JABG
Journal of Animal Breeding & Genomics

JABG. 2023 December, 7(4)167-178
https://doi.org/10.12972/jabng.20230018

Print ISSN 1226-5543
Online ISSN 2586-4297

Research Article

한우 전염성 질병 저항성 연관 QTL 검출을 위한 유전체 모의실험 연구

정운영[1], 김영국[2], 김윤식[1], 정윤지[1], 이동재[1], 강지민[1], 이두호[1], 구양모[3], 이승환[1]

[1]충남대학교 농업생명과학대학 동물자원과학부

[2]퀀토믹 리써치 앤 솔루션

[3]한국종축개량협회

Genomic breeding simulation study to identify QTL associated with cattle disease traits in Hanwoo

WoonYoung Jeong[1], YeongKuk Kim[2], Yoonsik Kim[1], Yoonji Chung[1], Dong Jae Lee[1], Ji Min Kang[1], Dooho Lee[1], Yang Mo Koo[3], Seung Hwan Lee[1,*]

[1]Division of Animal & Dairy Science, Chungnam National University, Daejeon, 34134, South Korea

[2]Quantomic research & solution, Daejeon, 34134, South Korea

[3]Korea Animal Improvement Association, 88 Myeongdal-ro, Seoul, South Korea

***Corresponding author: Seung Hwan Lee,** Division of Animal & Dairy Science, Chungnam National University, Daejeon, Korea, E-mail: slee46@cnu.ac.kr

ABSTRACT

This research used the simulation program to identify individuals resistant to foot-and-mouth disease (FMD) using the QMSim program. The simulation program was utilized to generate genetic and phenotypic data for individuals with and without FMD immunity. Subsequently, based on the simulated data, a genome-wide association study (GWAS) was performed to detect quantitative trait loci (QTL) associated with FMD immunity. Additionally, the QTLs identified by GWAS were compared with Random Forest (RF) and XGBoost. Out of the 41,461 SNPs, which included QTLs generated from the simulation, a total of 20 markers were found to be associated with FMD immunity. When comparing the performance of GWAS, RF, and XGBoost, RF identified the highest number of QTLs (7), followed by GWAS (6) and XGBoost (3). Furthermore, GBLUP, RF, and XGBoost were employed to classify individuals as either having or lacking FMD immunity. The classification accuracy, sensitivity, and specificity were evaluated using a confusion matrix, and the results were compared. The overall accuracy of the classification was as follows: XGBoost 0.53, RF 0.52, GBLUP 0.51. Sensitivity values were RF 0.98, XGBoost 0.97, GBLUP 0.19, and specificity values were GBLUP 0.83, XGboost 0.08, RF 0.05. XGBoost consistently outperformed the other methods in the overall accuracy and sensitivity, while GBLUP exhibited the lowest performance. Therefore, the research suggests that combining various methods in an ensemble approach, rather than relying solely on GBLUP, can lead to better predictions of FMD-resistant individuals. This approach has the potential to help mitigate the damages caused by future FMD outbreaks.

Key words: FMD resistance, GBLUP, RandomForest, XGBoost

Received November 14, 2023 **Revised** December 26, 2023 **Accepted** December 26, 2023

INTRODUCTION

구제역은 소, 돼지, 양 등의 발굽이 갈라진 동물들이 걸리는 매우 전염성이 높은 질병으로, 구제역에 감염된 가축은 입술, 잇몸, 구강, 코, 유두 및 발굽사이에 물집이 발생하고 잘 걷지 못하거나 유량이 감소한다. 또한 식욕저하가 일어나며 심한 경우 폐사될 수 있다(Yu, 2014). 이 질병은 피코르나 바이러스인 구제역 바이러스에 의한 감염으로 발생하며, 구제역 바이러스는 변이가 쉽게 일어나기 때문에 현재까지는 치료 방법이 개발되지 않아 백신을 통한 예방이 중요하다(Grubman & Baxt, 2004; Suh, 2000). 농림축산식품부의 구제역 발생 현황표에 따르면 대한민국에서는 2000년대 이후로 총 11회 424건의 구제역 발생이 있었다. 특히 2010년부터 2011년까지는 구제역이 가장 크게 발생한 시기였으며, 이 시기에 구제역에 감염된 개체의 살처분이 진행되어 구제역 감염의 유행을 제어했다(Han, 2021). 이렇게 진행한 구제역에 감염된 가축의 살처분은 한 번에 많은 수를 토양에 묻기 때문에 가축에서 흘러나온 핏물이 지하수 오염시키는 것과 같이 2차 피해가 발생함으로 농가의 경제적 피해분만 아니라 환경적인 피해를 야기한다(Kim, 2011). 이에 더하여 가축의 살처분에 참여한 연구자나 이해당사자가 정신적 스트레스 및 외상 후 스트레스를 초래할 수 있다(Heomin Park & Jin, 2017).

이러한 경제적, 정신적 피해를 줄이기 위해 구제역을 예방하고 관리하기 위한 노력이 필요할 것이다. 세계적으로도 철저한 백신을 통한 예방과 관리, 구제역에 대한 연구가 수행되고 있다(Naderi et al., 2016). 국내에서는 2010~2011년 구제역의 발생 이후, 구제역의 전파과정을 지리역학적 관점에서 예측하기 위한 전파 확산 시뮬레이션 모델 개발과 적용 연구가 진행되었고(Pak & Bae, 2012), 구제역바이러스에 대한 저항성을 일으키는 유전자를 탐색하는 연구도 진행되었다(Lee et al., 2015). 또한, 구제역은 생체내 면역반응에 대한 생물학적 기전으로 개별 가축에 있어서 구제역 저항성에 대한 분산성분(variance component)분석을 통한 구제역에 대한 유전력이 확인되고 있다(Gowane et al., 2013). 특히 Gowane 등은 구제역 백신접종에 따른 항체 형성율에 대한 유전력을 추정했다. 구제역 바이러스는 O, A, Asia1 등과 같은 7가지의 혈청형으로 구분할 수 있는데(Yeo-Joo Lee et al., 2011), serotype O의 경우 17%의 유전력을 보였고, A type의 경우 약 3%, 그리고 Asia1의 경우 약 5%의 유전력을 확인하였다. 동물 육종의 관점에서 보면 소의 건강이 크게 개선되고 유전적 개량이 증가한다는 것은 건강 형질을 전체 육종 목표 또는 선발 지표에 직접 포함시키는 육종프로그램을 설계한다는 것을 의미한다(Egger-Danner et al., 2012). 하지만, 이러한 질병 형질을 후대검정과 같은 육종계획에 포함하여 진행한다는 것은 매우 어렵다. 따라서 구제역 저항성 표현형과 유전체자료를 이용하여 유전체선발 모델을 적용하는 것이 보다 효율적 일 것이다(Meuwissen et al., 2001).

