

Research Article

Effect of the number of QTL of carcass weight on the accuracy of genomic estimated breeding values in Hanwoo

Youngjae Choi¹, Yeongkuk Kim², Doocho Lee³, Dong Jae Lee³, Seung Hwan Lee^{3*}¹Department of Bio-AI Convergence, Chungnam National University, Daejeon 34134, Republic of Korea²Quantomic Research & Solution, Daejeon 34134, Republic of Korea³Division of Animal & Dairy Science, Chungnam National University, Daejeon 34134, Republic of Korea*Corresponding author: slee46@cnu.ac.kr

ABSTRACT

The objective of this study was to investigate the relationship between the number of QTL (Quantitative Trait Loci) affecting the carcass weight of Hanwoo and the accuracy of GEBV (Genomic Estimated Breeding Values). The study was performed simulation study to generate three populations comprising 100, 500 and 1,000 QTL. Each population consisted of 13,500 individuals which were selected from the 6th to 10th generations of all 10 generations, while keeping litter size constant (N=1) in population parameters. The genetic parameters for these populations were set at 0.45 trait heritability, 0.3 QTL heritability and 48,578 total markers. In order to confirm the number of significant QTL under the threshold of Bonferroni correction, both GWAS (Genome-wide association study) and GWAS based on Random Forest (RF) were performed. In each population, there were 33, 50 and 32 QTL identified by only GWAS, with 100, 500 and 1,000 QTL. Additionally, RF-based GWAS detected 65, 260 and 491. These results showed that the number of significant QTL is not related to the overall number of QTL in the population. Furthermore, machine learning based GWAS is more accurate than solely using GWAS. To estimate the accuracy of GEBV (Genomic Best Linear Unbiased Prediction) in each population, 5-fold cross-validation was employed using SNP effect of the QTL over p-value. In the result of the estimation, GEBV rose with an increase in the number of QTL in the population, and the correlation between TBV (True Breeding Values) followed the same tendency. As a result, this study suggested that the number of QTL related to carcass weight in Hanwoo could exert a significant influence on genetic improvement.

Key words: Hanwoo, QTL, GWAS, Random Forest, GEBV

INTRODUCTION

농업이 국가 주요 산업이었던 과거에는, 한우 사육목적은 농사에 활용하는 역우(Draft animal)로서의 기능이 주요 목적이었다(Lee et al., 2014). 그러나 경제 발전과 국민 생활 수준이 향상됨에 따라서 쇠고기 수요가 증가하게 되고, 한우 사육의 목적이 고기 생산의 목적으로 바뀌게 되었다(KOSIS, 2022). 이에 따라 한우 개량은 농가의 소득과 관련이 있는 도체중(CWT, Carcass weight), 근내지방도(Marbling score), 등지방두께(Backfat thickness)와 같은 경제형질 위주의 개량이 시작되었다.

한우 개량은 우수한 유전자를 갖는 개체의 육종가(Breeding value)를 기반으로 개체를 선발하고 이들 우수한 유전자를 갖는 개체를 다음 세대의 부모(parent)로 이용함으로써 우수한 유전자를 다음 세대에 확산시켜 집단 전체를 개량한다. 한우개량 초기 1980년대 이전에는 전국 단위 품평회를 통해 개체의 표현형을 근거로 개체를 선발하였고, 1983년부터 한우 순수개량을 위하여 후대검정사업이 시작되었다(Kim et al., 2014). 1980년대부터 시작된 한우검정사업은 한우개량의 확고한 체계를 만들었고 한우 유전적 개량의 초석이 되었다(Park et al, 2013). 한우검정 사업은 한우개량에 있어서 매우 정확하고 효과가 높은 방법이지만, 도체형질 개량에 대해서는 검정 기간이

길다는 단점이 있다(Lee et al., 2014). 또한, 혈통정보를 이용하여 검정집단에서 생산된 표현형을 기초로 육종가를 추정하기 때문에 집단의 근친이 증가한다는 문제점이 있어 장기적인 관점에서 유전적 개량 효과를 감소시킬 수 있다(Quinton et al., 1992). 이러한 문제를 해결하고 더 정확한 육종가 추정 및 세대간격을 줄이면서 근친을 제어할 수 있는 새로운 육종 방법인 유전체 선발 방법이 개발되었다(Meuwissen et al., 2001)

