



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 박사 학 위 논 문

**Estimating genetic marker effects in
population-based genomic study using
regression model**

집단 유전체학에서 회귀 모형을 이용한

유전 표지자 효과 추정

2017년 2월

서울대학교 대학원

생물정보학 협동과정

이 영 섭

**Estimating genetic marker effects in
population-based genomic study using
regression model**

By

Young-Sup Lee

Supervisor: Professor Heebal Kim

February, 2017

Interdisciplinary Program in Bioinformatics

Seoul National University

집단 유전체학에서 회귀 모형을 이용한

유전 표지자 효과 추정

2017년 2월

지도교수 김 희 발

이 논문을 이학박사 학위 논문으로 제출함

2017년 2월

서울대학교 대학원

생물정보학 협동과정

이 영 섭

이영섭의 이학박사 학위논문을 인준함

2017 년 2월

위 원 장 원 성 호 (인)

부위원장 김 희 발 (인)

위 원 한 재 용 (인)

위 원 윤 숙 희 (인)

위 원 조 서 애 (인)

Abstract

**Estimating genetic marker effects
in population-based genomic study using
regression model**

Young-Sup Lee

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

After various DNA (deoxyribonucleic acid) markers at the genomic DNA level had been discovered, scientists paid attention to DNA sequencing and genotyping. Genotyping is to uncover the genetic variants as one of the molecular markers. Single nucleotide polymorphisms (SNPs) are undeniably one of the most important markers. Especially, population-based SNP can possess the characteristics of an individual that may be different from others.

To reveal the causes of an individual's characteristics, one of the

possible ways is to employ established statistical models. Regression analysis has frequently been used in the bioinformatics area. I analyzed the data using the regression models such as linear, nonlinear regression and mixed models.

This doctoral dissertation comprises five chapters. In chapter 1, overviews of the required population genetics theories, effective population size estimation, best linear unbiased prediction (BLUP) and genome-wide association study (GWAS) is introduced. To estimate the effective population size, two methods have been employed: classical Sved's equation and Kimura 2-Parameter (K2P) model and Watterson theta estimator. Sved's equation is based on nonlinear regression, computationally and K2P uses the number of SNPs. The BLUP is used to estimate the random effects in linear mixed models. Moreover, GWAS is used to find causal genetic variants associated with a trait. As one of the methods to predict random marker effects, I propose the Single Nucleotide Polymorphism – Genomic Best Linear Unbiased Prediction (SNP-GBLUP). This new BLUP is based on Genomic Relationship Matrix (GRM) in theory.

In chapter 2, effective population size of Korean Thoroughbred horses (TB horses) has been estimated. TB breeds have been beloved because of those breeds' great racing capability. I tried to examine the

genetic diversity and stability of Korean TB population using by estimating effective population size. I used two methods as mentioned earlier: Sved's equation as basic approach, K2P and Watterson theta estimator as the second approach. I estimated TB horses' effective population size as 79 (Sved's equation) and 77 (K2P). This is rather weak when compared to other countries' TB effective population size. For instance, Corbin et al. estimated Irish TB effective population size as 100. The author used Sved's equation which is based on linkage disequilibrium (LD).

In Chapter 3, I introduced SNP-SNP Relationship Matrix (SSRM) which deals with the pairwise relationships between SNPs. This relationship matrix can be considered more advanced and differentiated notion than the Genomic Relationship Matrix (GRM) which is important in Genomic-Best Linear Unbiased Prediction (G-BLUP). GRM extracts individual relationships that are crucial concepts of mixed model or BLUP. In the BLUP area, to deal with the random effects effectively, GRM is one of the requisites. SSRM is a novel concept, although it is based on multivariate normal distribution (MVN) and GRM. The difference of SSRM from GRM is grounded on the different definition of the relationships since it is defined at the individual or SNP level. The SSRM is certainly more difficult and not-easily-validated one. Despite this, the bioinformatic information contained in

SSRM is sufficient because it can contain extensive information. I think that SSRM is the hidden information and GRM may be disguised or processed one by SNP information. By introducing SSRM, I analyzed the human height data using mixed model. Korean Association Resource Phase 3 (KARE3), Ansan-Ansung cohorts data contains each individual's traits and SNP information. The main objective was to check SSRM's usefulness in mixed model and compare SSRM-based SNP-GBLUP with SNP-BLUP (Single Nucleotide Polymorphism-Best Linear Unbiased Prediction) which is based on IID (independent & identically distributed) between SNP relationships. First, I introduced the theoretical derivation of SSRM based on probability density function (PDF) of the model and linear algebra. Second, I compared SNP-GBLUP with SNP-BLUP and G-BLUP by using human height and SNP data. The genetic values between SNP-GBLUP and SNP-BLUP were very disparate along with the SNP effects.

In chapter 4, I tried to solve "Missing heritability problem" in BLUP. Missing heritability problem is a problem that the associations cannot fully explain heritability that are estimated from correlations between relatives. This is important in association like GWAS or BLUP. BLUP deals with global genetic variants and complex traits. The traits were Berkshire eight pork quality traits (fat, carcass weight, shear force, Minolta color L, A,

protein content, water holding capacity, backfat thickness). These traits are very important economic traits in the pork meat production industry and therefore those breeding values (BV) must be predicted with better accuracy as breeding strategies. First, using the GWA study, the putative quantitative trait loci (QTL) for traits of interest were scanned at the SNP level. I chose the criteria of the QTL as unadjusted P-value (<0.01) arbitrarily. Then I analyzed the Berkshire traits with the SNPs using the BLUP. The heritability estimated from BLUP was close to the known heritability estimates. The results showed better results than the results from using total SNPs (original data) in terms of genomic estimated breeding values (GEBVs) and heritability estimates.

In chapter 5, the selection coefficient in F1 generation (if borrowed from genetics) –the next generation of the current generation) – was predicted using Fisher’s fundamental theorem of natural selection and BLUP. Fisher’s fundamental theorem of natural selection states: “The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time”. The selection is one of the major driving forces to be able to change allele frequency. Thus not only to reveal the history of selection but also to predict future selection trends is very imperative. The statistical model was additive linear model like BLUP. I calculated the

additive genetic variance of each SNP using SNP effects (from SNP-GBLUP) and using the Fisher's theorem, calculated the selection coefficient. Then the gene ontology of significant SNP-containing genes was surveyed. The phenotypes were three Holstein milk-related traits (milk yield, fat and protein contents). These traits are very crucial to dairy farmers. The heritability estimates from the BLUP were not bad (milk yield, fat and protein content 0.39, 0.45 and 0.40, respectively). The gene catalogue was retrieved from Ensembl server (www.ensembl.org). The theorem links the genetic variance to selection coefficient. The features of selection coefficient were the next generation, "expected", "relative". The "expected" implies that the selection coefficient of this kind of approach is just predicted one and "relative" means that the predicted values were recalibrated using the maximum values because the order of the values are dependent on the units of phenotypic values. The gene ontology contained in highly selective SNPs predicted from milk protein traits was dendritic spine morphogenesis, nitric oxide biosynthetic process, etc. Specially, dendritic spine morphogenesis was the most significant gene ontology. The dendritic spine is the major sites of excitatory synaptic transmission in the mammalian brain and is very imperative in synaptic development and plasticity. Thus the related genes of the dendritic spine morphogenesis are expected to be important target of future artificial selection trends of Holstein cattle in Korea. The gene

ontology of milk yield and fat did not have any significant ontologies.

Key words: Regression analysis, SNP-SNP relationship matrix (SSRM), Missing-heritability problem, Selection coefficient, Genome-wide association study (GWAS), Best linear unbiased prediction (BLUP), Effective population size (N_e)

Student number: 2012-30909

Contents

Abstract	i)
Contents	viii)
List of Tables	xii)
List of Figures.....	xiv)
Abbreviations	xvi)
Chapter 1. LITERATURE REVIEW	1
1.1 Overview of population genetics	2
1.2 Effective population size estimation	3
1.3 Best linear unbiased prediction (BLUP)	6
1.4 Genome-wide association study (GWA Study).....	12
Chapter 2. Estimating effective population size of Thoroughbred (TB)	

horses using linkage disequilibrium and theta ($4N\mu$) value	14
2.1 Abstract	15
2.2 Introduction	16
2.3 Materials and Method.....	17
2.4 Results	25
2.5 Discussion	34
Chapter 3. The usage of an SNP-SNP relationship matrix for best linear unbiased prediction (BLUP) analysis using a community based-cohort study	39
3.1Abstract.....	42
3.2Introduction.....	44
3.3 Materials and Method.....	45
3.4 Results.....	50
3.5 Discussion.....	64

Chapter 4. BLUP-based analysis using GWA candidate markers improve GEBV in accord with narrow-sense heritability	65
4.1 Abstract.....	66
4.2 Introduction.....	66
4.3 Materials and Method	68
4.4 Results	70
4.5 Discussion	79
Chapter 5. Combined Analysis of Fisher’s theorem of natural selection and BLUP can expect current selection coefficient of SNP.....	82
5.1Abstract	83
5.2 Introduction.....	84
5.3 Materials and Method.....	86
5.4 Results.....	90
5.5 Discussion	102

GENERAL DISCUSSION AND CONCLUSION.....105

REFERENCES108

국문초록123

List of Tables

Table 2.1. Conversion ratios of the recombination rate of the each chromosome based on the physical length (Mb) and the genetic distance (cM)	20
Table 2.2. Description of the generation of the binning processes.....	25
Table 3.1. The predicted genetic values of 1 ~ 57 th human individuals according to the BLUP method	55
Table 3.2. The SNP-SNP relationship matrix of 1 ~ 8 th SNPs using the relationship of $G_u^{-1} = Z^T G^{-1} Z$ and Sherman-Morrison-Woodbury lemma.....	59
Table 3.3. Each SNP effect in IID and SSRM cases	60
Table 4.1. The narrow-sense heritability of best linear unbiased prediction (BLUP) using total SNPs and trimmed SNPs (P-value < 0.01 in GWAS).....	72
Table 5.1. F ₁ generation's (the next in the current generation) allele frequency change according to the single nucleotide polymorphism (SNP) effect under linear additive model.....	93

Table 5.2. Highly selective SNPs with any P-value <0.001 (top 1% SNPs) in the analysis of milk yield, fat and protein phenotypes and the genes containing it.....95

List of Figures

Figure 2.1. Number of single nucleotide polymorphisms (SNPs, 1/16 scale) per chromosome and each chromosome length.....27

Figure 2.2. Plot of observed linkage disequilibrium (LD) against physical distance.....29

Figure 2.3 Plot of the decrease in the predicted and the observed with no kinship matrix linkage disequilibrium (LD) against physical distance (Mb).....30

Figure 2.4. Manhattan plot showing the mutation rate against chromosome number when R (Transition/Transversion ratio) is 1.5 in Kimura 2-parameter model.....32

Figure 2.5 The historically estimated effective population size (N_e) against T generations ago.....34

Figure 3.1. The histogram of genetic value variances of G-BLUP, SNP-BLUP and SNP-GBLUP
.....52

Figure 3.2. The histogram of the SNP effects of SNP-BLUP and SNP-GBLUP.....53

Figure 3.3. The histogram of the genetic values of G-BLUP, SNP-BLUP and

SNP-GBLUP.....	54
----------------	----

Figure 4.1. The plot of breeding values against 8 Berkshire pork quality traits.....	75
---	----

Figure 4.2. Manhattan plot of $-\log_{10}(\text{P-value})$ across chromosomes.....	76
---	----

Figure 4.3. Manhattan plot of $-\log_{10}(\text{P-value})$ across chromosomes.....	77
---	----

Figure 4.4. 1 st and 2 nd discriminant functions of Berkshire 8 pork quality traits and corresponding GEBVs of total SNPs and trimmed SNPs.....	78
--	----

Figure 5.1. The flow chart of the analysis. It is categorized as method and theory.....	954
--	-----

Figure 5.2. Plot of selection coefficient against SNP effects.....	100
---	-----

Figure 5.3. Pie chart of gene ontology of the significant genes which contain highly selective SNPs in milk protein traits.....	101
--	-----

Abbreviation

SNP: Single Nucleotide Polymorphism

CNV: Copy Number Variation

BLUP: Best Linear Unbiased Prediction

GWAS: Genome-Wide Association Study

SNP-GBLUP: Single Nucleotide Polymorphism-Genomic Best Linear Unbiased Prediction

GRM: Genomic Relationship Matrix

K2P: Kimura 2-Parameter model

SUB: Subscapular

SUP: Suprailiac

TB: Thoroughbred

LD: Linkage Disequilibrium

SSRM: SNP-SNP Relationship Matrix

G-BLUP: Genomic-Best Linear Unbiased Prediction

MVN: Multivariate Normal distribution

KARE3: Korean Association Resource phase 3

SNP-BLUP: Single Nucleotide Polymorphism-Best Linear Unbiased Prediction

IID: Independent & Identically Distributed

PDF: Probability Density Function

BV: Breeding Value

QTL: Quantitative Trait Loci

GEBV: Genomic Estimated Breeding Value

EBV: Estimated Breeding Value

GS: Genomic Selection

SCAD: Smoothly Clipped Absolute Deviation Penalty

LASSO: Least Absolute Shrinkage and Selection Operator

BLUE: Best Linear Unbiased Estimation

P-BLUP: Pedigree Best Linear Unbiased Prediction

SS-BLUP: Single-Step Best Linear Unbiased Prediction

HWE: Hardy-Weinberg Equilibrium

DAVID: Database for Annotation, Visualization and Integrated Discovery

GO: Gene Ontology

Tr/Tv ratio: Transition/Transversion ratio

JC model: Jukes-Cantor model

CLT: Central Limit Theorem

KRA: Korean Racing Authority

MAF: Minor Allele Frequency

NCBI: National Center for Biotechnology Information

CSH: Chromosome Segment Homozygosity

MSA: Multiple Sequence Alignment

Indel: Insertion/deletion

NGS: Next-Generation Sequencing

MLE: Maximum Likelihood Estimation

IBD: Identity By Descent

EM: Expectation-Maximization

REML: REstricted Maximum Likelihood

ML: Maximum Likelihood

GLS: Generalized Least Squares

SMW: Sherman-Morrison-Woodbury

CWT: Carcass Weight

BF: Backfat

SF: Shear Force

WHC: Water Holding Capacity

MC_L: Minolta L color

MC_A: Minolta A color

Chapter 1. LITERATURE REVIEW

1.1 Overview of population genetics

The population genetics can be classified as variation in genomic and phenotypic level, neutral theory, selection theory, genomics, population substructure, quantitative genetics, etc in the population level. I focused on the variation, neutral theory and selection theory and quantitative genetics in my PhD thesis.

The variation can be classified as genomic variation and phenotypic variation in population genetics. The related topics are Hardy-Weinberg equilibrium (HWE), linkage disequilibrium (LD), etc. It assess the numerical value of genomic variation such as single nucleotide polymorphism (SNP), copy number variation (CNV), microsatellites, and transposable elements, etc. My analysis was focused mainly on SNP. Phenotypic variation is mainly dealt with complex traits. Its variation was mainly due to the infinite causal sites of any markers. I focused on the association of SNP with traits such as Berkshire 8 pork quality traits, Holstein milk-related traits. My main association test method was genome-wide association study (GWAS) and best linear unbiased prediction (BLUP). GWAS and BLUP is the main analysis method in complex traits.

The neutral theory was originated by Kimura (Kimura 1984). It generally states that most of genetic variation is neutral, not advantageous

and deleterious to the organisms and the evolution may be caused by neutral mutation. One of the important concepts in neutral theory is population mutation parameter, θ ($4N\mu$) where N is the relevant effective population size and μ is the mutation rate. I used this to estimate the effective population size.

On the contrary, the selection theory is related to breeding science, deeply because of livestock domestication and breeding strategy in modern science. We used the selection theory and BLUP to estimate the selection coefficient of next generation. Fisher, as one of the prominent statisticians and population geneticists, contributed the quantification and mathematical induction of quantitative genetics. And Henderson developed the mixed model (or BLUP model) to predict the estimated breeding values (EBVs).

1.2 Effective population size estimation

1.2.1. Overview of effective population size (N_e)

Effective population size (N_e) is the number of individuals that an idealized population would need to have and in general is equal to the number of breeding individuals in the population. The census population size is usually larger than N_e . The allele frequencies can fluctuate according to

the random genetic drift and inbreeding. Thus estimating effective population size is very important in this regards.

Random genetic drift uses Wright-Fisher model, generally and the effective population size arises naturally in random genetic drift. Also, the effective population size concept comes naturally in diffusion equation of population genetics category. In fact, effective population size has various definitions such as inbreeding effective population size, variance effective size and eigenvalue effective size, etc. The fluctuations in population size can make it important to estimate the current and ancestral effective population size estimation. The founder effect and bottlenecks in population can cause the fluctuations in population size. Unequal sex ratio, sex chromosomes, organelle genes can alter the accurate estimation of effective population size.

In every fields in population genetics, the effective population size concept arises in nature. In LD, the effective population size estimation can be a very crucial factor to decaying patterns of LD. Also, coalescence theory, homozygosity and heterozygosity and fixation index concepts can have effective population size factor.

1.2.2. The estimation of effective population size (N_e)

Although estimating effective population size (N_e) is very important,

accurate estimation is not very easy. Because there are various definitions as mentioned above and various methods exist, it cannot be said that specific method is prevalent and supreme. Despite this, the generally used method is LD-based one. LD can be measured between two loci. The markers are usually SNPs which can easily calculate LD. The Sved's equation (1.1) states that the expected value of correlation coefficient of LD can link the effective population size and recombination frequency.

$$E(r^2)=1/(1+4N_e c) \tag{1.1}$$

Where $E(r^2)$ is the expected value of correlation coefficient of LD, N_e is the effective population size to estimate and c is the recombination frequency. N_e can be estimated through the nonlinear regression using Sved's equation (1.1).

I originally proposed the other method to estimate the effective population size (N_e). Its method was based on Kimura 2-parameter model (K2P) and Watterson theta estimator. The sequence-based evolutionary distance and substitution rate can be calculated using K2P. Watterson theta estimator is widely-used method to estimate theta. I estimated the genomic regions' substitution rate and N_e . K2P is based on sequences in two species. However, though its estimated evolutionary distance values are dependent in

sample size, Watterson theta estimator can be a complement in this method. Watterson theta estimator includes sample size concept. Also, I further assert that evolutionary distance in one species scale can be defined in the population level. Although K2P is originally designed for two species' evolutionary model, I considered that there are no hinder to my assertion because evolutionary distance can be defined in the polymorphism level and it cannot be differentiated with between polymorphism such as SNPs in the population and sequence substitution, insertion and deletion in the sequence alignment.

1.3. Best linear unbiased prediction (BLUP)

1.3.1 Overview of BLUP

With the advent of new sequencing technologies, genomic selection is revolutionizing livestock's breeding. Genomic selection (GS) refers to selection decisions based on genomic estimated breeding values (GEBVs). GEBVs are the sum of the effects of dense genetic markers, thereby potentially capturing all the quantitative trait loci (QTL) effects that contribute to variation in the trait. In these days, marker information is the prerequisites to estimate GEBVs (Hayes, Bowman et al. 2009). The genomic selection evolution has been accomplished through the sequencing technologies with lower cost and powerful analysis method like best linear

unbiased prediction (BLUP).

BLUP estimates of random marker effects are called “best”, “linear”, “unbiased” and “prediction”. It means that the estimators are the best predictors that find the best solution and its solution is linear and BLUP solution is unbiased one. In some methods like “ridge regression”, “elastic net”, “SCAD (Smoothly Clipped Absolute Deviation Penalty)” and “LASSO (Least absolute shrinkage and selection operator)”, the unbiased solution are discarded to improve the precision.

