Dimensionality of genomic information and its impact on GWA and variant selection: a simulation study

Sungbong Jang¹*, Shogo Tsuruta¹, Natalia Galoro Leite¹, Ignacy Misztal¹, Daniela Lourenço¹

¹Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602

*Corresponding author

E-mail addresses:
SJ: jsbng@uga.edu
ST: shogo@uga.edu
NL: nataliagaloro@uga.edu
IM: ignacy@uga.edu
DL: danilino@uga.edu
Abstract

Background

Identifying true-positive variants in genome-wide associations (GWA) depends on several factors, including the number of genotyped individuals. The limited dimensionality of the genomic information may give insights into the optimal number of individuals to use in GWA. This study investigated different discovery set sizes in GWA based on the number of largest eigenvalues explaining a certain proportion of variance in the genomic relationship matrix (G). An additional investigation included the change in accuracy by adding variants, selected based on different set sizes, to the regular SNP chips used for genomic prediction.

Methods

Sequence data were simulated containing 500k SNP with 200 or 2000 quantitative trait nucleotides (QTN). A regular 50k panel included one every ten simulated SNP. Effective population size (Ne) was 20 and 200. The GWA was performed with the number of genotyped animals equivalent to the number of largest eigenvalues of G (EIG) explaining 50, 60, 70, 80, 90, 95, 98, and 99% of the variance. In addition, the largest discovery set consisted of 30k genotyped animals. Limited or extensive phenotypic information was mimicked by changing the trait heritability. Significant and high effect size SNP were added to the 50k panel and used for single-step GBLUP with and without weights.

Results

Using the number of genotyped animals corresponding to at least EIG98 enabled the identification of QTN with the largest effect sizes when Ne was large. Smaller populations required more than EIG98. Furthermore, using genotyped animals with higher reliability (i.e., higher trait heritability) helped better identify the most informative QTN. The greatest prediction accuracy was obtained when the significant or the high effect SNP representing twice the number of simulated QTN were added to the 50k panel. Weighting SNP differently did not increase prediction accuracy, mainly because of the size of the genotyped population.

Conclusions

Accurately identifying causative variants from sequence data depends on the effective population size and, therefore, the dimensionality of genomic information. This dimensionality can help identify the suitable sample size for GWA and could be considered for variant selection. Even when variants are accurately identified, their inclusion in prediction models has limited implications.
Background

Several factors influence the statistical power to identify causative variants in genome-wide associations (GWA), including the number of quantitative trait nucleotides (QTN) affecting the trait, the number of single nucleotide polymorphisms (SNP) in the discovery panel, the number of genotyped individuals [1], and the size of the genome blocks segregating in the population [2], among others. Those genome blocks are chromosome segments inherited from founders and are subject to recombination every generation. Stam [3] showed that segments are of different sizes but with a mean size of $1/4N_e$, where $N_e$ is the effective population size. Given a species with a genome length equal to $L$ Morgans, the number of independent chromosome segments ($M_e$) segregating in a population can be calculated as $4N_eL$.

Animal populations have lower $N_e$ than human populations, implying smaller $M_e$. Pocrnic et al. [4] showed that although millions of individuals can be genotyped, non-redundant information is finite, which means the genomic information has a limited dimensionality; therefore, the additive genetic information in a population is contained in a limited $M_e$. The same authors related the limited dimensionality to $M_e = 4N_eL$ and observed that this quantity corresponds to the number of largest eigenvalues explaining 98% (EIG98) of the variance of the genomic relationship matrix ($G$). In cattle populations, EIG98 varies from 10K to 14K and is about 4K in pigs and chickens. The minimum number of SNP needed to cover those segments is approximately $12M_e$ [5].

With the availability of sequence information, causal variants are expected to be in the data, generating more opportunities for discovery than with mid-density SNP panels [6]. When the causal variants are known and included in the usual SNP panels, accuracy of predicting genomic estimated breeding values (GEBV) should increase. This is clearly observed in
simulated studies where the QTN and their effects are known [7, 8]. However, the accuracy increased by using significant variants from the sequence in real populations is almost inexistent [9-11]. This raises a question on the effectiveness of GWA in real populations. Although most traits of economic importance in farm animal populations are polygenic, very few peaks are usually statistically associated with traits of interest.