질병 관련 형질은 일반적으로 범주형 이고, 여러 유전자의 영향을 받으며, 멘델 유전에서 벗어나고, 유전자 및 환경과의 상호작용이 뚜렷하게 나타나기 때문에 GEBV 추정에 통계적 문제가 발생할 수 있다(Blazer & Hernandez, 2006). 고전적으로 Meuwissen 등(2001)의 주요 논문에서 소개된 바와 같이, 혼합 모델 방정식은 유전체 BLUP(GBLUP) 모델 그리고 일부 베이지안 모델을 적용하고 있다. 그러나, 새로운 형질의 경우, SNP와 같은 유전 마커의 수(m)보다 현저히 적은 개체의 수(n)로 유전 마커의 과추정이 문제가 된다(Kramer et al., 2014). 이러한 문제를 해결하기 위해서 RandomForest(RF)와 같은 기계학습모델이 사용되고 있으며, 이와 같은 모델들은 분석 수행시 중요도가 높은 특성들을 기반으로 feature selection이 가능하여 마커 선별에 효과적이라는 보고가 발표되었다(Jungmin Choi et al., 2022).

양식어종의 질병에 대한 저항성을 예측하기 위하여 시뮬레이션 프로그램과 머신러닝을 이용한 연구(Palaiokostas, 2021)와 양돈의 PRRS virus에 대한 저항성을 시뮬레이션을 통해 유전데이터을 이용하여 다른 선발 전략을 비교하는 연구(Schaeffer, 2014) 등과 같은 시뮬레이션을 이용하여 유전체 모의실험을 실행한 선행연구가 진행되었다. 앞선 이 연구 중 특히 양식어종의 질병에 대해 예측한 연구에서는 genomic best linear unbiased prediction for threshold traits backend by Markov chain Monte Carlo (GBLUP-MCMC)와 다양한 머신러닝들의 결과를 비교했는데 XGBoost와 SVM, RF가 특정 범위에서 GBLUP-MCMC보다 약간의 이점을 갖는다는 결과를 발표했다.따라서, 본 연구는 앞서 언급한 선행연구들을 기초로 모의실험 집단을 구성하여 GBLUP 모델 및 다양한 기계학습 모델을 적용하여 유전평가 및 유전자 탐색을 진행했고 선행연구와 같이 GBLUP 및 다른 머신러닝과의 성능을 비교하는 것을 진행했다. 또 다른 선행연구에서도 Genomic prediction에 있어서 SVM, XGBoost, RF,그리고 CNN을 이용하여 비교했을 때 머신러닝 모델들이 genomic

prediction을 크게 향상시킨다는 것을 입증했고, XGBoost는 중요한 SNP을 선택하는 기능이 포함되어 있기 때문에 다른 머신러닝 사용시에 함께 사용하는 것을 추천한다고 발표했다(Xiang et al., 2023). 이에 근거하여 이번 연구에서 QTL 검출을 위해 다양한 머신러닝 중 feature selection 기능을 가진 XGBoost와 RF를 이용하여 연구를 진행했다.

본 연구의 목표는 (1) GWAS 및 RF와 XGBoost와 같은 기계학습 모델을 통해 구제역 면역과 관련된 QTL을 탐색하여 원인유전자인 QTL을 정확하게 검출할 수 있는지 조사했고, (2) GBLUP 및 RF, XGBoost를 사용하여 유전체육종가(GEBV)를 구하여 구제역 면역력이 있는 개체를 예측하고 이 방법들의 정확성을 비교하여 어떤 방법이 유용한지 확인했다.

MATERIALS AND METHODS

1. 모의실험

이번 연구에서는 FMD (Foot-and-Mouth Disease) 저항성 개체의 표현형과 유전정보, TBV (True Breeding Value)를 구하기 위해서 QMSim 프로그램을 사용했다(Sargolzaei & Schenkel, 2009). QMSim 프로그램을 이용하여 모의실험 집단 구성의 시나리오는 Table 1과 같다. 본 연구에서 모수(parameters)는 Global parameters, Historical population, Population, Genome의 총 4 종류로 구성되어 있다. 먼저 Global parameter에서 heritability와 QTL heritability를 설정할 수 있다. 형질에 대한 전체 유전력(heritability)은 0.36으로 설정 하였다(Leach et al., 2010). 원인유전변이인 QTL이 설명하는 QTL heritability는 0.30, 그리고 QTL이 설명하지 못하는 polygenic은 0.06으로 설정했다. 다음으로 historical population은 3가지 집단 형태 중 하나로 구성할 수 가 있는데, 개체 수가 변하지 않는 집단과 개체 수가 변하는 집단, 그리고 병목현상이 일어난 집단이다. 이 때, 개체의 수가 변하는 historical population 집단이 가장 근친율이 낮다는 결과(Rachwicha et al., 2022)와 시뮬레이션 데이터로 질병저항성이 있는 개체에 관해 연구한 선행연구(S. Naderi et al., 2016)의 시나리오를 참고하여서 작성했다. 먼저 5200마리로 시작해서 1000세대가 지나 10,400마리를 구성하게 설정하였고, 마지막 세대의 male의 수를 400 마리로 고정했다. Population은 총 8세대로 구성했고, founder의 female은 10,000마리, male은 400마리로 historical population에서 임의적(random)으로 선발하는 것으로 지정하였다. 개체를 소(cattle)로 지정했기 때문에 자손의 수를 1마리로 정했고, 자손의 성비는 0.5, sire로의 대체율은 0.5, dam으로의 대체율은 0.2로 지정하였다. 개체 선발은 EBV가 높은 순으로, 도태는 나이가 높은 순으로 설정했고 교배계획(mating design)은 완전확률화(random)로 지정했다. 유전체(Genome) 부분에서는 이번 연구에서 필

Table 1. QMSim simulation parameter(continued)

Parameter	
Global parameter	
Heritability of trait	0.36
QTL heritability	0.30
Polygenic heritability	0.06
Historical population	
No. of generation / population size	1,000/10,400
No. of males in the last generation	400
Recent population	
No. of founder male / female	400/10,000
No. of generation	8
No. of offspring	1
No. of proportion of male progeny	0.5
Replacement ratio for sires/dams	0.5/0.2
Mating design	Random
Selection design	EBV/high
Culling design	Age