The accuracy and reliability of GEBVs has already been evaluated in United States, New Zealand, Australia, and the Netherlands. Many of the countries have been tried to find the accurate GEBVs and well-defined QTL regions which are associated with the traits. However, the accurate marker effects calculation has been ignored in most of the previous studies. I tried to find out the random marker effects in SNP-GBLUP.

BLUP is involved in complex traits association in general. Each complex trait has a definite and well-measured heritability like narrow-sense heritability (h^2) and broad-sense heritability (H^2). Broad-sense heritability is defined by the proportion of genetic variance of phenotypic variance. If normal distribution fits in phenotypic values of traits of interest, H^2 exactly reflects the genetic proportional effects in quantitative genetics theory. h^2 is

defined by the proportion of additive genetic variance of phenotypic variance. Additive genetic variance is generally smaller than total genetic variance. Total genetic variance can contain the information of epistasis (gene-gene interactions), gene-environment interactions, additive and dominant effects. One of the main goal of BLUP is to estimate the heritability accurately and predict the breeding values of livestock animals.

Best linear unbiased prediction (BLUP) is used in linear mixed models as one of the solution of mixed models. Henderson (1950) argued that BLUP solution of random effects are similar to the best linear unbiased estimation (BLUE) (Henderson 1975). Gauss-Markov theorem in BLUE can be the general solution of regression, but Henderson's BLUP must estimate the fixed effects and random effects, simultaneously. Furthermore, variance components of random effects and residuals must be estimated simultaneously, although it is related to phenotypic variance.

There are various BLUP such as P-BLUP, G-BLUP, SNP-BLUP, etc. Originally, the random effects have been estimated through pedigree information like Pedigree-BLUP (P-BLUP). With the advent of sequencing technologies, SNP information can be the alternative of pedigree information. G-BLUP and SNP-BLUP usually uses SNP information. G-BLUP uses genomic relationship matrix (GRM) and SNP-BLUP assumes the IID

(independent and identically distributed) among markers. GRM describes the relationships between individuals and one of the variance-covariance matrices in statistics. And there are various subclasses of BLUP. Single-step BLUP (SS-BLUP) uses the pedigree and marker information and calculates the genomic estimated breeding values (GEBVs) and through computational iteration of GEBVs, it can estimate the marker random effects. R package “rrBLUP” provides the GRM and BLUP solution. I used this package to analyze Berkshire 8 pork quality traits and Holstein milk-related traits and human height.

1.3.2 BLUP analysis: SNP-GBLUP

I used SSRM (SNP-SNP Relationship Matrix) to predict the SNP random effects. SSRM can be derived from GRM (Genomic Relationship Matrix) which is used in G-BLUP. G-BLUP uses GRM to complement individual’s relationship to predict the breeding values. Our usage of SSRM was to complement the SNPs’ relationships and was a success despite the accuracy of SNP effects cannot be valid. The correlation of breeding values between G-BLUP and SNP-GBLUP in human height was 0.99, although some of those values were different from each other. In single-step BLUP (SSBLUP), other than using pedigree information, it computes the SNP effects iteratively and finds optimal SNP effects and breeding values. Despite

this, SNP effects which should be constant, can be different in the sample by different investigators in SS-BLUP because of this iteratively optimal finding solution. On the contrary, SNP-GBLUP's SNP effects would be rather unchangeable if investigator's sample were fixed. This can be the power of SNP-GBLUP.

I used the human height, Berkshire pork quality traits and Holstein milk yield, fat and protein contents as the BLUP's phenotypic values. In human height, I proved the good usage of SNP-GBLUP. In Berkshire pork quality traits, I used GWAS to complement the 'missing heritability problem'. Missing heritability problem states that estimated heritability from GWAS or BLUP are usually lower than the known heritability and it hinders the accuracy of those. GWAS results ($P\text{-value} < 0.01$) can complement missing heritability problem in BLUP analysis. I showed that GWAS results which can choose the significant SNPs can improve the genomic estimated breeding values (GEBVs) in BLUP, although the fact that chosen significant SNPs were in quantitative trait loci (QTL) regions, was not guaranteed.

In Holstein article, I focused on the expected current relative selection coefficients of SNPs using BLUP analysis. "Expected" means that selection coefficients are trait-dependent and next generation's coefficients. "Current" means that selection coefficients are predicted next generation's

coefficients. “Relative” implies that selection coefficient are adjusted values by maximum selection coefficients and this reflects that the selection coefficients are dependent on the units of phenotypic values. Because Holstein genomic selection is conducted by mainly milk-related traits, I regarded that the selection coefficient in large population can be viable. Our genetic variance model was additive model and BLUP (esp. SNP-GBLUP). And Fisher’s fundamental theorem of natural selection can link the additive genetic variance to the selection coefficient. I found out that the main governing factor of selection coefficient is SNP effect itself. The allele frequency can be the factor of selection coefficient but SNP effect governs the additive genetic variance. Especially, under HWE at current generation, the selection coefficient is identical $2 \times$ SNP effects. The fitness change across generations was determined by linear additive model which was established by population geneticists.

1.4. Genome-wide association study (GWA study)

1.4.1. Brief introduction for GWA study

Genome-wide association study (GWA study or GWAS) is a test using genetic variants to find any variants associated with traits. Traditionally, GWA study was used to search causal variants for patients (cases) and normal individuals (controls). For example, GWA study can be applied to

various diseases like abnormal blood pressure, diabetes, diverse types of carcinoma, cardiovascular disease and other clinically important ones. Although there are several issues and difficulties of accurate interpretation of GWAS results, a variety of the analysis techniques have been developed. The main difficulties can be sampling method, insufficient sample size, multiple testing and population substructure problem. I tried to use GWA study to estimate the random marker effects like SNP effects. In GWA study, the causal variants associated with the traits can be found. Thus the random marker effects can be classified to '0' effects and nonzero effects. Like Bayes $C\pi$, most of the random marker can be viewed to '0' effects because the quantitative trait loci (QTL) related to the specific traits is finite and limited. It can be stated that the random marker effects related to QTL is nonzero effects, GWAS can find it and the random marker effects can be predicted using mixed model like BLUP. Random marker effects is related to the specific genes due to LD and QTL. Thus predicting random marker effects are the representatives to some genes and QTL.

*This chapter
was published in livestock science as a partial fulfillment
of Young-Sup Lee's Ph.D program.*

**Chapter 2. Estimating effective population size of
Thoroughbred (TB) horses using linkage disequilibrium and
theta ($4N\mu$) value**

2.1 Abstract

Single nucleotide polymorphisms (SNPs) have been widely used in the polymorphic study. Particularly, SNP can be used to estimate the effective population size (N_e), theta (θ) and the substitution rate in population-based data. To estimate N_e , we used the two methods: Sved's equation based on linkage disequilibrium (LD) and Kimura-2-Parameter (K2P) model based on the evolutionary theory. Sved's equation is based on nonlinear regression and K2P model exploits the evolutionary distance using transition/transversion (tr/tv) ratio. Using these two methods, I estimated N_e of Korean Thoroughbred (TB) horses to be 79 and 77, respectively. I also computed the historical effective population size of TB horses which showed a gradual decrease in size from 100 generations ago to the current generation. The average substitution rate was estimated to be 1.24×10^{-9} /bp/year.

2.2 Introduction

Linkage disequilibrium (LD) denotes the nonrandom association of alleles at different loci and can result from processes such as migration, genetic drift and mutation in finite populations (Wang 2005). By using the single nucleotide polymorphisms (SNPs), genome-wide association studies (GWAS), genomic selection and other genomic techniques have been exploited. Especially, the effective population size estimation can be calculated using LD in the population SNP data. This can be possible with the emerging techniques like Illumina SNP BeadChip. LD structure can provide insights into the evolutionary population history. The strength of LD between two loci can be used to infer ancestral effective population size (N_e). Effective population size, N_e is the idealized number of individuals that would cause the same rate of inbreeding as observed in the actual breeding population (Mackay 2001). The pattern of historical N_e in domestic livestock populations can help us to understand selective breeding strategies on the genetic variation and can provide an insight in the level of inbreeding in populations (Corbin et al. 2010).

Kimura 2-paramter (K2P) model represents the evolutionary distance between two species using the transition/transversion (tr/tv) ratio. It is the developed model of Jukes-Cantor (JC) model which assumes the tr/tv

is equal to 1 (Srivathsan and Meier 2012). In this study, we used K2P model in the population data. The estimated evolutionary distance can represent the population parameter by central limit theorem (CLT), if the sample size is large enough (Rosenblatt 1956). Because the evolutionary distance is defined in the given sample, substitution rate estimated from evolutionary distance thereof belongs to one of the sample parameter values, not population values.

I estimated the substitution rate using K2P model in the genome scale and effective population size of Korean Thoroughbred horses (TB horses) using Watterson estimator and substitution rate (Wright et al. 2005, Felsenstein 2006). TB horse in Korea have been imported from diverse countries. The TB horse is one of the fastest breeds with pedigree records spanning three centuries. This breed was generated in England during (Watterson 1959) the 18th century from native Celtic and Oriental horses. A concern exists regarding the loss of genetic variation since the population is essentially closed (Cunningham et al. 2001). TB horses have been produced since 1990 in Korea and approximately 1,400 TB horses have been produced annually. Like other domesticated species, it is very important for TB horses to be preserved against various environmental risks threatening survival. The effective population size is the parameter that yields the population dynamics of genes and the strength to endure diverse environmental risks can be

assessed by the magnitude of N_e .

As mentioned earlier, I used two methods: LD-based and evolutionary distance-based estimation. LD-based method uses the relationship between LD and recombination frequency, which was first suggested by Sved and revised by Hill (Sved 1971, Hill 1981). Evolutionary distance-based method uses the substitution rate estimated from evolutionary model like JC, K2P and population mutation parameter (θ) from Watterson estimator (Watterson 1959). Watterson estimator is the widely used theta estimator. The objective of the study was to estimate N_e based on LD and theta and examine the stability of Korean TB horses' population.

2.3. Materials and Method

2.3.1 Genotypic data

A 240 Thoroughbred horses (TB horses) from the Korean Racing Authority (KRA) were sampled. 60 were sires and 180 individuals were randomly sampled. The owners of TB horses from KRA granted permission for blood extraction (Korea Racing Authority Act, Article 11, 12, 36). KRA has established an Animal Experimentation Ethics Committee according to Korea's Animals Protection Act 14. KRA performed all experimental procedures according to international guidelines, which is guaranteed by an affiliate association of the Korean government (Korea Racing Authority Act,

Article 44).

A complete blood samples (10ml) were collected from the carotid artery of each horse and treated with heparin to prevent clotting. Genomic DNA quality check was performed with agarose gel and fluorescence-based quantification. Electrophoresis on a 0.6% agarose gel and Pulse Field Gel Loading (200ng) were performed, also. The manufacturer's protocol followed for pair library construction 500bp fragment, which involved purifying genomic DNA, and isolating fragments of less than 800bp. The fragments were blunt-ended with 5'-phosphorylated ends, and a 3'-dA overhang, (adaptor-modified ends). The ligation products, were purified and the genomic DNA library was constructed as the protocol. Illumina's HiSeq platform was used to generate SNP data (Illumina Equine SNP 50K Bead chip).

Some markers were excluded using Hardy-Weinberg equilibrium (HWE P-value < 0.0001), minor allele frequency (MAF < 0.05) because incorrect LD estimates can bias the effective population size (N_e) estimation. I also excluded the physical distances less than 100 bp (Purcell, Neale et al. 2007, Tenesa, Navarro et al. 2007, Corbin, Blott et al. 2010). Additionally, only autosomal markers were used because sex chromosomes have different deterministic equations (Hill 1981).

2.3.2 Effective population size estimation using linkage disequilibrium (LD)

I used the SNP data to compute effective population size (N_e). LD-based approach is based on the relationship between correlation coefficient of LD and recombination frequency (Morgan units). Sved suggested that expected r^2 and recombination rate (c) are related to N_e as the coefficient at the autosomal loci.

$$E(r^2) = 1/(1+4N_e c) \quad (2.1)$$

It was shown that equation (2.1) was only applicable when ignoring mutations and it should be amended when considering mutations. Hill demonstrated that $E(r^2) = 1/(2+4N_e c)$ in the autosomal loci in the presence of mutations (Hill 1981). Practically, the following model was used:

$$y_i = 1/(a+4bd_i) + e_i \quad (2.2)$$

Where y_i is the value of r^2 for SNP pair i , and d_i is the value of linkage distance in Morgans. The symbol b represents the effective population size of N_e and a represents 2 but in reality depicts a value close to 2. These two values (a , b) were originated from (2.1). Computationally, the equation (2.2)

is the nonlinear regression model which minimizes the residual error e_i value using least squares. I used the LDcorSV package to determine the correlation coefficient of LD between two markers. LDcorSV is a package based on R and provides set of functions to measure the r^2 (the correlation coefficient of LD). r^2 was calculated for the syntenic marker pairs within 100 SNPs apart. I computed r^2 using three methods: a kinship matrix for all chromosomes, a kinship matrix for each chromosome, without a kinship matrix. The kinship matrix were introduced to complement the LD structure bias. (Flury, Tapio et al. 2010, Mangin, Siberchicot et al. 2012). The recombination frequency was calculated using physical distances (Mb), genetic distance (cM) and its cM/Mb conversion ratio which was obtained from NCBI website (see Table 2.1). The total chromosome lengths was obtained from equine linkage map (Swinburne et al. 2006). We used this information to perform nonlinear regression over total chromosomes.

Table 2.1. Conversion ratios of the recombination rate of the each chromosome based on the physical length (Mb) and the genetic distance (cM).

Chromosome	Physical Length (Mb)	Genetic distance (cM)	Conversion Ratio (cM/Mb)
------------	----------------------	-----------------------	-----------------------------

1	186	194	1.04
2	121	129	1.07
3	119	120	1.01
4	109	123	1.13
5	100	100	1.00
6	85	127	1.49
7	99	102	1.03
8	94	109	1.16
9	84	105	1.25
10	84	106	1.26
11	61	65	1.07
12	33	58	1.76
13	43	58	1.35
14	94	153	1.63
15	92	97	1.05
16	87	111	1.28
17	81	71	0.88
18	82	88	1.07
19	60	56	0.93
20	64	81	1.27
21	58	76	1.31
22	50	80	1.6
23	56	56	1.00
24	47	47	1.00
25	40	49	1.23
26	42	24	0.57
27	40	93	2.33

28	46	63	1.37
29	34	75	2.21
30	30	50	1.67
31	25	41	1.64

2.3.3 Effective population size estimation using evolutionary distance

I calculated the substitution rate using evolutionary distance and then effective population size, N_e . Evolutionary distance model was Kimura 2-parameter model (K2P model). It is based on the transition/transversion (tr/tv) ratio (Srivathsan and Meier 2012). Transition is the mutation between purines and transversion between pyrimidines. And we calculated the substitution rates of single nucleotide polymorphisms (SNPs). Then using Watterson theta estimator, we estimated the N_e . The following equations were used to estimate N_e :

$$d_{K2P} = -0.5\ln(1-2P-Q) - 0.25\ln(1-2Q)$$

(2.3)

Where P and Q are fractions of aligned sites whose two bases are related to transitions and transversions, respectively and d_{K2P} is the evolutionary

distance of the K2P model.

$$R=\alpha/2\beta \quad (2.4)$$

$$d=4\lambda t \quad (2.5)$$

$$\hat{\theta}=K/a_n \quad (2.6)$$

$$\theta=4N_e\mu \quad (2.7)$$

Where R is the transition-transversion ratio, λ is the mutation rate of aligned sequences, d is the evolutionary distance. The t is the divergence time of the population, $\hat{\theta}$ is the estimated theta using Watterson theta estimator, K is the number of segregating sites and a_n is the (n-1) harmonic number. In my data, the mutation rate was principally based on the neutral substitution rate (Watterson 1959, Purcell et al. 2007, Srivathsan and Meier 2012).

I obtained P and Q values from the SNP data. From the number of SNPs and the length of TB SNP sequences, we calculated the P and Q value. Although there are various sources of mutation, the SNP is usually the major source and the neutrality of SNPs can be guaranteed after HWE and MAF pruning. Generation time of TB breeds was set to be 10.7 years (Taveira, Mota et al. 2004). The tr/tv ratio was set to be 1, 1.5 and 2. It is often a good approximation for a majority of mammalian nuclear genes (a primer to the

phylogenetic analysis using the PHYLIP packages, J. Tuimala, 4th ed). I arbitrarily chose the R value as 1, 1.5, 2 and calculated θ of the correspondent sequences using Watterson theta estimator. I used $d=4\lambda t$ because I assumed the origin of TB horses were following horses: Byerly Turk (Turkey), Darley Arabian (Arab region), Godolphin Barb (Morocco region) and UK native horses. I assumed that the TB horses diverged 5000 years ago (Outram et al. 2009). My horse sample were assumed to be derived from these horses. The distinct regions of 4 TB horses supports it. From θ and λ , we estimated N_e .

2.3.3. Ancestral effective population size

The ancestral effective population size T generations ago can be calculated assuming linear growth of the recombination rate. Based on the assumption, recombination frequency $c = 1/ (2T)$ were calculated and then equation (2.2) were used. The binning based on $c = 1/ (2T)$ were performed using average LD and average linkage distance. The ancestral generation had 10 years step. The historical effective population size were computed from each binning.

Table 2.2. Description of the generation of the binning processes.

Generation	Generation Range	Distance range (cM)
10	5-15	10-33
20	15-25	2-33
30	25-35	1.43-2
40	35-45	1.1-1.43
50	45-55	0.91-1.1
60	55-65	0.77-0.91
70	65-75	0.66-0.77
80	75-85	0.58-0.66
90	85-95	0.5-0.58
100	95-105	0.47-0.5

2.4 Results

2.4.1 Pruning of genotypic data

I pruned 54,602 SNPs by screening Hardy-Weinberg equilibrium (HWE $P < 0.0001$), minor allele frequency (MAF < 0.05) and excluding

SNPs on sex chromosomes. After filtered, 41,371 SNPs remained. The number of SNPs on each chromosome and chromosome length are shown in Figure 2.1. The average linkage disequilibrium (LD) between SNP pairs after pruning and excluding the physical distance for 100bp or less 0.1467 ± 0.2035 (mean \pm SD). The average physical distance was 2.03 ± 1.49 Mb and the average recombination rate was 2.45 ± 0.04 cM. To calculate the substitution rate in SNP-based population data, I used the PLINK homozygosity option (Purcell et al. 2007). 13,521 loci data was obtained with the physical distance (kb) and number of SNPs to calculate θ and the substitution rate. And then I used these results to determine effective population size, N_e .

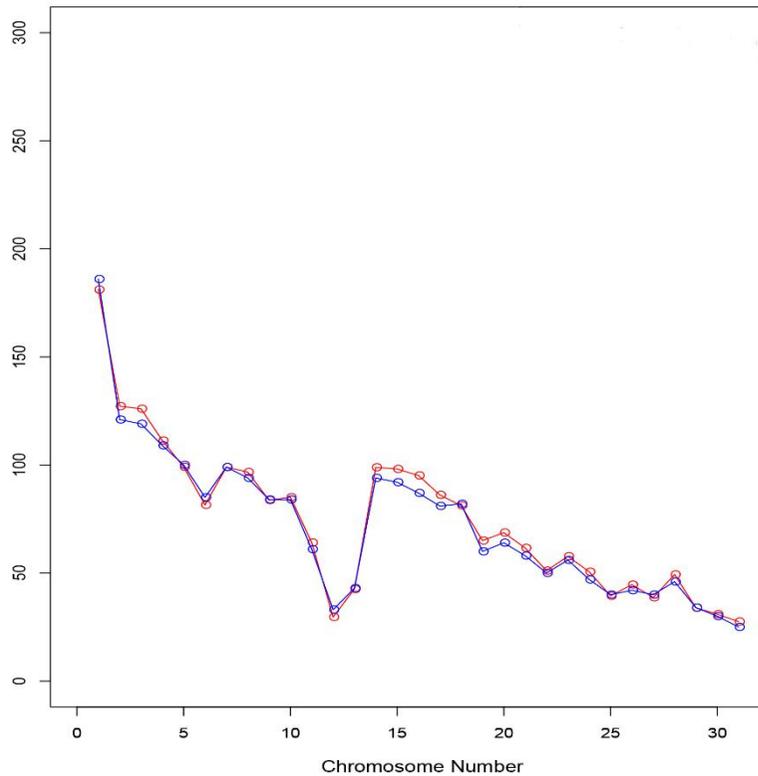


Figure 2.1. Number of single nucleotide polymorphisms (SNPs, 1/16 scale) per chromosome (red) and each chromosome length (Mb; blue) in the SNP dataset of TB horses. It represents that the number of SNPs is roughly proportional to each chromosome length.