Misztal et al. [12] investigated the distribution of SNP around the QTN and the ability to identify QTN depending on the \( N_e \) in simulated populations. They found that identifying QTN in populations with small \( N_e \) (i.e., 60) required three times more genotyped animals with phenotypes than in populations with large \( N_e \) (i.e., 600). However, not all simulated QTN were identified, independently of the \( N_e \) or amount of data. Distinguishing between noise and the true signal is more difficult in small populations because of longer chromosome segments and the uncertainty about the exact QTN location. Additionally, the level of noise may mask the signal, preventing associations. With sequence data, a clear GWA resolution for small populations may be even harder to achieve due to the reasons mentioned above.

Although it is well-known that increasing the sample size for GWA improves the resolution, the links among the number of genotyped individuals, \( N_e, M_e \), and GWA resolution are missing. Additionally, understanding the appropriate sample size for variant discovery, especially with sequence data, can help to alleviate both the economic and computational costs for practical applications. Based on the limited dimensionality of the genomic information, there may be an optimal number of animals that carry all the independent chromosome segments segregating in the population, and consequently, all the genomic information available in the population [4]. When animals have lots of information, GEBV are estimated with high accuracy. Knowing that GEBV can be backsolved to SNP effects raises the question on whether GWA resolution is high when \( M_e \) animals with high accuracy GEBV are used.
Therefore, we hypothesize that the ability to identify causative variants is high when the sample size for GWA approaches $M_e$, and using a larger sample size may not further improve GWA resolution. Here we used the number of eigenvalues explaining different proportions of the variance in $G$ to assess the dimensionality of the genomic information and used this number as the sample size in GWA. We used simulated populations with varying $N_e$, number of QTN, and amount of information on genotyped individuals. We also evaluated the impact of incorporating the pre-selected variants, from GWA with different sample sizes based on dimensionality, to a 50k SNP chip for genomic prediction using weighted and unweighted single-step GBLUP.

**Methods**

**Data simulation**

QMSim [13] was used to simulate a quantitative trait with 0.3, 0.9, and 0.99 heritability. Different heritabilities mimicked limited or extensive phenotypic information. The historical population was simulated for 2,000 non-overlapping generations with an increase in size from 1,000 (generation -2,000) to 50,000 (generation -1,000), and a decrease from 50,000 (generation -999) to 20,000 (generation 0) to create LD and mutation–drift equilibrium. Random mating and no selection or migration were assumed in the historical population. Recent populations of $N_e$ equal 20 ($N_e$20) and 200 ($N_e$200) were simulated by changing the number of breeding males from 5 to 50 but keeping the number of females at 15,000. The founders of the recent populations came from generation 0 of the historical population. Twenty generations of random mating were carried out, considering a replacement rate of 80% for sires and 30% for dams. Animals were randomly selected and culled based on age. A total of 315,005 and 315,050 animals were generated in the recent population for $N_e$20 and $N_e$200, respectively.
However, only animals from generations 11–20 had phenotypic and pedigree information that was used for the current study. Of those, 75,000 animals from generations 16-20 were genotyped (N = 15,000 in each generation). The phenotype was the sum of an overall mean equal to 1.0, true breeding value (TBV), and random residual effect. The phenotypic variance was set to 1.0, whereas the additive genetic variance was 0.3, 0.9, or 0.99, all explained by the simulated QTN.

To mimic the bovine genome, we simulated 29 chromosomes with a total length of 23.19 Morgans. The overall number of SNP was 500,000, all with minor allele frequency greater than 0.05, whereas QTN numbers were 200 and 2,000 for the scenarios Q200 and Q2000, respectively. Biallelic SNP and QTN were randomly placed on each chromosome, with numbers varying from 9,000–35,000 SNP and 8–31 (Q200) or 80–320 (Q2000) QTN. The QTN effects were sampled from a gamma distribution with shape parameter 0.4 and scale parameter calculated internally for a genetic variance of 0.3, 0.9, and 0.99, depending on the scenario. A recurrent mutation rate of $2.5 \times 10^{-5}$ was assumed for both SNP and QTN. A regular 50k panel was created for genomic predictions (GP) that included one every ten simulated SNP. Because different simulation replicates would generate different QTN positions and effects, no replicate was used to obtain consistent GWA results.