Table 1. QMSim simulation parameter

Parameter	
Genome	
No. of chromosomes	29
Marker positions	Random
Marker allele frequency	Equal
QTL position	Chr 4 : 56.7 Chr 6 : 6.4, 7.3, 62.4, 127.0, 127.1 Chr 7 : 85.4 Chr 9 : 77.2 Chr 16 : 80.4 Chr 20 : 66.5 Chr 29 : 0.294, 8.7
QTL allele frequency	Equal
QTL allele effect	Gamma (0.4)
LD	Marker and QTL

요한 다양한 유전정보를 설정하였다. 집단의 유전적 구성과 관련하여 넣은 QTL은 FMDVCP(Foot-and-Mouth Virus peptide-induced cell proliferation)로 총 12개이며, 염색체 4번에 1개, 6번에 5개, 7번에 1개, 9번에 1개, 16번에 1개, 20번에 1개, 29번에 2개 위치해 있다. 이 QTL에 따라 구제역의 면역이 되는 유무가 달라진다고 가정했다. QTL allele frequency는 모두 동일하게, QTL allele effect는 shape이 0.4인 감마분포에서 나온다고 설정했다. 이러한 모의실험을 한우로 가정하기 위하여 Illumina Hanwoo ver1(company)에서 활용되고 있는 유전 마커의 개수, 염색체의 길이, 염색체의 개수를 적용했다. 유전자 마커의 총수는 41,449개이며, SNP의 수와 길이(cM)는 한우의 유전 데이터를 참고하여 1 ~ 29번까지의 염색체에 지정했다(Lee et al., 2011). 이 marker들의 위치는 random하게 지정했고, maker allele frequency는 모두 동일하게 설정했다. 또한 이 marker들과 QTL사이에 Linkage Disequilibrium (LD)가 형성되도록 설정했다. LD란 두 marker들 사이에 무작위가 아닌 상관관계를 갖고 있는 것을 의미한다(Porto-Neto et al., 2014). 즉, LD가 높게 측정된다면 다음 세대로 유전될 때 독립적으로 유전되는 것이 아니라 연관단위로 함께 유전된다고 가정한다. 본 연구에서는 marker의 거리에 따른 LD 형성 간의 관계를 확인하고 시뮬레이션에서 설정한 maker와 QTL간의 LD가 잘 설정되었는지 확인하기 위해 PLINK프로그램(Purcell et al., 2007)을 이용하여 LD를 r^2 로 측정했다. 또한, GWAS를 이용하여 구제역 면역과 연관 되어있는 QTL과 SNP을 구하고 RF, XGBoost와 비교했다. 본 연구에서 8세대 중 마지막 세대인 10,000마리를 연구집단으로 이용하였다. 표현형 데이터를 기준으로 상위 5,000마리는 구제역 바이러스에 면역을 가지고 있는 개체, 나머지 5,000마리는 구제역 바이러스에 면역을 갖고 있지 않는 개체로 지정하였고, 면역을 가지고 있는 개체(구제역에 걸리지 않은 개체) = 0, 면역이 없는 개체(구제역에 걸린 개체) = 1로 설정하였다.

2. 질병 저항성 QTL 검출 통계모델

1) Genome wide association (GWAS)

GWAS는 여러 게놈에 있는 수십만개의 유전적 변이를 테스트하여 특정한 특성이나 질병과 연관되어진 변이를 통계적으로 찾는 방법이다(Uffelmann et al., 2021). 대부분의 GWAS에 대한 방법은 case-control 설정을 하는 방법이며, 이번 연구에서도 구제역 면역 반응에 있어서 저항성을 갖는 개체와 저항성을 갖지 않는 개체로 나누어 설계했다. GWAS 분석을 위하여 GCTA(Yang et al., 2011) 소프트웨어를 사용했으며, 다중검정(Multiple-testing)에 대한 위양성(false positive rate; FDR)을 보정하기 위하여 Bonferroni 교정을 사용했다. GWAS가 규명하고자 하는 특성에 연관된 유전적 변이를 발견할 수 있는 강력한 도구이지만, 원인이 되는 변이와 유전자를 정확하게 밝히지 못한다(Tam et al., 2019). 이를 보완하기 위하여 GWAS가 관찰한 SNP과 실제 원인 돌연변이 간의 강력한 linkage disequilibrium (LD)가 있다는 가정으로 진행되는 것을 이용하여(Liang et al., 2020), 이번 연구에서는 SNP간의 LD를 확인함으로 구제역 저항성에 영향을 주는 QTL을 얼마나 검출 할 수 있는지 확인했다.

2) Genomic BLUP (GBLUP)

GBLUP은 유전체 전체에 균일하게 위치하고 있는 대량의 SNP 정보를 활용하여 표현형과의 연관성을 통하여 SNP의 effect를 추정하고, 아직 능력을 알 수 없는 개체의 SNP 유전자형을 바탕으로 GEBV를 추정한다(Lee et al., 2022). 이를 진행하기 위해서 RENUMF90과 BLUPF90 프로그램을 사용했다. BLUPF90의 매뉴얼에 따르면, BLUPF90은 동물 육종을 위한 mixed-model 프로그램이며 BLUPF90에 사용할 실행 파일은 RENUMF90으로 만들 수 있다. GBLUP을 위해 사용되는 수학적 모델은 아래와 같다.

$$y = 1_n\mu + Zg + e$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1'n1n & 1'nZ \\ Z'1n & Z'Z + G^{-1}\,\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} 1'ny \\ Z'y \end{bmatrix}$$

y는 관측치에 대한 벡터이고, 1_n 은 1로 구성된 벡터 μ는 overall mean, Z는 SNP효과를 표현형 자료에 연결 시키는 계획행렬이며, g는 추정할 DGV의 벡터이고, e는 잔차 효과이다(Gao et al., 2012). GBLUP에서는 A행렬 대신 개체 간의 유전적 관계를 나타내는 G행렬을 이용하여 육종가를 계산한다. BLUPF90을 실행할 때, train 집단 8,000마리, test 집단 2,000마리로 설정하고 구제역에 면역이 있는 집단과 면역을 가지고 있지 않은 집단에서 무작위로 나눴으며, train 집단에는 유전 정보와 표현형 정보가 모두 있고 test 집단에는 유전 정보만 있다고 가정했다. 이렇게 계산한 GEBV 값의 정확도를 구하기 위해서 TBV와 GEBV의 상관관계를 확인했다. 이 때, 면역 유무에 대한 기준은 시뮬레이션 프로그램의 결과 파일에서 표현형 수치로 정했기 때문에 다시 GEBV를 통해 기준을 설정하기 위해서는 새로운 기준이 필요했다. 따라서, GEBV에 대한 구제역 면역 여부의 기준을 정하기 위해 GBLUP 방법을 가지고 예측한 train 집단의 값을 사용했다. train 집단의 GEBV를 질병의 면역 여부 즉, 0과 1로 나누면 중복되는 부분이 존재하게 된다. 여기서 0의 GEBV의 최소값과 1의 GEBV의 최대값의 중간 값으로 임의의 기준을 정했다. 다시 말하면, 중간 값보다 낮은 GEBV의 값을 가진 개체는 면역을 갖지 않은 개체로, 높은 값을 가진 개체는 면역을 갖고 있는 개체로 분류했다.

