical significance values of the association of SNPs across 18 autosomal chromosomes and the X chromosome with IMF are plotted as $-\log_{10} p$ values. The horizontal dotted line indicates the cutoff p =0.001.

were included in reported pig IMF QTL regions. 7 genes on chromosome 9, 6 genes on both chromosome 2 and 6, and 2 genes on chromosome 4, 7, and 17 each were mapped on QTLs (Figure 2-2). This suggests a considerable

The 365 significant SNPs found from GWAS were annotated to the nearest genes within 100kb. 153 SNPs among the 365 significantly identified SNPs were annotated across 120 genes. There were some SNPs annotated to same genes and none of the significant SNPs on chromosome 8 and 15 had genes within 100kb distance. Full information of significant SNPs, their chromosome number, position, closest gene, distance from the closest gene, raw p-value is listed on Supplementary Table S2-1.

Mapping on QTL database

Identified genes were compared with IMF QTL regions base on the Pig QTLdb. Total 25 genes from the 120 significant genes, which is 20.8%, were included in reported pig IMF QTL regions. 7 genes on chromosome 9, 6 genes on both chromosome 2 and 6, and 2 genes on chromosome 4, 7, and 17 each were mapped on QTLs (Figure 2-2). This suggests a considerable part of the genes identified from this study was consistent with the previous QTL studies, and those genes can be considered as genes that are located on the section of the genome having high correlation with IMF of pigs.

Functional classification

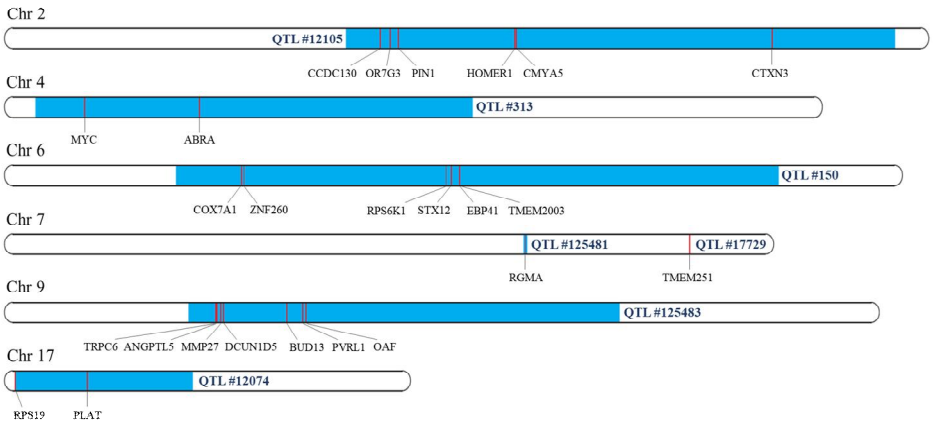

Figure 2-2. Location of significant genes mapped on QTLs. The red areas indicate where genes are located and the blue areas indicate QTL regions.

Identified genes were classified by their biological function and ontology. Regulation of MAPK cascade was the most significant gene ontologies from GOTERM_BP_5. The full result of functionally annotated genes are listed on Supplementary Table S2-2. Especially, BMPER, FOXO1, EDAR, RNF149, CD40, PTPN1, SOX9, MYC, MIF were categorized as genes related to both MAPK cascade and the regulation of MAPK cascade. Also, GDF7 and BMP6 were related to the regulation of MAPK pathway. FOXO1, RNF149, PTPN1, MYC were additionally annotated to negative regulation of MAPK cascade and regulation of stress-activated MAPK cascade.

Estimated heritability

Heritability of IMF was estimated by GCTA. The total variance of the sample was 1.020818 and the genetic variance was 0.526911. The genetic variance was estimated by the variance of genome-wide SNPs. The estimated heritability was 0.516166, approximately 0.52, and the standard deviation of the estimated heritability was 0.061655.

In previous studies, heritability of IMF was estimated at 0.39 (Suzuki, Irie et al. 2005), 0.44 (Larzul, Lefaucheur et al. 1997), 0.52 (Lo, McLaren et al. 1992), 0.65 (Newcom, Baas et al. 2004). Estimated heritability of IMF in the referred studies had values between 0.39 and 0.65. The heritability estimated from the SNP and phenotype data in this study,

0.52, was in the range of reported estimations and was according with the previous studies.

2.5 Discussion

Genes related to MAPK cascade and adipocyte differentiation

MAPK cascade was the most significant GO term from the functional annotation results of significant genes and other GO terms related to MAPK cascade appeared multiple times as well. MAPK pathway regulates various cell functions such as proliferation, differentiation and mitosis (Pearson, Robinson et al. 2001). Moreover, MAPK pathway is closely related to the differentiation of preadipocytes to adipocytes (Sakaue, Ogawa et al. 2004), (Aouadi, Jager et al. 2007). Some of the proteins involved in MAPK pathway also regulate adipocyte differentiation. For example, MAPK phosphatase-1 (MKP-1) downregulates the expression of p42/p44 MAPK and plays an important role in adipocyte differentiation (Sakaue, Ogawa et al. 2004). In addition, inhibition of p38MAPK decreases adipocyte differentiation in human and therefore p38MAPK activation can be seen as a requirement for primary human adipocyte differentiation (Aouadi, Jager et al. 2007). Since IMF is determined by the amount of adipocytes, genes related to MAPK pathway could affect IMF by regulating the amount of adipocyte differentiation.

Some of the significant genes related to MAPK cascade or the regulation of MAPK cascade (BMPER, FOXO1, SOX9, PTPN1, CD40) are previously reported to have influence on adipocyte differentiation. BMPER (bone morphogenetic protein [BMP]-binding endothelial cell precursor-

derived regulator) directly interacts with BMPs (Moser, Binder et al. 2003). Some BMPs activate p38MAPK pathway through the MAPK kinase kinase (MAPKKK) cascade (Tseng and He 2007) and BMPER could be needed for adipocyte differentiation to activate p38MAPK. Furthermore, BMP4 has an effect on lipid accumulation as well as expression of adipocyte markers (Bächner, Ahrens et al. 1998). Also, BMP2 and BMP7 induces adipocyte differentiation at low concentrate in C3H10T1/2 cell line (Wang, Israel et al. 1993). FOXO1 is expressed in the early stages of adipocyte differentiation and act as a preadipocyte differentiation preventing substituent (Nakae, Kitamura et al. 2003). Epidermal growth factor [EGF] repeat containing transmembrane protein (pref1) activates MAPK and upregulates SOX9 resulting in inhibition of adipocyte differentiation (Wang and Sul 2009). CD40 is related to the activation of MAPK (Shirakata, Ishii et al. 1999) and PTPN1 is a negative regulator of CD40 (Medgyesi, Hobeika et al. 2014). PTPN1 polymorphisms is reported to be associated with adipocyte related measures such as body fat percentage (Ukkola, Rankinen et al. 2005).

Heritability of IMF and selection

The estimated heritability of IMF, 0.52, was relatively high. This means that a substantial part of the total phenotypic variance of IMF is explained by the genetic variance. Here, the variance is that of the population, and thus high heritability suggests high genetic influence in the

population on the whole (Griffiths, Wessler et al. 2005). Heritability is an important parameter for predicting the response to selection (Visscher, Hill et al. 2008). Since the Breeder's equation is given as $R=h^2 S$, where R is the response to selection, S is the selection differential and h is the heritability (Falconer 1960), higher heritability can result in stronger response of selection and effective selection. Therefore, the phenotype information of IMF can be useful information for selecting pigs to breed pigs having high level of IMF.

Pork containing more than 3% IMF tends to have higher sensory qualities including juiciness, tenderness and taste (Daszkiewicz, Bąk et al. 2005). As IMF of pork increased from a range of 2.01~3.0% to higher than 3%, juiciness, tenderness, and both the intensity and desirability of taste increased. Since the current average of IMF measured from the Berkshire sample was 2.82%, if we increase IMF up to 3% by selection and breeding, we would be able to produce pork with improved juiciness, tenderness and taste.

Limitations of results

The tool used for association analysis in this study, GEMMA, adjusts the effect of sex by using sex as a covariate and use genetic information from the X chromosome in the same way as those from autosomal chromosomes while computing the genetic relationship matrix

(Zhou and Stephens 2012). However, since females carry two copies of X chromosomes while males carry a single copy, different methods should be used to estimate the genetic relationship for female-female pairs, male-male pairs and female-male pairs respectively in GWAS analyses (Yang, Lee et al. 2011). Furthermore, among the 704 samples used in this study, the sex of 133 was unknown. Approximately 19% of the sample had unknown sex. Also, to balance the allele dosage between sexes, one of the female X chromosome is silenced by random X chromosome inactivation (Ahn and Lee 2008). Therefore, additional information coding which allele was inactivated is needed to adjust GWAS analyses. In this study, information about which allele was inactivated was not provided, and this might together cause inaccuracy in the results from the X chromosome (Tukiainen, Pirinen et al. 2014). However, the proportion of significant SNPs on the X chromosome was 2.47% (9 out of 365) which was relatively low. Thus some part of inaccuracy in the results from the X chromosome may have not affect the overall results of the study.

Owing to the small sample size of animals, the overall significance of the study was low. Small sample size makes the effect size to be estimated low and consequently lowers the power of the study. The estimated power of the study was only 0.21 (Visscher, Hemani et al. 2014). To detect significantly associated SNPs in a study with low power, we had to use liberal statistics and a liberal cutoff (raw p-value and $p < 0.01$). This

might cause some significant SNPs to be false positive, but still the SNPs detected in this study can be suggested as candidates for SNPs related to IMF of pig. Besides, we could pick out some SNPs more likely to be actually related to IMF of pig by comparing them with known QTLs or searching their biological pathways. The genes mapped on QTLs or related to MAPK cascade may be stronger candidates for genes that are associated with IMF of pig than others.

Supplementary Table S2-1. SNPs and their chromosome number (Chr.), position on the chromosome, annotated gene name, distance from the gene, raw p-value are listed. 'match' indicates that the SNP was included in the gene and '-' indicates that no gene was in 100kd distance of the SNPs.

SNP	Chr	Position	Gene	Distance	p-value
MARC0018001	1	13478198	-	-	2.04E-03
ASGA0001052	1	13677927	CLDN20	85316	8.32E-03
M1GA0000785	1	16529867	ESR1	12501	7.76E-03
ALGA0001232	1	17130048	CCDC170	match	5.20E-04
ASGA0001260	1	17256411	RMND1	match	2.22E-03
ALGA0001244	1	17279855	ZBTB2	16864	1.76E-03
DRGA0001141	1	79571312	-	-	6.11E-03
INRA0002909	1	83308522	SEC63	match	9.40E-03
ALGA0005610	1	1.25E+08	-	-	8.69E-03
H3GA0003313	1	1.86E+08	-	-	5.19E-03
ALGA0007007	1	1.91E+08	-	-	9.84E-03
INRA0005278	1	1.92E+08	LRFN5	51070	3.96E-03
MARC0080275	1	1.92E+08	LRFN5	22518	3.96E-03
ASGA0005303	1	1.92E+08	LRFN5	2176	3.96E-03
MARC0033468	1	1.92E+08	LRFN5	13558	4.74E-03
H3GA0003345	1	1.92E+08	LRFN5	31112	4.81E-03
MARC0002276	1	1.92E+08	LRFN5	48118	6.86E-03
ALGA0007015	1	1.92E+08	LRFN5	63056	2.69E-03
ALGA0115186	1	1.92E+08	-	-	3.79E-03
MARC0036886	1	2.13E+08	NTRK3	match	9.80E-03
ASGA0005512	1	2.14E+08	-	-	6.56E-03
M1GA0001554	1	2.97E+08	SNORD90	44608	2.30E-03
ASGA0008803	2	6896014	GPR137	match	5.42E-03
M1GA0002584	2	9336635	MYRF	4509	7.74E-03
MARC0005659	2	20249153	HSD17B12	37203	5.24E-03