유전체 선발(GS, Genomic Selection)은 유전체 정보를 이용하여 기존 혈통 정보가 갖는 단점을 개선하고 유전력이 낮거나 측정이 어려운 형질에 대한 육종가를 더 정확히 계산할 수 있어, 혈통 정보를 기반으로 한 기존 BLUP 방법 보다 개체 선발에 효율적일 것이다(Meuwissen et al., 2016). 유전체 선발 방법에서 이용하는 유전체 정보는 집단 내 개체 간 DNA에서 변형이 일어난 SNP (Single Nucleotide Polymorphism)를 이용한다(Zimin et al., 2009). 한우 유전체에 흩어져 있는 SNP들은 한우의 경제 형질에 영향을 주는 원인 유전자(Causal mutation)들 주변에 위치하고 있고, 강한 연관불평형(LD, Linkage Disequilibrium)을 이루고 있어 원인 유전변이를 모르더라도 SNP정보만으로도 원인 유전변이의 효과를 추정하고 그 효과를 이용하여 유전체 육종가를 추정할 수 있다(Meuwissen et al., 2001; Heyes and Goddard, et al 2010). 소의 참조 유전체 정보 서열이 개선되고, 상용화 SNP chip 기술의 비약적인 발전으로 소 1 마리의 유전체 정보를 보다 값싸게 활용하게 되면서 유전체 선발 기술뿐만 아니라 형질과 연관되어 있는 QTL을 탐색하는 전장유전체 연관분석이 활발하게 연구되고 있다(Goddard and Hayes, 2009). 이를 통해 소의 체중에 영향을 미치는 2개의 후보 QTL이 소의 14번 염색체에서 780kb 간격으로 존재하는 PLAG1-CHCHD7 유전자라는 것을 밝혔고, 일본 흑우의 체중에 영향을 미치는 PLAG1-CHCHD7을 포함하여 도체중에 영향을 미치는 3가지 주요 QTL이 6, 8, 14번 염색체에 존재함을 밝혀냈다(Lee et al., 2013).

이처럼 표현형에 영향을 미치는 QTL이 다수 존재하지만, 개체의 경제 형질과 관련된 대부분의 QTL은 형질에 영향을 미치는 정도가 적어, 유전체 내 정확한 위치와 기능을 확인하기 어렵다(Daetwyler et al., 2010). 다양한 유전적 효과를 보이는 QTL과 그 주변에 조밀하게 위치해 있는 마커 간의 연관불평형(LD, Linkage Disequilibrium) 정보를 이용하여 QTL을 정확히 검출할 수 있다면 보다 효율적으로 유전체 육종가(GEBV, Genomic Estimated Breeding Values)를 추정할 수 있을 것이다(Rolf et al., 2010; Quanbari et al., 2010). 양적 형질은 QTL의 수가 많고 그 효과가 작고, 환경적인 영향도 많아서 다유전성 인자(polygenic trait)로 정의한다(Hayes et al., 2009). 또한 집단의 규모 및 환경적인 영향 등과 같은 요인들은 QTL 탐지에 영향을 준다(Lillehammer et al., 2008).

따라서 본 연구의 목적은 효율적인 유전평가 방법론을 개발하기 위해서 한우의 경제형질에 영향을 미치는 QTL의 수를 시뮬레이션하고 기존 GWAS방법과 기계학습 방법론 중 하나인 Random Forest 방법을 적용하여 QTL검출의 능력(power)과 유전체 육종가의 추정 정확도를 평가하였다.

MATERIALS AND METHODS

모의실험을 통한 데이터 생성

본 연구를 위하여 모의실험을 진행하였다. 모의 실험을 위한 집단은 시뮬레이션 프로그램 QMSim Ver.2.0 (University of Guelph & HiggsGene Solutions Inc.)을 이용하였다(Sargolzaei and Schenkel, 2009). 모의 집단의 표현형, 유전자형 등이 시뮬레이션 되고 난 후, 형질을 조절하는 QTL의 수를 서로 다르게 하여 집단을 구성하였다(Ju and Kim, 2002).

시뮬레이션으로 생성된 집단의 구조(Population structure)와 모수(Parameter)들은 Table 1에 나타났다. 모의 집단들의 유전체는 29 쌍의 염색체로 구성되었고, 유전체의 길이는 143,872cM로 설정됐다. 각 집단들은 QTL의 개개수별로 100개, 500개, 1000개로 설정하여 Model 1, Model 2, Model 3로 나누었다. 형질에 대한 유전력과 QTL에 대한 유전력은 실제 한우의 도체중에 대한 유전력과 동일하게 각각 0.45와 0.35로 설정하였고, 표현형에 대한 분산은 1.0으로 설정했다. 집단들은 처음에 1,000두에서 시작해 LD와 Mutation-Drift Equilibrium을 형성하기 위해 800세대를 걸쳐 세대마다 2,000두씩, 1,000세대에 걸쳐 3,000두씩 시뮬레이션을 했다. 3,000두씩 시뮬레이션을 하는 세대의 수컷과 암컷 구성은 각각 300두와 2,700두로 설정했다. 성별에 따라 개체 수는 동일하고, 선발은 표현형 값이 높은 것을 기준으로 설정했다. 자손의 암수 비율은 각각 0.5로 고정하여 각 세대마다 1마리씩 자손을 생산하였고, 집단 내 EBV

는 BLUP을 통해 산출해, EBV가 높은 개체를 선발해 교배를 하고 낮은 개체는 도태되는 방식으로 진행했다. 개체들의 각 염색체 당 QTL 개수는 집단별로 35개, 17개, 4개로 설정했다. 그러나 9번과 15번 염색체에 QTL이 집중적으로 분포되게 설정했고, MAF (Minor Allele Frequency)는 0.1로 설정했다. QTL은 유전자좌당 2개의 대립유전자로 구성했고, 상가적 유전 효과는 0.4 형태인 Gamma Distribution으로부터 sampling을 하였다. QTL의 빈도는 동일하게 설정했고, 마커와 QTL의 missing과 돌연변이 비율은 설정하지 않았다. 표현형은 QTL 효과에 무작위 잔차를 추가하여 생성하였다.