2.4.2. LD decay with physical distance

The correlation coefficient of LD (r^2) declines smoothly with physical distance which can be predicted by Sved's formula (Equation 2.1). Figure 2.2 indicates that observed r^2 decays against the physical distance (Mb) in three cases: without the kinship matrix, with the kinship matrix of total chromosomes and with the kinship matrix of each chromosome. The kinship matrix contains the information of the pedigree relationship and it can be calculated from SNP information (LANGE, WESTLAKE et al. 1976). The first and second cases showed the similar declining patterns. However, the third case showed a significant discrepancy in the range of 6 – 10 Mb. I asserted that the use of kinship matrix could not lead to estimate effective population size (N_e) with better accuracy in our data. Thus we chose the case without kinship matrix to estimate N_e .

Figure 2.3 demonstrates the patterns of the predicted and observed r^2 declines without the kinship matrix. It shows the similarity at most physical distances, but dissimilarity near 0 Mb and above 6 Mb. The predicted plot was based on Sved's equation and observed plot was based on the result of LDcorSV. This result coincides with Corbin which was not identical below 0.01 cM (0.01Mb) (Corbin, Blott et al. 2010).

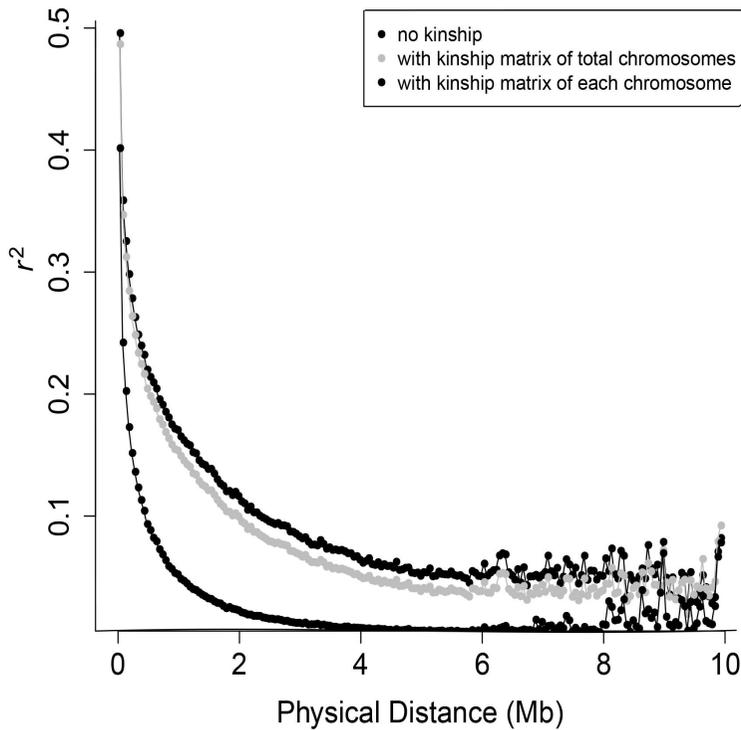


Figure 2.2. Plot of observed linkage disequilibrium (LD) against physical distance with the kinship matrix (middle position: with kinship matrix of total chromosomes, lowest position: with kinship matrix of each chromosome) and without the kinship matrix (highest position). The decay in LD shows the discrepancy between with and without the kinship matrix. I chose the decay of LD without the kinship matrix to estimate the effective population size (N_e) because I compared N_e of TB horses with another method of Kimura-2-parameter model (K2P).

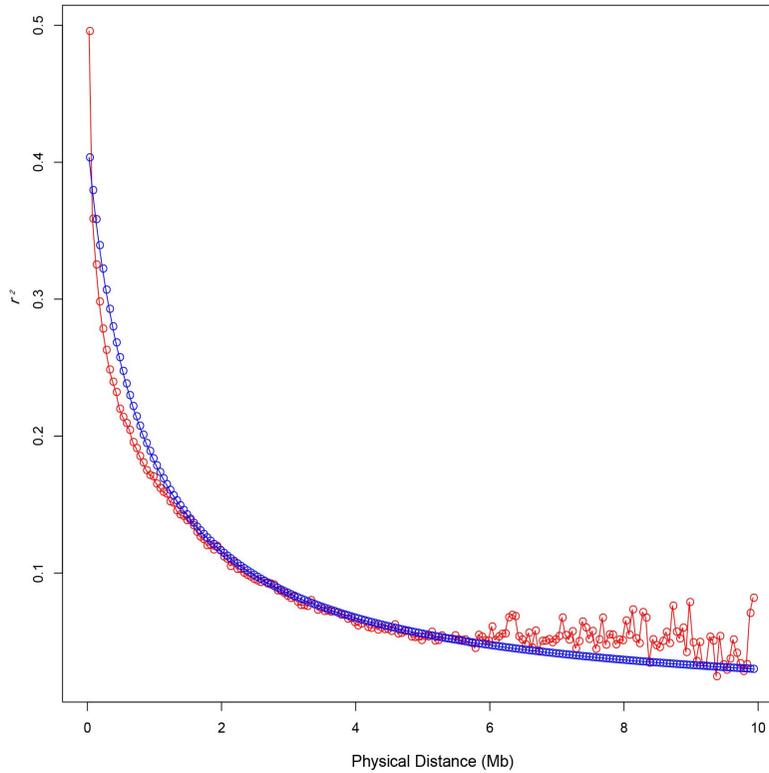


Figure 2.3 Plot of the decrease in predicted (blue) and observed (red) with no kinship matrix linkage disequilibrium (LD) against physical distance (Mb). The predicted LD plot was based on the Sved's equation and observed LD plot was based on the result of R package "LDcorSV". This suggests that the estimation of effective population size (N_e) of Korean TB horses in the dataset be reliable.

Coefficient b (Equation 2.2) can be interpreted as N_e of TB horses,

as seen in Sved's formula (Equation 2.1). The mean and interval estimates for parameter a , were 2.40 ± 0.004 , 2.35 ± 0.004 , 2.06 ± 0.003 and those for parameter b , were 79 ± 0.17 , 99 ± 0.21 , and 461 ± 0.97 without the kinship matrix and with the kinship matrix of total chromosomes, with the kinship matrix of each chromosome, respectively.

2.4.3 Substitution rate and effective population size estimation

The estimated substitution rates were 1.55×10^{-9} , 1.24×10^{-9} , and 1.24×10^{-9} per base pairs per year when tr/tv ratio were 1, 1.5 and 2, respectively. The divergence time of the domesticated horse breeds from Przewalski's horse was estimated to be 5,000-6,000 years ago (Outram, Stear et al. 2009). I assumed the divergence time of TB horses to be 5,000 years ago. A Manhattan plot of the mutation rates (in this case substitution rates in the SNP data) against chromosome number is shown in Figure 2.4. Each dot represents the regions which was obtained from PLINK homozygosity option. The substitution rate were similar to the average mutation rate of mammalian genomes (2.2×10^{-9} per base pair per year) (Kumar and Subramanian 2002). The estimated N_e (\pm SD) were 62 ± 0.07 , 77 ± 0.003 , and 77 ± 0.003 , respectively. N_e was similar to LD without the kinship matrix when tr/tv ratio were 2 or 1.5. Thus, we chose N_e to be 79 (LD-based)

and 77 (theta value-based).

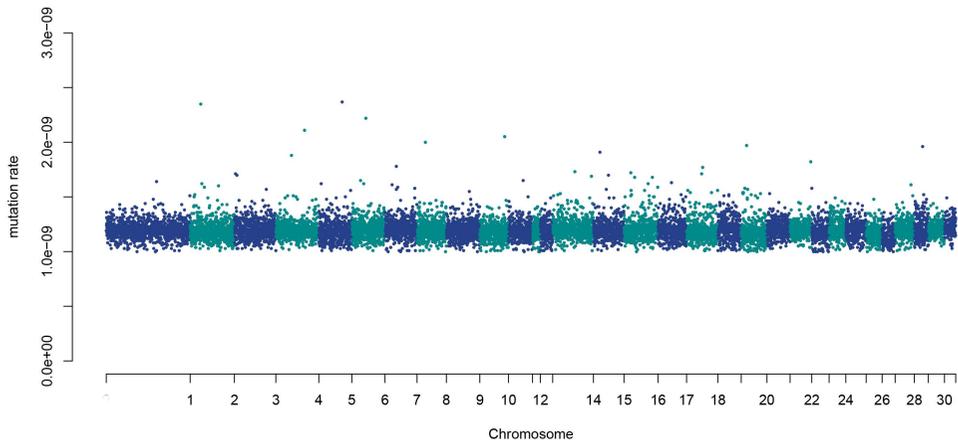


Figure 2.4. Manhattan plot showing the mutation rate against chromosome number when R (Transition/Transversion ratio) is 1.5 in Kimura 2-parameter model. The substitution rate (in the same context, mutation rate by SNPs) was obtained using the binning by PLINK homozygosity option. Chromosomes are enumerated from 1 to 31. This shows uniform substitution rate of binning regions across the board.

2.4.5 Ancestral effective population size estimation

Figure 2.5 shows the estimated $N_e T$ generations ago. It shows that

the historical N_e had decreased gradually until 10 generations ago and has increased slightly until the current generation since that time. This can be explained by the fact that TB horses in Korea are imported from diverse countries.

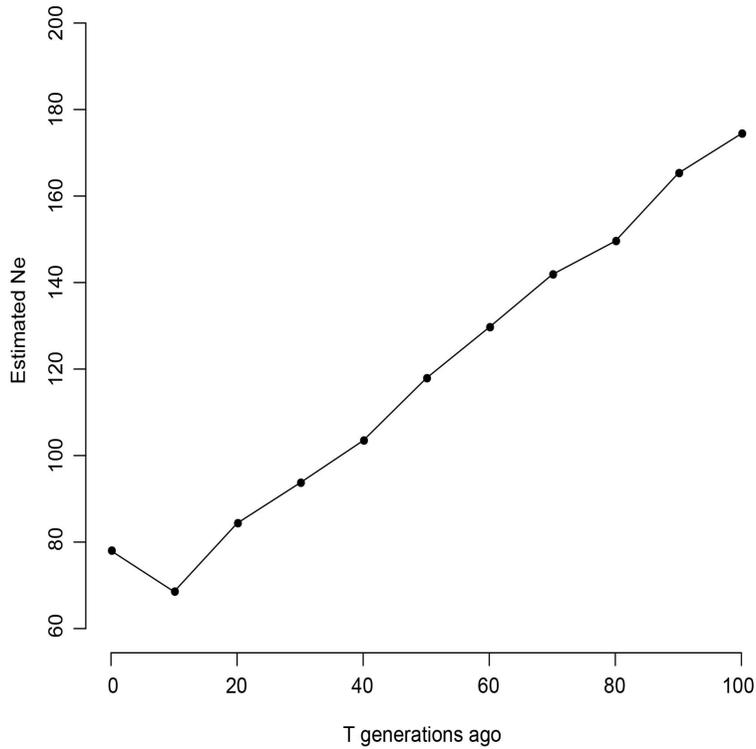


Figure 2.5 The historically estimated effective population size (N_e) against T generations ago. The historical N_e was obtained using $c = 1/2T$ where T is T generations ago and c is the recombination frequency. The current N_e was set to be 78 as the analysis. This figure showed the gradual decrease of effective population size N_e up to 10 generations ago and gradual increase up to the present.

2.5 Discussion

I used the two methods to estimate the effective population size (N_e): linkage disequilibrium (LD) - based and theta value - based methods. These two methods are possible in population-based data. The grounds of theta value-based methods was evolutionary theory. The theory is mainly based on the sequences divergence between two distance species. However, we used the theory in population-based data. Population-based data is based on the sampling and its theories. Although population-based evolutionary distance measure can be a controversy, I believe that enough sample size can guarantee the population-based evolutionary distance model and evolutionary distance can be defined at the population level. This means that the sequence data can be the polymorphic data such as single nucleotide polymorphisms (SNPs).

LD-based method is based on Sved's equation. Sved's equation denotes that the pattern of average LD decays according to the physical distance and its linkage distance. The pattern of decreasing LD with no kinship matrix was faster than that of with the kinship matrix of total chromosomes and with the kinship matrix of each chromosome. It decays with smooth shape at 0 – 6 Mb. Any significant discrepancies were not shown between predicted and observed LD decreases. The coefficients b and a (Equation 2.2) describe the expected correlation coefficient of LD and the

regression line passed through the y-axis, i.e. the linkage distance was effectively zero between SNP pairs, respectively.

The estimated N_e in Korean TB horses was smaller than those reported in previous studies of other countries. N_e is likely to represent a conceptual average of N_e over the period inferred from the marker distance (Toosi, Fernando et al. 2010). Corbin (2010) estimated the effective population size of Irish TB breeds to be 100, while Cervantes estimated N_e of TB in Spain to be 322 (Corbin, Blott et al. 2010, Cervantes, Goyache et al. 2011). This implies that Korean TB horses has been imported around the world and has weaker stability than other countries. Mares and sires of TB breeds were imported to Korea from various countries including United States, Australia, Russia, New Zealand, United Kingdom, India, Ireland, China, Japan, France and Canada. The imported horses are allowed to mate naturally in Korea (no artificial insemination, though). Thus I measured the N_e of TB breeds globally, which are genetically related to the imported Korean TB breeds.

The evolutionary model was Kimura-2-Parameter (K2P). K2P model assumes the different mutation rate between purine and pyrimidine bases, respectively. Mutation between purines is called transition and mutation between pyrimidines is called transversion. The transition/transversion ratio

(tr/tv ratio) assumption is the developed version of Jukes-Cantor evolutionary model (Steel and Fu 1995). Watterson theta estimator is widely-used theta estimator and population-based concept. It links the number of polymorphic sites to the effective population size (Felsenstein 2006). I calculated the substitution rate of SNPs and estimated N_e of TB horses in Korea using Watterson estimator.

Hayes et al. used the notion of chromosome segment homozygosity (CSH) to estimate the ancestral N_e . CSH is the probability that two segments of chromosomes of same size and location drawn at random from the population are from a common ancestor, without intervening recombination (Hayes, Visscher et al. 2003). Hayes et al. validated $c = 1/T$ using CSH where c is the recombination frequency and T designates T generations ago. Using this, the historical N_e was estimated and it showed the gradual decrease until 10 generations ago and increased slightly since 10 generations ago. The increase of N_e around the present implies that TB horses have been imported into Korea around the world.

The substitution rates of the regions containing SNPs was calculated and those averages were used to estimate N_e . Because this method is based on one species' population, it needs not to align the sequences and can calculate the genetic substitution rates. The multiple sequence alignment

(MSA) is based on the contiguous sequence data and evolutionary distance can be calculated via this alignment method. The result of this method includes insertion/deletion (indel) and substitution (SNP in population data) and disparateness between two species. However, the polymorphic data measures only differences between individuals of population and thus reflects the evolutionary distance in one species. It seems that the advent of Next-generation sequencing (NGS) can make it feasible to calculate mutation rate at the gene's level.

*This chapter was published in Genomics
& Informatics as a partial fulfillment of
Young-Sup Lee's Ph.D program.*

**Chapter 3. The usage of an SNP-SNP relationship matrix for
best linear unbiased prediction (BLUP) analysis using a
community based-cohort study**

3.1 Abstract

One of the complex traits' analyses is mixed-model. It includes best linear unbiased prediction (BLUP) and best linear unbiased estimation (BLUE). BLUP has been widely used to predict the random genetic effects of single nucleotide polymorphisms (SNPs). Traditionally, SNP-based BLUP analysis has been focused to predict the breeding values in animal or genetic values (human; synonym of breeding values in animal). My main purpose is to test so-called single nucleotide polymorphism – genomic best linear unbiased prediction (SNP-GBLUP) which can estimate the random genetic effects such as SNP effects directly and genetic values, also. The genomic – best linear unbiased prediction (G-BLUP) uses the genomic relationship matrix (GRM) and I used GRM to deduce the SNP-SNP relationship matrix (SSRM) to predict the SNP effects in SNP-GBLUP. I used R package “rrBLUP” to analyze the BLUP using human height trait. The SNP-GBLUP result was nearly identical to G-BLUP in the prediction of genetic values. However, I observed the discrepancies of SNP-GBLUP and SNP-BLUP which assumes IID (identical & independent distributed) between SNP markers. The predicted SNP effects as well as genetic values were very disparate between two BLUPs. I concluded that SSRM-based SNP-GBLUP can be an alternative of SNP-BLUP to predict the accurate

SNP effects.

3.2 Introduction

The best linear unbiased prediction (BLUP) in animal model have been used widely to predict and assess the livestock's commercial quality. The BLUP was originally proposed by Henderson (1975). Henderson' BLUP suggested that livestock could be rated according to the breeding values which could be decomposed of genetic components (Henderson 1975). This was the beginning of BLUP to analyze the complex traits of animal. And this method is now applicable to animals, plants and humans.

Many important human traits are complex traits and are moderately to highly heritable (de los Campos et al. 2013). Human height is one of the complex traits and Yang et al. asserted that roughly 45% of the genetic variance of human height can be explained by common single nucleotide polymorphisms (SNPs). Dark matter of the genome associated with human height can be explained by genic effects as well as gene-by-gene and gene-by-environment interactions (Maher 2008, Hindorff et al. 2009, Manolio and Collins 2009, Manolio et al. 2009). Although causal variants are not in complete linkage disequilibrium (LD) with the analyzing SNPs and SNP-SNP interactions do exist, which exerts the violation of IID (identical & independent distributed) assumption between SNP markers, the assumption has been used in the previous study. Single nucleotide polymorphism-best

linear unbiased prediction (SNP-BLUP) analysis has the merits to analyze the random effects like SNP effects (Koivula et al. 2012, Shen et al. 2013). However, I considered that its IID assumption should be altered into the usage of an SNP-SNP relationship matrix (SSRM). This is SNP-genomic best linear unbiased prediction (SNP-GBLUP) which uses SSRM. In fact, SSRM exploits the genomic relationship matrix (GRM) to deduce the formula of it.

To estimate the numerical values of genetic factors, I used BLUP which can analyze the genetic effects that reside in DNA information. The expression of diverse and complicated genetic factors, molecular biological networks and biochemical pathways can lead to phenotypic quantitative values and SNP can be the representative of those knowledge. Despite the complexity of biological knowledge about the genetic mechanisms of genes, the phenotypic values of complex traits of interest is concise and one-dimensional vector quantities. Thus BLUP can simply analyze the numerical values of SNP effects and those linked genic effects, and SNP can be the representative of genetic effects. The invisible but virtual SNP effects can be translated into numerical values.