3) Random Forest

Random Forest(RF)은 분류, 군집분석, 회귀모형, 생존분석 등에 적용되는 모형으로, 의사결정나무 모형을 다수 만들어 더 정확한 예측을 하는 것을 목적으로 한다(Yoo, 2015). RF는 최근 유전 정보를 이용한 다수의 분석에서 꾸준히 사용되고 있다. RF는 많은 수의 의사결정나무를 생성해서 최종 클래스를 결정하는 방법인데, 그 클래스를 나누기 위한 수학적 공식은 아래와 같다.

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

$$h_l(\oslash) = h(\theta) < T$$

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

θ는 feature parameter이고, T은 한계선, h은 전체 데이터의 수, h_l은 왼쪽 노드, h_r은 오른쪽 노드이다. 파라미터의 최적값(θ*)은 정보 획득량이 가장 큰 파라미터 값을 노드의 분할 조건으로 결정한다(Kim, 2017). 여기서 정보 획득량은 어떤 사건이 얼마만큼의 정보를 줄 수 있는지 수치화 한 값을 의미한다. 정보 획득량을 알기 위해서는 엔트로피를 알아야 하는데, 엔트로피는 무질서도를 의미하며, 엔트로피가 높다는 것은 그 집단에서의 패턴, 특징을 찾기 어렵다는 것을 의미한다. 인용한 논문에 따르면, 정보 획득량은 섀넌 엔트로피로 결정된다고 한다(Hongduk Seo & Kim, 2019). 이 섀넌 엔트로피에 따라 가장 적합한 파라미터가 되는 조건을 찾을 수 있다. 이번 연구에서는 RF을 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 SNP과 QTL 유전 정보는 시뮬레이션 데이터에서 만들어진 유전 정보 데이터를 0,1,2로 바꾸어 진행했기에 개체의 수(n) X SNP과 QTL의 수(p) 크기의 데이터 프레임으로 만들어 분석했다. 이번 연구에 맞게 RF 하이퍼 파라미터인 결정트리의 개수, 노드를 분할하기 위한 최소한의 샘플 수, 리프노드

가 되기 위한 최소한의 샘플 데이터 수, 트리의 최대 깊이를 조정하기 위해서 GridSearchCV을 사용했다. 이후 변수 중요도(Variable Importance)을 통해서 어느 변수가 예측 성능에 중요한 역할을 하는지를 추정했다. 즉, 어떤 SNP과 QTL이 질병 저항성 유무를 예측하는데 중요한지 추정한 것이고, 이를 GWAS에서 추정한 SNP과 QTL과 비교했다. 또한 RF로 구한 GEBV의 정확도는 GBLUP과 마찬가지로 TBV와의 상관관계로 구했다.

4) XGBoost

XGBoost는 트리 기반 앙상블 기계학습 알고리즘이며, 약한 예측 모형들의 학습 오차에 가중치를 두고, 순차적으로 다음 학습 모델에 반영하여 손실을 최소화하는 강한 예측모형을 만든다(Woojin Yoon et al., 2021). 인용한 논문에 따르면 학습기 A가 Z를 예측할 확률은 아래와 같다.

$$z = A(x) + Error$$

Error에 대해서 정밀하게 분류하는 학습기가 B라고 한다면(Error > $Error_2$)

$$Error = B(x) + Error_2$$

다시 $Error_2$에 대해서 정밀하게 분류하는 학습기를 C라고 한다면($Error_2$ > $Error_3$)

$$Error_2 = C(x) + Error_3$$

이 식들을 다시 Z를 구하는 식에 대입하게 된다면

$$Z = A(x) + B(x) + C(x) + Error_3$$

이와 같아진다고 발표했으며 이 방법을 적용했을 때, 학습기 A만 단독으로 사용했을 때보다 더 높은 정확도를 보인다고 설명했다. 하지만 모든 모델이 같은 비율을 하고 있기 때문에 오류가 높아질 수 있다. 따라서, 각 모델에 가중치를 두고, 최적의 비중을 찾으면 아래의 식과 같이 좋은 분류 모델이 된다(Youngjin Han & Cho, 2021). 이렇게 해서 XGBoost는 데이터를 분류하게 된다.

$$Z = \alpha * A(x) + \beta * B(x) + r * C(x) + Error_4$$

XGBoost의 train과 test 집단은 위의 분석방법과 동일하게 진행했다. XGBoost에서도 유전정보, 면역 유무의 데이터와 표현형 수치 데이터를 가지고 분석을 했으며, RF을 분석했을 때 사용했던 데이터 형태로 진행했다. 분석을 진행하기 위해서 XGBoost의 하이퍼파라미터인 트리 모델의 개수, 학습 단계 반영률, 트리의 최대 깊이, 각 트리 마다 feature 샘플링 비율, child에서 필요한 모든 관측치에 대한 가중치의 최소합을 GridSearchCV를 이용해서 조정했다. 이후, RF와 동일하게 변수 중요도를 통해서 어느 SNP과 QTL이 질병 저항성 유무를 예측하는데 중요한지 추정했다. 또한 XGBoost를 통하여 구한 GEBV의 정확도는 RF와 동일한 방법으로 구했다.

5) 정확도 추정(Confusion Matix)

혼동행렬은 분류 모델의 성능을 측정하기 위해서 사용된다. 혼동행렬은 TP(True Positive), FP(False Positive), TN(True Negative), FN(False Negative)로 나뉘며 정확도와 민감도, 특이도를 아래와 같은 식으로 구할 수 있다.

$$\text{정확도} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$\text{민감도} = \frac{TP}{(TP + FN)}$$

$$\text{특이도} = \frac{TN}{(TN + FP)}$$

정확도는 전체 분류된 데이터 중 Positive가 Positive로, Negative가 Negative으로 바르게 분류된 정도를 의미하며, 일반적으로 민감도는 Positive(0)가 Positive(0)로 제대로 분류된 정도, 특이도는 Negative(1)가 Negative(1)로 바르게 분류된 정도를 의미한다. 이번 연구에서 혼동행렬은 python의 scikit-learn으로 구현했고 positive의 값을 구제역 면역이 있는 개체(0)로 설정했다. 다시 말하면 민감도는 구제역 면역이 있는 것을 있다고 한 확률, 특이도는 구제역 면역이 없는 것을 없다고 한 확률을 나타낸다. GBLUP, RF, XGBoost를 사용하여 GEBV를 구한 후, 설정한 기준에 의하여 구제역 면역을 가진 개체(0)와 없는 개체(1)로 분류했고, 이를 올바르게 분류되었는지 혼동행렬을 이용해 정확도, 민감도, 특이도를 비교했다.