Table 1. The structure and parameters of simulated populations

Parameter	Values		
	Model 1	Model 2	Model 3
Global Parameter			
Trait heritability		0.45	
QTL heritability		0.3	
Phenotype variance		1.0	
Historical Population			
Number of generation(size) - phase 1		0 (1,000)	
Number of generation(size) - phase 2		200 (2,000)	
Number of generation(size) - phase 3		800 (3,000)	
Recent Population			
Number of founder males from HP		300	
Number of founder females from HP		2700	
Number of generations		10	
Number of offspring per dam		1	
Ratio of males		0.5 / fix	
Mating system		Selective	
Replacement ratio of sires		0.4	
Replacement ratio of dams		0.2	
Selection		EBV / positive assortative	
Genome			
Number of Chromosome		29	
Number of total markers		50634	
Marker position		Random	
QTL distribution		Random	
MAF for markers		Equal	
MAF for QTL		Equal	
QTL allele effect		Gamma distribution (shape = 0.4)	
QTL mutation rate		0	
Number of total QTLs	100	500	1000
Total length of Chromosomes		143,872	
Number of total markers	35	17	4
	(9 chr - 135)	(9 chr - 46)	(9 chr - 7)
	(15 chr - 143)	(15 chr - 50)	(15 chr - 11)

HP, Historical Population; EBV, Estimated Breeding Values; QTL, Quantitative Trait Loci

MAF, Minor Allele Frequency

QTL 검출을 위한 통계 모형

Genome-wide association (GWAS)

본 연구에서는, 10세대에 걸쳐 시뮬레이션을 진행했고, 마지막 6~10세대에 해당하는 13,500두의 표현형 정보와 유전자형 정보를 이용하여 QTL 검출력을 비교하였다. 시뮬레이션 된 유전체 정보는 총 48,578개의 SNP으로 구성되었고, SNP QC(Quality

Control)는 Plink ver1.90 (Purcell, 2019) 프로그램을 통해 MAF 1% 미만, HWE(Hardy-Weinberg Equilibrium 오류 p-value 10^{-5})에 해당하는 SNP 마커들을 제거하였다(Marees et al., 2018). 전장 유전체연관분석은 혼합선형모형(Linear Mixed Model)을 이용하여 유전 마커와 표현형 간의 연관성을 분석하였다. 선형회귀분석은 독립변수인 유전변이와 종속변수인 표현형 간의 연관분석을 수행한다. 여기에서 혼합선형기법은 표현형 및 유전체 자료뿐만 아니라 개체 간의 유전적 관계를 설명하는 GRM(Genetic Relationship Matrix)를 계산하여 임의효과로 적합함으로 두 유전 마커 사이의 연관정보를 상쇄하고 유전 마커와 표현형 간의 연관성을 보다 정확하게 분석할 수 있다.

GWAS(Genome-Wide Association Study) 모델식은 아래와 같다.

$$Y = a + bX + G + E$$

여기서, Y는 표현형 벡터를 의미하며, a는 평균값을 의미함. b는 개체들의 유전자형에 따른 상가적 유전자 효과를 의미하며, X는 SNPs들의 유전자형을 의미함. G의 경우 모든 SNPs들을 통한 개체들의 유전적 혈연관계(GRM, Genomic Relationship Matrix)효과를 의미하며, E의 경우 환경적효과 및 error를 의미한다. 위의 모델식을 이용하여 표현형의 연속형 자료와 유전체자료를 통합한 혼합선형회귀모델을 적용한 전장 유전체연관분석을 수행하였다. 전장 유전체연관 분석결과로 도출된 각 유전좌위의 P-값을 $-\log(pvalue)$ 로 변환하여 각 유전체 좌위에 대해서 $-\log(pvalue)$ 를 플롯하여 맨하튼 플롯을 도출하였다. 통계적 유의성으로는 본페로니 검정(Bonferroni correction)을 적용하였다.

Random Forest

Random Forest(RF)은 분류, 군집분석, 회귀모형, 생존분석 등에 적용되는 모형으로, 다수의 의사결정나무 모형을 만들어 높은 정확도로 예측을 하는 것을 목적으로 한다(Yoo, 2015). RF는 최근 유전 정보를 활용한 분석에서 꾸준히 사용되는데, 많은 수의 의사결정나무를 생성해서 최종 클래스를 결정하는 방법으로 활용된다. 클래스를 나누기 위한 공식은 아래와 같다.