ALGA0112320	2	20842615	-	-	4.44E-03
ALGA0109169	2	22336683	-	-	5.67E-03
ALGA0012462	2	22590108	-	-	7.23E-03
ALGA0012504	2	23230160	-	-	2.98E-03
ASGA0009568	2	23271214	-	-	3.68E-04
ALGA0012515	2	23297131	-	-	7.60E-03
ALGA0113768	2	32333658	DCDC1	72522	3.06E-03
ALGA0012891	2	32621630	MPPED2	match	4.27E-03
ASGA0009858	2	32649983	MPPED2	match	4.42E-03
ALGA0012897	2	32887535	FSHB	22530	8.46E-03
MARC0113797	2	33100405	-	-	6.46E-03
H3GA0006477	2	33751046	-	-	9.35E-03
MARC0061061	2	35054445	-	-	6.69E-03
MARC0049526	2	35068263	-	-	9.62E-03
MARC0057893	2	35069583	-	-	7.40E-03
ALGA0012954	2	35089353	-	-	7.40E-03
MARC0067928	2	65512888	CC2D1A	match	6.55E-03
ALGA0013819	2	65741376	CCDC130	41960	4.27E-03
ALGA0013817	2	65761499	CCDC130	62083	4.52E-03
MARC0097970	2	67524684	OR7G3	94155	4.94E-03
DIAS0000914	2	69067481	PIN1	8231	1.23E-03
ALGA0014210	2	88585407	-	-	9.87E-03
ASGA0010795	2	89804886	HOMER1	39474	7.47E-03
H3GA0007086	2	90190571	CMYA5	3578	9.71E-03
ALGA0123643	2	1.36E+08	CTXN3	66969	3.84E-03
ALGA0015985	2	1.37E+08	-	-	1.71E-03
H3GA0007722	2	1.38E+08	-	-	7.32E-03
DRGA0003625	2	1.38E+08	-	-	7.26E-03
MARC0063459	2	1.39E+08	-	-	7.08E-03
ALGA0118500	3	12373502	-	-	6.27E-03

MARC0113856	3	18690752	ATXN2L	1178	2.09E-03
ASGA0090908	3	18873907	SBK1	6056	7.01E-03
ALGA0102956	3	18876394	SBK1	8543	7.72E-03
ALGA0116808	3	18943557	XPO6	60677	2.51E-03
ALGA0114914	3	32478883	-	-	5.23E-03
ALGA0107071	3	35759370	-	-	8.58E-03
ALGA0018852	3	49511934	EDAR	91461	1.76E-03
ALGA0018859	3	49555847	EDAR	47548	3.38E-03
H3GA0009479	3	49619062	EDAR	match	1.40E-03
ALGA0018869	3	49634868	CCDC138	match	2.27E-03
MARC0016359	3	49652312	CCDC138	match	4.49E-03
ALGA0018856	3	49698069	CCDC138	12421	2.79E-03
ASGA0014485	3	49720172	CCDC138	34524	2.54E-03
ALGA0104619	3	49867263	GCC2	20143	2.27E-03
CASI0006979	3	49869950	GCC2	17456	2.27E-03
MARC0009789	3	55205882	-	-	8.46E-03
MARC0100326	3	55496229	CREG2	match	9.63E-03
ALGA0019011	3	55561820	RNF149	16604	5.46E-03
MARC0065978	3	55717403	TBC1D8	59678	2.81E-03
ALGA0019168	3	58986683	TMEM131	match	4.92E-03
ASGA0094490	3	1.26E+08	GDF7	15667	5.33E-03
ASGA0103683	3	1.35E+08	ITGB1BP1	match	5.83E-03
H3GA0011795	4	8845519	-	-	1.98E-03
ALGA0023171	4	11921004	-	-	6.29E-03
ASGA0018288	4	11933016	-	-	4.40E-03
ALGA0023189	4	12065855	-	-	9.08E-03
ASGA0018338	4	12257838	-	-	7.63E-03
ASGA0018370	4	12821695	MYC	38214	2.85E-03
ASGA0018384	4	12941813	-	-	4.70E-03
DRGA0004465	4	13201853	-	-	8.34E-03

ALGA0023289	4	13259971	-	-	6.89E-03
ASGA0018422	4	13393209	-	-	6.96E-03
ALGA0023303	4	13453436	-	-	6.17E-03
ASGA0019158	4	33419642	ABRA	57931	9.51E-03
ASGA0019164	4	33447340	ABRA	85629	9.51E-03
M1GA0005832	4	33879367	-	-	8.95E-03
ALGA0026965	4	1.04E+08	KCNN3	match	8.73E-03
INRA0015741	4	1.04E+08	KCNN3	match	8.73E-03
MARC0109265	4	1.19E+08	DDX20	match	1.99E-03
INRA0016754	4	1.23E+08	-	-	2.72E-03
ALGA0123355	4	1.24E+08	-	-	7.76E-03
H3GA0014317	4	1.25E+08	-	-	6.27E-03
ALGA0028566	4	1.28E+08	-	-	4.13E-03
ASGA0022428	4	1.28E+08	-	-	8.52E-03
H3GA0014451	4	1.31E+08	-	-	9.21E-03
MARC0071918	5	6772267	DDX17	match	3.73E-03
ASGA0106044	5	6784209	DDX17	425	4.92E-03
ALGA0105509	5	6793058	KDELR3	match	4.09E-03
M1GA0007422	5	6800467	KDELR3	3641	3.48E-03
ALGA0032587	5	69669256	WASH1	24816	8.91E-03
MARC0019446	6	1829369	JPH3	match	8.52E-03
ASGA0084674	6	1838412	JPH3	match	7.89E-03
ALGA0115499	6	11401848	-	-	9.15E-03
ALGA0119163	6	40937246	COX7A1	14345	6.24E-03
MARC0005493	6	41367001	ZNF260	47542	6.24E-03
MARC0061190	6	41583185	ZNF829	match	6.69E-03
ALGA0115443	6	76356357	-	-	3.82E-03
ALGA0105228	6	76712823	MTFR1L	match	6.78E-03
ASGA0097645	6	77180997	PDIK1L	match	3.98E-03
ASGA0098887	6	77540317	RPS6KA1	60318	5.72E-03

ALGA0105183	6	77561790	RPS6KA1	81791	6.35E-03
ALGA0035761	6	78173716	CD164L2	match	8.02E-03
ASGA0099240	6	78503422	STX12	1399	8.27E-03
MARC0018089	6	79884592	EPB41	8507	9.02E-03
ALGA0103867	6	79924229	TMEM200B	5615	7.62E-03
ASGA0105794	6	1.27E+08	-	-	9.13E-03
M1GA0008912	6	1.27E+08	-	-	9.98E-03
ASGA0029775	6	1.39E+08	-	-	5.94E-03
ASGA0104725	6	1.46E+08	TTC4	26930	6.12E-03
ALGA0116372	6	1.55E+08	ERI3	match	5.26E-03
H3GA0010900	7	5120430	BMP6	41987	8.90E-03
H3GA0020119	7	17178315	CDKAL1	match	9.43E-03
MARC0024047	7	18074254	-	-	5.79E-03
H3GA0021382	7	49132338	-	-	3.14E-03
ASGA0033396	7	49191515	-	-	4.06E-03
ALGA0041186	7	49212847	-	-	2.86E-03
ASGA0033431	7	49785873	-	-	1.57E-03
H3GA0021402	7	50260219	CRISP1	15588	6.28E-03
H3GA0021406	7	50518152	-	-	9.02E-03
ALGA0041468	7	52319155	-	-	3.51E-03
H3GA0021745	7	56152592	-	-	8.15E-04
ASGA0034040	7	56199606	-	-	6.76E-03
MARC0001031	7	57317276	TM6SF1	match	4.43E-03
ALGA0042134	7	58188114	PDE8A	98766	5.61E-03
ALGA0042187	7	59301837	-	-	9.27E-03
H3GA0021903	7	62283996	LINGO1	93788	4.09E-03
M1GA0010442	7	62310234	LINGO1	67550	5.84E-03
H3GA0021941	7	63940639	ISLR	27268	3.03E-03
ASGA0034428	7	70929327	NPAS3	match	5.46E-03
M1GA0010466	7	71050476	NPAS3	79545	2.53E-03

MARC0006751	7	71727031	-	-	4.90E-03
ASGA0034705	7	88646106	-	-	7.80E-03
ALGA0043433	7	91741766	RGMA	73267	7.33E-03
ALGA0043428	7	91766941	RGMA	48092	2.95E-03
ALGA0043424	7	91780068	RGMA	34965	4.17E-03
ALGA0043415	7	91825127	RGMA	match	4.17E-03
ALGA0043406	7	91890551	CHD2	match	1.60E-03
ALGA0043404	7	91922539	CHD2	match	1.60E-03
DRGA0007977	7	92008579	CHD2	5988	1.60E-03
ALGA0043403	7	92021174	CHD2	18583	2.39E-03
ALGA0043398	7	92050130	CHD2	47539	2.84E-03
ALGA0043393	7	92089815	CHD2	87224	1.45E-03
ALGA0043388	7	92162122	FAM174B	49639	2.39E-03
H3GA0022349	7	92212775	FAM174B	match	3.85E-03
M1GA0010535	7	92231886	FAM174B	match	1.80E-03
MARC0095879	7	95409946	RAB15	7121	5.13E-03
MARC0098820	7	1.2E+08	SLC24A4	1550	7.88E-03
H3GA0023236	7	1.21E+08	TMEM251	48069	9.42E-03
ALGA0045073	7	1.22E+08	-	-	6.11E-03
MARC0044680	7	1.23E+08	GSC	26063	2.37E-03
ALGA0120902	9	11173743	MOGAT2	40781	9.37E-03
ALGA0119045	9	15055605	ssc-mir-708	64872	1.68E-03
ASGA0106225	9	22588132	ME3	1498	7.76E-04
ASGA0099198	9	22748344	ME3	56081	4.19E-03
ALGA0118782	9	22793075	-	-	2.20E-03
MARC0019308	9	25110616	TYR	54202	4.53E-03
H3GA0026870	9	30522282	PIWIL4	17025	4.76E-03
ALGA0052237	9	32169995	-	-	5.64E-03
MARC0018577	9	36364394	TRPC6	255	1.25E-03
ALGA0114399	9	36476161	ANGPTL5	82952	3.32E-03

ASGA0096889	9	37305665	MMP27	6788	6.09E-03
MARC0086124	9	37323349	MMP27	match	6.90E-03
ALGA0102606	9	37337886	MMP8	match	6.08E-03
ASGA0042475	9	37614741	DCUN1D5	5328	3.93E-03
DIAS0004102	9	37622654	DCUN1D5	match	3.36E-03
ASGA0042913	9	49127680	BUD13	81254	9.20E-03
H3GA0027281	9	51825334	PVRL1	55932	5.33E-03
ASGA0043018	9	52486044	OAF	53775	3.61E-03
MARC0057714	9	1.32E+08	-	-	9.38E-03
ALGA0055453	9	1.42E+08	SMYD2	50732	7.71E-04
ALGA0055456	9	1.42E+08	SMYD2	69072	9.58E-03
ALGA0056924	10	11062696	LYPLAL1	1973	7.47E-03
ASGA0046818	10	17948351	AKT3	17715	9.55E-03
ASGA0046812	10	18176414	AKT3	match	9.23E-03
MARC0064247	10	19258175	ZBTB18	97827	6.37E-03
ALGA0058975	10	50080838	PTER	match	2.86E-03
H3GA0030974	11	2923443	-	-	1.06E-03
ALGA0060411	11	4117930	USP12	46528	2.25E-04
ALGA0060892	11	12388831	CCNA1	24088	2.99E-03
MARC0031054	11	15537141	FOXO1	7719	8.32E-03
MARC0065987	11	16153984	NEK5	match	4.06E-03
ASGA0091162	11	18394930	ARL11	18017	1.53E-03
ALGA0061166	11	18613210	SETDB2	14124	4.01E-03
INRA0035578	11	19434517	CYSLTR2	31003	8.76E-04
H3GA0031500	11	19458211	CYSLTR2	7309	6.58E-03
DRGA0010995	11	28499617	-	-	2.04E-03
DRGA0011044	11	30843049	-	-	2.90E-04
ALGA0111608	11	33518592	-	-	6.97E-03
MARC0058247	11	40557571	-	-	8.83E-03
MARC0038885	11	42076772	-	-	9.26E-03

DRGA0011147	11	42683281	-	-	7.20E-04
ALGA0061971	11	43178260	-	-	5.91E-03
ALGA0062095	11	48231247	-	-	4.59E-03
ASGA0050759	11	49418020	RPS3A	60162	7.71E-03
DRGA0011263	11	51624925	-	-	7.89E-03
MARC0064023	11	52484631	UCHL3	7498	1.34E-03
MARC0057778	11	52507508	UCHL3	30375	2.14E-03
H3GA0031956	11	53021827	-	-	5.27E-03
ALGA0062282	11	53081698	-	-	3.95E-03
INRA0036447	11	53204913	-	-	1.28E-03
MARC0112524	11	53343454	-	-	5.63E-04
DRGA0011285	11	53365158	-	-	1.46E-03
INRA0036471	11	54041663	MYCBP2	7359	2.48E-03
INRA0036473	11	54059956	MYCBP2	25652	3.99E-03
ALGA0062309	11	54125788	MYCBP2	91484	7.84E-03
ASGA0102962	11	75272010	-	-	2.09E-03
ALGA0063437	11	76088356	ZIC5	match	3.77E-03
M1GA0015266	11	76187291	ZIC2	79290	2.13E-03
MARC0021953	11	76512208	GGACT	37247	1.01E-03
ASGA0051686	11	76688874	TMTC4	match	5.72E-03
ASGA0051750	11	77617889	-	-	8.06E-03
ASGA0098350	12	8826038	-	-	2.58E-03
ASGA0052986	12	8952780	SOX9	74859	9.47E-03
ALGA0065691	12	24829127	HOXB9	13910	6.16E-03
MARC0013292	12	24901251	HOXB13	2806	9.43E-03
MARC0065078	12	42958510	-	-	7.96E-03
ALGA0066582	12	44523799	RHOT1	23380	9.54E-03
MARC0110796	12	48693970	GOSR1	10842	1.48E-03
H3GA0034708	12	54636449	BCL6B	71720	7.59E-03
M1GA0017107	12	59863471	-	-	9.96E-03