BLUP is a standard method for predicting random effects and fixed effects of a mixed model. This method was originally designed in the field of

animal breeding to estimate breeding values of those but is now applicable to many areas of research (Piepho et al. 2008). The advent of DNA sequencing and SNP chip technology has made it possible to apply BLUP to predict the SNP effects. In fact, polygenic effects are the replacement of Quantitative Trait Loci (QTL) effects of traits of interest and in other words, randomly distributed genome-wide effects. BLUP model is the following form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad (3.1)$$

Where \mathbf{y} is the vector of phenotypic values and \mathbf{b} and \mathbf{u} are vectors of fixed and random effects, respectively. \mathbf{X} and \mathbf{Z} are the design matrices. Random effects and residual error effects are assumed to have a multivariate normal distribution as $\mathbf{u} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G}_u)$, $\mathbf{e} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{R})$ where \mathbf{MVN} indicates a multivariate normal distribution and $\mathbf{E}(\mathbf{y}) = \mathbf{Xb}$, $\mathbf{cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$. The solution using Maximum Likelihood Estimation (MLE) is the following linear system of equation:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_u^{-1} \end{bmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{pmatrix} \quad (3.2)$$

Where \mathbf{G}_u^{-1} is the inverse of the variance-covariance matrix of random effects, which indicate SNP-SNP variance-covariance matrix, or SNP-SNP relationship matrix (SSRM). I used the genomic relationship matrix to deduce SSRM. Genomic relationship matrix (GRM), \mathbf{G} , can be calculated

using R package “rrBLUP” (Endelman 2011).

My purpose was to compare SNP-GBLUP with G-BLUP and SNP-BLUP. G-BLUP uses GRM and SNP-BLUP assumes IID as mentioned earlier. The tabular method developed by Fernando and Grossman (1989) uses identity by descent (IBD) to calculate SSRM, which is complicated and vulnerable and can lead to incorrect calculations of SSRM (Fernando et al. 2014). However, GRM can be calculated by using SNP information. We proposed a simple method to calculate SSRM from GRM. Human height is a classical complex trait and has high heritability (~ 0.8) (Visscher 2008, Manolio et al. 2009, Yang et al. 2010). I analyzed height traits to test SNP-GBLUP. Despite the use of abundant SNPs, I did not achieve the satisfactory heritability. However, I identified that genetic values of SNP-GBLUP and G-BLUP was nearly identical and those are different from SNP-BLUP.

3.3 Materials and Methods

3.3.1 Data preparation

The data was Korean Ansan-Ansung cohort (Ver. 2.1.). This dataset was established to study Korean chronic diseases in the region of Ansan city and Ansung rural area. It constituted human aged 40-69 who had been the

residents of those regions at least 6 months. It was surveyed basically from 2000 to 2001 years and my study was based on the third Ansan-Ansung cohort dataset. The phenotypic data was height and fixed effect was sex (men and women). The Affymetrix Genome-wide Human SNP Array 5.0 was used to construct the single nucleotide polymorphism (SNP) data. The mean call rate was 99.01% and the genetic analysis result was 99.934% accuracy, which was proved by SNPstream UHT 12 plex. The number of genotyped SNPs was 352,228. They were filtered using Minor Allele Frequency (MAF < 0.01), Hardy-Weinberg Equilibrium (HWE $P < 0.00001$) and missing genotyping (missing > 0.2). It was conducted using PLINK, leaving 35,675 SNPs (Purcell et al. 2007). The individuals was 997.

Genomic relationship matrix (GRM) was calculated using R package “rrBLUP” with the option Expectation-Maximization (EM) algorithm (Endelman 2011). Then I calculated SNP effects, genetic values, genetic value variance and error variance using the restricted maximum likelihood (REML) method of the same package. The EM imputation algorithm was used for the GRM because we used the high-density SNPs. The REML method was chosen because we used smaller sample size than the number of markers instead of maximum likelihood (ML) method.

3.3.2 The derivation of statistical SNP-SNP relationship matrix (SSRM)

Henderson used maximum likelihood estimation (MLE) to derive BLUP equation which is the following form (Henderson 1975, Aldrich 1997):

$$\mathbf{f}(\mathbf{y}, \mathbf{u}) = \mathbf{g}(\mathbf{y}|\mathbf{u})\mathbf{h}(\mathbf{u}) = \mathbf{g}(\mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} | \mathbf{u})\mathbf{h}(\mathbf{u}) = \mathbf{g}(\mathbf{e})\mathbf{m}(\mathbf{u}) \quad (3.3)$$

Where $\mathbf{f}(\mathbf{y}, \mathbf{u})$ is the joint probability density function (joint pdf) of the BLUP model and it can resolved into $\mathbf{g}(\mathbf{e})\mathbf{m}(\mathbf{u})$ which is the product of pdf of residual errors and random effects.

I assumed the normality condition of the genetic values $\mathbf{Z}\mathbf{u}$ and random effects \mathbf{u} in the BLUP model. On these assumptions, because the variance-covariance matrix of genetic values, $\mathbf{Z}\mathbf{u}$, can be GRM, Equation (3.7) can be deduced as processed from Equation (3.3) ~ (3.6).

$$\mathbf{f}(\mathbf{y}, \mathbf{Z}\mathbf{u}) = \mathbf{g}(\mathbf{y}|\mathbf{Z}\mathbf{u})\mathbf{h}(\mathbf{Z}\mathbf{u}) = \mathbf{g}(\mathbf{e})\mathbf{m}(\mathbf{Z}\mathbf{u}) = \mathbf{g}(\mathbf{e})\mathbf{k}(\mathbf{u}) \quad (3.4)$$

Where $\mathbf{f}(\mathbf{y}, \mathbf{Z}\mathbf{u})$ is the joint probability density function (joint pdf) of the BLUP model and it can resolved into $\mathbf{g}(\mathbf{e})\mathbf{m}(\mathbf{Z}\mathbf{u})$ which is the product of pdf of residual errors and genetic values.

From Equations (3.3) and (3.4), we can infer Equation (3.5) and (3.6).

$$\mathbf{f}(\mathbf{y}, \mathbf{u}) = c \mathbf{e}^{-\frac{1}{2} \mathbf{e}' \mathbf{R}^{-1} \mathbf{e}} \mathbf{e}^{-\frac{1}{2} \mathbf{u}' \mathbf{G}_u^{-1} \mathbf{u}} \quad (3.5)$$

$$\mathbf{f}(\mathbf{y}, \mathbf{Zu}) = g(\mathbf{e})k(\mathbf{u}) = c \mathbf{e}^{-\frac{1}{2} \mathbf{e}' \mathbf{R}^{-1} \mathbf{e}} \mathbf{e}^{-\frac{1}{2} (\mathbf{Zu})' \mathbf{G}^{-1} \mathbf{Zu}} = c \mathbf{e}^{-\frac{1}{2} \mathbf{e}' \mathbf{R}^{-1} \mathbf{e}} \mathbf{e}^{-\frac{1}{2} \mathbf{u}' \mathbf{Z}' \mathbf{G}^{-1} \mathbf{Zu}} \quad (3.6)$$

$$\mathbf{G}_u^{-1} = \mathbf{Z}' \mathbf{G}^{-1} \mathbf{Z} \quad (3.7)$$

Where c is the constant and \mathbf{R} is the variance-covariance matrix of residual errors, \mathbf{G} is GRM and \mathbf{G}_u is the SSRM. By comparing Equation (3.5) and (3.6), we could derive the relationship between SSRM and GRM. Through these considerations, we can calculate SSRM for predicting SNP random effects.

3.3.3 Generalized least squares (GLS) for solving best linear unbiased prediction (BLUP)

I used R package “rrBLUP” to calculate GRM and predict the genetic values of cohort’s height complex traits in G-BLUP, SNP-BLUP and

SNP-GBLUP. In fact, “rrBLUP” package uses a generalized least squares (GLS) solution to solve BLUP equation. GLS in BLUP produces identical solution in MLE solution of BLUP and in fact GLS solution was derived from MLE in BLUP. The solution of fixed and random effects is the following form:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad \text{for } \mathbf{V} = \mathbf{Z}\mathbf{G}_u\mathbf{Z}' + \mathbf{R} \quad (3.8)$$

$$\hat{\mathbf{u}} = \mathbf{G}_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad (3.9)$$

3.3.4 Sherman-Morrison-Woodbury lemma (SMW lemma)

To use SSRM, as seen in Equation (3.7), the inverse of \mathbf{G}_u^{-1} should be calculated. Because the number of markers was larger than sample size, computation time of the inverse matrix can take very long (at least 1 days longer in the case of 10^4 order in marker size). Thus we used the Sherman-Morrison-Woodbury lemma (SMW lemma) (Sherman and Morrison 1950, Woodbury 1950). The formula is in the following form:

$$(\mathbf{A} + \mathbf{Y}\mathbf{G}\mathbf{Z}^*)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{Y}(\mathbf{G}^{-1} + \mathbf{Z}^*\mathbf{A}^{-1}\mathbf{Y})^{-1}\mathbf{Z}^*\mathbf{A}^{-1} \quad (3.10)$$

Where \mathbf{A} and \mathbf{G} are both invertible, and $\mathbf{A} + \mathbf{Y}\mathbf{G}\mathbf{Z}^*$ are invertible if and only if $\mathbf{G}^{-1} + \mathbf{Z}^*\mathbf{A}^{-1}\mathbf{Y}$ are invertible. Practically, we used \mathbf{A} as the identity matrix and the formula used in our analysis to find SSRM was the following form:

$$(\mathbf{I} + \mathbf{G}_u^{-1})^{-1} = (\mathbf{I} + \mathbf{Z}^T\mathbf{G}^{-1}\mathbf{Z})^{-1} = \mathbf{I} - \mathbf{Z}^T(\mathbf{G} + \mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z} \quad (3.11)$$

Where the notation was given the same as stated previous section. SSRM could be calculated from inverse matrix of it and the computation time of SSRM was less than 30 minutes in my analysis.

3.4 Results

3.4.1 SNP-GBLUP and genetic value prediction

Single nucleotide polymorphism – genomic linear unbiased

prediction (SNP-GBLUP) is based on genomic relationship matrix (GRM). Thus the genetic values in human in GRM-based G-BLUP were nearly identical to SNP-GBLUP, although some genetic values were significantly different. Figure 3.1 shows that the genetic values are nearly identical between G-BLUP and SNP-GBLUP but dissimilar to SNP-BLUP as shown in Table 3.1. Table 3.1 shows the individual's height phenotypic values and the genetic values prediction of three BLUPs. Figure 3.2 demonstrates the histogram of SNP effects in SNP-BLUP and SNP-GBLUP. Despite the normality of SNP effects of two BLUPs, the histogram was dissimilar. Thus we concluded that SNP-SNP relationship matrix (SSRM)-based SNP-GBLUP leads to different genetic values and SNP effects compare to SNP-BLUP. The estimated residual error variance was 21.39 in three BLUPs. The fixed effects were 166.6 (men) and 153.3 (women) in G-BLUP and SNP-GBLUP and 167.4 (men) and 154.1 (women) in SNP-BLUP.

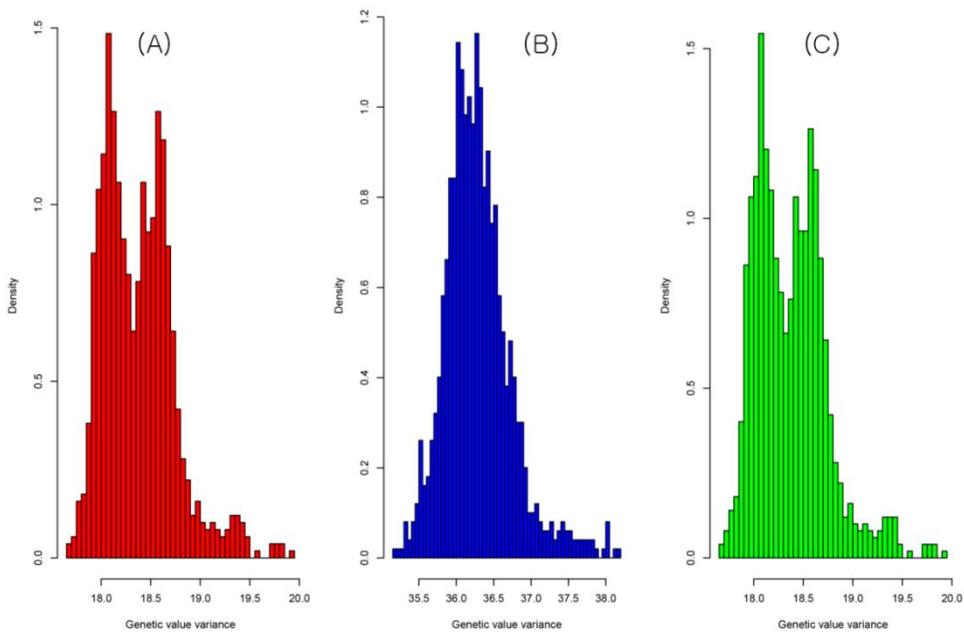


Figure 3.1. The histogram of genetic value variances of G-BLUP (left), SNP-BLUP (middle), and SNP-GBLUP (right). The shapes of histogram of the G-BLUP and the SNP-GBLUP were nearly identical. However, the shape of histogram of the SNP-BLUP was much disparate with other two BLUPs. These plots illustrate the similarity between G-BLUP and SNP-GBLUP in terms of genetic value variances.

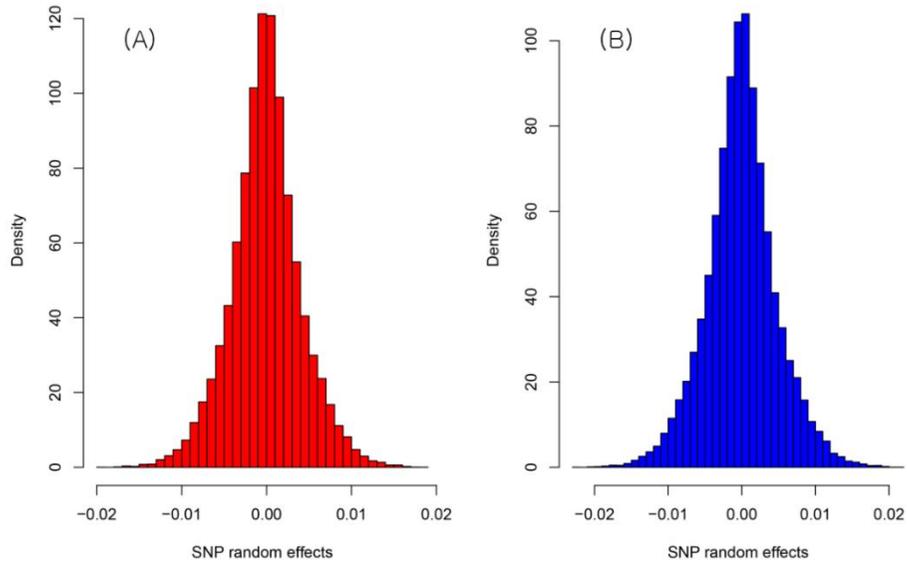


Figure 3.2. The histogram of the SNP effects of SNP-BLUP (left) and SNP-GBLUP (right). They were approximately distributed normally. However, the predicted genetic values and SNP effects as seen here were very dissimilar.

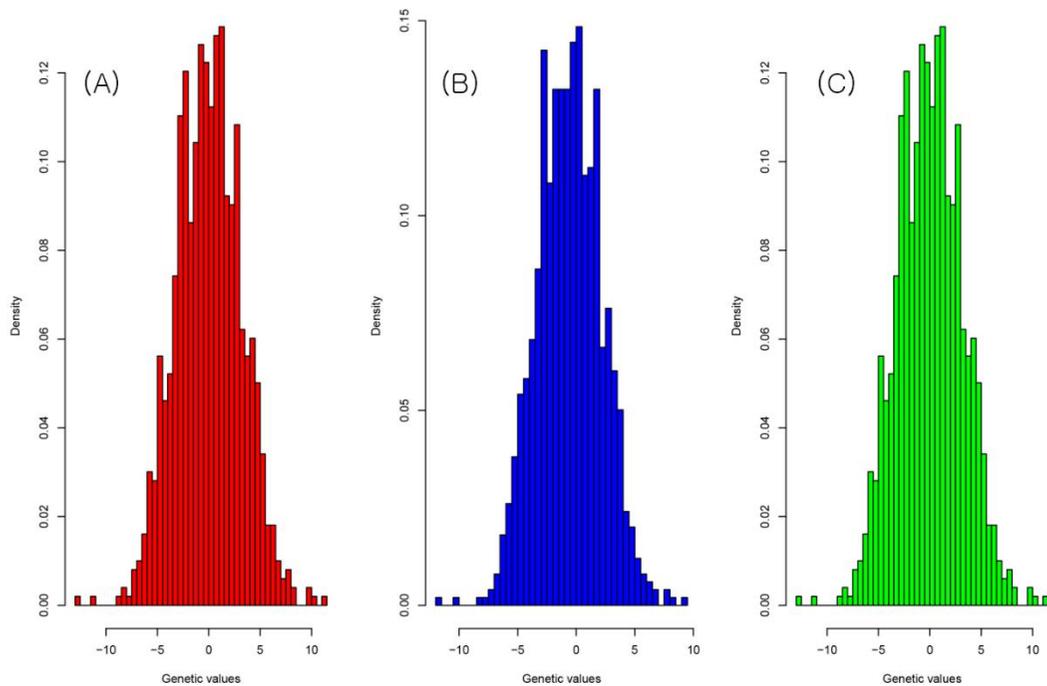


Figure 3.3. The histogram of the genetic values of G-BLUP (left), SNP-BLUP (middle), and SNP-GBLUP (right). Genetic values' distribution was similar to the normal distribution. As seen in this histogram, the genetic values of G-BLUP and SNP-GBLUP were nearly identical.

Table 3.1. The predicted genetic values of 1 ~ 57th human individuals according to the BLUP methods. The fixed effect was sex (men, women) and the phenotype was height. The results of genetic values were nearly identical between G-BLUP and SNP-GBLUP. However, that of SNP-BLUP was dissimilar.