$$\phi = (\theta, T)$$

$$h_1(\phi) = h(\theta) < T$$

$$h_r(\phi) = \frac{h}{h_l(\phi)}$$

θ 는 feature parameter이고, T 는 한계선, h 은 전체 데이터의 수, h_l 은 왼쪽 노드, h_r 은 오른쪽 노드이다. 파라미터의 최적값($\hat{\theta}$)은 정보 획득량이 가장 큰 파라미터 값을 노드의 분할 조건으로 결정한다. 여기서 정보 획득량은 어떤 사건이 정보를 얼마만큼 줄 수 있는지 수치화 한 값이다. 이를 알기 위해서는 무질서도를 의미하는 엔트로피를 알아야 하며, 엔트로피가 높다는 것은 그 집단에서의 패턴, 특징을 찾기 어렵다는 것을 의미한다. 정보 획득량은 Shannon entropy로 결정된다고 한다(Seo and Kim, 2019). 이 Shannon entropy에 따라 가장 적합한 파라미터가 되는 조건을 찾을 수 있다. 이번 연구에서는 RF를 파이썬 scikit-learn을 통해서 구현했다. train과 test 집단은 GBLUP을 했을 때와 동일한 집단으로 했으며, 그들의 유전정보와 구제역 면역의 유무 데이터, 표현형 수치 데이터를 사용했다. RF에 사용한 SNP와 QTL 유전 정보는 시뮬레이션 데이터에서 만들어진 유전 정보 데이터를 0, 1, 2로 바꾸어 진행했기에 개체의 수(n) X SNP와 QTL의 수(p) 크기의 데이터 프레임으로 만들어 분석했다. 이번 연구에 맞게 RF 하이퍼 파라미터인 결정트리의 개수, 노드를 분할하기 위한 최소한의 샘플 수, 리프노드가 되기 위한 최소한의 샘플 데이터 수, 트리의 최대 깊이를 조정하기 위해서 GridSearchCV를 사용했다. 이후 변수 중요도를 통해서 어느 변수가 예측 성능에 중요한 역할을 하는지를 추정했다. 즉, 어떤 SNP와 QTL이 질병 저항성 유무를 예측하는데 중요한지 추정한 것이고, 이를 GWAS에서 추정한 SNP와 QTL과 비교했다. 또한 RF로 구한 GEBV의 정확도는 GBLUP과 마찬가지로 TBV와의 상관관계로 구했다.

유전체 육종가 추정 통계모형

Genomic BLUP (GBLUP)

유전체 정보를 활용한 불편추정방법(GBLUP, Genomic Best Linear Unbiased Prediction)은 개체의 유전체 정보를 이용하여 유전체 혈연관계 행렬을 구성하고 기존 불편추정모형에서 NRM(Numerator Relationship Matrix), 즉 혈통을 이용한 혈연관계 행렬을 유전체 혈연관계 행렬로 대체하여 유전체 육종가(Genomic Breeding Value; gEBV)를 계산하였다. 유전체 혈연관계 행렬은 아래의 공식으로 계산하였다.

$$G = \frac{MM'}{\sum_{i=1}^m 2p_i(1-p_i)}$$

$$Var(u) = G\sigma_u^2$$

여기에서 G는 유전적 혈연관계 행렬(GRM, Genomic Relationship Matrix), M은 개체의 유전자형 행렬이고, m은 전체 SNP마커의 수이다. 그리고 p_i 는 i번째 SNP의 대립유전자형의 빈도이다. 일반화 선형혼합모형은 아래와 같다.

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Z는 개체의 유전자 효과(a)의 incidence matrix이고, X는 고정효과에 대한 incidence matrix를 의미한다. σ_e^2 는 잔차 분산이고, 그리고 개체 유전자효과인 $a \sim N(0, G\sigma_a^2)$. $\lambda = \sigma_e^2/\sigma_a^2$. 그러므로, GBLUP은 기존 BLUP에서 A 행렬을 G행렬로 대체한 모델과 같다.

GEBV 정확도 예측

추정된 육종가의 정확도는 BLUPF90 (University of Georgia) 프로그램을 통해 추정됐다. 정확도는 BLUPF90을 통해 계산된 예측 오차 분산값(PEV, Predicted Error Variance)과 REMLF90 (University of Georgia) 프로그램을 통해 나온 형질에 대한 유전분산값을 이용해 추정했다(Misztal et al., 2018). 정확도 계산식은 다음과 같다.

$$r = \sqrt{1 - \left(\frac{PEV}{\sigma_a^2}\right)}$$

0r은 EBV의 정확도; PEV는 EBV에 대한 예측 오차 분산값; σ_a^2 는 형질별 유전분산값을 의미한다(Misztal and Wiggans, 1988).

RESULTS AND DISCUSSION

QTL 검출

가축에서 형질의 유전분산에 관여하는 양적 형질 좌위(QTL, Quantitative Trait Loci)의 수는 알려지지 않고 매우 많은 수의 QTL이 경제형질을 조절한다고 알려져 있다. 따라서 가축 개량에서는 양적 형질 좌위(Quantitative traits loci; QTL)와 연관된 연관불평형(LD, Linkage Disequilibrium)인 유전 마커 정보를 이용하여 개체의 유전능력을 추정한다(Meuwissen et al, 2001).

본 연구는 형질을 조절하는 QTL의 수에 따라서 전장유전체 연관분석 방법과 Random Forest모형이 QTL을 검출하는 statistical power를 조사하였고, 아울러 추정 육종가의 정확도를 계산하였다. QMSim 프로그램을 사용하여 Table 1과 같이 시뮬레이션 데이터

를 생성하였다. 총 10세대 중에서 6~10세대에 해당하는 13,500개의 자료를 활용하여 QTL검출한 결과, Figure 1에서 Bonferroni 임계 수준(genome-wide suggestive level)를 충족하는 QTL 개수는 1000QTL 집단에서 33개(3.3%), 500QTL에서는 50개(10%), 100QTL에서