MARC0093419	12	60329152	-	-	1.84E-03
ASGA0100802	12	60335124	-	-	7.49E-04
ASGA0101283	12	60445578	-	-	2.12E-04
H3GA0052996	12	60451706	-	-	4.63E-05
ALGA0120651	12	60466957	-	-	5.28E-05
H3GA0034965	12	60524116	-	-	9.13E-04
M1GA0017119	12	60549103	-	-	2.07E-04
M1GA0017151	12	60577246	-	-	1.61E-03
MARC0030180	12	61535474	U6atac	30049	5.75E-03
MARC0050410	12	61605585	-	-	5.69E-03
ASGA0055602	13	2180354	-	-	4.49E-03
ALGA0067480	13	3433598	GALNT15	86918	7.86E-03
MARC0015921	13	4575905	TBC1D5	match	8.01E-03
ASGA0089913	13	5308771	-	-	7.47E-03
ALGA0067602	13	6000731	SATB1	16413	1.97E-03
MARC0037054	13	7650575	EFHB	77580	6.36E-04
ASGA0055807	13	7699209	EFHB	28946	2.55E-03
M1GA0025009	13	25577420	SCN5A	match	7.19E-03
MARC0004520	13	1.41E+08	-	-	7.37E-03
MARC0093228	13	1.41E+08	-	-	5.62E-03
H3GA0037291	13	1.41E+08	-	-	6.88E-03
MARC0093203	13	1.44E+08	TNK2	8716	6.88E-03
ALGA0073790	13	2.1E+08	-	-	3.90E-03
ALGA0109869	13	2.1E+08	SIM2	24130	8.40E-03
ALGA0073982	13	2.14E+08	-	-	1.13E-03
ALGA0073987	13	2.14E+08	-	-	7.95E-04
M1GA0017861	13	2.14E+08	-	-	7.25E-04
H3GA0038523	14	5972387	GFRA2	92072	4.20E-03
ALGA0074628	14	6363534	-	-	8.00E-03
ASGA0061144	14	10777706	CDCA2	46175	1.36E-03

ALGA0075064	14	10905579	-	-	3.09E-03
MARC0025520	14	30477143	-	-	6.00E-03
MARC0080850	14	41317882	RPH3A	match	8.00E-03
MARC0016119	14	44298981	UNG	12724	6.56E-03
ALGA0077164	14	44313339	UNG	match	3.07E-03
ALGA0077178	14	44528915	SSH1	match	3.10E-03
INRA0043787	14	44548710	SSH1	match	2.19E-03
ASGA0063107	14	46665998	-	-	7.46E-03
ASGA0063110	14	46685558	-	-	8.05E-03
ASGA0063368	14	53273094	MIF	9458	4.02E-03
ALGA0077597	14	54540355	SLC25A1	1209	3.65E-03
ASGA0063385	14	54569925	SLC25A1	25470	2.77E-03
ALGA0077602	14	54595487	HIRA	617	5.78E-03
ALGA0077603	14	54680716	UFD1L	42	4.44E-03
H3GA0040220	14	54744216	UFD1L	match	2.77E-03
ASGA0063388	14	54791586	CLDN5	12252	2.77E-03
ASGA0063392	14	54837568	U3	25135	5.78E-03
MARC0059175	14	54867497	SEPT5	24594	3.65E-03
ASGA0063418	14	56633326	-	-	7.32E-03
ALGA0077635	14	56741944	-	-	6.38E-03
MARC0066981	14	56894113	-	-	7.70E-03
ASGA0063487	14	58847222	LGALS8	match	9.39E-03
ASGA0063736	14	62108085	MAP10	13196	3.47E-03
MARC0008126	14	65033626	URB2	6270	9.33E-03
ASGA0063956	14	65870468	ZNF37A	24351	8.90E-03
MARC0059823	14	65951741	-	-	9.87E-03
ASGA0063978	14	66119152	RET	42564	1.56E-03
MARC0097527	14	66239886	CSGALNACT2	match	9.08E-03
H3GA0040656	14	66437562	FXYD4	27335	3.18E-03
ALGA0078253	14	66455930	FXYD4	8967	8.41E-03

ASGA0064014	14	66708189	BICC1	match	4.02E-03
H3GA0040682	14	67337783	-	-	5.75E-03
MARC0009517	14	67360532	-	-	3.82E-03
ALGA0078315	14	67556695	FAM13C	match	4.79E-03
ASGA0064046	14	67586214	FAM13C	match	4.59E-03
ASGA0064057	14	67938907	SLC16A9	61736	4.09E-03
ALGA0078325	14	67973979	SLC16A9	26664	3.65E-03
H3GA0040698	14	68241945	CCDC6	match	8.48E-03
ALGA0078353	14	68492083	ANK3	match	8.69E-03
MARC0047133	14	68911064	-	-	8.13E-03
DRGA0013970	14	69219663	PSMA5	35898	7.51E-03
DRGA0013984	14	70168690	-	-	3.25E-03
ALGA0078438	14	70180492	-	-	4.19E-03
ALGA0078447	14	70318427	-	-	8.01E-03
ASGA0064171	14	70424031	-	-	5.89E-03
M1GA0018867	14	81218454	DNAJB12	224	1.23E-03
H3GA0042409	14	1.35E+08	-	-	4.89E-03
ASGA0066628	14	1.35E+08	ADRB1	96737	4.06E-03
ASGA0066777	14	1.37E+08	ATRNL1	28528	3.48E-03
ASGA0068236	14	1.52E+08	-	-	9.40E-03
ASGA0092166	15	2613630	-	-	3.24E-03
MARC0083940	15	4198197	-	-	9.19E-03
DRGA0014803	15	5072478	-	-	7.10E-03
ASGA0068402	15	7398646	-	-	4.63E-03
MARC0035392	15	53557172	-	-	3.16E-03
DRGA0015125	15	53593754	-	-	7.49E-03
ALGA0090697	16	52371543	MAP1B	20273	3.25E-03
H3GA0046642	16	55928776	-	-	8.26E-03
ALGA0090781	16	57686358	-	-	3.61E-03
ASGA0073900	16	71043011	-	-	8.27E-03

ALGA0091318	16	71070479	-	-	9.07E-03
ASGA0074891	17	15174	-	-	5.08E-03
H3GA0052370	17	439303	RPS19	13887	3.25E-03
ASGA0106200	17	494222	RPS19	68806	8.38E-03
ASGA0097925	17	536580	-	-	8.11E-03
ASGA0075356	17	11848326	SFRP1	match	1.59E-03
MARC0096794	17	13254722	PLAT	57497	2.32E-03
ALGA0095128	17	46668799	KIAA1755	12352	1.66E-03
ASGA0077339	17	53943853	CD40	519	5.52E-03
INRA0054309	17	58229635	-	-	4.60E-04
INRA0054308	17	58236218	-	-	7.96E-04
INRA0054314	17	58294982	PTPN1	60151	6.44E-03
H3GA0050343	18	9979868	-	-	4.04E-03
MARC0069211	18	10393660	TBXAS1	match	3.38E-03
ALGA0103897	18	10395342	TBXAS1	1259	3.51E-03
ALGA0102027	18	10404529	TBXAS1	10446	3.38E-03
ALGA0114284	18	10521082	-	-	3.62E-03
INRA0055202	18	13299138	ssc-mir-490-1	62842	2.33E-03
ALGA0097246	18	17876779	-	-	9.88E-03
ASGA0079081	18	18911412	-	-	5.24E-03
ASGA0090721	18	19225959	TSGA13	match	7.22E-03
ASGA0079728	18	43556023	BMPER	93939	5.49E-03
ASGA0079726	18	43567691	-	-	7.29E-03
H3GA0050905	18	47853853	-	-	2.42E-03
ALGA0098358	18	48005642	CPVL	match	2.36E-03
MARC0061468	18	48297065	-	-	6.70E-03
MARC0103241	18	48620026	-	-	2.00E-03
DBWU0000577	18	49563919	-	-	3.13E-04
ASGA0080057	18	49673542	-	-	3.07E-04
M1GA0023271	18	50032307	ssc-mir-196b-1	5195	1.31E-03

MARC0033103	18	50615451	-	-	2.39E-03
ALGA0111601	23	1182388	-	-	8.24E-03
ASGA0102465	23	2108185	CH242-227G12.1	17804	7.74E-03
ASGA0096921	23	2174682	CH242-227G12.1	34046	1.78E-03
ALGA0103241	23	3024536	NLGN4X	match	3.00E-03
ALGA0105315	23	3059017	NLGN4X	match	2.41E-04
ASGA0101131	23	3744915	-	-	3.62E-03
MARC0114252	23	3766084	-	-	6.24E-03
ALGA0124535	23	3876374	-	-	1.58E-03
ASGA0093054	23	4325817	STS	70802	6.87E-03

Supplementary Table S2-2. Gene ontologies with p <0.05, count >3 and the genes involved in the ontology. Count is the number of genes involved the ontology and % is (involved genes)/(total genes).

GO ID	GO term name	Count	%	p	Genes
GO:0043408	regulation of MAPK cascade	10	8.93	0.0018	BMPER, GDF7, FOXO1, EDAR, RNF149, CD40, PTPN1, MYC, MIF, BMP6
GO:0006355	regulation of transcription, DNA-templated	22	19.64	0.0050	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:2001141	regulation of RNA biosynthetic process	22	19.64	0.0054	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:1902531	regulation of intracellular signal transduction	15	13.39	0.0058	GDF7, ESR1, FOXO1, SMYD2, CD40, EDAR, SOX9, MIF, BMPER,

					ADRB1, ABRA, PTPN1, RNF149, MYC, BMP6
GO:0000165	MAPK cascade	9	8.04	0.0062	BMPER, FOXO1, EDAR, RNF149, CD40, PTPN1, SOX9, MYC, MIF
GO:0023014	signal transduction by protein phosphorylation	9	8.04	0.0065	BMPER, FOXO1, EDAR, RNF149, CD40, PTPN1, SOX9, MYC, MIF
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	14	12.5	0.0072	DDX17, ADRB1, RPS6KA1, GDF7, UNG, MYRF, ESR1, ABRA, FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0019220	regulation of phosphate metabolic process	15	13.39	0.0072	WASH1, GDF7, FOXO1, EDAR, CD40, SOX9, MIF, BMPER, ADRB1, CDCA2, RNF149, TNK2, PTPN1, MYC, BMP6
GO:0031399	regulation of protein modification process	15	13.39	0.0076	WASH1, GDF7, FOXO1, EDAR, CD40, SOX9, MIF, BMPER, CDCA2, RNF149, TNK2, PTPN1, MYC, DCUN1D5, BMP6

GO:0035556	intracellular signal transduction	20	17.86	0.0081	GDF7, ESR1, FOXO1, EDAR, CD40, SMYD2, SOX9, MIF, ARL11, BMPER, ADRB1, RPS6KA1, RHOT1, ABRA, RAB15, RNF149, PTPN1, MYC, GFRA2, BMP6
GO:0045893	positive regulation of transcription, DNA-templated	12	10.71	0.0091	DDX17, RPS6KA1, GDF7, MYRF, ESR1, ABRA, FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0051252	regulation of RNA metabolic process	22	19.64	0.0092	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:0010604	positive regulation of macromolecule metabolic process	20	17.86	0.0095	GDF7, UNG, ESR1, FOXO1, EDAR, CD40, SOX9, MIF, DDX17, BMPER, RPS6KA1, MYRF,

					CDCA2, ABRA, HOXB9, TNK2, PTPN1, MYC, DCUN1D5, BMP6
GO:1902680	positive regulation of RNA biosynthetic process	12	10.71	0.0096	DDX17, RPS6KA1, GDF7, MYRF, ESR1, ABRA, FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0045937	positive regulation of phosphate metabolic process	11	9.82	0.0112	ADRB1, BMPER, GDF7, CDCA2, EDAR, CD40, PTPN1, TNK2, SOX9, MIF, BMP6
GO:0010562	positive regulation of phosphorus metabolic process	11	9.82	0.0112	ADRB1, BMPER, GDF7, CDCA2, EDAR, CD40, PTPN1, TNK2, SOX9, MIF, BMP6
GO:0006351	transcription, DNA-templated	20	17.86	0.0115	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, SMYD2, CD40, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, ABRA, HOXB9, MYC, BMP6, SIM2
GO:0051254	positive regulation of RNA metabolic process	12	10.71	0.0127	DDX17, RPS6KA1, GDF7, MYRF, ESR1, ABRA,

					FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0022612	gland morphogenesis	4	3.57	0.0140	GDF7, HOXB13, EDAR, SOX9
GO:0032774	RNA biosynthetic process	22	19.64	0.0140	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:0031401	positive regulation of protein modification process	11	9.82	0.0141	BMPER, GDF7, CDCA2, EDAR, CD40, PTPN1, TNK2, SOX9, DCUN1D5, MIF, BMP6
GO:0051247	positive regulation of protein metabolic process	13	11.61	0.0145	BMPER, GDF7, CDCA2, FOXO1, EDAR, CD40, PTPN1, TNK2, SOX9, MYC, DCUN1D5, MIF, BMP6
GO:0007167	enzyme linked receptor protein signaling pathway	9	8.04	0.0161	PLAT, RGMA, BMPER, GDF7, FOXO1, PTPN1, TNK2, SOX9, BMP6
GO:0010468	regulation of gene expression	24	21.43	0.0163	SATB1, GSC, GDF7, ESR1,

					HIRA, FOXO1, HOXB13, EDAR, CD40, SMYD2, SOX9, MIF, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, SIM2, BMP6
GO:0010628	positive regulation of gene expression	13	11.61	0.0175	DDX17, RPS6KA1, GDF7, MYRF, ESR1, ABRA, FOXO1, HOXB9, EDAR, CD40, SOX9, MYC, BMP6
GO:0048732	gland development	6	5.36	0.0207	TYR, GDF7, HOXB13, HOXB9, EDAR, SOX9
GO:2000112	regulation of cellular macromolecule biosynthetic process	22	19.64	0.0208	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:0043409	negative regulation of MAPK cascade	4	3.57	0.0211	FOXO1, RNF149, PTPN1, MYC
GO:0031328	positive regulation	13	11.61	0.0218	DDX17,

	of cellular biosynthetic process				ADRB1, RPS6KA1, GDF7, MYRF, ESR1, ABRA, FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0032270	positive regulation of cellular protein metabolic process	12	10.71	0.0233	BMPER, GDF7, CDCA2, EDAR, CD40, PTPN1, TNK2, SOX9, MYC, DCUN1D5, MIF, BMP6
GO:0030509	BMP signaling pathway	4	3.57	0.0253	RGMA, BMPER, GDF7, BMP6
GO:0010556	regulation of macromolecule biosynthetic process	22	19.64	0.0272	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, RPS6KA1, MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, BMP6, SIM2
GO:0034654	nucleobase-containing compound biosynthetic process	23	20.56	0.0282	SATB1, GSC, GDF7, ESR1, HIRA, FOXO1, HOXB13, CD40, SMYD2, SOX9, RGMA, DDX17, BMPER, ADRB1, RPS6KA1,

					MYRF, BCL6B, CHD2, ABRA, HOXB9, MYC, SIM2, BMP6
GO:0035148	tube formation	4	3.57	0.0284	RGMA, GDF7, EDAR, SOX9
GO:0071772	response to BMP	4	3.57	0.0292	RGMA, BMPER, GDF7, BMP6
GO:0060627	regulation of vesicle-mediated transport	6	5.36	0.0300	SEPT5, TBC1D8, RAB15, GOSR1, PTPN1, TNK2
GO:0010557	positive regulation of macromolecule biosynthetic process	12	10.71	0.0313	DDX17, RPS6KA1, GDF7, MYRF, ESR1, ABRA, FOXO1, HOXB9, CD40, SOX9, MYC, BMP6
GO:0015031	protein transport	12	10.71	0.037	KDELR3, STX12, TBC1D8, XPO6, ABRA, EDAR, GOSR1, CD40, PTPN1, GCC2, MIF, BMP6
GO:0032268	regulation of cellular protein metabolic process	16	14.29	0.042	PLAT, WASH1, GDF7, FOXO1, EDAR, CD40, SOX9, MIF, BMPER, CDCA2, RNF149, TNK2, PTPN1, MYC, DCUN1D5, BMP6
GO:1902532	negative regulation	6	5.36	0.0451	ESR1, FOXO1,

	of intracellular signal transduction				RNF149, PTPN1, MYC, MIF
GO:0032872	regulation of stress-activated MAPK cascade	4	3.57	0.0463	FOXO1, EDAR, PTPN1, MYC
GO:0036211	protein modification process	21	18.75	0.0473	SATB1, WASH1, GDF7, FOXO1, EDAR, CD40, SMYD2, SOX9, MIF, BMPER, USP12, GALNT15, CDCA2, UCHL3, RNF149, LYPLAL1, PTPN1, TNK2, MYC, DCUN1D5, BMP6
GO:0006464	cellular protein modification process	21	18.75	0.0473	SATB1, WASH1, GDF7, FOXO1, EDAR, CD40, SMYD2, SOX9, MIF, BMPER, USP12, GALNT15, CDCA2, UCHL3, RNF149, LYPLAL1, PTPN1, TNK2, MYC, BMP6, DCUN1D5
GO:0070302	regulation of stress-activated protein kinase signaling cascade	4	3.57	0.0476	FOXO1, EDAR, PTPN1, MYC

This chapter will be published in elsewhere

as a partial fulfillment of Sohyoung Won's Master program.

Chapter 3. Genomic prediction accuracies using haplotypes defined by different methods in Hanwoo

3.1 Abstract

Genomic prediction is an effective way to measure the breeding values from genetic information based on statistical methods such as best linear unbiased prediction (BLUP). Using haplotypes, clusters of linked single nucleotide polymorphism (SNP), as markers instead of individual SNPs can improve the accuracy of genomic prediction, since the probability of a quantitative trait loci to be in strong linkage disequilibrium (LD) with markers is higher. To efficiently use haplotypes in genomic prediction, finding optimal ways to define haplotypes is needed.

In this study, 770K SNP chip data was collected from Hanwoo (Korean cattle) population consisted of 3498 cattle. Haplotypes were first defined in three different ways using 770K SNP chip data: haplotypes were defined based on 1) length of haplotypes (bp), 2) the number of SNPs included, and 3) k-medoids clustering based on LD. To compare the methods in parallel, haplotypes defined by all methods were set to have comparable sizes; in each method, haplotypes defined to have an average number of 5, 10, 20 or 50 SNPs were tested respectively. A modified genomic BLUP (GBLUP) method using haplotype alleles as explanatory variables was implemented for testing the prediction accuracy of each haplotype set. Also, GBLUP using individual SNPs were tested to evaluate the performance of the haplotype sets on genomic prediction. Carcass weight was used as the phenotype for testing.

As a result, using haplotypes defined by all three methods showed increased accuracy compared to GBLUP using individual SNPs. The prediction accuracy was highest when the average number of SNPs per haplotype was 20 in all three methods, implying that haplotypes including around 20 SNPs can be optimal to use as markers for genomic prediction. When the number of alleles generated by each haplotype defining methods was compared, clustering by LD generated the least number of alleles. This suggests that defining haplotypes based on LD can reduce computational costs and allows efficient prediction. Finding optimal ways to define haplotypes and using the haplotype alleles as markers can provide improved performance and efficiency in genomic prediction.

3.2 Introduction

Genomic prediction is an effective way to measure the abilities of livestock for breeding based on their genetic information. In practical, the genomic estimated breeding values (GEBV) of animals is calculated by using their single nucleotide polymorphism (SNP) chip genotype data and statistical prediction methods such as the best linear unbiased prediction (BLUP) or Bayesian methods. The accuracies of these methods generally rely on degree of linkage disequilibrium (LD) between the SNP markers and real quantitative trait loci (QTL) (Goddard 2009). Here, linkage disequilibrium is the nonrandom association between different loci in a certain population, which can be calculated by measuring the frequencies of alleles and the haplotype frequencies of the pair of alleles at the loci (Slatkin 2008).

By using clusters of related SNPs as markers instead of individual SNPs, the probability that a QTL is in strong LD with a marker becomes higher (Goddard and Hayes 2007). Thus, the accuracy of genomic prediction can be improved by using clusters of SNPs, which are referred to as haplotypes. To efficiently use haplotypes in genomic predictions, many studies have focused on finding optimal ways to define a cluster of SNPs as a haplotype. The simplest ways proposed were considering segments of equal sizes in the genome as haplotypes (Ferdosi, Henshall et al. 2016), (Hess, Druet et al. 2017), (Sun, Fernando et al. 2015), (Villumsen, Janss et

al. 2009). Here, size can be defined as the physical length in basepairs (Ferdosi, Henshall et al. 2016), (Hess, Druet et al. 2017), or the length in centimorgans (Sun, Fernando et al. 2015), or the number of SNPs in one haplotype (Villumsen, Janss et al. 2009). Along with, methods combining information about identity by descent (IBD) with clusters of adjacent SNPs to define haplotypes (Calus, De Roos et al. 2008), (Calus, Meuwissen et al. 2009), and using predicted genealogy to define haplotypes (Edriss, Fernando et al. 2013) were studied. Also, setting minimum pairwise LD cutoffs to group SNPs into haplotypes was considered (Cuyabano, Su et al. 2014).

Some of the methods to define haplotypes for genomic prediction attempts to incorporate the LD structure of the genome (Calus, De Roos et al. 2008), (Cuyabano, Su et al. 2014). An advantage of defining haplotypes based on LD is that the number of haplotypes alleles, which is the number of explanatory variable used for computation, can be reduced compared to other methods (Cuyabano, Su et al. 2014). Recently, to more precisely represent the LD structure while defining haplotypes, some clustering methods originated in the data mining field have been applied (Dehman 2015). One of them is hierarchical clustering, which produces a tree that has nodes representing clusters in a hierarchical order from, where each element being each cluster is the leaf the all the elements being one cluster is the root. Applying hierarchical clustering to make SNP clusters based on LD was

implemented by Alia Dehman in 2015 (Dehman 2015). Another popular clustering method is partitioning clustering, which splits the data into a given number of clusters. One of the most known methods of partitioning clustering is the k-medoids clustering. k-medoids clustering make clusters so the distance between the data in a cluster and the center of the cluster is minimized.

In this study, k-medoids clustering was used to construct haplotypes based on LD from phased genotypes of 770K SNP chips. In addition, haplotypes were alternatively defined as segments with given sizes. The length of a haplotype in basepairs and the number of SNPs within a haplotype were respectively used as criteria of sizes. The genomic prediction accuracies using haplotypes defined based on 1) length of haplotypes (bp), 2) the number of SNPs, and 3) LD clustering by k-medoids clustering in Hanwoo population were tested and compared with the accuracy of using individual SNPs to find out whether these methods can bring improvement in genomic prediction. Also, to find out the optimal size of haplotypes, various sizes of haplotypes defined by each method were tested. To compare the methods in parallel, haplotypes defined by all methods were set to have comparable sizes.

3.3 Materials and method

Genotypic and phenotypic data

The genotypic and phenotypic data used in this study were collected from the Hanwoo (Korean cattle) population consisted of 3498 cattle. Among them, samples with their sex and slaughter age available were used for the study. The carcass weight of the samples was measured after slaughter and was used as the phenotypic value of the genomic prediction model. Genomic DNA was extracted from blood samples. Genotyping was performed by using Illumina BovineHD Genotyping BeadChip in 1166 samples and Illumina BovineSNP50 Genotyping BeadChip in 2332 samples. 50K genotype data was imputated to 770K by Minimac3.

Total 732225 SNPs were genotyped and used after quality control. SNPs having low minor allele frequency (<0.01), low genotyping rate (<0.95), significant deviation from Hardy-Weinberg equilibrium ($p<0.001$) were discarded and only one SNP was left if multiple SNPs were on the same site. Individuals with low genotype call rate (<0.95) were excluded from the study. Two-sided Grubb's test with alpha=0.05 was performed to check whether there were outliers in phenotypic data and was not significant ($p=0.07$).

Consequently, 555678 SNPs and 2506 individuals including 831 males and 1675 females remained to be used for the study. The total genotyping rate of 0.9971. Genotypes were phased using SHAPEIT2 with

200 states and window size of 0.5Mb for haplotyping.

Defining haplotypes

Three methods to define haplotypes were considered respectively in this study. First, segmentations of the genome with equal sizes in basepairs were regarded as haplotypes (method 1). Second, segments of the genome containing constant number of SNPs were treated as haplotypes (method 2). Third, k-medoids clustering based on LD was used to construct haplotypes (method 3). In the three methods, the start points and end points of haplotypes were designated accordingly and the SNPs within the point formed haplotypes.