RESULTS AND DISCUSSION

1. 구제역 면역에 영향을 미치는 True QTL 검출

모의실험을 통해 생성한 개체의 SNP 및 True QTL과 설정된 파라미터 간의 LD 형성 여부를 확인하였다. 먼저, 마커들 간의 거리와 LD 형성 간의 관계를 시각적으로 나타내는 그래프를 생성하였으며, 결과적으로 거리가 가까워질수록 LD가 잘 형성되는 것을 확인하였다(Figure 2). 또한, SNP과 QTL 간의 LD 형성은 Haploview 프로그램(Barrett et al., 2005)을 사용하여 확인하였으며, 일부 영역에서는 LD가 약하지만 Q3, Q10, Q11, Q12와 같이 강한 LD가 형성된 영역이 검출되었다(Figure 3).

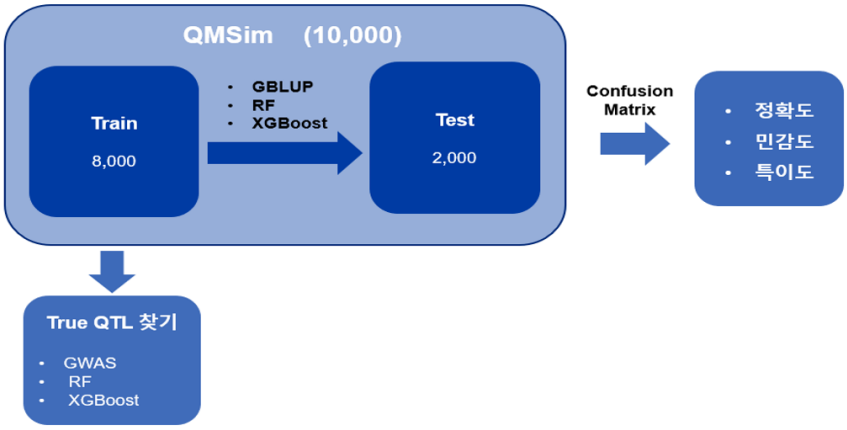

Figure 1. A schematic representation of this study

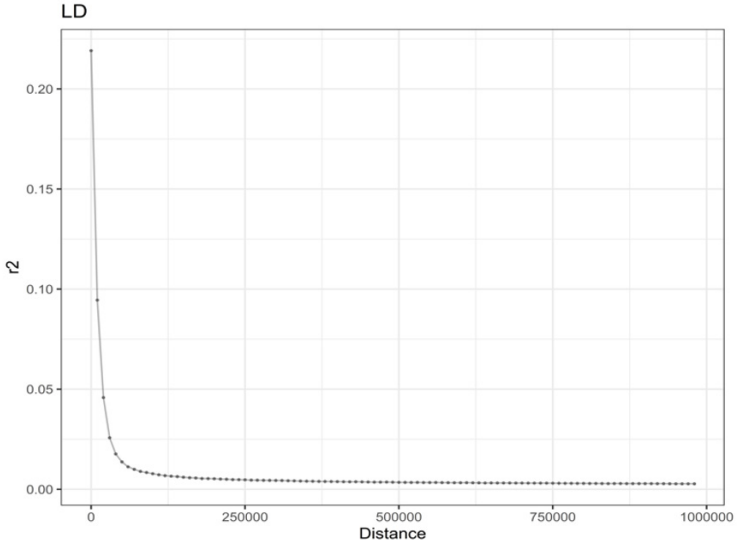

Figure 2. LD plot based on marker distance in this study.

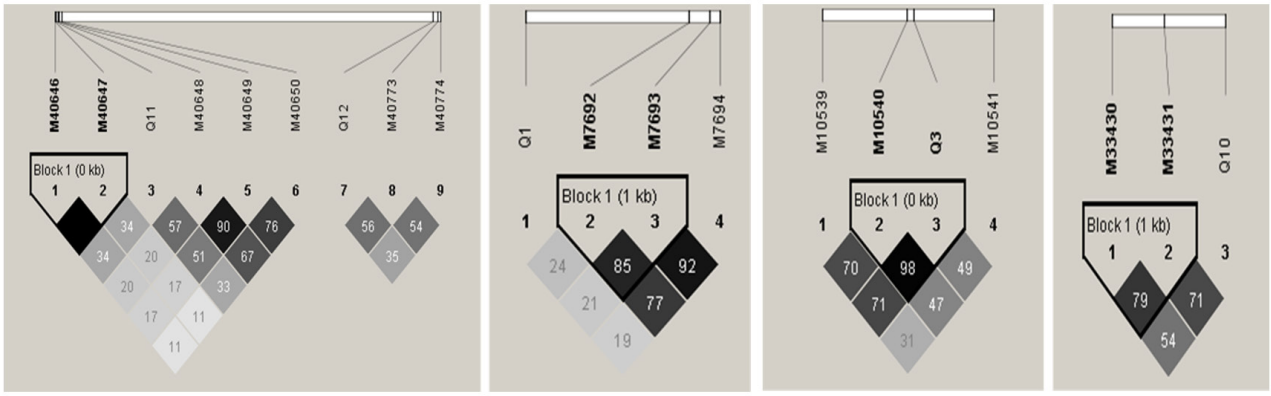

Figure 3. Linkage disequilibrium (LD) formation of true QTL and adjacent SNP marker in historical population.