ID	Height	sex	G-BLUP	SNP-BLUP	SNP-GBLUP
1	156	2	1.648225	0.911914	1.648219
2	170.5	1	1.483278	0.517589	1.483253
3	158.1	2	2.633289	1.538064	2.633256
4	157	2	2.68686	1.768591	2.68684
5	145.5	2	-4.0131	-4.092	-4.01308
6	161.9	2	4.469241	3.264376	4.469209
7	151.5	2	-0.87395	-1.28892	-0.87394
8	178	1	5.852538	4.40246	5.852492
9	158.3	2	2.504963	1.669991	2.504954
10	153.5	1	-6.7398	-6.38447	-6.73975
11	156	2	1.666214	0.745763	1.666191
12	174	1	3.948101	2.843941	3.948075
13	167.4	1	0.540129	0.040505	0.540138
14	156.1	2	1.46345	0.696352	1.46344

15	154	2	0.561605	-0.20975	0.561589
16	143.6	2	-5.29632	-5.15257	-5.29628
17	162.1	2	4.577998	3.353439	4.577965
18	153	2	0.157351	-0.33828	0.157358
19	159.7	2	3.193719	2.180766	3.193697
20	147.5	2	-3.39569	-3.38528	-3.39565
21	159	1	-4.56443	-4.61144	-4.56441
22	165.8	1	-0.81186	-1.23908	-0.81185
23	172	1	3.11896	2.130818	3.118938
24	171.5	1	2.717489	1.7975	2.717472
25	157	2	1.609231	0.8463	1.609225
26	145.4	2	-4.91548	-4.92232	-4.91545
27	164	1	-0.72305	-1.09665	-0.72304
28	155.4	1	-5.93336	-5.70991	-5.93332
29	162.1	2	4.770904	3.433848	4.77086
30	155.3	2	1.315899	0.686963	1.315901
31	175.3	1	4.465203	3.322152	4.465176
32	159.7	2	3.825022	2.804944	3.825001
33	175	1	4.696189	3.519737	4.696158
34	151.3	2	-1.21501	-1.65395	-1.215
35	157.1	2	2.2699	1.415556	2.269886
36	164.7	1	-0.91886	-1.32805	-0.91885

37	157.2	1	-5.21908	-5.01209	-5.21904
38	153.3	2	0.424467	-0.21125	0.424463
39	172.6	1	2.67969	1.77551	2.679677
40	168	1	0.902939	0.232925	0.902934
41	166.5	1	-0.53252	-1.12386	-0.53252
42	148.1	2	-2.87264	-2.95374	-2.87261
43	156.5	2	1.940328	1.190233	1.940322
44	151.2	2	-0.96373	-1.50465	-0.96373
45	161	2	3.697817	2.604254	3.697792
46	151.4	2	-0.7596	-1.27732	-0.7596
47	171	1	2.718487	1.759407	2.718465
48	150.5	2	-1.68384	-2.10578	-1.68383
49	149.8	2	-2.13181	-2.2962	-2.13178
50	143.4	2	-5.64759	-5.5306	-5.64755
51	153.7	2	-0.01213	-0.54869	-0.01213
52	143	2	-5.58965	-5.37865	-5.58961
53	160.5	2	3.823257	2.661133	3.823224
54	161.5	1	-2.77593	-3.10262	-2.77592
55	156.2	2	1.853024	1.139107	1.85302
56	157.6	2	2.481937	1.621284	2.481923
57	153.6	2	0.603728	0.009288	0.603728

3.4.2 SNP-SNP relationship matrix (SSRM)

Table 3.2 shows 1-8th SNP components of the \mathbf{G}_u square matrix by using Sherman-Morrison-Woodbury (SMW) lemma (Sherman and Morrison 1950, Woodbury 1950). The variance components (diagonal) were values close to 1. It suggests that the \mathbf{G}_u matrix can be interpreted as the SNP-SNP relationship matrix (SSRM). The off-diagonal terms of SSRM represent normalized covariance term which can be called relationship between SNPs. The exploitation of SSRM in SNP-GBLUP can lead to predict the SNP effects. Because the genetic values between G-BLUP and SNP-GBLUP, the SNP effects are more reliable to the case of SNP-BLUP. SNP effects was very disparate between IID and SSRM assumption in Table 3.3.

Table 3.2. The SNP-SNP relationship matrix of 1 ~ 8th SNPs using the relationship of $G_u^{-1} = Z^T G^{-1} Z$ and Sherman-Morrison-Woodbury lemma. G matrix is the numerator relationship matrix and G_u matrix is the SNP-SNP relationship matrix (SSRM).

9.66E-01	-1.63E-03	5.52E-04	-1.12E-03	2.20E-03	3.49E-04	2.29E-03	-5.66E-04
-1.63E-03	9.82E-01	-2.10E-03	3.94E-04	3.73E-04	5.74E-04	4.23E-04	5.69E-04
5.52E-04	-2.10E-03	9.59E-01	-2.93E-04	-4.02E-03	-1.23E-03	2.22E-03	3.34E-03
-1.12E-03	3.94E-04	-2.93E-04	9.81E-01	1.43E-03	1.62E-03	9.87E-05	-1.46E-04
2.20E-03	3.73E-04	-4.02E-03	1.43E-03	9.61E-01	-1.99E-03	1.47E-03	1.31E-03
3.49E-04	5.74E-04	-1.23E-03	1.62E-03	-1.99E-03	9.57E-01	1.88E-03	2.15E-04
2.29E-03	4.23E-04	2.22E-03	9.87E-05	1.47E-03	1.88E-03	9.60E-01	2.39E-03
-5.66E-04	5.69E-04	3.34E-03	-1.46E-04	1.31E-03	2.15E-04	2.39E-03	9.60E-01

Table 3.3. Each SNP effect in IID and SSRM cases. As shown below, the SNP effects was very disparate between two cases.

IID	SSRM
-1.21E-03	-1.36E-03
-5.46E-03	-6.73E-03
2.07E-03	2.65E-03
1.24E-03	1.49E-03
2.43E-03	2.80E-03
2.02E-03	2.58E-03
-2.90E-03	-3.53E-03

3.4.3 Estimated heritability

Using simple regression between the genetic values and phenotypic values of height, I estimated the narrow-sense heritability (h^2). The estimated heritability was 0.24 in G-BLUP and SNP-GBLUP and 0.20 in SNP-BLUP. According to Jian Yang, only 45% of genetic variance in human height can be explained by common SNPs (Yang et al. 2010). My main initiative was to compare SNP-GBLUP with other BLUPs. Thus I used rather small SNPs and our results about estimated heritability was not disappointing. I expect that

larger and well-selected SNPs will elevate the estimated heritability and more accurate genetic values and SNP effects can be obtained.

3.5 Discussion

3.5.1 The applicability of SNP-GBLUP

Genomic relationship matrix (GRM) contains the information between individuals' relationships. With the advent of SNP chips, it uses the SNP information to designate these relationships. Single nucleotide polymorphism – best linear unbiased prediction (SNP-BLUP) assumes being IID (independent and identically distributed) between random markers like SNPs. This assumption is good but needs to be corrected because it ignores the relationships between random markers. The factor of the relationships can be interactions, epigenetic mechanism and epistasis between genes. SNP-GBLUP considers this factor numerically in the model.

Probably, SNP effects through SNP-BLUP can be used in Bayesian BLUP. Some Bayesian BLUPs exclude low or nearly zero effects of SNPs (Schenkel et al. 2002). The classified SNP effects as being low and high can be used as Bayesian BLUP. Bayesian BLUPs have diverse merits which is shown in their assumptions (Verbyla et al. 2009). One of the merits is

Bayesian prior of SNP effects. It assumes different weights between SNPs. Another is the assumption of Bayesian parameter distribution. These two merits are inherited from Bayesian statistics. The further extension of SNP-BLUP to Bayesian BLUP will be expected to predict accurate SNP effects and genetic values.

3.5.1 GLS approach in BLUP and heritability

The heritability which is estimated from the regression between genetic values and phenotypic values must be smaller than generally accepted heritability. Narrow-sense heritability reflects the additive effects of quantitative trait loci (QTL). Because trait-associated all QTL might not be found in any prediction programs or QTL-related experiments, scientists usually use SNP information in BLUP. Because causal variants is not in complete linkage disequilibrium (complete LD), SNP information cannot contain all information of polygene or QTL. Also, incomplete LD might occur if causal variants have a lower minor allele frequency (lower MAF) than genotyped. The SNP effects can be treated as random, statistically and each SNP has a small effect on the trait of interest (Yang et al. 2010). However, I achieved better heritability than in common GWAS, i.e., only 5% of phenotypic variance in human height (Gudbjartsson et al. 2008).

Before the advent of SNP chip, numerator relationship matrix which uses the pedigree information has been widely used. The usage of this matrix is called traditional G-BLUP. Then SNP-chip technology can make it possible to generate design matrix of random marker effects and SNP-GBLUP. GRM and \mathbf{Z} matrix can lead to generate SSRM. GRM uses the frequency of genotyped SNPs and \mathbf{Z} matrix. In previous study, SNP-SNP covariance structures were assumed IID between markers. However, SSRM does not assume IID. SSRM contains the information of relationships of SNPs and can make it feasible to predict the SNP effects. The use of Sherman-Morrison-Woodbury lemma (SWM lemma) could reduce the computation time of SSRM. I expect that SSRM can predict the SNP effects that are linked to QTL more accurately.

The accuracy of genomic prediction depends on many factors such as population structure, genetic architecture of trait of interest, methodology of BLUP, degree of LD and distribution of random markers (Bennewitz et al. 2009, Meuwissen et al. 2009). This can be important for BLUP analysis. Next, there are two approaches in BLUP which can lead to identical solutions. These are generalized least squares (GLS) and maximum likelihood estimation (MLE) approaches. The R package “rrBLUP” uses GLS approach. GRM, the imputation of markers and prediction of genetic

values can be calculated in this package (Endelman 2011). The restricted maximum likelihood (REML) method is used in BLUP with rather smaller sample size than markers. The use of REML for variance component estimation avoids the small sample bias and seems to selection bias (Henderson 1975, McGilchrist and Yau 1995, Lark et al. 2006). Henderson (1986) asserted the use of REML in BLUP solver (Henderson 1975). Thus I used REML method schemes.

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

**Chapter 4. BLUP-based analysis using GWA candidate
markers improve GEBV in accord with narrow-sense
heritability.**

4.1 Abstract

The heritability estimated from best linear unbiased prediction (BLUP) has been a main problem because its estimate was lower than expected. This is called missing heritability problem. As a trial of resolving it, I introduced the genome-wide association study (GWAS) into BLUP. My data was eight pork quality traits of Berkshire pigs. GWAS detects the putative quantitative trait loci (QTL). The single nucleotide polymorphisms (SNPs) obtained from GWAS results with P-value < 0.01 was treated as significant SNPs and these SNPs were used BLUP analysis. The heritability as well as breeding values in Berkshire pigs were improved in this situation rather than when using total SNPs. The GWAS model was linear regression using PLINK and BLUP model was G-BLUP (total SNPs) and SNP-GBLUP (significant SNPs). The SNP-GBLUP uses the SNP-SNP relationship matrix (SSRM). I used SNP-GBLUP rather than G-BLUP in significant SNPs' cases because SNP-GBLUP improved the estimated heritability than G-BLUP in significant SNPs' analyses. The BLUP using preprocessing of GWAS can be one of the alternatives of solving missing heritability problem and it can improve the genomic estimated breeding values (GEBVs).

4.2 Introduction

Genome-wide association study (GWAS) tests each marker locus associated the traits of interest. It assumes normal distribution of phenotypic values. Its results can contain the beta effects (slopes in the association, roughly) and those P-values (Cantor et al. 2010, Bolormaa et al. 2011, Visscher et al. 2012). To analyze each single nucleotide polymorphism (SNP) P-value, I used the PLINK linear regression model(Purcell et al. 2007). And then I used single nucleotide polymorphism – genomic best linear unbiased prediction (SNP-GBLUP) in significant SNPs and genomic – best linear unbiased prediction (G-BLUP) in total SNPs.

Pork accounts for 50% of daily meat protein intake, globally (Davis and Lin 2005). Genetic selection using best linear unbiased prediction (BLUP) have resulted in a lot of successes improving pork quality parameters. A lot of studies have been dedicated to estimate the genetic parameters of pork quality traits to use selection programs (Sellier et al. 1998, Leeds 2005). After invention of BLUP model by Henderson around 1950, genetic selection have been accelerated the performance of the breeding programs of livestock animals. However, the missing heritability problem have been one of the obstacles in BLUP analysis. It states that heritability estimated from BLUP is scanty than expected. Total genotyped SNPs have been used generally but we used the significant SNPs. We combined the

GWAS and BLUP method to solve the missing heritability problem and partly solved it.

4.3 Materials and Methods

4.3.1 Ethics statement

The 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.

4.3.2 Data preparation

The sampled data was 702 Berkshire pigs (365 male, 204 female and 133 castrated male). The animals were raised with the same commercial diet from Dasan experimental farm in Namwon, Korea. Genomic DNAs of 702 individuals were genotyped using Illumina Porcine 60K SNP Beadchip (Illumina, San Diego, CA, USA) following the standard protocol. Total number of 44,345 genotyped SNPs were filtered using quality-control processes with MAF (< 0.05), HWE ($P < 0.001$) and missing data (> 0.01 missing) that resulted in 36,896 autosomal SNPs.

A total of 8 meat quality traits were used to analyze BLUP. The traits were carcass weight (CWT), back fat thickness (BF), intramuscular fat

content (fat), protein contents, Shear force (SF), water holding capacity (WHC) and color (L* and A*). Carcass weight was measured immediately after slaughter. BF and color were measured from the longissimus dorsi muscle between 10th and 11th rib. Intramuscular fat content (fat) was measured using a chemical fat extraction procedures. WHC (%) was measured as a difference between moisture content (%) and expressible water (EW; %). Shear force (SF) was measured using the Warner-Bratzler shear force meter (G-R Elec. Mfg. Co., USA). General indication of lightness and degree of green-redness of meat color were measured referred to MC_L [Minolta L, Commission Internationale de l'Eclairage (CIE) L* color space] and MC_A [Minolta A, Commission Internationale de l'Eclairage (CIE) a* color space], respectively. In each sex group, we standardized the phenotypic values to z-scores for GWAS.

4.3.3 Linear regression: GWAS

I used linear regression model in PLINK software (additive option) for the genome-wide association (GWA) study (Purcell et al. 2007). The data was preprocessed with sex-adjustment. The SNPs with P-value less than the level of 0.01 were selected for genome-wide significant SNPs.

4.3.4 BLUP solution

The mixed model including fixed effects and random marker effects to estimate GEBVs was SNP-GBLUP (significant SNPs; P-value < 0.01 in GWAS) and G-BLUP (total SNPs). The model of SNP-GBLUP was described in Chapter 3. I used R package “rrBLUP” to analyze BLUP (Endelman and Endelman 2014). I used SNP-GBLUP in trimmed cases because its estimated heritability was slightly better than G-BLUP of trimmed cases. This is the dissimilar result as shown in Chapter 2 (the narrow-sense heritability in SNP-GBLUP and G-BLUP was nearly identical).

4.4 Results

I analyzed genome-wide association study (GWAS) and chose significant single nucleotide polymorphisms (SNPs P-value < 0.01) associated with 8 pork quality traits. 859 (MC_L), 1,028 (CWT), 2,014 (Protein), 1,478 (BF), 2,580 (SF), 3,659 (Fat), 5,830 (WHC) and 3,210 (MC_A) SNPs were selected and involved in the BLUP analysis (Table 4.1). Generally, the results of SNP-GBLUP analyzed with significant SNPs mentioned above, have higher narrow-sense heritability than those of total SNPs' cases as shown in Table 4.1. On the contrary, SF, fat, MC_A and WHC cases did not achieve the satisfactory results that showed increasing

smaller than any other cases. I regarded that the reason was to fail finding the appropriate number of SNPs. Maybe, the traits may have more quantitative trait loci (QTL) regions than those predicted in $P\text{-value} < 0.01$ and thereof less stringent level of $P\text{-value}$ may be required in SF, fat, MC_A and WHC cases. Because the most crucial part of our analysis was choosing the number of SNPs, I considered that the criteria of $P\text{-value}$ in GWAS should be modified in SF, fat, MC_A and WHC to fulfill the estimated heritability and genomic estimated breeding values (GEBVs).

Table 4.1. The narrow-sense heritability of best linear unbiased prediction (BLUP) using total SNPs and trimmed SNPs (P-value < 0.01 in GWAS). The number of analyzed SNPs in trimmed cases are represented in the second row and the fixed effects designated as male, female, castrated male are represented in the 3~5th rows. The heritability (%) of trimmed highly significant SNPs (P < 0.01) was greater than that of total SNPs' cases in all traits. The method was single nucleotide polymorphism-genomic best linear unbiased prediction (SNP-GBLUP) in trimmed SNPs' cases and genomic best linear unbiased prediction (G-BLUP) in total SNPs' cases.

SNP- GBLUP	MC_ L	CW T	Protei n	BF	SF	Fat	WH C	MC_ A
					258	365		
# SNPs	859	1028	2014	1478	0	9	5830	3210
				25.2				
male	48.73	86.31	24.00	6	2.89	2.80	59.29	6.15
				23.0				
female	48.15	86.00	24.00	3	3.14	2.41	57.84	6.10
castrated				28.1				
male	48.59	85.26	23.86	0	2.51	3.51	60.48	6.35
mean	48.53	85.96	23.96	25.1	2.89	2.82	59.1	6.17

				5				
σ^*	2.82	5.49	0.75	5.22	0.7	1.17	3.05	1
h^2								
(trimmed;%)	32	24	42	37	29	39	47	35
h^2								
(total;%)	6	9	26	20	20	37	43	29

* σ is standard deviation

Figure 4.1 shows the plot of GEBVs and phenotypic values of 8 pork quality traits. The black dots refer to total SNPs' BLUP (G-BLUP) and colored dots refer to significant SNPs' BLUP (SNP-GBLUP). Because the slopes of colored ones was higher than those of black ones (in terms of linear regression coefficients), I concluded that the GEBVs and narrow-sense heritability was performed better in significant SNPs than total SNPs. Figure 4.2 and Figure 4.3 show Manhattan plot of SNP effects (in $-\log_{10}$ scale) across chromosomes. The plots crudely show the aggregates of SNPs on each chromosome and this may imply the putative QTLs. Specifically, Figure 4.2 shows the MC_L, CWT, protein, BF traits cases and Figure 4.3 shows the SF, fat, WHC and MC_A traits cases which showed increase of

heritability. Figure 4.4 indicates K-means clustering (K=4) of the phenotypic values and BLUP results of the Berkshire eight pork quality based on the 1st and 2nd discriminant functions. I used the R package “fpc” (Hennig 2010). These sorts of plots can assist the breeders who wish to select better-performed livestock animals.

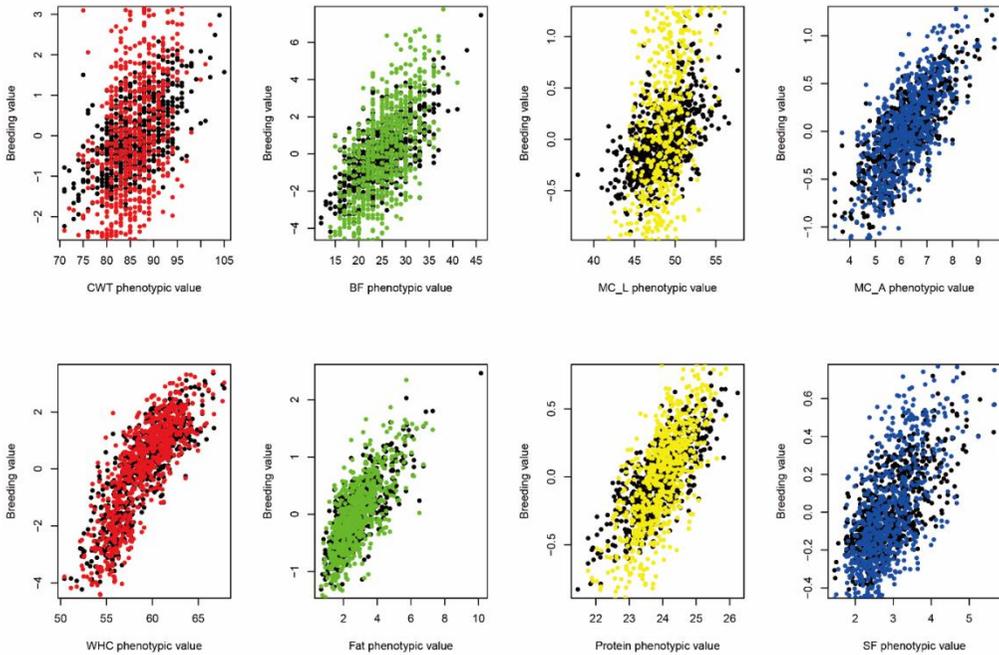


Figure 4.1. The plot of breeding values against 8 Berkshire pork quality traits. The black dots represent the results using total SNPs and colored dots represent the results using trimmed SNPs obtained from GWA study (P -value < 0.01). In view of heritability (considered the slopes of dots aggregates), the trimmed cases have better estimates than total SNPs' cases. Each phenotypes are the following: CWT (carcass weight), BF (backfat thickness), MC_L (Minolta L color), MC_A (Minolta A color), WHC (water holding capacity), SF (shear force).