In each method, the sizes of haplotypes were set variously to find out the optimal size of haplotypes for accurate genomic prediction. To compare the three methods in a parallel way, the average number of SNPs per block were balanced to be approximately 5, 10, 20, or 50. In brief, three haplotype defining methods with four average size criteria, making twelve kinds of haplotype were tested. The lengths of haplotypes in method 1 was calculated by the total number of SNPs and the total length of the genome. In method 3, the number of clusters, k was set as the total number of SNPs divided by 5, 10, 20, or 50. The lengths of haplotypes in method 1 and numbers of clusters in method 3 are later shown in Table 3-1.

k-medoids clustering based on LD

In k-medoids clustering based on LD of SNPs were calculated and the pairwise LD measured as D' was set as the proximity measure of two SNPs. Here, D' is calculated as the following equation.

$$D_{AB} = p_{AB} - p_A p_B$$

$$D_{max} = \begin{cases} \max(-p_A p_B, -(1-p_A)(1-p_B)) & when\ \ D < 0 \\ \min(p_A(1-p_B), (1-p_A)p_B) & when\ \ D > 0 \end{cases}$$

$$D' = D_{AB}/D_{max}$$

In other words, (1-D') was defined as the distance of two SNPs. k-medoids clustering includes two parts; finding the center SNPs of clusters and assigning SNPs to the cluster with the closest center. However, since there were cases that the pairwise LD doesn't match the physical distance of SNPS, by naively assigning SNPs to clusters of the closest center it was impossible to make non-overlapping and mutually exclusive clusters.

To make non-overlapping and linear clusters using all the SNPs for haplotype defining, we instead set boundaries of clusters and regarded all the SNPs within the boundary to be a cluster. The boundaries of clusters were defined as where they could minimize the sum of distances between the SNPs in each cluster and the center. A center of a cluster was defined as the SNP which minimize the sum of distances between itself and other SNPs in the cluster. First, SNPs of a given number as the number of clusters were randomly chose to be center SNPs. Then an iterative process of finding the boundaries of clusters and finding the centers of clusters was proceeded.

The process was repeated while the total sum of distances was decreasing.

Halotyping

After defining the start points and end points of haplotypes throughout the genome, the phased genotype was re-coded according to the haplotype alleles from the haplotype definition. All present alleles of the haplotypes were found from the phased genotype and the data for each allele was coded into 0, 1, 2 to represent the number of that allele each sample carried. R package 'GHap' was used for this procedure (Utsunomiya, Milanesi et al. 2016).

Genomic prediction

The BLUP model was used to perform genomic predictions using the haplotype markers defined in the previous stage. The BLUP model was described as:

$$y = Xb + Zu + c,$$

where y is the vector of CWT of the bull, b is vector of fixed effects including sex and slaughter age, u is the vector of additive genetic effects, and c is the vector of residual errors. X is the design matrix for fixed effects, Z is the design matrix associating haplotype alleles and effects to appropriate observations, which is N×H where N is the number of animals and H is the total number of haplotype alleles. The additive genetic effects g

and residual errors ϵ were estimated as random effects assuming that they follow the distributions bellow:

$$\mathbf{u} \sim \mathrm{N}\left(0, \mathrm{G}\sigma_g^2\right)$$

$$\mathrm{e} \sim \mathrm{N}(0, \mathrm{R}\sigma_e^2)$$

The BLUP solution for the model $\hat{\mathrm{a}}$ was computed using the equation $\hat{\mathrm{a}} = \mathrm{qDM'K^{-1}}\hat{g}$, where M and g are as in the model, q is the inverse weighted sum of variances in the columns of M, $\mathrm{D} = \mathrm{diag}(\mathrm{d}_i)$, d_i is the weight of each haplotype allele, and K is the haplotype-based kinship matrix with R package 'GHap' (Utsunomiya, Milanesi et al. 2016). Then, the GEBVs were obtained as the following equation:

$$\mathrm{GEBV}_i = \sum_j m_{ij}\hat{a}_j$$

Finally, the performances different haplotype defining methods were compared based on the accuracy of the models, which was calculated as the correlation of the GEBVs and EBVs. 5-fold cross validation was used to obtain the accuracies of different methods. To compare with the conventional GBLUP method, the same model using single SNPs as haplotypes was also tested.

3.4 Results

Haplotype construction

The statistics of haplotypes constructed by different haplotype defining methods and different average SNP number criteria of each method are presented in Table 3-1. The actual average numbers of SNPs per haplotype were also obtained to check whether the haplotypes were constructed with intended sizes. The average numbers of SNPs were consistent with the intended numbers in LD clustering-based haplotypes and length-based haplotypes with sizes of 44.5kb, 89kb and 222.5kb, while larger than intended in length-based haplotypes of 22.25kb.

The total number of haplotype alleles were computed to compare the number of explanatory variables used for genomic prediction (Figure 3-1). The number of alleles increased as the average number of SNPs per haplotype increased. However, the numbers of alleles from haplotypes of similar sizes were least when LD clustering was used to define haplotypes. In addition, the differences of haplotype alleles from other methods and LD clustering increased as the average number of SNPs per haplotype increased. The average number of alleles per haplotypes showed similar tendencies with total number of alleles.

Genomic prediction accuracy

The accuracy of genomic prediction using haplotypes was higher

Table 3-1. Haplotype and allele statistics of each haplotype defining method at different sizes. K is the number of clusters and N is the number of total SNPs.

SNP count-based haplotypes	**5 SNPs**	**10 SNPs**	**20 SNPs**	**50 SNPs**
Number of SNP markers	555678	555678	555678	555678
Number of haplotype alleles	1303861	1877160	2713296	3710659
Number of haplotypes	111123	55554	27768	11099
Average number of SNPs per haplotypes	5.000567	10.00248	20.01145	50.06559
Average number of alleles per haplotypes	11.73349	33.78983	97.71305	334.3237
Length-based haplotypes	**22.25kb**	**44.5kb**	**89kb**	**222.5kb**
Number of SNP markers	555678	555678	555678	555678
Number of haplotype allele markers	1364861	1867261	2621574	3581059
Number of haplotypes	97061	54163	27797	11196
Average number of SNPs per haplotypes	5.725038	10.25936	19.99057	49.63183
Average number of alleles per haplotypes	14.06188	34.47484	94.31140	319.8516
LD clustering-based haplotypes	**K=N/5**	**K=N/10**	**K=N/20**	**K=N/50**
Number of SNP markers	555678	555678	555678	555678
Number of haplotype alleles	1277525	1764074	2472637	3358562
Number of haplotypes	111123	55554	27768	11099
Average number of SNPs per haplotypes	5.000567	10.00248	20.01145	50.06559
Average number of alleles per haplotypes	11.49649	31.75422	89.04628	302.6004

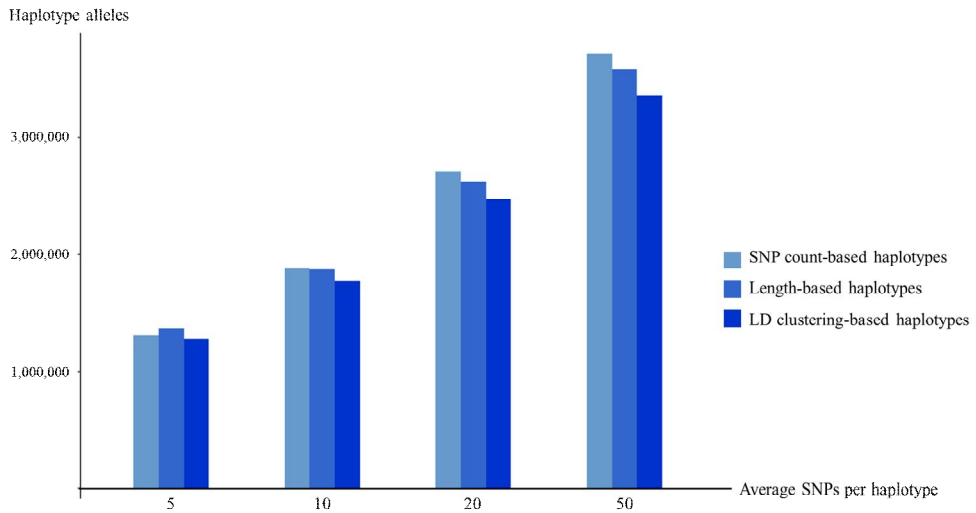

Figure 3-1. Number of haplotype alleles generated by different haplotype defining methods and sizes. Position on the horizontal axis indicates haplotype sizes as the number of average SNPs included and different colors indicate haplotype defining methods.

compared to GBLUP using individual SNPs in all haplotype defining methods (Figure 3-2, Table 3-2). By using haplotype alleles, genomic prediction accuracy increased at least 0.47%. SNP count-based haplotypes with an average of 20 SNPs yielded the highest accuracy, 0.5959, which is 1.32% higher than using individual SNPs. Different haplotype defining methods performed best depending on the size of haplotypes. When haplotypes contained average 5 SNPs, length-based haplotype performed best, probably because it actually contacting more than 5 SNPs. LD clustering-based haplotypes had the highest accuracies at average 10 SNPs and SNP count-based haplotypes at average 20 and 50 SNPs. From the view of haplotype size, containing average 20 SNPs showed highest accuracies in all haplotype defining methods.

Paired t-tests were performed in order to test whether the increases in prediction accuracies by using haplotypes compared to using individual SNPs were statistically significant. Tests were respectively performed for different haplotype defining methods with different sizes. As a result, all of the tests except length-based haplotypes of average 50 SNPs per haplotype were significant at significance level 0.05, suggesting that using haplotypes defined by any of the three methods for all sizes bring about a statistically meaningful increase in prediction accuracy generally (Table 3-3). Differences in accuracies were significant at significance level 0.01 at haplotypes having average 5 SNPs of all methods and all sizes of haplotypes

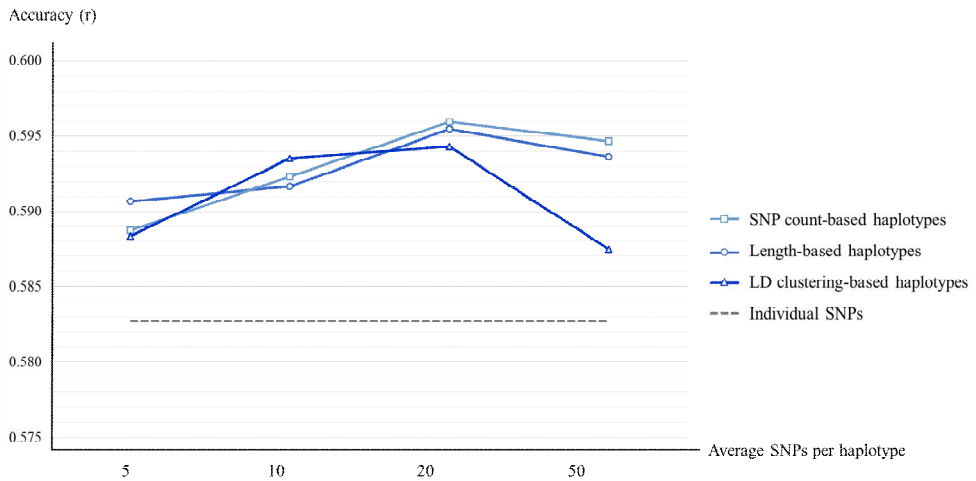

Figure 3-2. Genomic prediction accuracies of using various sizes of haplotypes defined by different methods compared with accuracy using individual SNPs. Straight lines of different colors indicate accuracies from different haplotype defining methods and the dashed line shows the accuracy from individual SNPs. Position on the horizontal axis indicates haplotype sizes as the number of average SNPs included. Accuracies were calculated as the correlation coefficients of GEBVs and true phenotypes.

Table 3-2. Genomic prediction accuracies of using various sizes of haplotypes defined by different methods and accuracy using individual SNPs.

	Average number of SNPs per haplotype				
	1	**5**	**10**	**20**	**50**
SNP count-based haplotypes		0.588737	0.592317	0.595922	0.594643
Length-based haplotypes		0.590697	0.591648	0.592344	0.593607
LD clustering-based haplotypes		0.588335	0.593527	0.594325	0.587476
Individual SNPs	0.582764				

Table 3-3. P-values of paired t-tests comparing prediction accuracies using individual SNPs and haplotypes defined by different methods and sizes. *, **, and *** indicates significant at α=0.05, 0.01, 0.0001 respectively.

	Average number of SNPs per haplotype			
	5	**10**	**20**	**50**
SNP count-based haplotypes	0.007471**	0.041925*	0.036644*	0.041693*
Length-based haplotypes	0.004286**	0.002871**	0.004691**	0.066436
LD clustering-based haplotypes	0.005356**	0.025304*	0.007696***	0.001526**

defined using LD clustering.

3.5 Discussion

Genomic prediction accuracy using haplotypes designed in this study was always higher than using individual SNPs and mostly statistically significant. The increased accuracy by using haplotypes may be due to higher LD between alleles and QTLs, better detection of ancestral relationships (identity-by-descent), and capturing of short range epistatic effects (Hess, Druet et al. 2017). Haplotyping and constructing genomic prediction models using haplotype alleles can improve prediction accuracy without any additional cost for data production though it may cause some more computational cost.