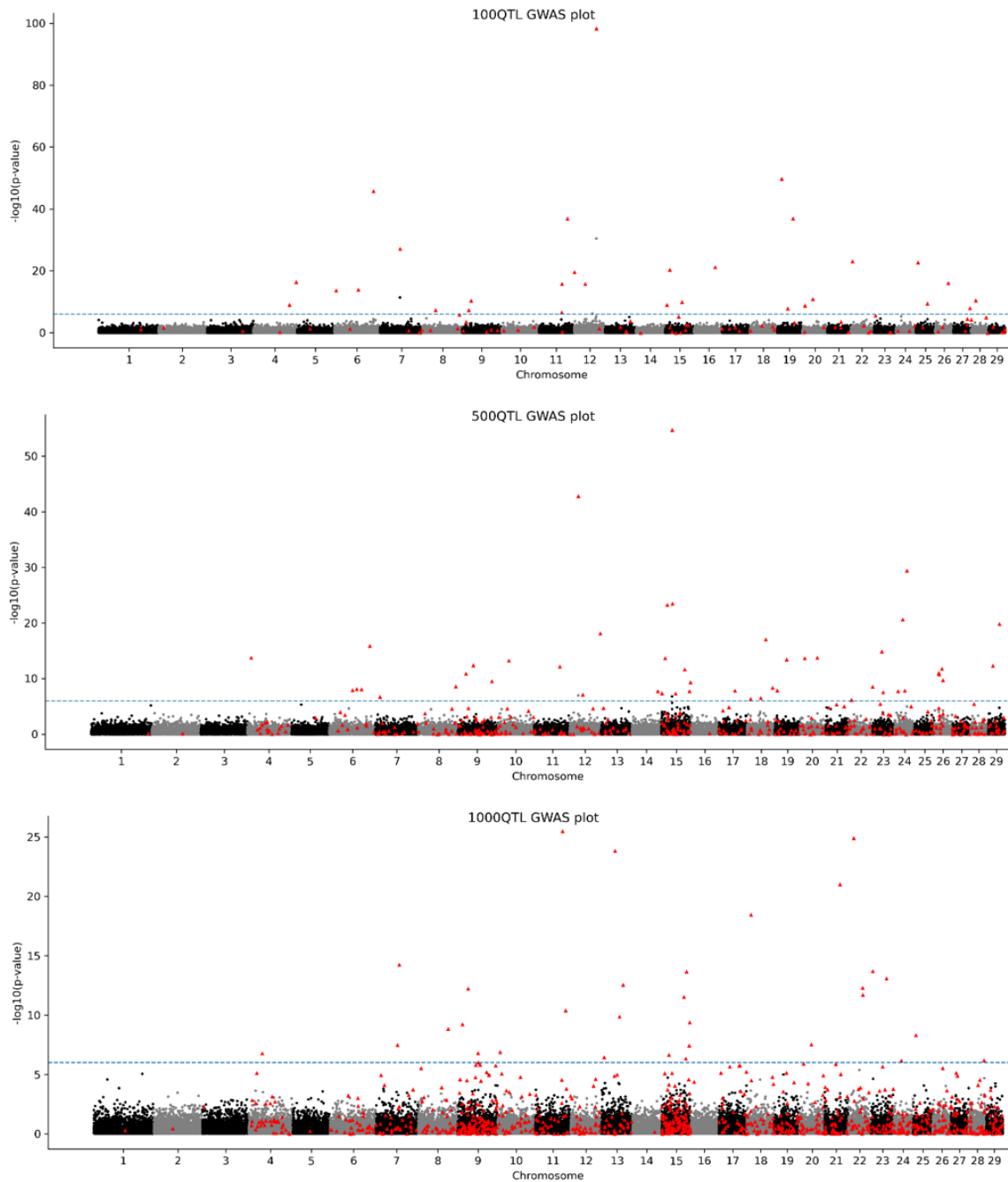


Figure 1. The Manhattan plot for GWAS is the QTL search for a population with 100, 500 and 1,000 QTL, showing a $-\log_{10}(p\text{-value})$ across 29 chromosomes with a heritability of 0.45. The red triangle and dotted line indicated the QTL location and threshold. QTL, Quantitative Trait Loci