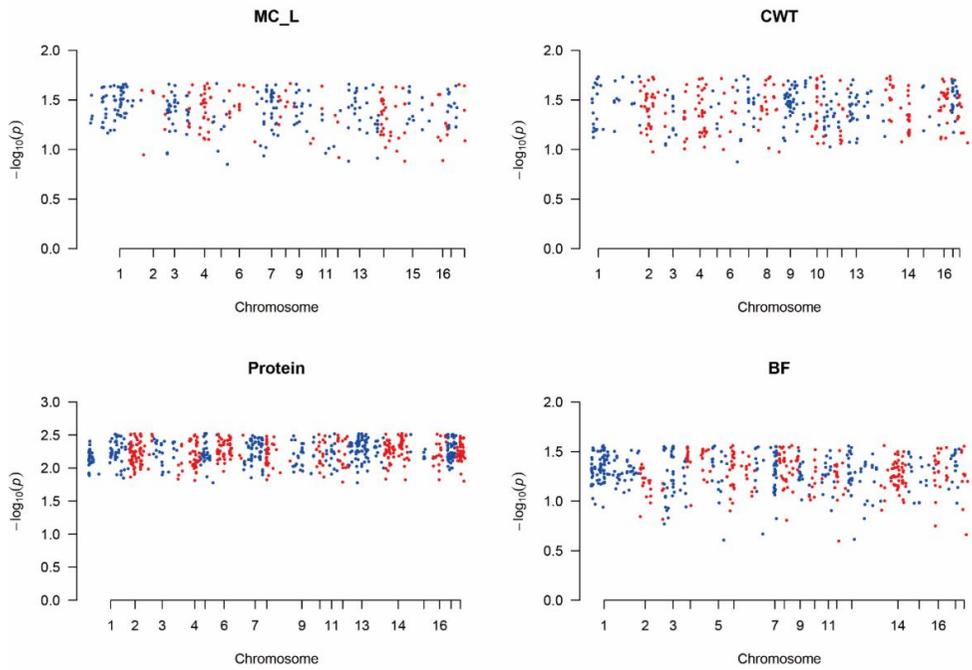


Figure 4.2. Manhattan plot of $-\log_{10}(P\text{-value})$ across chromosomes. The P-value was obtained from GWA study ($P\text{-value} < 0.01$). It shows the aggregates of SNP which may imply the putative quantitative trait loci (QTL) regions. The method used in BLUP analysis was SNP-GBLUP and the dots across each chromosome were used in BLUP analysis as trimmed cases.

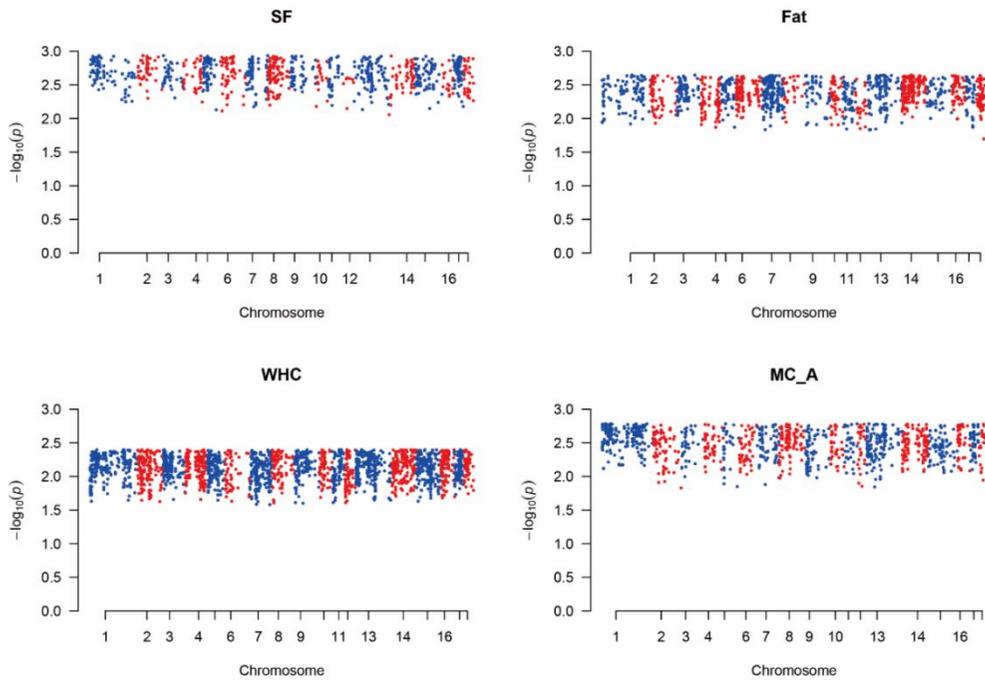


Figure 4.3. Manhattan plot of $-\log_{10}(\text{P-value})$ across chromosomes. The P-value was obtained from GWA study (P-value <0.01). It shows the aggregates of SNP which may imply the putative quantitative trait loci (QTL) regions. The method used in BLUP analysis was SNP-GBLUP and the dots across each chromosome were used in BLUP analysis as trimmed cases.

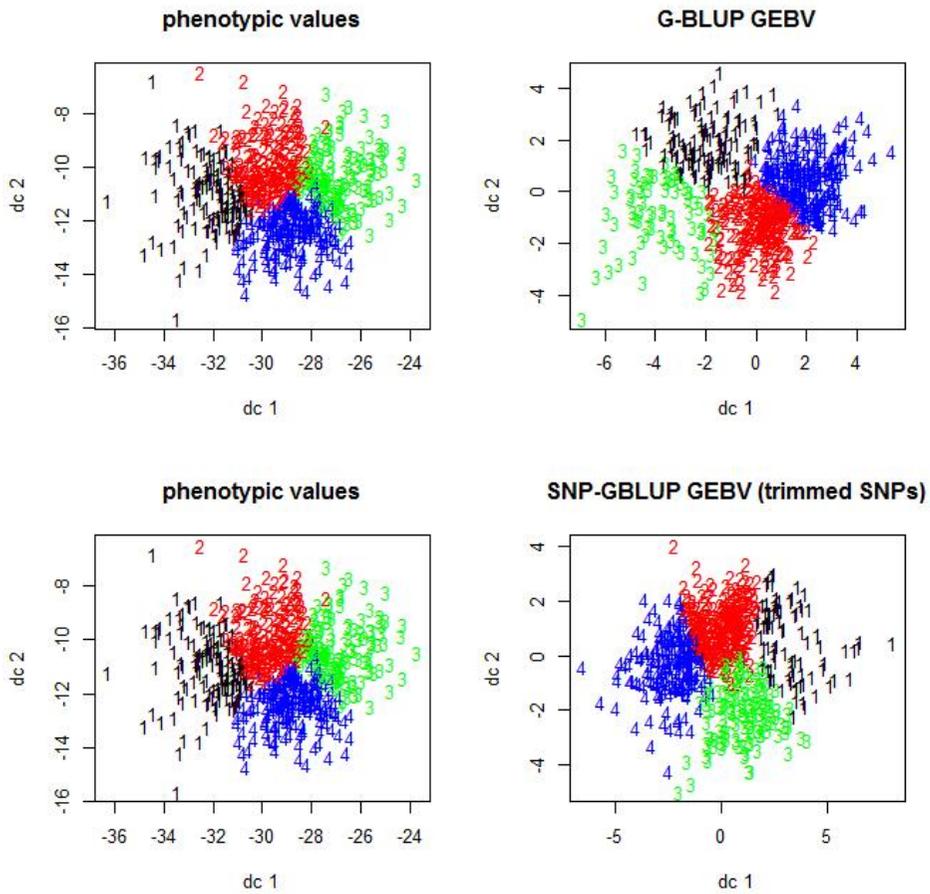


Figure 4.4 1st and 2nd discriminant functions of Berkshire 8 pork quality traits and corresponding GEBVs of total SNPs and trimmed SNPs. The method of BLUP of trimmed SNPs was SNP-GBLUP and that of total SNPs was G-BLUP.

4.5 Discussion

4.5.1 Genome-wide association study (GWAS)

Mapping of quantitative trait loci (QTL) has been used to detect genetic variation which is responsible for economically important traits in the field of livestock science. However, it has been difficult to identify genetic variation affecting complex traits, due to the low density of markers and its confidence interval of QTL mapping (Wang et al. 1999, Seaton et al. 2002, Collard et al. 2005). Genome-wide association study (GWAS) typically has been focused on the association between genetic variants and traits especially developed in human disease study (McCarroll and Altshuler 2007, Hardy and Singleton 2009). GWAS has been extended to exploit in domestic animals since genomic sequences and large scale of genomic variants of livestock had been available. GWAS can detect the causal variants which is responsible for economic traits underlying QTL. This was the basis of the study because GWAS can detect the putative QTL regions, causal SNPs. It may improve the genomic estimated breeding values (GEBVs) in best linear unbiased prediction (BLUP) analysis. Non-QTL regions' markers can be assumed to effect 0. Thus, this GWAS approach was the strategy which was assumed the zero effect of non-selected SNPs.

4.5.2 The application of GWAS to BLUP

As mentioned earlier, my strategy was to choose putative QTL SNPs and Zhang et al. reported that GWAS could improve the accuracy of genomic selection (GS) and they used GWAS and QTL database (DB) knowledge (Zhang, He et al.). They asserted that the superiority of BLUP|GA model can improve the GEBVs. However, because they used the QTL DB information, BLUP|GA model needed the QTL knowledge of livestock of interest. I further asserted that only GWAS results sorted by P-value could contain all the information we required to analyze BLUP. For sure, the results could contain false-positive discoveries. On the contrary, by using GWAS results, the BLUP analysis can be simplified and computational method because we did not require QTL DB. The estimated heritability of MC_L and CWT in our analysis were highly improved with 3~ 5 fold increase. This can be possible because of putative detection of associated QTL regions using GWAS.

I chose P-value rather than beta effects in GWAS results. This was because we considered that the significance (as P-value) is much more important in improving GEBVs rather than beta effects. The number of selected SNPs was 2~10% of total SNPs. As the number of QTL changes, the required SNPs which can fulfill the heritability varies. I adopted putative

QTL regions' markers with highly significant P-values. BLUP of significant SNPs was better than that of total SNPs in terms of heritability and GEBVs

4.5.3 Missing heritability problem

Missing heritability problem has been occurred in several association studies. In GWAS, complex diseases and human height have been one of the representatives of missing heritability problem. In BLUP, there have been always missing heritability problem (Manolio, Collins et al. 2009). Thus the beta effects in GWAS and GEBVs in BLUP could not be predicted with high accuracy. However, in my study, the application of GWAS to BLUP was a success in part although the number of SNPs can be a controversy to predict GEBVs better.

Genomic relationship matrix (GRM) is a statistically variance-covariance matrix which uses whole SNP information. Partial GRM which was used in my analysis is GRM with partly selected SNPs. The partial GRM was constructed using SNP information in part (P-value < 0.01 in GWAS) cannot matter because it can be a variance-covariance matrix. And SNP-SNP relationship matrix (SSRM) was calculated from GRM and \mathbf{Z} matrix.

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

Chapter 5. Combined Analysis of Fisher's theorem of natural selection and BLUP can expect the current selection coefficient of SNP.

5.1 Abstract

Milk-related traits (milk yield, fat and protein contents) is crucial to genomic selection of Holstein. It is also essential to find current selection trends of Holstein. However, finding the current selection trends have been ignored in the previous studies, although there have been various breeding studies. My approach was first to determine the single nucleotide polymorphisms' (SNPs) effect from best linear unbiased prediction (BLUP) of Holstein milk-related traits. Then I calculated the genetic variance of SNP from the effect. Using Fisher's fundamental theorem of natural selection, we predicted current relative selection coefficients. I assumed that the current selection trends could be determined using these selection coefficient of SNPs because population size is large in Holstein of Korea. Despite nearly 100% correlation of SNP effects and selection coefficients, selection coefficients can be used as current gene ontology of highly selected SNPs in the genes. Identified significantly selective SNPs with P-value < 0.01 (nearly top 1% SNPs) in all traits and P-value < 0.001 (nearly top 0.1%) in any traits was 14. They were PDE4B (phosphodiesterase 4B), STK40 (serine/threonine kinase 40), COL11A1 (collagen, type XI, alpha 1), EFNA1 (ephrin-A1), NTN4 (netrin 4), NSG1 (neuron specific gene family member 1), ESR1 (estrogen receptor 1), NRXN3 (neurexin 3), SPTBN1 (spectrin, beta, non-

erythrocytic 1), ARFIP1 (ADP-ribosylation factor interacting protein 1), MLH1 (mutL homolog 1), TMC7 (transmembrane channel-like 7), CPXM2 (carboxypeptidase X, member 2) and ADAM12 (ADAM metallopeptidase domain 12). These genes may take the responsibility of selection of Korean Holstein. Also, I found out that SNP effect might be the governing factor to determine selection coefficient of SNP rather than allele frequency. The selection coefficient of SNP is mathematically equivalent to $2 \times \text{SNP effect}$ under Hardy-Weinberg equilibrium (HWE).

5.2 Introduction

After breeding programs had been started in the 1960s, Holstein-Friesian cattle have been selected intensively during the last millennia, especially in the last five decades (Skjervold and Langholz 1964). Because Holstein cattle have been selected to produce more milk and better milk composition, the development of reproductive technologies such as artificial insemination, embryo transfer and pedigree evaluation of bulls and the like was very important. Especially, recent genomic selection have accelerated the selection progress (Hayes et al. 2009). This selection progress has increased the favorable allele frequency affecting the traits of interest and also increased neutral markers' allele frequency in linkage disequilibrium

(LD) with favorable alleles (genetic hitchhiking). The identification of genomic regions which is subject to selection is necessary but the study about the current selection trends have been short. This is because there have been no adequate mathematical model. I developed new formula which used Fisher's fundamental theorem of natural selection and BLUP. This is the general formula which uses phenotypic association to the genetic components like single nucleotide polymorphisms (SNPs).

Fisher's fundamental theorem of natural selection states that the rate of fitness increase of any organism at any time is equal to its genetic variance at that time (Hartl 1988). We could compute the current relative selection coefficient of SNPs based on the theorem and linear additive model. The linear additive model was best linear unbiased prediction (BLUP) and the predicted SNP effects using BLUP results could be used the genetic variance of those SNPs. This genetic variance computation was based on the population genetics. Then because Fisher's theorem can link the genetic variance in the linear additive model to selection coefficient of SNP. This selection coefficient is dependent on the phenotypic values as anyone considers this fact obviously. Not only natural selection but also artificial genomic selection must be traits-dependent and despite the human arbitrary choice of selection, the large population size can assure the allele frequency

change which can be expected in the selection coefficient. Additionally, although only highest breeding value-owned groups will be selected in F_1 generation, it seems that predicting the highest effects-owned SNPs' allele frequency change patterns will be very important and useful.

The name of selection coefficient in our mathematical proposal is “expected relative current selection coefficient”. “expected” means that it is the expected value in the F_1 generation. “relative” implies that it is dependent on the unit of phenotypic values. Thus the computed selection coefficient was recalibrated by the maximum absolute value. “current” means that it is predicted in the F_1 generation (Mendelian genetics notation).

The genes containing highly significant SNPs were obtained from Ensembl (Flicek et al. 2011). I considered that because SNPs were in LD with those-containing genes, the gene ontology research were available. I surveyed gene ontology containing the highly significant SNPs (top 1% or P-value < 0.01) which P-values were calculated from the normality assumption of selection coefficient.

5.3 Materials and Methods

5.3.1 Materials

Holstein cattle sample whose sex was female was collected in Korea and the traits were milk yield, fat and protein contents with parity 1. The number of Holstein individuals with phenotypic values was 462. Genomic DNAs were genotyped using Illumina 50K SNP Beadchip (Illumina, San Diego, CA, USA). Total number of 41,099 genotyped SNPs were imputed using BEAGLE version 4.0 (Browning and Browning 2009). Then the SNPs were filtered using minor allele frequency (MAF <0.05), Hardy-Weinberg equilibrium (HWE P-value < 0.001) and missing genotype (>0.1) and I excluded SNPs on the sex chromosome. The remaining SNPs were 37,854. A number of SNPs were filtered out because individuals with missing phenotypic values were excluded before filtering.

5.3.2 Prediction of selection coefficients in milk production traits

The prediction model of SNP effects was single nucleotide polymorphism – genomic best linear unbiased prediction (SNP-GBLUP) and the analysis tool was R package “rrBLUP” (Endelman 2011). The delicate information about SNP-GBLUP was mentioned in Chapter 3.

I used Fisher’s fundamental theorem of natural selection to calculate the selection coefficients of SNPs (Hartl 1988). Fisher’s theorem states that fitness change of any organism per unit time is equal to its genetic variance

at that time. I could calculate the genetic variance in the additive linear model of SNP-GBLUP and then computed the selection coefficients based on Fisher's theorem (Price 1972, Ewens 1989). The selection coefficient was computed using the following formula:

$$\begin{aligned}
 \sigma_a^2 &= \frac{d\bar{w}}{dt} = 2pq[p(w_{AA} - w_{AA'}) + q(w_{AA'} - w_{A'A'})]^2 \\
 &= 2pq \left[p * \left(-\frac{s}{2} \right) + q * \left(-\frac{s}{2} \right) \right]^2 \\
 &= \frac{pqs^2}{2}
 \end{aligned}
 \tag{5.1}$$

Where fitness per genotype is $(w_{AA}, w_{AA'}, w_{A'A'}) = (1-s, 1-s/2, 1)$ and s is the selection coefficient symbol.

$$\begin{aligned}
 &s_j^2 \\
 &= \frac{2\sigma_{a,j}^2}{p_j q_j} \text{ (according to Fisher's theorem (by Equation (5.1) and if } dt \\
 &= 1)
 \end{aligned}$$

$$= \frac{2 * \text{var}(Z_{ij} * u_j)}{p_j q_j} \text{ according to } \sigma_{a,j}^2 = \text{var}(Z_{ij} * u_j) \quad (5.2)$$

Where i represents i^{th} individual, j represents j^{th} marker or SNPs, Z_{ij} represents the i^{th} individual and j^{th} SNP's coding. u_j represents the SNP effect. The additive genetic variance computation is data-driven method which uses the Z matrix.

Equation (5.2) is based on the Fisher's theorem (Frank and Slatkin 1992). The relative selection coefficient of a given locus is in the range of from -1 to 1 because we recalibrated the selection coefficient with maximum absolute value. I presumed the normality of relative selection coefficient, computed the P-values and set the criteria of highly selective SNPs as P-value < 0.01 (nearly top 1% SNPs). Especially, if the SNP markers are under HWE in current generation, $s^2 = 4u^2$ according to $\text{var}(Z_i) = 2pq$. If we pay heed on the expected relationship of the sign between selection coefficient s and SNP effect u , we can infer that $s = 2u$ (sign is the same).

$$s_j = 2u_j \text{ (if HWE in current generation)}$$

(5.3)

5.3.3 Characterization of candidate genes under highly selective regions

The genes containing highly selective SNPs (P-value < 0.01) were used to analyze gene ontology. The gene ontology program was ClueGo plugin of Cytoscape program (Bindea, Mlecnik et al. 2009). The gene catalog was retrieved from Ensembl site (www.ensembl.org). I used the default parameter expect 2 minimum number of genes in GO term/Pathway selection in ClueGo plugin and I used Benjamini-Hochberg P-value correction (Benjamini and Hochberg 1995).

5.4 Results

5.4.1 Highly selective single nucleotide polymorphisms (SNPs)

The mean and standard deviation of Holstein milk yield, fat and protein records for parity 1 were 8845; 1425, 339; 58 and 283; 44, respectively. The fixed effects (season) of milk yield, fat and protein (kilograms) were (8655, 8847, 8935, 8907), (325, 342, 344, 343) and (275, 286, 286, 283) for spring, summer, autumn, and winter, respectively. The narrow-sense heritability estimates of the milk yield, fat and protein using

SNP-GBLUP method were 0.39, 0.45 and 0.40, respectively.