GWAS는 질병과 관련된 유전자의 위치를 파악하기 위한 효과적인 방법임으로, 모의실험에서 생성한 구제역 면역을 가진 개체와 가지지 않은 개체 간의 대립 유전자 빈도를 비교함으로써 어떤 SNP과 QTL이 구제역 면역과 관련이 있는지 확인했다(Chung, 2012). 먼저, PLINK 프로그램을 사용하여 Quality Control(QC)를 수행하고, GCTA 프로그램을 사용하여 Genetic Relationship Matrix(GRM)을 생성한 후 GWAS를 수행했다(Yang et al., 2011). GWAS 결과로 나온 QTL과 SNP 중 Bonferroni correction을 기준으로 유의적인 QTL과 마커를 약 20개를 확인하였다(Figure 4). GWAS를 통하여 검출한 QTL과 SNP은 p-value가 낮은 순서로 정리하였고, RF와 XGBoost로 검출한 QTL과 SNP은 변수 중요도가 높은 순서로 정렬한 후 세 가지의 방법을 통하여 나온 각각의 유의미한 QTL과 마커를 비교했다(Table2). 그 결과, 각 방법에 따라서 QTL과 SNP이 갖는 구제역 저항성 형질에 영향을 주는 정도와 개수에서 차이가 발생했다. GWAS에서는 12개의 QTL 중 6개인 Q1, Q2, Q3, Q10, Q11, Q12를 연관성을 가진 것으로 검출했으며, 검출 정확도는 0.5였다. 반면에 RF에서는 가장 많은 7개 Q1, Q2, Q3, Q6, Q10, Q11, Q12을 검출했으며, 정확도는 0.58였다. 마지막으로 XGBoost에서는 3개인 Q7, Q10, Q11만이 검출되어 검출 정확도는 0.25로 나타났다. 이러한 결과를 통해 GWAS, RF, XGBoost의 결과에 차이가 있으며, 모든 방법에서 12개의 QTL을 전부 검출하지 못했으므로 성능 향상이 필요함을 확인할 수 있었다. 특히, QTL 검출에서 가장 적은 결과를 얻은 XGBoost의 분석 방법을 개선함으로써 QTL을 더 정확하게 추정할 수 있을 것으로 판단된다.

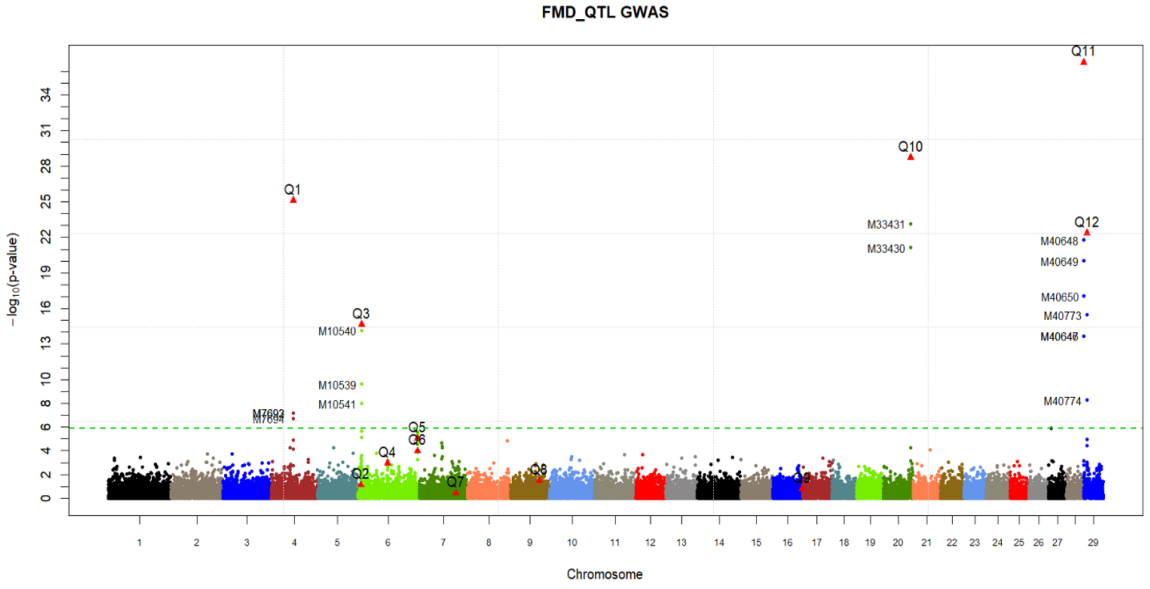

Figure 4. Manhattan plot of genome-wide association study for foot-and-mouth disease. Q1~Q12 indicates true QTL and LD markers around true QTL

Table 2. Comparison of detection performance for true QTL using GWAS, RandomForest(RF), XGBoost.

	GWAS	RF	XGBoost
1	Q11	Q11	M40648(Q11)
2	Q10	Q10	M33431(Q10)
3	Q1	Q33431	M20412
4	M33431	Q40648	M12614
5	Q12	Q33430	M29863
6	M40648	M40649	M37953
7	M33430	MQ12	M7405
8	M40649	Q1	M10245
9	M40650	M40647	M14499(Q7)
10	M40773	M40646	M3248
11	Q3	M40773	M19707
12	M10540	M40650	M36172
13	M40646	M40773	M2614
14	M40647	Q3	M14931
15	M10539(Q2)	M10540(Q2)	M5083
16	M40774	M10541	M24702
17	M10541	M5591	M21928
18	M7692	M12873(Q6)	M12738
19	M7693	M12874	M7018
20	M7694	M373	M13349
Acc	0.5	0.58	0.25

2. 구제역 면역 관련 개체 예측 정확도 비교

이산형인 구제역 표현형을 이용하여 개체의 유전능력을 예측하였다. 예측 방법은 GBLUP, RandomForest(RF), 그리고 XGBoost를 이용하여 개체의 능력을 예측하였고, GEBV의 정확도는 실제 육종가(True breeding value; TBV)와의 상관관계를 통해 평가했다. GBLUP, RF, XGBoost로 구한 각 GEBV와 TBV와의 상관관계 값은 각 0.32, 0.35, 0.33이다(Table3). 이렇게 구한 GEBV값으로 질병 저항성을 가지는 개체를 선발하기 위해서는 이 특성을 가지는 GEBV값에 대한 기준이 필요한데, 이에 대해 연구된 바 없으므로 구제역 면역의 유무를 나누는 새로운 기준이 필요하다. 전체집단의 표현형 수치를 기준으로 구제역 면역 여부를 설정했으므로 질병 저항성을 갖는 개체와 갖지 않는 개체의 GEBV 값 중 중복되는 부분이 존재하게 된다. 이 중복 구간을 중간으로 나누는 임의의 값을 설정하여, 이 값보다 낮으면 면역이 없는 개체로, 높으면 면역이 있는 개체로 분류하였다. 그 이후, GBLUP, RF, XGBoost를 사용하여 예측한 전체 정확도, 민감도, 특이도를 비교하였다. 전체 정확도는 XGBoost가 0.53, RF가 0.52, GBLUP가 0.51 순으로 비슷하게 나타났으나, 민감도는 RF가 0.98, XGBoost가 0.97, GBLUP가 0.19 순으로 나타났다. 특이도는 GBLUP이 0.83, XGBoost가 0.08, RF가 0.05 순으로 나타났다(Table 4). GBLUP이 전반적으로 우수하게 예측할 것이라고 예상한 것과 달리, 전체 정확도, 민감도에서 비교적 낮