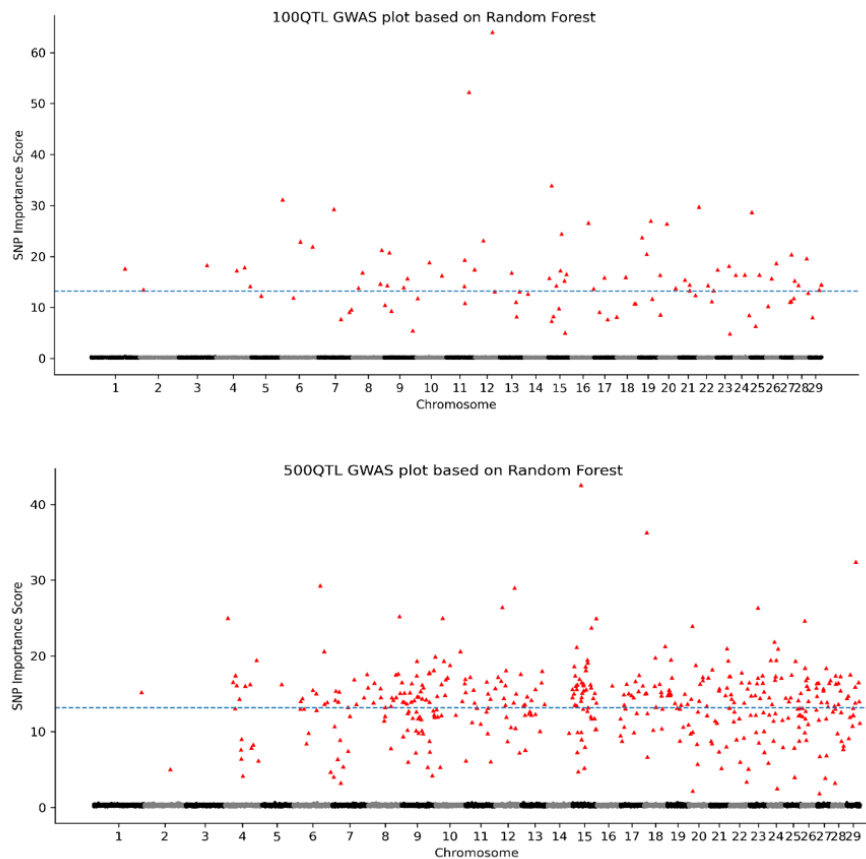
Table 2. Results of GWAS and GWAS based on RF. Threshold of GWAS was estimated by Bonferroni correction, and GWAS based on RF was calculated by mean of feature importance of SNPs.

	GWAS(LSM)			GWAS based on RF		
	100QTL	500QTL	1,000QTL	100QTL	500QTL	1,000QTL
Threshold	5.987	5.991	5.994	13.215	13.151	60.808
Num of Significant QTL	33	50	32	65	260	491
ΣSNP_i	-0.042	0.427	0.265	-0.042	0.426	0.265

RF, Random Forest; QTL, Quantitative Trait Loci.

는 32개(32%)로, QTL수가 적게 시뮬레이션 된 집단에서 검출력이 증가하였다. 이러한 결과는 고정된 13,500개의 표본 수에서 QTL의 수가 많다는 것은 개개의 QTL효과가 작다는 것을 의미하고, 반대로 QTL의 수가 작다는 것은 개개의 QTL이 보다 큰 효과를 갖는다는 것을 의미한다. 따라서 100개의 QTL이 시뮬레이션 집단에서 QTL검출력이 보다 높게 나타난 것을 Table 2에서 볼 수 있다.

Random Forest 방법을 이용한 QTL 검출은 각 SNP의 feature importance score를 계산하여 이들의 평균값을 QTL검출 threshold로 설정하였다(Alves et al., 2022). 그 결과 Figure 2에서 볼 수 있듯이, threshold를 충족하는 QTL의 개수는 100QTL 집단에서 65개(65%), 500QTL 집단에서 260개(52%), 1000QTL 집단에서 491개(49.1%)이다. Random Forest(RF) 방법에서 역시 QTL수가 100개인 집단에서 검출력이 65%로 가장 높았으며, 1000개 집단에서는 49.1%로 낮았다. LSM(Least Square Method)을 기반으로 한 회귀분석으로 QTL을 검출하는 방법보다 Random Forest 방법이 QTL 검출 능력에서 약 30~40% 높게 나타나는 것으로 분석이 되었다. 즉, RF를 적용한 GWAS에서 유의적인 QTL들이 일반적인 GWAS보다 검출이 많이 된 것으로 보아, 머신 러닝을 적용했을 때 더 QTL 검출 정확도가 더 높은 것으로 판단된다. 이는, 이미 알려져 있는 QTL 효과를 시뮬레이션 한 집단에서 GRAMMAR-CG 방법과 Random Forest를 이용하여 QTL 검출 능력을 평가한 결과, RF 방법이 QTL검출 능력이 뛰어난 것(Minozzi et al., 2014)으로 분석된 결과와 매우 유사하다.



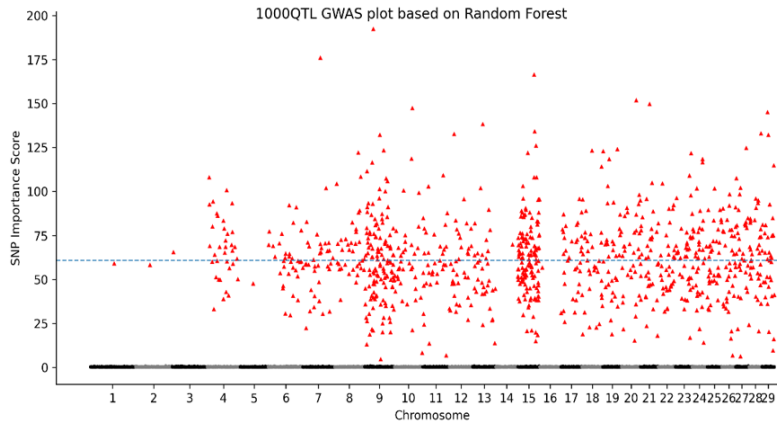


Figure 2. The Manhattan plot for GWAS based on Random Forest is the QTL search for a population with 100, 500 and 1,000 QTL, showing a $-\log_{10}(p\text{-value})$ across 29 chromosomes with a heritability of 0.45. The red triangle and dotted line indicated the QTL location and threshold.
QTL, Quantitative Trait Loci

고 볼 수 있다. 한우의 전장유전체 서열로부터 형질과 연관이 높은 SNP 마커를 선별해 유전능력평가를 실시할 경우 높은 정확도가 나오므로(Lopez et al., 2021), 이러한 결과는 QTL과 연관된 SNP 정보를 이용하여 유전능력을 예측하는 GBLUP에 활용한다면 효율 적일 것이다.