Genotype scenarios – heritability and the presence of QTN in the data

Trait heritabilities of 0.3, 0.9, and 0.99 were simulated to represent the animals with low reliability of EBV (H30), high reliability of EBV (H90), and very high reliability of EBV (H99), respectively. Therefore, higher heritabilities mean more information was added to the simulated animals without directly changing the number of records assigned to them [14]. Further, sequence data scenarios were created after the simulation. We assumed that QTN were in the
data (with QTN), following the general assumption for sequence data; therefore, the SNP and QTN files were combined based on the corresponding maps. Descriptions for all the scenarios and combinations are in Table 1.

**Discovery, training, and test sets**

Before the GWA analyses, all genotyped animals were separated into three non-overlapping data sets: discovery, training, and test. The test set was composed of genotyped animals from the last generation (N = 15,000), whereas the remaining genotyped animals (N = 60,000) were randomly assigned to the discovery and training sets (N = 30,000, respectively). To test the possible bias in GP by using the same data set for discovery and training, two different schemes were designed: 1) discovery = training: genotyped animals used for discovery were further used for training, 2) discovery ≠ training: a different set of genotyped animals were used for discovery and training.

**EIGx scenarios for discovery and training**

Different scenarios were made based on the dimensionality of the genomic information to investigate the effect of sample size for discovery and training. The number of genotyped animals in each discovery and training set (EIGx) was equivalent to the number of largest eigenvalues explaining x percent of the variance in \( G \), where \( x \) assumed the values 50, 60, 70, 80, 90, 95, 98, or 99. For example, the number of largest eigenvalues explaining 50% of the variance in \( G \) was 530 in the \( N_e200 Q2000 H30 \) scenario (Table 2); thus, the size of discovery and training sets in scenario EIG50 was set to 530. In addition, one extra scenario (ALL) in which the discovery and training sets consisted of all available genotyped animals (N = 30,000)
was also evaluated. The number of largest eigenvalues explaining x percent (50, 60, 70, 80, 90, 95, 98, 99) of the variance in $G$ was computed by squaring the singular values from the matrix of genotypes centered for current allele frequencies ($M$), using all the simulated genotyped animals ($N = 75,000$). The singular value decomposition was done in preGSf90 [15]. For that, $G$ was constructed using all simulated SNP without QTN. All genotyped animals for each discovery and training set were randomly selected beginning from the scenario explaining the least proportion of variance (EIG50). Ensuring consistent results involved keeping all the animals from a previous scenario when moving to the next one, e.g., genotyped animals in EIG60 contained the ones from EIG50. The number of genotyped animals for all scenarios used as discovery and training sets is described in Table 2.

Preselection of variants (TOPv scenarios)

Different numbers of variants were selected from GWA to be included in the 50k SNP panel for GP. Each QTN scenario had a specific number of selected SNP based on the order of the p-values (TOPv) or statistical significance using Bonferroni corrected p-values (SIG). For Q200, $v$ corresponded to 10, 50, 100, 200, and 400, whereas for Q2000 $v$ assumed the values of 10, 100, 500, 1000, 2000, and 4000.

Association among number of significant QTN, sample size, and EBV reliability

In the current study, we approximated sample size based on the total proportion of variance explained by significantly identified QTN from the results of different heritability scenarios. This approximation was done by local polynomial regression [16] using the ‘loess’ and ‘approx’ function in the R, and the resulting sample size was represented by $SS_{pol}$. 
The main purpose of comparing scenarios with different heritabilities was to investigate the effect of using genotyped animals with low to high reliability of EBV on GWA performance; however, not all the genotyped animals have high reliability of EBV. Therefore, in this study, we investigated the corresponding sample size from low to high heritability scenarios at the point where the same percentage of variance was explained; we estimated the sample size using H30 as a benchmark. This helped us identify how many samples are needed for GWA given the average reliability of breeding values for the animals in the population and the benchmark reliability. For that, derived an equation to estimate the sample size relating the total number of samples, $M_e$, $N_e$, proportion of additive genetic variance explained by significant QTN ($\%Var$), and reliability of EBV as:

$$SS_{rel} = \frac{N_s\text{rel}(\log(\frac{\sigma_{QTN}^2}{N_s}))\lambda\sigma_{QTN}^2}{\text{rel}_t(\frac{\sigma_{QTN}^2}{\text{rel}_t} + \ln(M_e))}$$

in which $SS_{rel}$ is the approximated sample size for target reliability, $N_s$ is the benchmark sample size, $\text{rel}$ and $\text{rel}_t$ are the benchmark and target reliabilities of EBV (heritability), $\lambda$ is a constant equal to 0.4, and $\sigma_{QTN}^2$ is the $\%\text{Var}$ explained by identified QTN. The total proportion of genetic variance explained by the identified QTN was calculated as the sum of the genetic variance explained by each QTN. As QTN effects were given by the simulation, the percentage of genetic variance explained by individual QTN was calculated as:

$$\%\text{Var} = \frac{2pq(\beta)^2}{\sigma_a^2}$$

where the $p$ and $q$ are the major and minor allele frequency of the QTN, $\beta$ is the QTN effect, $\sigma_a^2$ is the total additive genetic variance of the model.