In all haplotype defining methods, prediction accuracy was highest when haplotypes of average 20 SNPs were used. Average 20 SNP, which is approximately 89kb, appears to be the optimal haplotype size from the results of this study. The optimal size to define haplotypes for genomic prediction depends on the distance between SNPs and the LD structure of the population (Calus, Meuwissen et al. 2009). The mean distance between SNPs was 4118.24bp and the mean LD (r^2) was 0.43 in the Hanwoo population used for the study. As high density of SNPs was used and the LD between SNPs are high, the optimal size of haplotypes tend to be larger than other studies, where optimal numbers of SNPs per haplotype were 4~10 (Hess, Druet et al. 2017) (Calus, Meuwissen et al. 2009) (Villumsen and Janss 2009).

The number of haplotype alleles indicates the number of explanatory variables used for genomic prediction. As the number of explanatory variables increases, the dimension of the design matrix in equation 1 becomes larger and it takes more time and memory to solve the mixed model equation. Thereby, reducing the number of haplotype alleles enable more efficient calculation of GEBVs. In this study, two methods are possible to reduce the number of haplotype alleles. The first is using LD clustering to define haplotypes and the second is using smaller sizes of haplotypes. To obtain best prediction accuracy, using SNP count-based haplotypes with 20 SNPs is optimal. Considering both accuracy and computational cost, LD clustering-based haplotypes with average 10 SNPs seem reasonable.

The estimation of GEBV from haplotype alleles depends on the phasing results from genotypes. Errors from phasing may produce allele which is not actually present. Especially in haplotypes defined by LD clustering, inaccurate phasing may cause haplotype boundaries to be differently defined resulting in lower accuracy. Therefore, finding more accurate phasing methods can further improve the prediction accuracy by using haplotypes. In addition, discarding haplotype alleles of low frequencies can be considered, since generation of alleles having extremely low frequency (e.g. only one in the population) can be a cause of overfitting, potentially lowering the prediction accuracy. Also, this and reduce the

computational cost by lessening explanatory variables.

This chapter will be published in elsewhere

as a partial fulfillment of Sohyoung Won's Master program.

Chapter 4. A height prediction model using selected genetic markers and parental heights in Korean

4.1 Abstract

Human height is a polygenic trait with high heritability, which can be estimated from genetic markers with high accuracy. Since the Genomic Best Linear Unbiased Prediction (GBLUP) is an efficient method to predict breeding values or phenotypes from genotype data, was applied for the prediction of adult height of 490 Koreans. In addition, a GBLUP model adjusted with mid-parental height was fitted for height prediction, in which GBLUP was employed to predict the residuals from the mid-parental height model. Genetic markers explaining largest parts of the residuals were selected from bootstrap resampling and linear mixed models. Then, models using different numbers of selected markers were tested. As a result, the prediction accuracy of the GBLUP model adjusted with mid-parental height was higher than both the mid-parental model and the GBLUP predicting raw height. Also, the predictive performance was improved when selected markers were used for the model. A model using 10,000 SNPs showed the highest prediction accuracy, 0.9330. The approach of this study can generally be applied to improve the accuracy of genomic prediction in other complex traits or other species, by fitting GBLUP models adjusted with the phenotypes of parents and sex.

4.2 Introduction

Human height is a polygenic trait, which has been studied well as a model trait for studying the genetic background of complex traits. The narrow-sense heritability of height is about 0.8 (Fisher 1919) (Silventoinen, Sammalisto et al. 2003) (Visscher, Medland et al. 2006), suggesting that about 80% of the variation in adult height is accounted for additive genetic effects. As a substantial part of human height is explained by genetic effects, it can be estimated using genetic markers accurately.

A classical model for human height estimation is the mid-parental model designed by Tanner, defining the target height of a person as the average height of the parents plus or minus 6.5 according to sex (Tanner 1986). In addition, as single nucleotide polymorphisms (SNPs) account for a large portion of the variance in height (Yang, Benyamin et al. 2010), height can be precisely predicted from SNP genotype data of individuals. A method to predict phenotypes such as height from genotypes is using the genomic best linear unbiased prediction (GBLUP). In GBLUP, a genomic relationship matrix representing the relatedness of individuals is constructed and utilized to estimate phenotypes.

As GBLUP showed high predictive performances in other studies (Meuwissen, Hayes et al. 2016), it was applied to predict the adult height of Koreans. Also, a model combining the mid-parental model and genomic prediction was designed to exploit the advantages of both models. In

addition, SNPs best explaining the residuals from the mid-parental models were selected and used for prediction in the combined model. SNP sets of various sizes were tested to find the most efficient model.

4.3 Materials and Method

Data preparation

The genotypic and phenotypic data for this study were collected from Korean Association Resource (KARE) project. The Institutional Review Board of the Korea National Institute of Health approved this study. Data from a cohort consisted of 1,188 parents and 615 offspring was obtained. Among the samples, 492 individuals with their parental height available were kept. Genomic DNA was extracted using the Affymetrix Genome-Wide Human SNP array 6.0, and total 516610 autosomal SNPs were genotyped. For quality control, SNPs with low genotyping rate (<0.95), low minor allele frequency (<0.1), significant deviation from Hardy-Weinberg equilibrium ($p<0.0001$) were excluded and samples with low genotyping rate (<0.9) were removed. Finally, 368,813 SNPs of 490 individuals were used for the study.

Prediction models

Three prediction models were fitted and their prediction accuracies were compared. The first model is Tanner's mid-parental model,

$$Height_{Mid} = \frac{Height_M + Height_F + 13(2k-1)}{2} \quad (1)$$

, where $Height_M$ and $Height_F$ are the heights of the mother and father respectively, and k is 1 when male and 0 when female. The second model is the GBLUP model predicting raw heights.

$$Height_{raw} = Zu + X\beta + e, \quad u \sim N(0, G\sigma_g^2) \text{ and } e \sim N(0, I\sigma_e^2) \quad (2)$$

, where u is the additive genetic effect, β is the fixed effect which was sex, e is random error. Z and X are design matrices, G is the genomic relationship matrix, and I is the identity matrix. The third model is a GBLUP model adjusted with mid-parental height. In this model, first the residual heights were calculated from the mid-parental model.

$$Height_{res} = Height_{raw} - Height_{Mid} \quad (3)$$

, where $Height_{res}$ is the residual heights, $Height_{raw}$ is the raw height, and $Height_{Mid}$ is the height estimated from equation (1). Then, a GBLUP model predicting the residuals was fitted.

$$Height_{res} = Z'u' + e', \quad u' \sim N\left(0, G\sigma_{g'}^2\right) \text{ and } e' \sim N(0, I\sigma_{e'}^2) \quad (4)$$

, where u' is the additive genetic effect of residuals from the mid-parental model, e' is random error, Z' is a design matrix, G is the genomic relationship matrix, and I is the identity matrix. Finally, the predicted height was calculated as the sum of $Height_{res}$ predicted from equation (4) and $Height_{Mid}$ calculated in equation (1).

$$Height_{final} = Height_{res} + Height_{Mid} \quad (5)$$

LDAK software was used for the GBLUP models (Speed, Hemani et al. 2012).

Prediction accuracy measure

The prediction accuracy was measured as the correlation of

predicted heights and observed heights. In GBLUP, the effects of the SNP markers were estimated from the model and the phenotype is predicted as a linear combination of the genotypes and marker effects.

$$phenotype_i = \sum_j m_{ij}\hat{a}_j$$

, where $phenotype_i$ is the phenotype of the ith individual, m_{ij} is the genotype of the jth SNP of the ith individual, and $\hat{a}_j$ is the estimated effect of the jth SNP.

Ten-fold cross validation was used to measure the prediction abilities of the models. The data was randomly divided into 10 separate folds, then one of the folds was assigned as a test set and the remaining 9 folds were assigned as a training set. The effects of SNPs were estimated from the training set. Also, in the model using selected SNPs, SNPs were selected according to the results from the training set, then the effects of the selected SNPs were estimated again from the training set. The predicted height was calculated using the estimated effects from the training set and the genotypes of the test set. The prediction accuracy was measured as the correlation of predicted heights and observed heights in the test set. This was repeated 10 times and the final prediction accuracy was reported as the mean of the correlation coefficients. The square of the correlation coefficient was also calculated to see the proportion of variance explained from the model. In addition, the slope of the linear model regressing true values from predicted values was measured to evaluate whether the

predictions are inflated or depressed.

SNP selection

In the GBLUP model adjusted with mid-parental height, models using only selected SNPs were also tested. SNPs that explain the genetic effects of the residual heights well were extracted to be used in the model. For SNP selection, bootstrap resampling was applied. A subset containing half of the training set was randomly sampled. The effects of SNPs on residual heights were estimated as random effects using GCTA (Yang, Lee et al. 2011) from the subset. This was repeated for 100 times and the mean effects of SNPs were calculated. Then, SNPs were sorted according to the absolute value of the mean effects obtained. SNPs with the highest absolute mean effect were selected and used for the GBLUP model predicting residual heights. SNP sets consisted of 300000, 200000, 100000, 50000, 20000, 10000, 5000, 2000, 1000, 500, 200, 100 SNPs were extracted and tested.

4.4 Results and Discussion

Genomic prediction performances of three models

The prediction accuracies of the mid-parental model, GBLUP model, and GBLUP model adjusted with mid-parental height using all SNPs after quality control were 0.8287, 0.8091, 0.8399 respectively (Figure 4-1) and the proportions of phenotypic variance explained from the models were 0.6867, 0.6547, 0.7054 respectively (Figure 4-2). The slope of regression of the models in the same order were 0.9339, 1.0148, 0.9397 respectively (Figure 4-3). It is ideal if the slope of regression is 1, suggesting no inflation nor depression in the prediction results. The differences of the slopes from 1 were 0.0661 in the mid-parental model, 0.0148 in the GBLUP model, and 0.0603 in the GBLUP model adjusted with mid-parental height.

The prediction accuracy and proportion of variance explained from the mid-parental model was fairly high. However, this model has limitations since it cannot account for the differences in height among siblings. The prediction accuracy of GBLUP was the lowest, however the slope of regression was closest to 1. The GBLUP model adjusted with mid-parental height showed the best performance among the three models. The gain in prediction accuracy from additionally using genotypes compared to the mid-parental model was 0.0102 and the gain in prediction accuracy from adjusting parental height in the GBLUP model was 0.0308. Using both parental height and genetic information from genotype can explain

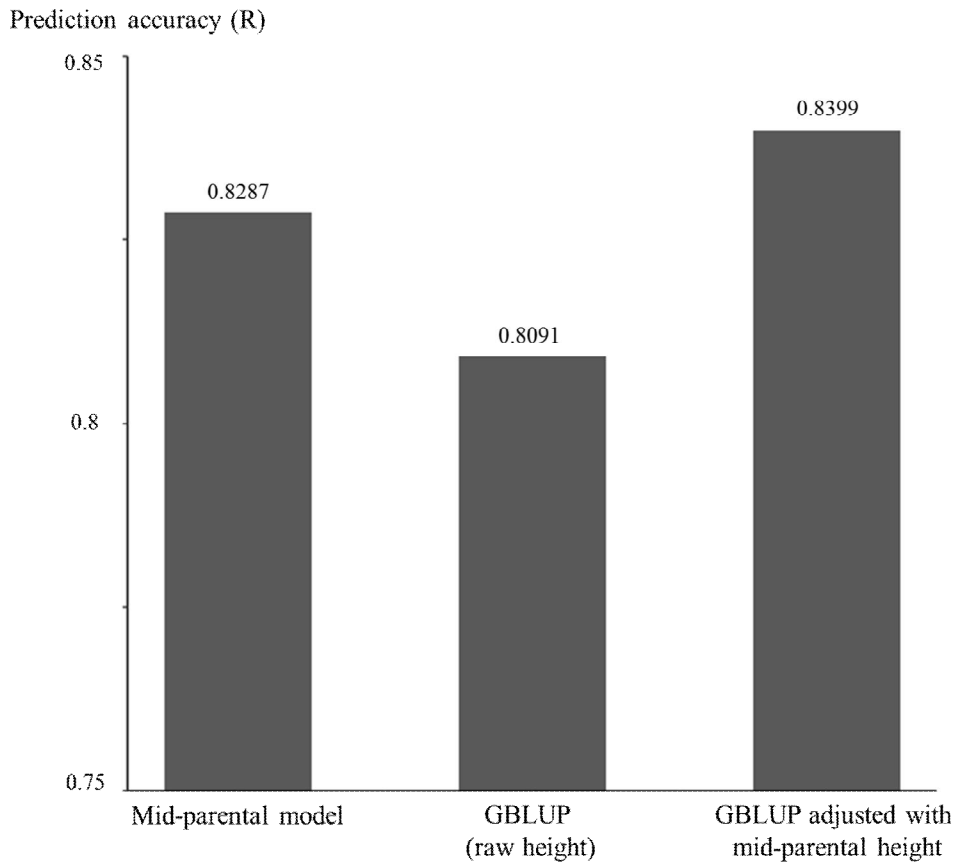

Figure 4-1. Prediction accuracies as the correlation coefficients of three models; the mid-parental model, the GBLUP model, and the GBLUP model adjusted with mid-parental height.