QTL 수에 따른 GEBV 추정 정확도

가축의 경제형질 뿐만 아니라 인간의 질병과 같은 형질도 원인성이 정확하게 알려져 있지 않은 수많은 유전자에 결정된다고 하여, 다유전인자이라고 알려져 있다(Yang et al., 2010). 소 유전체를 대표하는 참조 유전체 서열이 완성되어 있고, 약 25,000여개의 유전자 정보가 알려져 있으며, 한우에서 개체 간의 표현형 차이를 유발하는 유전 변이인 SNP 정보는 대략 2천5백만개가 보고되어 있다(Lee et al., 2023). 전장 유전체에 흩어져 있는 수많은 SNP들 중에서 형질과 연관 혹은 연관불평형 block내에 있는 SNP들을 선별하여 상업용 칩이 출시되어 있다. 이러한 상업용 칩을 이용하여 GBLUP 방법을 적용함으로써 개체의 유전능력을 예측한다.

본 연구에서 QTL 100, QTL 500, 그리고 QTL 1000개를 시뮬레이션 한 집단에서 유전능력을 평가하고, 그 정확도를 계산한 결과, Table 3에서와 같이 QTL이 100개인 집단에서는 정확도가 27%, 그리고 QTL이 500개인 집단에서는 그 정확도가 34%, 마지막으로 QTL이 1000개인 집단에서는 그 정확도가 40%로 분석되었다. 이러한 결과는 형질을 조절하는 QTL의 수가 많을 수록 개체의 QTL

Table 3. The number of QTL size statistics by 5-fold cross-validation according to GBLUP method

Population	EBV	Accuracy	Accuracy Correlation
	Mean	Mean	Mean
QTL = 100	0.4657	0.6447	0.2751
QTL = 500	0.4897	0.6549	0.3441
QTL = 1000	0.5036	0.6451	0.4008

EBV, Estimated Breeding Values; QTL, Quantitative Trait Loci; Var, Variance

효과는 작아지기 때문에 GBLUP 모델이 보다 정확한 육종가를 추정할 수 있는 것으로 판단된다. 이는, 브로일러 계통에서 무수히 작은 QTL보다 매우 큰 QTL 효과가 존재하는 집단에서 GBLUP보다 Bayesian B 모델이 보다 정확하게 육종가를 예측되기 때문이다(Wolc et al., 2016). 이러한 결과는 본 연구에서 살펴본 QTL 100개인 집단보다 QTL 1000개인 집단에서 예측의 정확도가 높게 나온 결과와 일치하는 결과이다. 즉 동일한 유전력에서 QTL 100개가 QTL 1000개보다 각 QTL 효과가 훨씬 크기 때문이다.

이전 연구에서는 세 가지 유효 모집단 크기와 다양한 수의 QTL에 대해 확률론적 시뮬레이션을 수행하여 유전체선택방법(GBLUP)과 베이지안 변수 선택 방법(Bayes B)을 비교하였다(Daetwyler et al., 2010). 그 결과 GBLUP은 주어진 유전력과 표본 크기에 대해서 QTL의 수에 관계없이 일정한 정확도를 보였다. 그러나, Bayes B는 $N(QTL)$ 이 낮을 때 GBLUP보다 정확도가 높았고, QTL의 수가 많아지면서 GBLUP이 Bayes B를 약간 상회하는 결과를 보고하였다.

결론적으로, GWAS와 RF 기반 QTL 검출력은 Random Forest 기반 유전체 연관분석이 기존 유전체 연관분석보다 목표 형질에 핵심적인 QTL을 보다 많이 검출하는 것을 확인하였다. 또한 QTL 수에 따라서 육종가의 정확도의 변화는 QTL의 수가 증가할수록 GBLUP모델이 추정 육종가를 보다 정확하게 추정하는 것으로 분석되었다.

CONFLICT OF INTERESTS

No potential conflict of interest relevant to this article is reported.

ACKNOWLEDGEMENTS

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00155857, 인공지능융합혁신인재양성(충남대학교)).