**Models and analyses**

**Genome-wide associations**
Efficient mixed-model association expedited (EMMAX) was performed using Gemma software [17], with the following model:

\[
y = \mathbf{1} \mu + \mathbf{X}_i b_i + \mathbf{Z}_u + \mathbf{e},
\]

where \( y \) is a vector of phenotypes, \( \mu \) is an overall mean, \( \mathbf{X}_i \) is a vector of genotypes for \( i \)th SNP, \( b_i \) is the substitution effect of the \( i \)th SNP, \( \mathbf{Z} \) is an incidence matrix for vector \( \mathbf{u} \), and \( \mathbf{u} \) is a vector of random additive genetic effects, with \( \mathbf{u} \sim N(0, \mathbf{G} \sigma_u^2) \), and \( \mathbf{e} \) is a vector of residuals, with \( \mathbf{e} \sim N(0, \mathbf{I} \sigma_e^2) \) and \( \mathbf{I} \) an identity matrix. The \( \mathbf{G} \) was computed as in Zhou [17]:

\[
\mathbf{G} = \frac{1}{n_s} \sum_{i=1}^{p} (x_i - 1/n_s \bar{x}_i) (x_i - 1/n_s \bar{x}_i)^T
\]

where the \( x_i \) is the \( i \)th SNP locus column, \( \bar{x}_i \) is the marker sample mean of the \( i \)th locus, \( n \) and \( n_s \) are numbers of genotyped animals and SNP.

**Genomic prediction**

A linear mixed model was used to compute GP:

\[
y = \mathbf{1} \mu + \mathbf{Z}_u + \mathbf{e},
\]

where \( y \) is a vector of phenotypes, \( \mu \) is an overall mean, \( \mathbf{Z} \) is an incidence matrix for vector \( \mathbf{u} \), and \( \mathbf{u} \) is a vector of random additive genetic effects, with \( \mathbf{u} \sim N(0, \mathbf{H} \sigma_u^2) \) and \( \mathbf{H} \) is the realized relationship matrix, and \( \mathbf{e} \) is a vector of residuals, with \( \mathbf{e} \sim N(0, \mathbf{I} \sigma_e^2) \). The GP was carried out with ssGBLUP and weighted ssGBLUP (WssGBLUP) [18] using the BLUPF90 family of programs [15]. For the mixed model equations in ssGBLUP and WssGBLUP, \( \mathbf{H}^{-1} \) combines both pedigree and genomic relationships [19]:

\[
\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - A_{22}^{-1} \end{bmatrix},
\]

where \( \mathbf{G}^{-1} \) is the inverse of the genomic relationship matrix and \( A_{22}^{-1} \) is the inverse of the pedigree relationship matrix for the genotyped animals. The \( \mathbf{G} \) was created as in VanRaden
\[ G = \frac{MDM'}{2 \sum p_i(1-p_i)}, \]

where \( M \) was defined before, \( p_i \) is the minor allele frequency of the \( i^{th} \) SNP, \( D \) is the diagonal matrix of SNP weights with dimensions equivalent to the number of SNP. In ssGBLUP, all SNP were assumed to have homogeneous weights, meaning that \( D \) was an identity matrix.

To avoid singularity issues, \( G \) was blended with 5% of \( A_{22} \).

For the WssGBLUP, SNP effects were back-solved from GEBV (\( \hat{u} \)) as described in Wang et al. [18]:

\[ \hat{a} = \sigma_a^2 \sigma_u^{-2} DM' G^{-1} \hat{u}, \]

where \( \hat{a} \) is a vector of estimated SNP effects, \( \sigma_a^2 \) is the SNP variance, \( \sigma_u^{-2} \) is the genetic variance, \( D \) is the diagonal matrix of SNP weights (\( I \) in ssGBLUP), and \( M \) is the centered matrix of genotypes. After SNP effects were estimated, the variance for the \( i^{th} \) SNP was calculated using the non-linearA method [20]:

\[ d_i = \frac{|\hat{a}_i|}{CT \sigma(\hat{a})^2}, \]

where CT is a constant that determines the departure from normality when deviating from 1; \( |\hat{a}_i| \) is the absolute estimated SNP effect for \( i^{th} \) SNP; and \( \sigma(\hat{a}) \) is the standard deviation of the vector of estimated SNP effects. This study used CT as 1.125, an empirical value based on polygenic traits in dairy cattle populations. Results from the second iteration of weights were used in this study to maximize the prediction accuracy, as suggested by Zhang et al. [21], Lourenco et al. [22].