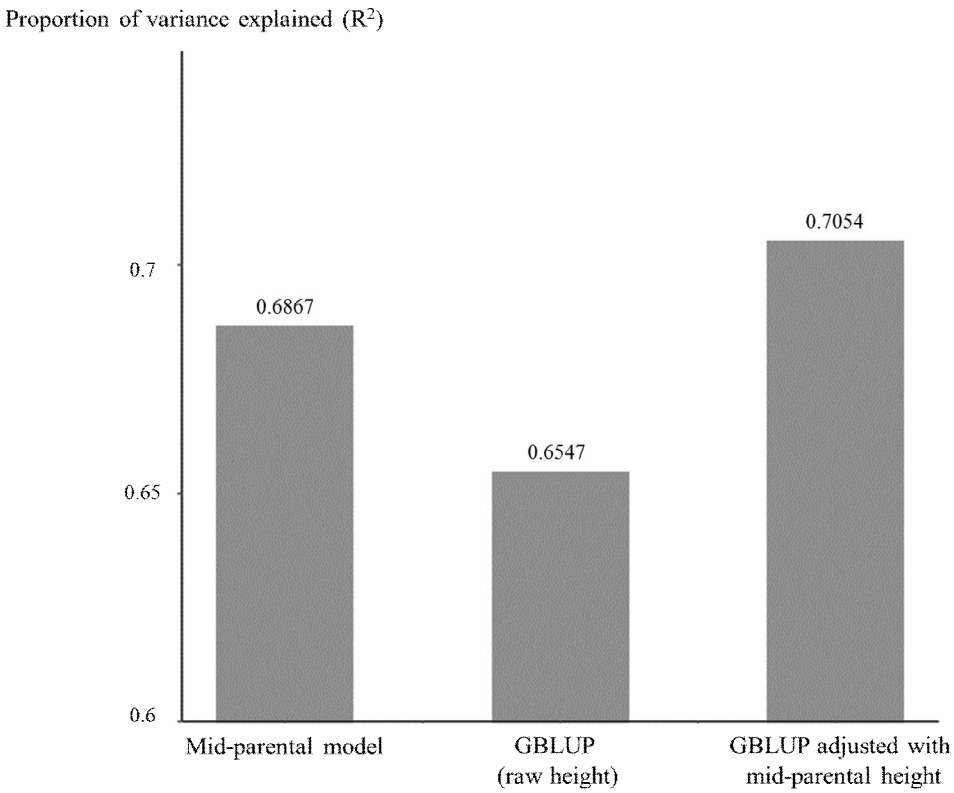

Figure 4-2. Proportions of phenotypic variance explained from the models as the squares of the correlation coefficients of three models; the mid-parental model, the GBLUP model, and the GBLUP model adjusted with mid-parental height.

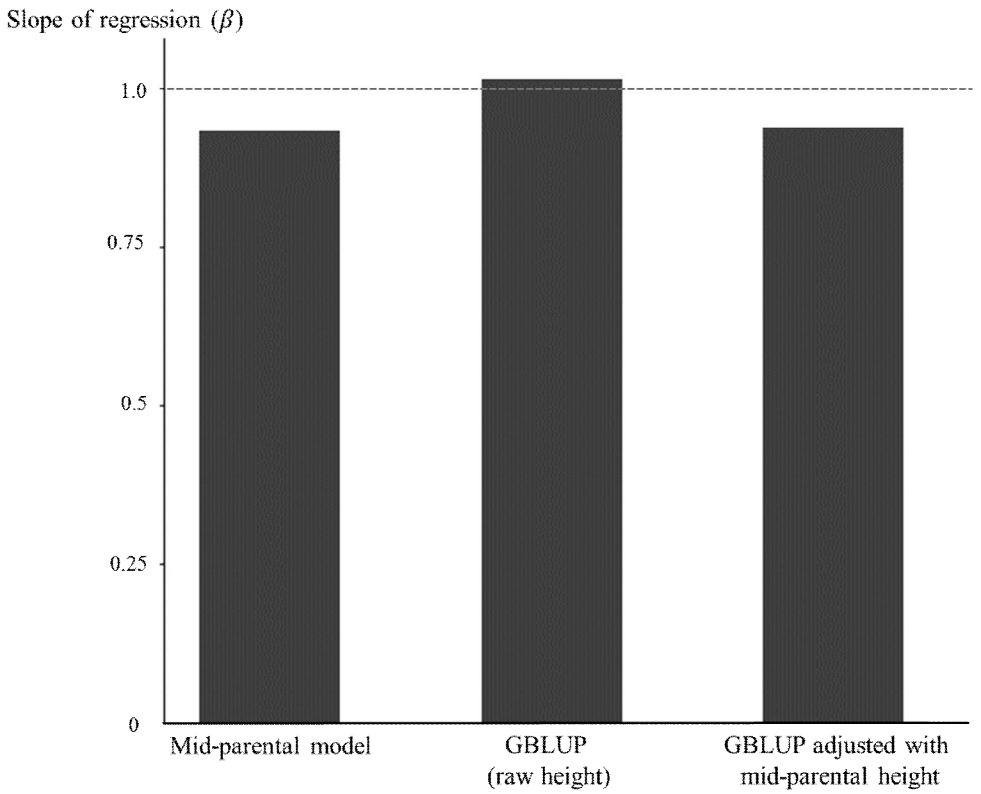

Figure 4-3. Slopes of regressions from linear models fitting true heights from predicted heights of three models; the mid-parental model, the GBLUP model, and the GBLUP model adjusted with mid-parental height.

approximately 70% of the total variance in height.

The mid-parental height may explain some part of the genetic effects in child height and also some part of the environmental effects since parents and children usually share common environmental factors. This may be the reason why the prediction accuracy was higher in the mid-parental model than the GBLUP model, even though GBLUP model better explains the genetic variances. In the GBLUP model adjusted with mid-parental height, some part of the environmental variance and some part of the genetic variance are explained from the parental heights, and GBLUP further accounts for the genetic variances explaining the genetic variances of the residuals.

Genomic prediction performances of GBLUP model adjusted with mid-parental height using selected SNPs

The predictive performance improved when only selected SNPs were used for the GBLUP model adjusted with mid-parental height. For all sizes of the selected SNP sets, from 300,000 to 100, the prediction accuracies and the proportions of variance explained were higher than when all SNPs were used. The prediction accuracy and the proportion of variance explained was highest, 0.9330 and 0.8705 respectively, when 10,000 SNPs were used for prediction (Figure 4-4A, 4-4B). The prediction accuracy of using 10,000 selected SNPs was 0.0931 higher than using all SNPs and the

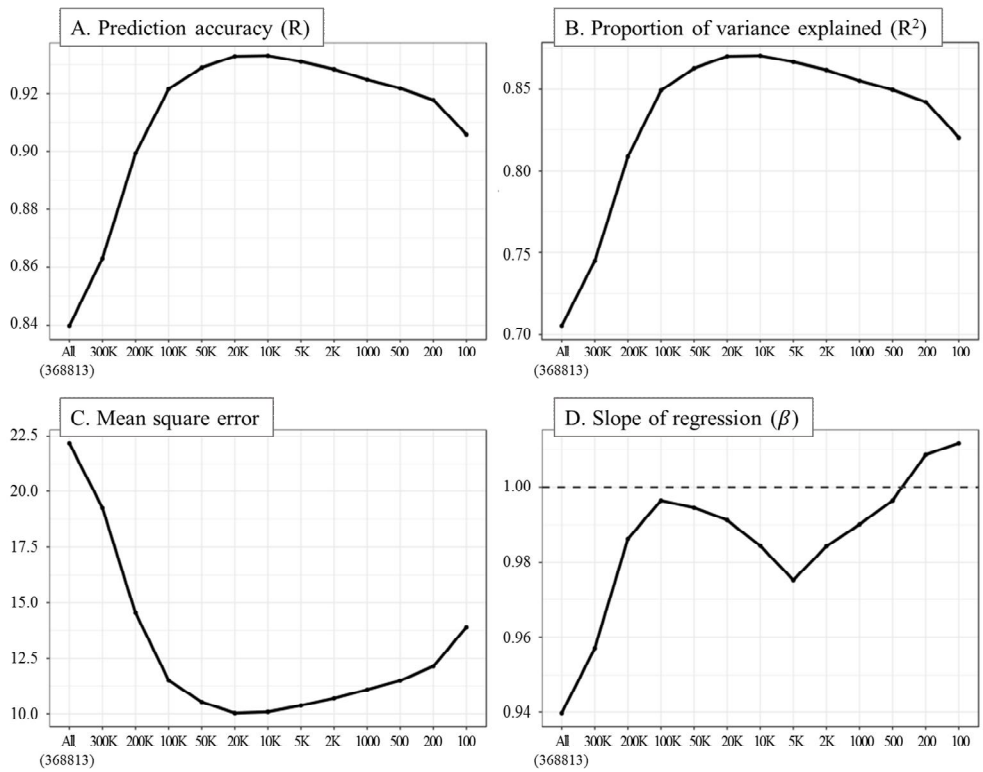

Figure 4-4. Predictive performances using different numbers of selected SNPs in the GBLUP model adjusted with mid-parental height (K=1000). A: prediction accuracies measured as the correlation coefficient (R), B: proportions of phenotypic variance explained from the model measured as R^2, C: mean square errors, D: slopes of regressions from linear models fitting true heights from predicted heights.

variance explained increased for 16.5%. The results from using 20,000 SNPs were as good as using 10,000 SNPs. The prediction accuracy was 0.9328 and the proportion of variance explained was 0.8702 showing little difference, while the mean square error, measuring both the bias and variance of a model, was lower (figure 4-4C). The predictive performances improved as smaller numbers of SNPs were used until 20,000 or 10,000, and decreased when less SNPs were used. Still, the prediction accuracies were above 0.9, even when only 100 SNPs were used. This indicates that height can be accurately predicted from the GBLUP adjusted with mid-parental height from low density of SNPs if the SNPs are properly selected.

The slope of regression was closest to 1 when 100,000 SNPs or 500 SNPs were used, which were both 0.0035. The differences of slope of regression and 1 were always smaller for using selected SNPs compared to using all SNPs. Among the models of highest prediction accuracies, using 10,000 SNPs and 20,000 SNPs, using 20,000 SNPs was better in the point of slope of regression (figure 4-4D).

Since mid-parental height and genotypes both accounts for genetic variances in height, there might be multicollinearity, where some of the SNPs are highly correlated with parental height. Multicollinearity of variables cause the unstableness in prediction and also overfitting since redundant variables are used in the model (Farrar and Glauber 1967). Therefore, if using large number of SNPs in the GBLUP model adjusted

with mid-parental height resulted in multicollinearity, it could be the reason why using all SNPs showed lower predictive performances than using selected SNPs. As SNPs that best explains the residuals of the mid-parental model instead of the raw height were selected, the selected SNPs and mid-parental height may have less correlation compared to other SNPs. The approach of this study can generally be applied to other complex traits or other species, by fitting GBLUP models adjusted with the phenotypes of parents and sex in which selected markers best explaining the residuals are used.

Reference

Ahn, J. and J. Lee (2008). "X chromosome: X inactivation." Nature Education **1**(1): 24.

Aouadi, M., et al. (2007). "p38MAP Kinase activity is required for human primary adipocyte differentiation." FEBS letters **581**(29): 5591-5596.

Bächner, D., et al. (1998). "Bmp-2 downstream targets in mesenchymal development identified by subtractive cloning from recombinant mesenchymal progenitors (C3H10T½)." Developmental dynamics: an official publication of the American Association of Anatomists **213**(4): 398-411.

Balding, D. J. J. N. R. G. (2006). "A tutorial on statistical methods for population association studies." **7**(10): 781.

Calus, M., et al. (2008). "Accuracy of genomic selection using different methods to define haplotypes." Genetics **178**(1): 553-561.

Calus, M. P., et al. (2009). "Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values." Genetics Selection Evolution **41**(1): 11.

Chen, J., et al. (2014). "Distribution of H-FABP and ACSL4 gene polymorphisms and their associations with intramuscular fat content and backfat thickness in different pig populations." Genet. Mol. Res **13**(3): 6759-6772.

Christensen, O. F. and M. S. Lund (2010). "Genomic prediction when some animals are not genotyped." Genetics Selection Evolution **42**(1): 2.

Cohen, J. C., et al. (2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." **305**(5685): 869-872.

Cuyabano, B. C., et al. (2014). "Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population." BMC genomics

15(1): 1171.