REFERENCES

- Lee SH, Choi BH, Lim DJ, Gondro C, Cho YM, Dang CG, Sharma A, Jang GW, Lee KT, Yoon DH et al. 2013. Genome-Wide Association Study Identifies Major Loci for Carcass Weight on BTA14 in Hanwoo (Korean Cattle). *PLOS ONE*. 8(10). DOI:10.1371/journal.pone.0074677
- Lee SH, Park BH, Sharma A, Dang CG, Lee SS, Choi TJ, Choy YH, Kim HC, Kim SD, Yeon SH et al. 2014. Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. *Journal of Animal Science and Technology* 56(2). DOI:10.1186/2055-0391-56-2
- KOSIS(Korean Statistical Information Service). 2022. Supply and demand performance of livestock products. https://kass.mafra.go.kr/statHtml/statHtml.do?orgId=114&tblId=DT_114_015_B164&conn_path=I2 (2023.11.01.) [in Korean]
- Kim NS, Lee JJ, Dang CK, Jung YH, Cheon HJ. 2014. Prospect and Retrospect on Improvement of Hanwoo(Korean Cattle) for 50 Years. *Bulletin of the Animal Biotechnology* 6:1-13. [in Korean]
- Park B, Choi T, Kim S, Oh SH. 2013. National Genetic Evaluation (System) of Hanwoo (Korean Native Cattle). *Asian-Australian Journal of Animal Sciences* 26(2):151-156.
- Quinton M, Smith C, Goddard ME. 1992. Comparison of selection methods at the same level of inbreeding. *Journal of Animal Science*. 70(4):1060-1067.
- Meuwissen T, Hayes B, Goddard M. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157:1819-1829.
- Meuwissen T, Hayes B, Goddard M. 2016. Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers* 6(1):6-14.
- Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS et al. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology* 10. DOI:10.1186/gb-2009-10-4-r42
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. 2009. Invited Review: Genomic selection in dairy cattle: Progress and Challenges. *Journal of Dairy Science*. 92(2):433-443.
- Hayes B, Goddard M. 2010. Genome-wide association and genomic selection in animal breeding. *Genome* 53(11):876-883.
- Goddard ME, Hayes BJ. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*. 10:381-391.

- Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics*. 185(3):1021-1031.
- Rolf MM, Taylor JF, Schnabel RD, McKay SD, McClure MC, Northcutt SL, Kerley MS, Weaber RL. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genomic Data*. 11(24). DOI:10.1186/1471-2156-11-24.
- Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics*. 41(4):346-356
- Lillehammer M, Goddard ME, Nilsen H, Sehested E, Gro Olsen H, Lien S, Meuwissen THE. 2008. Quantitative Traits Locus-by-Environment Interaction for Milk Yield Traits on Bos Taurus Autosome 6. *Genetics*. 179(3):1539-1546.
- Sargolzaei M, Schenkel FS. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25: 680-681.
- Ju JC, Kim NS. 2002. Development of Sumulation Model for Breeding Schemes of Hanwoo(Korean Cattle). *Journal of Animal Science and Technology* 44(5):507-518.
- Purcell S, Neale B, Todd-Brown Kathe, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Society of Human Genetics* 81(3):559-575. DOI:10.1086/519795.
- Marees A, Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks E. 2018. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*. 27(2). DOI:10.1002/mpr.1608.
- Yoo JE. 2015. Random forests, an alternative data mining technique to decision tree. *Computer Science. Journal of Education Evaluation* 28(2):427-448. [in Korean]
- Seo WD, Kim DW, Lee JS. 2019. Generalized information-theoretic criterion for multi-label feature selection. *IEEE Access* 7. DOI:10.1109/ACCESS.2019.2927400.
- Misztal I, Wiggans GR. 1988. Approximately of Prediction Error Variance in Large-Scale Animal Models. *Journal of Dairy Science* 71(2):28-32
- Misztal I, Tsuruta S, Daniela L, Masuda Y. 2018. Manual for BLUPF90 family of program. 2022. University of Georgia, Athens, USA.
- Alves AAC, Magalhaes da Costa R, Fonseca LFS, Carvalheiro R, Ventura RV, Mahalhaes Rosa GJ, Albuquerque LG. 2022. A Random Forest-Based Genome-Wide Scan Reveals Fertility-Related Candidate Genes and Potential Inter-Chromosomal Epistatic Regions Associated With Age at First Calving in Nellore Cattle. *Frontiers in Genetics*. DOI:/10.3389/fgene.2022.834724.
- Minozzi G, Pedretti A, Biffani S, Nicolazzi EL, Stella A. 2014. Genome wide association analysis of the 16th QTL-MAS Workshop dataset using the Random Forest machine learning approach. *BMC Proceedings*. DOI:10.1186/1753-6561-8-S5-S4.
- Lopez BIM, An N, Srikanth K, Lee SH, Oh JD, Shin DH, Park WC, Chai HH, Park JE, Lim DJ. 2021. Genomic Prediction Based on SNP Functional Annotation Using Imputed Whole-Genome Sequence Data in Korean Hanwoo Cattel. *Frontiers in Genetics* 11. DOI:10.3389/fgene.2020.603822.
- Yang J, Visscher PM, Wray NR. 2009. Sporadic cases are the norm for complex disease. *European Journal of Human Genetics* 18:1039-1043. DOI:10.1038/ejhg/2009.177
- Lee DJ, Kim YS, Dinh PTN, Chung YJ, Lee DH, Kim YK, Lee SH, Choi IC, Lee SH. 2023. Identification of Missense Variants Affecting Carcass Traits for Hanwoo Precision Breeding. *Genes (Basel)* 10:1839. DOI:10.3390/genes14101839.
- Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Dekkers JCM, Fernando R, Garrick DJ. 2016. Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions. *Journal of Animal Science and Biotechnology* 7. DOI:10.1186/s40104-016-0066-z.

AUTHORS INFORMATION

Youngjae Choi: <https://orcid.org/0000-0003-1540-6970>