**Validation of genomic predictions**

In each scenario, prediction accuracy was calculated as the correlation between TBV and
genomic estimated breeding value (GEBV). Besides, the regression coefficient ($b_1$) of TBV on GEBV was used as an indicator of inflation or deflation of GEBV. When $b_1$ is lower than one, it is indicative of inflation and deflation otherwise. As replicates were not used in this study, standard errors (SE) were computed using the bootstrapping method [23].

Results

Variant identification

The preliminary analysis showed similar results for both GWA and GP when QTN were included (withQTN) or not included in the data. Therefore, results of withQTN are only described. The results of GWA are shown in Fig. 1 ~ 4. As most of the quantitative traits are highly polygenic, only results of Q2000 with H30 and H99 are shown with two $N_e$ scenarios ($N_{20}$, $N_{200}$), respectively. In addition, the GWA results of EIG60, EIG70, and EIG80 were not included in those figures due to their insignificance. All other results with H90, Q200, and EIG60 ~ EIG80 are described in the Additional file 1: Fig. 1. The number of significantly identified QTN, SNP, and the variance explained by those QTN of all scenarios above is in Additional file 1: Table 1. Considering a population with $N_e$ equal to 20 and using EIG50, EIG90, EIG95, EIG98, and EIG99 as the sample size for GWA in Q2000 was not enough to detect significant QTN when the amount of information for genotyped animals was small (Fig. 1). However, when the sample size increased to 30,000 (i.e., ALL), three significant QTN were identified. In contrast, using genotyped animals with high reliability increased the ability to identify simulated QTN correctly (Fig. 2). EIG95 could capture three significant QTN, and as sample size increased to EIG98, EIG99, and ALL, 17, 33, and 142 QTN were identified, respectively.
Different patterns were observed for a population with $N_e$ equal to 200 when having contrasting EIGx as the sample size (Fig. 3 and Fig. 4). Although EIG50, EIG90, and EIG95 were not sufficient to capture the significant QTN in H30, a sample size of EIG98 allowed the identification of seven QTN (Fig. 3). Moreover, increasing the number of genotyped animals to EIG99 and ALL helped detect more QTN and improve the GWA resolution, even though genotyped animals had low reliability. When $N_e$ was 200, but the animals had high reliability, EIG90 is an adequate sample size to detect the QTN with the largest effect size (Fig. 4). For this scenario, EIG98 provided a clear resolution, similar to EIG99 and ALL. It is important to note that the number of largest eigenvalues explaining a certain proportion of the variance in $G$ was different for $N_e20$ and $N_e200$ (Table 2). With all available genotyped animals (i.e., ALL), $N_e200$ had more significant QTN discovered in GWA than $N_e20$. For example, in Q2000 and H30, the three significant QTN in $N_e20$ captured 3.9% of the additive genetic variance, whereas, in $N_e200$, 15 QTN were capturing 13.9%. In both Ne scenarios, using genotyped animals with high reliability helped better detect QTN. For a less polygenic trait (Q200), fewer genotyped animals were required to identify the simulated QTN than in a more polygenic trait (Q200).