Daszkiewicz, T., et al. (2005). "Quality of pork with a different intramuscular fat (IMF) content." Polish Journal of Food and Nutrition Sciences **14**(1): 31-35.

de Koning, D. J., et al. (1999). "Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (Sus scrofa)." Genetics **152**(4): 1679-1690.

Dehman, A. (2015). Spatial clustering of linkage disequilibrium blocks for genome-wide association studies, Université d'Evry Val d'Essonne; Université Paris-Saclay; Laboratoire de ….

Edriss, V., et al. (2013). "The effect of using genealogy-based haplotypes for genomic prediction." Genetics Selection Evolution **45**(1): 5.

Falconer, D. S. (1960). Introduction to quantitative genetics, Oliver And Boyd; Edinburgh; London.

Farrar, D. E. and R. R. Glauber (1967). "Multicollinearity in regression analysis: the problem revisited." The Review of Economic Statistics: 92-107.

Ferdosi, M. H., et al. (2016). "Study of the optimum haplotype length to build genomic relationship matrices." Genetics Selection Evolution **48**(1): 75.

Fernandez, X., et al. (1999). "Influence of intramuscular fat content on the quality of pig meat—1. Composition of the lipid fraction and sensory characteristics of m. longissimus lumborum." Meat Science **53**(1): 59-65.

Fisher, R. A. (1919). "XV.—The correlation between relatives on the supposition of Mendelian inheritance." Earth and Environmental Science Transactions of the Royal Society of Edinburgh **52**(2): 399-433.

Frayling, T. M. J. N. R. G. (2007). "Genome–wide association studies provide new insights into type 2 diabetes aetiology." **8**(9): 657.

Gelman, A. and J. Hill (2006). Data analysis using regression and multilevel/hierarchical models, Cambridge university press.

Goddard, M. (2009). "Genomic selection: prediction of accuracy and maximisation of long term response." Genetica **136**(2): 245-257.

Goddard, M. and B. Hayes (2007). "Genomic selection." Journal of Animal Breeding and Genetics **124**(6): 323-330.

Griffiths, A. J., et al. (2005). An introduction to genetic analysis, Macmillan.

Habier, D., et al. (2011). "Extension of the Bayesian alphabet for genomic selection." BMC bioinformatics **12**(1): 186.

Hayes, B. (2007). "QTL mapping, MAS, and genomic selection." A short-course. Animal Breeding Genetics Department of Animal Science. Iowa State University **1**(1): 3-4.

Hayes, B. and M. Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics **157**(4): 1819-1829.

Hayes, B. and M. Goddard (2010). "Genome-wide association and genomic selection in animal breeding." Genome **53**(11): 876-883.

Hayes, B. J., et al. (2009). "Invited review: Genomic selection in dairy cattle: Progress and challenges." Journal of dairy science **92**(2): 433-443.

Hess, M., et al. (2017). "Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population." Genetics Selection Evolution **49**(1): 54.

Hocquette, J., et al. (2010). "Intramuscular fat content in meat-producing animals: development, genetic and nutritional control, and identification of putative markers." Animal **4**(2): 303-319.

Jannink, J.-L., et al. (2010). "Genomic selection in plant breeding: from

theory to practice." Briefings in functional genomics **9**(2): 166-177.

Lander, E. S. J. S. (1996). "The new genomics: global views of biology." **274**(5287): 536-539.

Larzul, C., et al. (1997). "Phenotypic and genetic parameters for longissimus muscle fiber characteristics in relation to growth, carcass, and meat quality traits in large white pigs." Journal of animal science **75**(12): 3126-3137.

Legarra, A., et al. (2009). "A relationship matrix including full pedigree and genomic information." Journal of dairy science **92**(9): 4656-4663.

Lello, L., et al. (2018). "Accurate genomic prediction of human height." Genetics **210**(2): 477-497.

Li, Z., et al. (2017). "Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia." **49**(11): 1576.

Lo, L., et al. (1992). "Genetic analyses of growth, real-time ultrasound, carcass, and pork quality traits in Duroc and Landrace pigs: II. Heritabilities and correlations." Journal of animal science **70**(8): 2387-2396.

Medgyesi, D., et al. (2014). "The protein tyrosine phosphatase PTP1B is a negative regulator of CD40 and BAFF-R signaling and controls B cell autoimmunity." Journal of Experimental Medicine **211**(3): 427-440.

Meuwissen, T., et al. (2016). "Genomic selection: A paradigm shift in animal breeding." Animal frontiers **6**(1): 6-14.

Meuwissen, T. H. and M. E. Goddard (2001). "Prediction of identity by descent probabilities from marker-haplotypes." Genetics Selection Evolution **33**(6): 605.

Moser, M., et al. (2003). "BMPER, a novel endothelial cell precursor-derived protein, antagonizes bone morphogenetic protein signaling and endothelial cell differentiation." Molecular cellular biology **23**(16): 5664-5679.

Nakae, J., et al. (2003). "The forkhead transcription factor Foxo1 regulates adipocyte differentiation." Developmental cell **4**(1): 119-129.

Nature, G. P. C. J. (2010). "A map of human genome variation from population-scale sequencing." **467**(7319): 1061.

Nature, I. S. C. J. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." **460**(7256): 748.

Newcom, D., et al. (2004). "Relationship between backfat depth and its individual layers and intramuscular fat percentage in swine." Animal Industry Report **650**(1): 103.

Ovilo, C., et al. (2000). "A QTL for intramuscular fat and backfat thickness is located on porcine chromosome 6." Mammalian Genome **11**(4): 344-346.

Paszek, A., et al. (2001). "Interval mapping of carcass and meat quality traits in a divergent swine cross." Animal Biotechnology **12**(2): 155-165.

Pearson, G., et al. (2001). "Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions." Endocrine reviews **22**(2): 153-183.

Price, A. L., et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." **38**(8): 904.

Pryce, J., et al. (2010). "A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes." **93**(7): 3331-3345.

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.

Raven, L.-A., et al. (2014). "Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle." **15**(1): 62.

Rudan, I., et al. "Genomic Prediction of Health Traits in Humans: Demonstrating the Value of Marker Selection. ML Bermingham1, R. Pong-Wong2, A. Spiliopoulou1, C. Hayward1."

Sakaue, H., et al. (2004). "Role of MAP kinase phosphatase-1 (MKP-1) in adipocyte differentiation." Journal of Biological Chemistry.

Scott, L. J., et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants."

Shirakata, Y., et al. (1999). "Distinct subcellular localization and substrate specificity of extracellular signal-regulated kinase in B cells upon stimulation with IgM and CD40." The Journal of Immunology **163**(12): 6589-6597.

Silventoinen, K., et al. (2003). "Heritability of adult body height: a comparative study of twin cohorts in eight countries." Twin Research and Human Genetics **6**(5): 399-408.

Slatkin, M. (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future." Nature Reviews Genetics **9**(6): 477.

Speed, D., et al. (2012). "Improved heritability estimation from genome-wide SNPs." The American Journal of Human Genetics **91**(6): 1011-1021.

Sun, X., et al. (2015). "Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes." Animal Industry Report **661**(1): 86.

Suzuki, K., et al. (2005). "Genetic parameter estimates of meat quality traits in Duroc pigs selected for average daily gain, longissimus muscle area, backfat thickness, and intramuscular fat content." Journal of animal science **83**(9): 2058-2065.

Tanner, J. (1986). "Use and abuse of growth standards." Human growth **3**: 95-109.

Tseng, Y.-H. and T.-C. He (2007). "Bone morphogenetic proteins and adipocyte differentiation." Cellscience Rev **3**(3): 342-360.

Tukiainen, T., et al. (2014). "Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation." PLoS genetics **10**(2): e1004127.

Turner, S., et al. (2011). "Quality control procedures for genome-wide association studies." **68**(1): 1.19. 11-11.19. 18.

Ukkola, O., et al. (2005). "Protein tyrosine phosphatase 1B variant associated with fat distribution and insulin metabolism." Obesity research **13**(5): 829-834.

Utsunomiya, Y. T., et al. (2016). "GHap: an R package for genome-wide haplotyping." Bioinformatics **32**(18): 2861-2862.

VanRaden, P. M. (2008). "Efficient methods to compute genomic predictions." Journal of dairy science **91**(11): 4414-4423.

Villanueva, B., et al. (2005). "Benefits from marker-assisted selection under an additive polygenic genetic model." Journal of animal science **83**(8): 1747-1752.

Villumsen, T., et al. (2009). "The importance of haplotype length and heritability using genomic selection in dairy cattle." Journal of Animal Breeding and Genetics **126**(1): 3-13.

Villumsen, T. M. and L. Janss (2009). Bayesian genomic selection: the effect of haplotype length and priors. BMC proceedings, BioMed Central.

Visscher, P. M., et al. (2014). "Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples." PLoS genetics **10**(4): e1004269.

Visscher, P. M., et al. (2008). "Heritability in the genomics era—concepts and misconceptions." Nature Reviews Genetics **9**(4): 255.

Visscher, P. M., et al. (2006). "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." PLoS genetics **2**(3): e41.

Visscher, P. M., et al. (2017). "10 years of GWAS discovery: biology, function, and translation." **101**(1): 5-22.

Wang, E., et al. (1993). "Bone morphogenetic protein-2 causes commitment and differentiation in C3Hl0T1/2 and 3T3 cells." Growth factors **9**(1): 57-71.

Wang, X., et al. (2009). "Differential display of expressed genes reveals a novel function of SFRS18 in regulation of intramuscular fat deposition." International journal of biological sciences **5**(1): 28.

Wang, Y. and H. S. Sul (2009). "Pref-1 regulates mesenchymal cell commitment and differentiation through Sox9." Cell metabolism **9**(3): 287-302.

Yang, J., et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." Nature genetics **42**(7): 565.

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, J., et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." **38**(2): 203.

Zhou, X. and M. Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies." Nature genetics **44**(7): 821.

국문초록

전장유전체연관분석과 표현형 예측 연구를 위한 선형혼합모형

원소영

농생명공학부

서울대학교 대학원 농업생명과학대학

염기서열분석과 유전형질분석 기술의 발전으로 생물학적 연구에 이용할 수 있는 많은 양의 데이터가 축적되었다. 뿐만 아니라 통계적 모형이 발달하고 큰 데이터를 계산하는 능력이 향상되면서 방대한 양의 유전정보에 대한 보다 정밀한 분석이 가능해졌다. 통계적 계산을 활용한 유전정보의 분석으로부터 유전형질과 표현형 간의 관계를 밝혀낼 수 있다.

본 졸업논문에서는 유전형질의 차이가 어떻게 표현형과 관련이 있는지를 주로 다루고 있다. 우선, 전장유전체연관분석을 통해 표현형과 연관성이 높은 유전 변이를 찾아내고자 하였다. 또한, 유전

형질로부터 표현형을 정확하게 예측할 수 있는 모형을 개발하고자 하였다. 표현형에 대한 유전 변이들의 효과를 추정하기 위해서 다양한 선형혼합모형을 적용하였다.

2장에서는 돼지의 근내지방도에 대해서 전장유전체연관분석을 실시하였다. 이로부터 통계적으로 유의한 효과를 가지는 단일염기다형성들을 발견했고, 유의한 단일염기다형성이 포함되어 있거나 물리적으로 가까이 있는 유전자들을 찾아내었다. 찾아진 유전자들 중, 마이토겐 활성화 단백질 키나제 경로와 관련된 유전자들을 돼지의 근내지방도에 영향을 주는 후보 유전자들로 제시하였다.

3장에서는 한우의 유전자형으로부터 도체중을 예측하기 위해 반수체의 대립 형질을 이용한 유전체 예측을 진행하였다. 다양한 방법으로 반수체를 정의하였고, 이로부터 얻어진 대립 형질을 사용했을 때 유전체 예측의 정확도를 비교하였다. 이 때, 반수체를 이용하였을 때의 정확도가 개개의 단일염기다형성을 이용했을 때의 정확도보다 높게 나타났다.

4장에서는 사람의 유전자형으로부터 키를 예측하는 모형을 설계하였다. 예측 모형으로는 부모의 키로 보정된 최적선형불편추정 모형을 사용하였다. 더불어 부트스트랩 재추출을 활용하여 키에 미치는 영향이 큰 단일염기다형성을 선택하였다. 선택된 단일염기다형성만을 변수로 사용하는 모형을 검증한 결과, 예측력이 높게 나

타났다.

위의 연구들을 통해 유전자형에서의 변이와 표현형에서의 변이 사이의 관계를 설명하기 위해 선형혼합모형을 어떻게 적용할 수 있는지 이해할 수 있었다. 연구에서 얻어진 결과는 동물과 사람의 유전적 구조를 이해하기 위한 선형혼합모형의 적용을 확장하는 것에 활용될 수 있다.

주요어: 선형혼합모형, 유전체 예측, 전장유전체연관분석

학번: 2017-22852