Figure 5.1 shows flow chart of my analysis which is described by theory and method. Figure 5.2 shows plot of relative selection coefficient against SNP effect of SNPs. We identified that the selection coefficient is governed by SNP effect in my data and because the factor to determine selection coefficient of SNPs was both allele frequency and SNP effect, governance of selection coefficient by SNP effect may be obvious in general cases, also. The sign of selection coefficient was inferred from the sign of SNP effect. Figure 5.3 indicates the gene ontology of milk protein content. The genes which contained nearly top 1% SNPs in selection coefficient were selected and I analyzed these gene ontology. The milk yield and fat cases had no great information in gene ontology analyses.

Table 5.1 shows the illustration about F_1 generation's expected allele frequency change under linear additive model. It demonstrates that allele frequency change can be predicted through SNP effect. Table 5.2 shows very highly selective SNPs and the genes containing those (any P-value < 0.001; nearly top 0.1% SNPs). The genes containing very highly selective SNPs with P-value < 0.01 (nearly top 1% SNPs) in all traits and P-value < 0.001 (nearly top 0.1%) in any traits were ESR1 (estrogen receptor 1), NRXN3 (neurexin 3), SPTBN1 (spectrin, beta, non-erythrocytic 1), ARFIP1 (ADP-

ribosylation factor interacting protein 1), MLH1 (mutL homolog 1), PDE4B (phosphodiesterase 4B), STK40 (serine/threonine kinase 40), COL11A1 (collagen, type XI, alpha 1), EFNA1 (ephrin-A1), NTN4 (netrin 4), NSG1 (neuron specific gene family member 1), TMC7 (transmembrane channel-like 7), CPXM2 (carboxypeptidase X, member 2) and ADAM12 (ADAM metallopeptidase domain 12). I inferred the sign of relative selection coefficient using sign of SNP effect in the Table 5.2. The positive sign of SNP effect directs positive sign of selection coefficient and vice versa because allele frequency coded in '2' in Z matrix will be expected to increase in next generation in the large population pool.

Table 5.1. F_1 generation's (the next in the current generation) allele frequency change according to the single nucleotide polymorphism (SNP) effect under linear additive model. We assumed the Hardy-Weinberg equilibrium (HWE) in P (Parental) generation and depicted the SNP effect as selection coefficient according to Equation (4). The allele coded as "2" assumed to be $A'A'$ and u denoted SNP effect*.

Allele frequency	0.25 (AA)	0.5 (AA')	0.25 (A'A')
Fitness change	$1-2u$	$1-u$	1
SNP effect	Change of allele frequency		
0.5	0	0.5	0.5
0.25	0.24	0.5	0.26
0.05	0.17	0.5	0.33

*Note that the SNP effect is sensitive to the unit of phenotypic values and we assumed that the SNP effect would be the selection coefficient*2 And App indicates approximately.

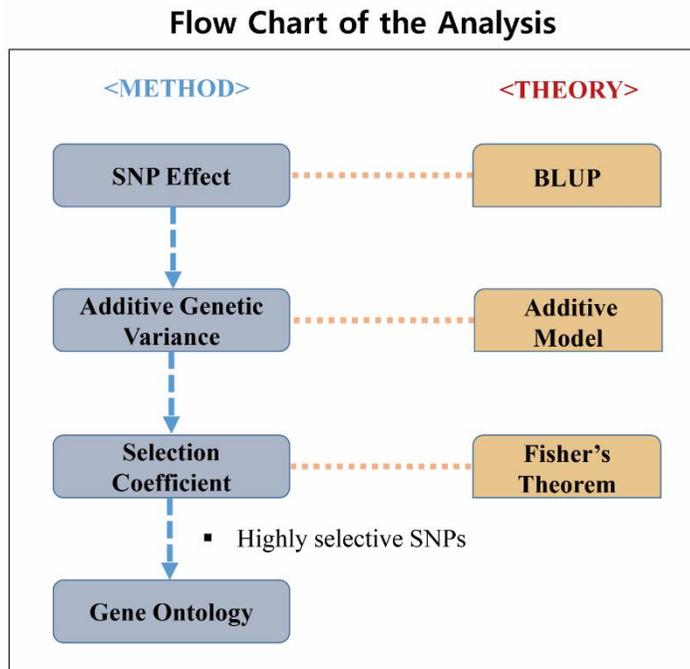


Figure 5.1. The flow chart of the analysis. It is categorized as method and theory. The SNP effect, additive genetic variance and selection coefficient were sequentially computed. The gene ontology was performed using Cytoscape program in ClueGo plugin.

Table 5.2. Highly selective SNPs with any P-value <0.001 (top 1% SNPs) in the analysis of milk yield, fat and protein phenotypes and the genes containing it. P-value was computed under the normality assumption of relative selection coefficient. The gene catalog was retrieved from Ensembl server.

CHRO	MOSO	SNP	SNP POSITI ON	M.R.S. C ¹	F.R.S.C 2	P.R.S. C ³	MILK ⁴	FAT ⁵	PROTE IN ⁶	Gene name
1		ARS- BFGL- NGS- 29472	138,65 0,000	0.444	0.769	0.530	0.0134 55	0.0000 53	0.0117 69	<i>KCNH8</i>
2		ARS- BFGL- NGS- 107330	116,82 9,914	0.653	0.305	0.783	0.0005 63	0.0639 54	0.0004 02	<i>DAWI</i>
3		BTA- 99819- no-rs	79,508, 402	-0.742	-0.698	-0.842	0.0000 98	0.0001 84	0.0001 42	<i>PDE4B</i>
3		BTB- 011554 79	79,378, 528	-0.754	-0.841	-0.957	0.0000 78	0.0000 09	0.0000 18	<i>PDE4B</i>
3		ARS- BFGL- NGS- 31953	79,480, 234	-0.757	-0.693	-0.866	0.0000 73	0.0002 05	0.0000 94	<i>PDE4B</i>
3		Hapma p39300 -BTA- 99855	79,333, 053	-0.803	-0.881	-0.994	0.0000 28	0.0000 04	0.0000 09	<i>PDE4B</i>
3		ARS- BFGL-	110,07 8,547	0.555	0.742	0.672	0.0028 13	0.0000 91	0.0020 19	<i>STK40</i>

	NGS- 102149									
3	BTB- 015823 89	40,625, 026	0.570	0.651	0.651	0.0022 35	0.0005 24	0.0026 82	COL11A1	
3	ARS- BFGL- NGS- 112442	40,588, 026	0.577	0.619	0.655	0.0020 23	0.0009 20	0.0025 40	COL11A1	
3	ARS- BFGL- NGS- 64215	15,525, 599	-0.622	-0.759	-0.802	0.0008 99	0.0000 54	0.0002 72	EFNA1	
4	BTB- 001722 04	31,172, 819	0.326	0.690	0.524	0.0524 24	0.0002 56	0.0126 14	RAPGEF5	
5	Hapma p53993 - rs29024 740	60,373, 086	0.499	0.678	0.567	0.0064 41	0.0003 16	0.0076 94	NTN4	
5	BTB- 002398 12	121,13 5,969	0.582	0.229	0.723	0.0018 52	0.1270 36	0.0009 90	MOV10L1	
6	ARS- BFGL- NGS- 4767	107,18 6,270	0.515	0.806	0.578	0.0051 27	0.0000 24	0.0067 13	NSG1	
6	Hapma p38694 -BTA- 76566	61,591, 415	0.533	0.401	0.783	0.0039 43	0.0221 42	0.0004 02	APBB2	
9	BTB- 004046 39	90,037, 629	1.000	0.605	0.708	0.0000 00	0.0011 72	0.0012 20	ESR1	
9	Hapma p47116 -BTA- 84683	90,002, 616	0.776	0.624	0.514	0.0000 54	0.0008 49	0.0141 35	ESR1	
10	ARS- BFGL-	81,459, 970	-0.790	0.060	-0.842	0.0000 37	0.3896 80	0.0001 42	GALNTL1	

	NGS- 113766									
10	ARS- BFGL- NGS- 82682	89,774, 836	0.698	0.095	0.775	0.0002 49	0.3223 71	0.0004 52	SPTLC2	
10	ARS- BFGL- NGS- 110578	91,602, 885	0.628	0.314	0.771	0.0008 63	0.0580 33	0.0004 83	NRXN3	
10	ARS- BFGL- NGS- 3900	89,804, 719	0.689	0.087	0.759	0.0002 96	0.3377 59	0.0005 83	SPTLC2	
11	ARS- BFGL- NGS- 51235	37,228, 325	0.627	0.368	0.835	0.0008 89	0.0326 57	0.0001 77	SPTBN1	
13	ARS- BFGL- NGS- 90758	35,352, 877	0.201	0.630	0.212	0.1594 41	0.0007 62	0.1850 64	JCAD	
13	Hapma p60259 - rs29016 362	34,887, 980	0.067	0.627	0.112	0.3732 41	0.0008 04	0.3196 34	SVIL	
13	Hapma p49926 -BTA- 24453	21,167, 068	-0.226	-0.740	-0.349	0.1264 67	0.0000 81	0.0650 57		
15	ARS- BFGL- NGS- 107160	75,065, 222	0.160	0.636	0.418	0.2143 12	0.0006 85	0.0371 64	ACS	
17	ARS- BFGL- NGS- 11818	4,393,2 29	0.745	0.449	0.775	0.0001 00	0.0121 79	0.0004 52	TRIM2	
17	BTB- 006687	4,827,0 67	0.707	0.670	0.775	0.0002 10	0.0003 71	0.0004 52	ARFIP1	

	97									
17	ARS- BFGL- NGS- 77442	63,480, 469	0.576	0.413	0.759	0.0020 35	0.0190 96	0.0005 76	<i>IQCD</i>	
21	ARS- BFGL- NGS- 104549	57,731, 221	0.689	0.291	0.738	0.0002 96	0.0730 74	0.0007 94	<i>SLC24A4</i>	
22	Hapma p38236 -BTA- 55228	10,502, 283	0.624	0.625	0.926	0.0009 25	0.0008 26	0.0000 37	<i>MLH1</i>	
23	Hapma p55007 - rs29021	13,484, 531	0.361	0.617	0.380	0.0362 46	0.0009 45	0.0524 79	<i>KIF6</i>	
25	986 ARS- BFGL- NGS- 93374	17,040, 004	0.602	0.654	0.906	0.0013 51	0.0004 97	0.0000 52	<i>TMC7</i>	
26	ARS- BFGL- NGS- 19663	43,933, 332	0.931	0.691	0.919	0.0000 02	0.0002 49	0.0000 41	<i>CPXM2</i>	
26	ARS- BFGL- NGS- 110497	45,870, 133	0.542	0.620	0.604	0.0034 40	0.0008 95	0.0049 04	<i>ADAM12</i>	
26	ARS- BFGL- NGS- 30392	44,539, 739	-0.526	-0.750	-0.464	0.0041 32	0.0000 65	0.0225 10	<i>LHPP</i>	
26	ARS- BFGL- NGS- 30060	45,983, 109	0.673	0.606	0.805	0.0003 94	0.0011 41	0.0002 82	<i>ADAM12</i>	
28	ARS- BFGL- NGS-	7,138,1 32	0.378	0.614	0.318	0.0300 58	0.0009 97	0.0878 74	<i>SLC35F3</i>	

1Milk yield Relative Selection Coefficient

2Fat Relative Selection Coefficient

3Protein Relative Selection Coefficient

4 Ensembl Gene ID

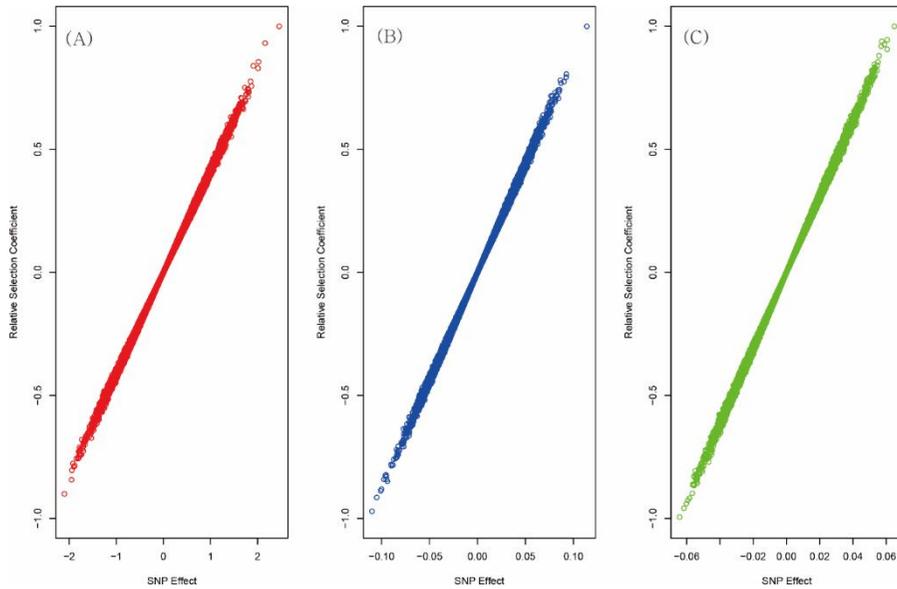


Figure 5.2. Plot of selection coefficient against SNP effects. The phenotypes were milk yield (A panel), fat (B panel) and protein content (C panel). It was estimated using single nucleotide polymorphism-genomic best linear unbiased prediction (SNP-GBLUP) and Fisher’s fundamental theorem of natural selection. The plot shows that the SNP effect is the major factor to determine the selection coefficient in Holstein dataset.

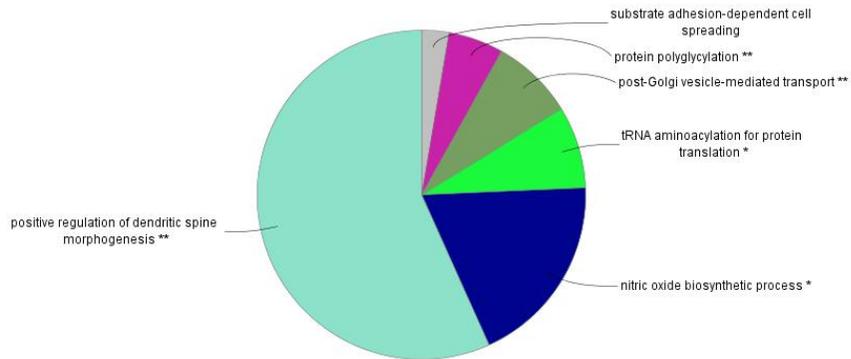


Figure 5.3. Pie chart of gene ontology of the significant genes which contain highly selective SNPs in milk protein trait. I selected the genes containing SNPs with P-value < 0.01 (nearly top 1% SNPs) and performed gene ontology. The condition was the default value except setting 2 minimum number of genes in the GO Term/Pathway selection item. Specially, positive regulation of dendritic spine morphogenesis was the most significant gene ontology. Dendritic spine morphogenesis is important in synaptic development and plasticity of the mammalian brain. I expect that this ontology can be the great concerning in Holstein breeding science because the related genes allele frequency may change drastically. The other two traits (milk yield, fat) did not have any significant gene ontology.

5.4.2 Gene ontology of highly selective SNPs in milk protein

The highly selective SNPs in protein case were analyzed to find gene ontology and those were dendritic spine morphogenesis, nitric oxide biosynthetic process and so on. Dendritic spine morphogenesis was the most highly significant gene ontology. Dendritic spine is the major site of excitatory synaptic transmission in the mammalian brain and is very crucial in synaptic development and plasticity generation (Penzes et al. 2003). The specific genes which yields milk protein and are related to dendritic spine morphogenesis, may be an important target of future genomic selection trends in Korean Holstein.

5.5 Discussion

5.5.1 SNP-GBLUP and selection

SNP-GBLUP can predict SNP effects directly by assigning SNP-SNP relationship matrix (SSRM). Accurate estimation of SNP effects is rudimentary to predict the selection coefficient and thus we used SNP-GBLUP rather than SNP-BLUP which assumes being IID (independent and identically distributed) between markers. The accurate prediction of SNP effect as well as sample size would be very important parameter to predict the allele frequency change of each SNP in the species population.

5.5.2 Fisher's fundamental theorem of natural selection and BLUP

One of Fisher's contributions to population genetics was a fundamental theorem of natural selection. The theorem states that the change of average fitness can be linked to markers' genetic variance in natural population. It can give lights to selection theory and subsequently breeding science (Frank and Slatkin 1992). Although this theorem is about the natural population, it can be viewed the artificial genomic selection in one generation as natural selection in one generation if the selection coefficient value is the predicted in next generation. And the linear additive model was best linear unbiased prediction (BLUP). Figure 5.2 shows that the larger SNP effects, the greater selection coefficients are. This finding that the selection coefficients are nearly proportional to the SNP effects, matches our common sense. Though allele frequency is the factor determining selection coefficient, those contributions were a little and thus the breeders of genomic selection can determine the allele frequency change in future generation. This reflects that if there were shortage of specific alleles and it might be a greatly important, breeders can increase the allele frequency of interests and gain the specific alleles in the population.

5.5.3 Sign of selection coefficient

Although the sign of SNP's selection coefficient is not explicit, the sign of SNP effect is definite. We inferred the sign of selection coefficient from that of SNP effect. The positive sign of selection coefficient reflect that of SNP effect because the frequency of allele coded as '2' would increase if sign of SNP effect were positive. And if the sign of it were negative, the situation would be vice versa. Reflected from this inference, the best breeding values' livestock would increase the '2'-coded allele frequency and decrease the '0'-coded allele frequency in the large population, obviously.

5.5.4 Features of my study

The best attributes of the study was first, that I found that SNP effect in BLUP model is equivalent to selection coefficient in scale and second, I used the Fisher's theorem and SNP-GBLUP. We predicted the SNP effects and SNP's selection coefficients from the theorem and SNP-GBLUP.

General Discussion and Conclusion

Using genomic information implied from mammalian SNP and CNV data, I tried to find out the implications of genetic markers and those genetic aspects. This can be important for finding causal variants.

The first was to estimate the effective population size (N_e) of Korean Thoroughbred horses (TB horses). In Chapter 2, I estimated N_e of TB horses using two approaches: linkage disequilibrium and Kimura 2-Parameter model. The effective population size is an important parameter to comprehend the genetic diversity of the population. There was no distinct difference in effective population size estimates between two approaches. The second was to estimate the marker effects of SNPs in BLUP. In Chapter 3, I introduced the novel concepts of SNP-SNP Relationship Matrix (SSRM) and estimated SNP effects in human height using BLUP model and SSRM. This one-step method can be called SNP-GBLUP [Single nucleotide polymorphism-genomic best linear unbiased prediction]. SNP-GBLUP uses SSRM and discards IID (independent & identically distributed) assumption between genetic markers (typical assumption in the association study). In chapter 4, I tried to solve “missing heritability problem” in BLUP. The phenotypic data was Berkshire pork quality traits which are crucial in evaluating the grade. I used genome-wide association study (GWAS) to find

SNPs in putative QTL regions. Probably, the regions can be the great factors to determine the phenotypic values and thus the SNPs in these regions can be expected to have large effects to determine the genomic estimated breeding values (GEBVs). Thus like Bayes C π , the null effect of markers was assumed in the analysis. The results were satisfactory in terms of heritability estimates and GEBVs.

Third, the selection coefficient was investigated. In chapter 5, I tried to determine the next generation' selection coefficient using Fisher's fundamental theorem of natural selection and BLUP. The selection coefficient was computed at the polymorphism level of SNP and it is a relative and expected value. I used the Korean Holstein data which comprises of milk-related traits like milk yield, fat and protein content. After the selection coefficient of SNPs had been computed, the gene ontology (GO) was surveyed. GO using the genes containing highly selective SNPs showed that milk protein – associated genes [PDE4B, GALNTL1, SPTLC2, NRXN3, TMC7, etc.] had significant GO terms. The most significant GO terms clustered with expected highly selective genes was dendritic spine morphogenesis. Dendritic spine morphogenesis is the major sites of excitatory synaptic transmission in the mammalian brain and is crucial in synaptic development and plasticity. The expected highly selective genes

which are related to milk protein production may undergo a rapid allele frequency change and may be an important target of future artificial selection trends in Holstein cattle, accordingly.