Table 3. Correlation between TBV and GEBV calculated using GBLUP, RF and XGBoost

	GBLUP	RF	XGBoost
Correlation	0.32	0.35	0.33

Table 4. Comparison the accuracy, sensitivity, and specificity of GBLUP, RF and XGBoost

	GBLUP	RF	XGBoost
전체 정확도	0.51	0.52	0.53
민감도	0.19	0.98	0.97
특이도	0.83	0.05	0.08

은 정확도를 보였다. 이번 연구와 비슷하게 QMSim 프로그램을 이용해 데이터를 만들고 이를 GBLUP과 RF 등의 방법으로 분석한 후, 정확도를 비교한 결과에서도 GBLUP보다 RF에서 높은 정확성을 보인 것을 확인 할 수 있었다(Naderi & Sadeghi, 2019). GBLUP이 혈통정보를 이용하여 추정한 BLUP 모델보다 높은 정확성을 갖고 있다고는 하지만(VanRaden et al., 2009), 다른 머신 러닝, 딥 러닝과 단일로 비교했을 때는 그 정확성은 독보적으로 높게 나오지는 않는다는 것을 확인했다. 오히려 GBLUP 하나만 사용했을 때보다, 딥 러닝의 방법을 앙상블해서 사용했을 때 더욱 좋은 정확도를 얻었다는 연구도 있다(Myungjin Jang et al., 2022). 이와 같이 비록 연구 결과에서 GBLUP의 결과가 예상했던 바와는 다르게 정확도가 낮게 나타나는 것을 확인했지만, 앞서 언급한 선행연구와 마찬가지로 GBLUP으로만 예측을 하는 것이 아닌, 더욱 잘 예측할 수 있는 방법을 활용하고, GBLUP 하나의 방법뿐 만 아니라 머신 러닝, 딥러닝과 함께 앙상블해서 사용한다면 구제역 저항성 개체를 선발함에 있어서 더욱 높은 정확성을 나타나며, 앞으로 구제역의 피해를 최소화하는데 도움이 되어질 것이라고 기대한다. 또한, GBLUP의 정확도를 개선하기 위해서 질병 저항성을 가지는 개체의 GEBV에 대한 연구와 따라서 질병 저항성 개체 선발을 위한 다양한 분야에서의 연구가 추가적인 필요하다고 생각된다.

요약

이번 연구는 구제역 저항성 개체를 선발하기 위해 시뮬레이션 프로그램을 이용하여 모의실험을 진행했다. 시뮬레이션 프로그램을 이용해 구제역 면역이 있는 개체와 없는 개체를 설정했으며, 이를 통해 유전 정보와 표현형 데이터를 생성했다. 이후, Genome Wide Association Study(GWAS)를 이용해 모의실험에 사용하는 유전 정보 중 구제역 면역에 연관이 있는 QTL을 선별했다. 또한, RandomForest(RF)와 XGBoost를 이용하여 GWAS로 선별한 QTL과 비교했다. 시뮬레이션으로 만든 41,461개의 QTL과 SNP 중에서 연관이 있다고 나온 것은 총 20개이다. GWAS, RF와 XGBoost을 비교했을 때 연관이 있다고 추정한 QTL은 RF가 7개로 가장 많이 추정되었고, 그 다음 GWAS가 6개, XGBoost가 3개 추정되었다. 그리고 Genomic Best Linear Unbiased Prediction (GBLUP)과 RF, XGBoost을 이용해 구제역 면역을 가진 개체와 갖지 못한 개체를 분류했고, 분류한 것에 대한 전체 정확도와 민감도, 특이도를 혼동행렬(confusionMatrix) 이용해 구했으며 각 값을 비교했다. 전체 정확도는 XGBoost 0.53, RF 0.52, GBLUP 0.51 순으로, 민감도는 RF 0.98, XGBoost 0.97, GBLUP 0.19 순으로 나왔다. 또한, 특이도는 GBLUP 0.83, XGBoost 0.08, RF 0.05 순으로, 전체 정확도와 민감도 부분에서 XGBoost가 가장 높게, GBLUP이 가장 낮게 나왔다. 따라서, GBLUP으로만 분석을 하는 것이 아니라, 다른 방법과 함께 앙상블해서 사용한다면 구제역 저항성 개체를 예측함에 있어서 더 좋은 성능을 예상하며, 이후 발생할 구제역의 피해를 최소화하는데 도움이 될 수 있기를 기대한다.

키워드: 구제역 저항성, GBLUP, Randomforest, XGBoost,

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Small and Medium-sized Enterprises(SMEs) and Startups(MSS), Korea, under the “Regional Specialized Industry Development Program+(R&D, S3370836)” supervised by the Korea Technology and Information Promotion Agency for SMEs(TIPA).

This work was partly supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) (321082-3).

REFERENCES

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21(2), 263-265.

Blazer, D. G., & Hernandez, L. M. (2006). Genes, behavior, and the social environment: Moving beyond the nature/nurture debate.

Chung, S. J. (2012). 전장 유전체 연관 연구. Program at a Glance, 30(1), 30.

Egger-Danner, C., Willam, A., Fuerst, C., Schwarzenbacher, H., & Fuerst-Waltl, B. (2012). Hot topic: Effect of breeding strategies using genomic information on fitness and health. Journal of dairy science, 95(8), 4600-4609.

Gao, H., Christensen, O. F., Madsen, P., Nielsen, U. S., Zhang, Y., Lund, M. S., & Su, G. (2012). Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics Selection Evolution, 44(1), 8. https://doi.org/10.1186/1297-9686-44-8

Gowane, G., Sharma, A., Sankar, M., Thirumurugan, P., Narayanan, K., Subramaniam, S., & Pattnaik, B. (2013). Evaluation of genetic and environmental parameters determining antibody response induced by vaccination against Foot and Mouth Disease. Agricultural Research, 2, 140-147.

Grubman, M. J., & Baxt, B. (2004). Foot-and-mouth disease. Clinical microbiology reviews, 17(2), 465-493.

Han, M. (2021). 가축 살처분의 보상 등 공법적 쟁점에 대한 소고. In 토지공법연구 (Vol. 94, pp. 237-258).