REFERENCES

Aldrich, J. (1997). "RA Fisher and the making of maximum likelihood 1912-1922." Statistical Science **12**(3): 162-176.

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

Bennewitz, J., T. Solberg and T. Meuwissen (2009). "Genomic breeding value estimation using nonparametric additive regression models." Genetics Selection Evolution **41**(1): 20.

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

Bolormaa, S., B. Hayes, K. Savin, R. Hawken, W. Barendse, P. Arthur, R. Herd and M. Goddard (2011). "Genome-wide association studies for feedlot and growth traits in cattle." Journal of Animal Science **89**(6): 1684-1697.

Browning, B. L. and S. R. Browning (2009). "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals." The American Journal of Human Genetics **84**(2): 210-223.

Cantor, R. M., K. Lange and J. S. Sinsheimer (2010). "Prioritizing GWAS results: a review of statistical methods and recommendations for their application." The American Journal of Human Genetics **86**(1): 6-22.

Cervantes, I., F. Goyache, A. Molina, M. Valera and J. Gutiérrez (2011). "Estimation of effective population size from the rate of coancestry in pedigreed populations." Journal of Animal Breeding and Genetics **128**(1): 56-63.

Cho, Y. S., M. J. Go, Y. J. Kim, J. Y. Heo, J. H. Oh, H.-J. Ban, D. Yoon, M. H. Lee, D.-J. Kim and M. Park (2009). "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits." Nature Genetics **41**(5): 527-534.

Collard, B., M. Jahufer, J. Brouwer and E. Pang (2005). "An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection

for crop improvement: the basic concepts." Euphytica **142**(1-2): 169-196.

Corbin, L. J., S. Blott, J. Swinburne, M. Vaudin, S. Bishop and J. Woolliams (2010). "Linkage disequilibrium and historical effective population size in the Thoroughbred horse." Animal Genetics **41**(s2): 8-15.

Cunningham, E., J. Dooley, R. Splan and D. Bradley (2001). "Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses." Animal Genetics **32**(6): 360-364.

Davis, C. G. and B.-H. Lin (2005). Factors affecting US pork consumption, US Department of Agriculture, Economic Research Service.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen (2013). "Prediction of complex human traits using the genomic best linear unbiased predictor." PLoS Genetics **9**(7): E1003608.

Endelman, J. and M. J. Endelman (2014). "Package 'rrBLUP'."

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP." The Plant Genome **4**(3): 250-255.

Ewens, W. J. (1989). "An interpretation and proof of the fundamental theorem of natural selection." Theoretical Population Biology **36**(2): 167-180.

Felsenstein, J. (2006). "Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci?" Molecular Biology and Evolution **23**(3): 691-700.

Fernando, R. L., J. C. Dekkers and D. J. Garrick (2014). "A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses." Genetics Selection Evolution **46**(1): 50.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley and S. Fitzgerald (2011). "Ensembl 2012." Nucleic Acids Research: gkr991.

Flury, C., M. Tapio, T. Sonstegard, C. Drögemüller, T. Leeb, H. Simianer, O. Hanotte and S. Rieder (2010). "Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium." Journal of Animal Breeding and Genetics **127**(5): 339-347.

Frank, S. A. and M. Slatkin (1992). "Fisher's fundamental theorem of natural selection." Trends in Ecology & Evolution 7(3): 92-95.

Glessner, J. T., J. P. Bradfield, K. Wang, N. Takahashi, H. Zhang, P. M. Sleiman, F. D. Mentch, C. E. Kim, C. Hou and K. A. Thomas (2010). "A genome-wide study reveals copy number variants exclusive to childhood obesity cases." American Journal of Human Genetics 87(5): 661-666.

Gudbjartsson, D. F., G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson, P. Zusmanovich, P. Sulem, S. Thorlacius, A. Gylfason and S. Steinberg (2008). "Many sequence variants affecting diversity of adult human height." Nature Genetics 40(5): 609-615.

Hardy, J. and A. Singleton (2009). "Genomewide association studies and human disease." New England Journal of Medicine 360(17): 1759-1768.

Hartl, D. L. (1988). A primer of population genetics, Sinauer Associates, Inc.

Hayes, B., P. Bowman, A. Chamberlain and M. Goddard (2009). "Invited review: Genomic selection in dairy cattle: Progress and challenges." Journal of Dairy Science 92(2): 433-443.

Hayes, B. J., P. M. Visscher, H. C. McPartlan and M. E. Goddard (2003). "Novel multilocus measure of linkage disequilibrium to estimate past effective population size." Genome Research **13**(4): 635-643.

Henderson, C. R. (1975). "Best linear unbiased estimation and prediction under a selection model." Biometrics **31**(2): 423-447.

Hennig, C. (2010). "fpc: Flexible procedures for clustering." R package version 2: 0-3.

Hill, W. G. (1981). "Estimation of effective population size from data on linkage disequilibrium." Genetical Research **38**(03): 209-216.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences **106**(23): 9362-9367.

Kim, H., S. Sung, S. Cho, T.-H. Kim, K. Seo and H. Kim (2014). "VCS: Tool for Visualizing Copy Number Variation and Single Nucleotide Polymorphism." Asian-Australasian Journal of Animal Sciences **27**(12):

1691.

Kimura, M. (1984). The neutral theory of molecular evolution, Cambridge University Press.

Koivula, M., I. Strandén, G. Su and E. A. Mäntysaari (2012). "Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP)." Journal of Dairy Science **95**(7): 4065-4073.

Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proceedings of the National Academy of Sciences **99**(2): 803-808.

LANGE, K., J. WESTLAKE and M. Spence (1976). "Extensions to pedigree analysis III. Variance components by the scoring method." Annals of Human Genetics **39**(4): 485-491.

Lark, R., B. Cullis and S. Welham (2006). "On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML." European Journal of Soil Science **57**(6): 787-799.

Lee, K. W., P. San Woon, Y. Y. Teo and K. Sim (2012). "Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt?" Neuroscience and Biobehavioral Reviews **36**(1): 556-571.

Leeds, T. D. (2005). Pork quality improvement: estimates of genetic parameters and evaluation of novel selection criteria, The Ohio State University.

Mackay, T. F. (2001). "The genetic architecture of quantitative traits." The character concept in evolutionary biology: Academic Press (Gunter P. Wagner), 391-411.

Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature News **456**(7218): 18-21.

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This and C. Cierco-Ayrolles (2012). "Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness." Heredity **108**(3): 285-291.

Manolio, T. A. and F. S. Collins (2009). "The HapMap and genome-wide association studies in diagnosis and therapy." Annual review of medicine **60**:

443.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon and A. Chakravarti (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

McCarroll, S. A. and D. M. Altshuler (2007). "Copy-number variation and association studies of human disease." Nature genetics **39**: S37-S42.

McGilchrist, C. and K. Yau (1995). "The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models." Communications in Statistics-theory and Methods **24**(12): 2963-2980.

Meuwissen, T., T. R. Solberg, R. Shepherd and J. A. Woolliams (2009). "A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value." Genetics Selection Evolution **41**(2).

Outram, A. K., N. A. Stear, R. Bendrey, S. Olsen, A. Kasparov, V. Zaibert, N. Thorpe and R. P. Evershed (2009). "The earliest horse harnessing and milking." Science **323**(5919): 1332-1335.

Outram, A. K., N. A. Stear, R. Bendrey, S. Olsen, A. Kasparov, V. Zaibert, N. Thorpe and R. P. Evershed (2009). "The earliest horse harnessing and milking." Science **323**(5919): 1332-1335.

Penzes, P., A. Beeser, J. Chernoff, M. R. Schiller, B. A. Eipper, R. E. Mains and R. L. Huganir (2003). "Rapid induction of dendritic spine morphogenesis by trans-synaptic ephrinB-EphB receptor activation of the Rho-GEF kalirin." Neuron **37**(2): 263-274.

Piepho, H., J. Möhring, A. Melchinger and A. Büchse (2008). "BLUP for phenotypic selection in plant breeding and variety testing." Euphytica **161**(1-2): 209-228.

Price, G. R. (1972). "Fisher's 'fundamental theorem' made clear." Annals of Human Genetics **36**(2): 129-140.

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

Rosenblatt, M. (1956). "A central limit theorem and a strong mixing

condition." Proceedings of the National Academy of Sciences of the United States of America **42**(1): 43.

Schenkel, F. S., L. R. Schaeffer and P. J. Boettcher (2002). "Comparison between estimation of breeding values and fixed effects using Bayesian and empirical BLUP estimation under selection on parents and missing pedigree information." Genetics Selection Evolution **34**(1): 41-60.

Seaton, G., C. S. Haley, S. A. Knott, M. Kearsley and P. M. Visscher (2002). "QTL Express: mapping quantitative trait loci in simple and complex pedigrees." Bioinformatics **18**(2): 339-340.

Sellier, P., M. Rothschild and A. Ruvinsky (1998). "Genetics of meat and carcass traits." The Genetics of the Pig: 463-510.

Shen, X., M. Alam, F. Fikse and L. Rönnegård (2013). "A novel generalized ridge regression method for quantitative genetics." Genetics **193**(4): 1255-1268.

Sherman, J. and W. J. Morrison (1950). "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix." Annals of Mathematical Statistics: 124-127.

Skjervold, H. and H. J. Langholz (1964). "Factors affecting the optimum structure of AI breeding in dairy cattle." Zeitschrift für Tierzüchtung und Züchtungsbiologie **80**(1-4): 25-40.

Srivathsan, A. and R. Meier (2012). "On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature." Cladistics **28**(2): 190-194.

Steel, M. A. and Y. Fu (1995). "Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model." Journal of Computational Biology **2**(1): 39-47.

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

Swinburne, J. E., M. Boursnell, G. Hill, L. Pettitt, T. Allen, B. Chowdhary, T. Hasegawa, M. Kurosawa, T. Leeb and S. Mashima (2006). "Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families." Genomics **87**(1): 1-29.

Taveira, R., M. Mota and H. Oliveira (2004). "Population parameters in Brazilian Thoroughbred." Journal of Animal Breeding and Genetics **121**(6): 384-391.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard and P. M. Visscher (2007). "Recent human effective population size estimated from linkage disequilibrium." Genome Research **17**(4): 520-526.

Toosi, A., R. Fernando, J. Dekkers and R. Quaas (2010). "Genomic selection in admixed and crossbred populations." Journal of Animal Science **88**(1): 32.

Verbyla, K. L., B. J. Hayes, P. J. Bowman and M. E. Goddard (2009). "Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle." Genetics Research **91**(05): 307-311.

Visscher, P. M. (2008). "Sizing up human height variation." Nature Genetics **40**(5): 489-490.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five

years of GWAS discovery." The American Journal of Human Genetics **90**(1): 7-24.

Wang, D., J. Zhu, Z. Li and A. Paterson (1999). "Mapping QTLs with epistatic effects and QTL× environment interactions by mixed linear model approaches." Theoretical and Applied Genetics **99**(7-8): 1255-1264.

Wang, J. (2005). "Estimation of effective population sizes from data on genetic markers." Philosophical Transactions of the Royal Society B: Biological Sciences **360**(1459): 1395-1409.

Watterson, G. (1959). "Linear estimation in censored samples from multivariate normal populations." The Annals of Mathematical Statistics: 814-824.

Wei, T. and M. T. Wei (2015). "Package 'corrplot'." Statistician **56**: 316-324.

Woodbury, M. A. (1950). "Inverting modified matrices." Memorandum Report **42**: 106.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, M. D.

McMullen and B. S. Gaut (2005). "The effects of artificial selection on the maize genome." Science **308**(5726): 1310-1314.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin and G. W. Montgomery (2010). "Common SNPs explain a large proportion of the heritability for human height." Nature Genetics **42**(7): 565-569.

Zhang, Z., J. He, H. Zhang, P. Gao, M. Erbe, H. Simianer and J. Li "Results of Genome Wide Association Studies Improve the Accuracy of Genomic Selection Zhe Zhang¹, Jinlong He¹, Hao Zhang¹, Ping Gao¹, Malena Erbe², Henner Simianer², Jiaqi Li¹."

회귀 모형을 이용한

집단 기반 게놈 데이터의 유전 표지자 효과 추정

이영섭

(협) 생물정보학 전공

서울대학교 대학원 자연과학대학

유전체 디옥시리보핵산 (deoxyribonucleic acid; DNA) 수준에서 다양한 DNA 마커들이 개발된 이후로, 과학자들은 DNA 시퀀싱과 유전형 분석(genotyping)에 관심을 보였다. 유전형 분석이란 분자 마커의 일종인 유전체 변이를 결정하는 것을 말한다. 단일 염기 다형성은 가장 중요한 표지자에 속한다. 특히, 집단에 기반한 단일 염기 다형성은 다른 개체와 구별 지을 수 있는 각 개체의 특징을 담고 있을 수 있다.

각 개체들의 특징의 원인을 규명하기 위해서, 가능한 한 가지 방법은 확립된 통계 모델을 이용하는 것이다. 통계학의 가장

중요한 분야 중 하나인 회귀 분석은 생물정보학적 연구에서 활용되어 왔다. 선형 회귀, 비선형 회귀, 혼합 모형 같은 회귀 모델을 이용하여 데이터 분석을 수행하였다.

이 학위 논문은 5 장으로 구성되었다. 1 장에서는 이 학위 논문의 기초가 될 수 있는 집단 유전학 이론, 유효집단 크기 추정, 최적 선형 불편 예측, 유전체 연관 분석에 대해 개괄적으로 소개하였다. 유효 집단수를 결정하기 위해서 두가지 방법이 사용되었다: 고전적 스비드 방정식 그리고 기무라 2-파라미터 모델과 와터슨 씨다 추정통계량. 스비드 방정식은 비선형 회귀를 사용하며, 기무라 2-파라미터 모델은 단일 염기 다형성의 개수를 이용한다. 최적 선형 불편 예측은 선형 혼합 모형에서 임의 효과를 추정하기 위해 사용한다. 그리고 유전체 연관 분석은 한 형질과 연관되어 있는 유전체 변이를 탐색하기 위한 것이다. 임의 표지자 효과를 예측하기 위한 하나의 방법으로서, 단일 염기 다형성 - 유전체 최적 선형 불편 예측 (Single nucleotide polymorphism - genomic best linear unbiased prediction; SNP-GBLUP) 방법을 도입하였다. 이 새로운 최적 선형 불편 예측은 이론적으로 유전체 관계 행렬에 기반한다.

제 2장에서는 한국 경주마의 유효 집단 크기를 추정하였다. 경주마 품종은 이 품종의 훌륭한 경주 능력 때문에 사랑을 받아왔다. 한국 경주마의 유효 집단 크기를 추정함으로써, 한국 경주마 집단의 유전적 다양성과 안정성을 가늠해 보았다. 한국 경주마의 유효 집단 크기는 79 (스비드 방정식), 77 (기무라 2-파라미터 모델)이었다. 이는 다른 국가의 경우와 비교했을 때 다소 작았다. 일례로, Corbin et al. 은 아일랜드 경주마 집단의 유효 집단 크기를 100으로 추정하였다. 코빈은 연관 불균형에 기반한 스비드 방정식을 이용하였다.

제 3장에서는, 단일 염기 다형성끼리의 관계를 다룬 단일 염기 다형성 관계 행렬 (SNP-SNP relationship matrix; SSRM)을 소개하였다. 유전체-최적 선형 불편 예측 (G-BLUP)에서 중요한 유전체 관계 행렬 보다 더 고급화되고 분화된 것으로 볼 수 있다. 유전체 관계 행렬이란 혼합 모형 또는 최적 선형 불편 예측에서 핵심적 개념인 개체 관계를 나타낸 행렬이다. 최적 선형 불편 예측에서 효과적으로 임의 효과를 다루기 위해서, 유전체 관계 행렬은 필수 요소이다. 단일 염기 다형성 관계 행렬은 이론적으로 다변량 정규 분포와 유전체 관계 행렬에 근거하였음에도 불구하고,

새로운 개념이다. 유전체 관계 행렬로부터 단일 염기 다형성 관계 행렬로의 분화는 관계가 개체 수준이나 또는 단일 염기 다형성 수준이나에 따라 어떻게 정의될 수 있는가에 대한 서로 다른 시각에 근거한다고 본다. 단일 염기 다형성 관계 행렬은 확실히, 어려울 뿐만 아니라 쉽게 검증될 수 있다고는 할 수 없다. 그러함에도 불구하고, 이 행렬에 담겨진 생물정보학적 정보는 풍부하다고 볼 수 있다. 단일 염기 다형성 관계 행렬은 은닉 정보이고, 은닉된 단일 염기 다형성 정보에 의해 가공된 것이 유전체 관계 행렬이라고 생각한다. 이 행렬을 도입함으로써, 인간 키 형질 데이터를 분석하였다. KARE3, 특히 안성-안산 코호트 데이터는 각 개체의 형질과 단일 염기 다형성 정보를 담고 있다. 이 분석의 주 목표는 혼합 모형에서 단일 염기 다형성 관계 행렬의 유용성을 검증하고, 단일 염기 다형성의 관계를 독립 동일 분포로 가정하는 모형과 비교하기 위한 것이었다. 첫째, 확률 분포 함수와 선형 대수학에 기초하여, 이 행렬을 이론적으로 유도하였다. 둘째, 인간 키 형질과 단일 염기 다형성을 이용하여, 이 행렬의 유용성을 검증하였다. 단일 염기 다형성의 관계를 독립 동일 분포로 가정했을 때보다 유전체가 (동물 육종학에서의 육종가)의

측면에서 보았을 때, G-BLUP에 거의 일치했다는 것을 확인 할 수 있었다.

제 4장에서는 잃어버린 유전력을 해결하려 하였다. 데이터는 버크셔 포크 육질 관련 8개의 형질이었으며, 유전체 연관 분석을 먼저 수행하였다. 유전체 연관 분석으로 관련 형질과 연관 되어 있을 수 있는 단일 염기 다형성을 선택한 뒤, 최적 선형 불편 예측을 수행하였다. 전체 단일 염기 다형성을 사용했을 때보다, 유전체 추정 육종가와 유전력 추정치에서 향상된 결과를 보였다.

제 5장에서는 현 세대 다음세대의 선택 계수를 피셔의 자연 선택 기본 정리와 최적 선형 불편 예측을 이용하여 예측했다. 피셔의 이 정리는 주어진 시간에서 생물체의 적합도의 증가는 적합도의 유전적 분산과 같다고 기술된다. 선택은 대립 유전자 빈도를 변화시키는 가장 중요한 요소 중 하나이다. 과거의 선택 트렌드 뿐아니라, 미래의 트렌드를 알아보는 것은 매우 중요하다. 각 단일 염기 다형성의 상가적 유전 분산을 계산한 후 피셔의 정리를 이용하여 선택 계수를 계산하였다. 그 후, 유의한 단일 염기 다형성이 포함된 유전자들의 gene ontology (GO)를 조사하였다. 형질은 한국 홀스타인 우유 관련 형질이었으며, 이는 비유량,

유지방, 유단백질이였다. GO 분석에서 가장 주목할 점은 유단백질과 관련된 GO였으며, 특히 수지상 척주 형태 발생이 가장 유의하였다. 수지상 척주 형태 발생과 관련된 유전자들은 한국 홀스타인 인위 선발에서 대립 유전자 빈도가 급격히 변화될 것이라고 예측할 수 있다.

주요어: 회귀 분석, 단일 염기 다형성 관계 행렬, 잃어버린 유전력, 선택 계수, 유전체 연관 분석, 최적 선형 불편 예측, 유효 집단 수

학번: 2012-30909