Hyomin Park, & Jin, B. (2017). 가축살처분 작업 트라우마와 작업자의 정신건강. In 한국사회학회 사회학대회 논문집 (pp. 877-878).

Jang, M., Lim, D., Park, W., & Park, J.-E. (2022). 한우 도체형질의 합성곱신경망을 이용한 유전체 예측 정확도 추정. In 한국산학기술학회 논문지 (Vol. 23, pp. 516-523).

Joung Min Choi, Jihyun Kim, Hyunkyung Choo, Chaelin Park, & Chae, H. (2022). 머신러닝 기반의 농업 유전자원 데이터 분석 플랫폼. In 정보과학회 컴퓨팅의 실제 논문지 (Vol. 28, pp. 57-62).

Kim, D. (2011). 우리에게 구제역은 무엇인가?: 국가 주도의 살처분 정책과 그 함의. In 민주사회와 정책연구 (pp. 13-40): 한신대학교 민주사회정책연구원.

Kim, J., Lee, K. B., & Hong, S. G. (2017). ECG-based biometric authentication using random forest. In (Vol. Journal of the Institute of Electronics and Information Engineers, Vol. 54, No. 6, pp. pp. 100-105).

Kramer, M., Erbe, M., Seefried, F. R., Gredler, B., Bapst, B., Bieber, A., & Simianer, H. (2014). Accuracy of direct genomic values for functional traits in Brown Swiss cattle. Journal of dairy science, 97(3), 1774-1781.

Leach, R. J., Craigmile, S. C., Knott, S. A., Williams, J. L., & Glass, E. J. (2010). Quantitative trait loci for variation in immune response to a Foot-and-Mouth Disease virus peptide. BMC genetics, 11(1), 107. https://doi.org/10.1186/1471-2156-11-107

Lee, B. Y., Lee, K. N., Lee, T., Park, J. H., Kim, S. M., Lee, H. S., Chung, D. S., Shim, H. S., Lee, H. K., & Kim, H. (2015). Bovine Genome-wide Association Study for Genetic Elements to Resist the Infection of Foot-and-mouth Disease in the Field. In Asian-Australas J Anim Sci (Vol. 28, pp. 166-170).

Lee, G. H., Lee, Y. S., Moon, S. J., & Kong, H. S. (2022). The Accuracy of Genomic Estimated Breeding Value Using a Hanwoo SNP Chip and the Pedigree Data of Hanwoo Cows in Gyeonggi Province. Journal of Life Science, 32(4), 279-284.

Lee, S., Cho, Y., Lim, D., Kim, H., Choi, B., Park, H., Kim, O., Kim, S., Kim, T., & Yoon, D. (2011). Linkage disequilibrium and effective population size in Hanwoo Korean cattle. Asian-Australasian Journal of Animal Sciences, 24(12), 1660-1665.

Liang, B., Ding, H., Huang, L., Luo, H., & Zhu, X. (2020). GWAS in cancer: progress and challenges. Molecular Genetics and Genomics, 295(3), 537-561. https://doi.org/10.1007/s00438-020-01647-z

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. genetics, 157(4), 1819-1829.

Naderi, S., Yin, T., & König, S. (2016). Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. Journal of dairy science, 99(9), 7261-7273.

Naderi, S., Yin, T., & König, S. (2016). Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. In Journal of dairy science (Vol. 99, pp. 7261-7273).

Naderi, Y., & Sadeghi, S. (2019). Assessment of the genomic prediction accuracy of discrete traits with imputation of missing genotypes. Animal Science Papers and Reports, 37(2), 149-168.

Pak, S. I., & Bae, S. H. (2012). A Space-Time Cluster of Foot-and-Mouth Disease Outbreaks in South Korea, 2010~ 2011. In Journal of the Korean association of regional geographers (Vol. 18, pp. 464-472).

Palaiokostas, C. (2021). Predicting for disease resistance in aquaculture species using machine learning models. In Aquaculture Reports (Vol. 20, pp. 100660).

Porto-Neto, L. R., Kijas, J. W., & Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genetics Selection Evolution, 46(1), 22. https://doi.org/10.1186/1297-9686-46-22

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

Rachwicha, C., Aryuman, S., Phatsara, C., & Chongkasikit, N. (2022). Impact of Historical Population Structure Different on Inbreeding Rate of Dairy Cattle Population. In The 1st International Conference on" Innovation for Resilient Agriculture (pp. 9).

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5), 680-681. https://doi.org/10.1093/bioinformatics/btp045

Schaeffer, L. (2014). Strategies for Using Genomics to Improve Swine Resistance to PRRS. In.

Seo, H. D., & Kim, E. M. (2019). 랜덤포레스트와 서포트벡터머신 기법을 적용한 포인트 클라우드와 실감정사영상을 이용한 객체분류. In 한국측량학회지 (Vol. 37, pp. 405-416).

Suh, C.-H., Kim, J., D, H., Kim, K., Chung, C., jeong, M.-k., & Im, J. (2000). 구제역의 파급 영향과 정책 과제. In.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 20(8), 467-484. https://doi.org/10.1038/s41576-019-0127-1

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1(1), 59.

VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., & Schenkel, F. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of dairy science, 92(1), 16-24.

Xiang, T., Li, T., Li, J., Li, X., & Wang, J. (2023). Using machine learning to realize genetic site screening and genomic prediction of productive traits in pigs. In The FASEB Journal (Vol. 37, pp. e22961).

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American journal of human genetics, 88(1), 76-82.

Yeo-Joo Lee, Jia-Qi Chu, Seo-Yong Lee, Su-Mi Kim, Kwang-Nyeong Lee, Young-Joon Ko, Hyang-Sim Lee, In-Soo Cho, Seok-Hyun Nam, & Jong-Hyeon Park. (2011). 구제역 Asia1 백신주의 전체 염기서열분석 및 특성. In 한국가축위생학회지 (KOJVS) (Vol. 34, pp. 95-102): 한국동물위생학회.

Yoo, J. E. (2015). 랜덤 포레스트. In 교육평가연구 (Vol. 28, pp. 427-448): 한국교육평가학회.

Young-Jin Han, & Joe, I.-W. (2021). XGBoost 기반의 조기 중지를 활용한 광고 클릭 예측 방안. In 한국통신학회논문지 (Vol. 46, pp. 993-1000).

Yu, H.-S. (2014). 주요 가축질병 발생현황과 예방대책-구제역의 백신접종 청정국 지위 획득 이후 방역대책. 사료, 23-25.

Yun, W., seo, d., Min, S., & Nam, H. (2021). RandomForest 와 XGBoost 를 활용한 유방암 종양 분류. In 한국통신학회 학술대회논문집 (pp. 113-114).