**Association between variance explained by identified QTN and sample size**

The proportion of variance explained by identified QTN is shown in Fig. 5. When the sample size increased, the proportion of variance explained by the identified QTN increased regardless of heritability, $N_e$, and number of QTN. Comparing the proportion of variance explained according to the heritability, high heritability scenarios better identified significant QTN. For example, using all genotyped animals in $N_e20$ and Q200 with H30 helped identify QTN explaining 44.9% of the variance (Fig. 5a). However, H90 and H99 identified QTN explaining 72.9% and 74.3% of the variance, respectively, with the same number of genotyped animals.
For a more polygenic trait (Fig. 5b), QTN explained 3.9% and 43.1% of the variance with H30 and H99, respectively. The latter was similar to the variance explained in Q200, H30. Similar patterns were observed between $N_e20$ and $N_e200$; however, QTN identified in $N_e200$ explained a greater proportion of variance than $N_e20$ for Q200 and Q2000 (Fig. 5c-d). For example, the maximum proportion of variance explained by QTN in $N_e20$ (Fig. 5a-b) was 74.3% and 43.1% for Q200 and Q2000 with H99, whereas those values were 96% and 65.1% in $N_e200$ (Fig. 5c-d). One remarkable discovery with a less polygenic trait was that the use of EIG98 and EIG99 showed a similar proportion of variance explained by identified QTN as in ALL (Fig. 5a and 5c). For H99, when EIG98, EIG99, and ALL were used, the variance explained by the identified QTN was 59.9%, 68%, and 74.3%, respectively (Fig. 5a). In the case of Q2000, the scenarios mentioned above identified QTN explaining 11.2%, 18.9%, and 43.1% of the variance (Fig. 5b); therefore, the proportion of variance explained increased by almost fourfold from using EIG98 to ALL, whereas this increase was only 25% with Q200. In a larger population ($N_e200$) and H99 (Fig. 5c), EIG98, EIG99, and ALL schemes detected QTN explaining 95.2%, 95.6%, and 96% of the variance for Q200. Even for the more polygenic scheme (Q2000), EIG98, EIG99, and ALL captured QTN explaining 52.5%, 59.7%, and 65.1% (Fig. 5d). Similar patterns were observed for the other two heritability scenarios in Fig. 5c-d.

To investigate the corresponding sample size from the use of low to high heritability scenarios at the point where the same percentage of variance was explained, we used H30 as a benchmark. Table 3 shows the approximated sample size when the largest discovery set (ALL, $N = 30,000$) was used. All approximated sample sizes based on other discovery sets as a benchmark are in Additional file 1: Table 2. The results of the $N_e20$ Q200 scenario showed that around 2223 and 1662 genotyped animals were required for H90 and H99 to reach the same magnitude of genetic variance explained by the identified QTN in H30 (Table 3). In $N_e20$
Q2000, using 3049 and 2378 genotyped animals with H90 and H99 helped identify QTN explaining 3.9% of the variance, which was accomplished using ALL in H30. Similar patterns were observed in $N_e$200 scenarios. One remarkable difference between $N_e$20 and $N_e$200 was that the sample size to reach the same proportion of variance explained by QTN using ALL in H30 was equivalent to EIG90 ~ EIG98 in $N_e$20 but EIG80 ~ EIG90 in $N_e$200 when considering H90 and H99.

In the current study, in addition to the approximated sample size by local polynomial regression, we derived the equation to estimate the sample size relating the number of samples, $M_e$, $N_e$, percentage of variance explained by the identified QTN, and reliability of EBV. For example, to approximate the sample size of H90 in $N_e$20 Q200 scenario when %var was 44.9, it would be

$$SS_{rel} = \frac{30000 \times 0.3 \times (log_{44.920} \times 0.4 \times 44.9)}{0.9 \times (44.9 + \ln (1840))}$$

Those sample sizes approximated by the proposed equation are described in Table 3. Approximating the sample size in Q2000 was more accurate than in Q200 for $N_e$20 and $N_e$200. Additionally, the formula and the polynomial regression provided sample sizes that were within the same EIGx range, except for one scenario ($N_e$20 Q200 H99). As an example, $N_e$20 Q2000 resulted in 3049 (H90) and 2378 (H99) samples in $SS_{pol}$ and 3007 (H90) and 2734 (H99) in $SS_{rel}$; however, the sample size in both cases laid within EIG95~98. Differences between $SS_{pol}$ and $SS_{rel}$ across more polygenic scenarios were not large.

**Genomic predictions**

Beforehand, we investigate the possible bias of GP by using the same set of genotyped animals for both discovery and training. Using different groups of genotyped animals for discovery and
training resulted in less inflation of GEBV than utilizing the same animals for both processes (results are not shown). Therefore, the GP analyses were done with training animals different from the discovery set.

We noticed very small standard errors for prediction accuracy (<0.005) and $b_1$ (<0.01) by bootstrapping. The results of prediction accuracy and inflation/deflation indicator of GEBV ($b_1$) are shown in Fig. 6 and Fig. 7, respectively. Those accuracies and $b_1$ were calculated as the average of all genotyped scenarios: 50k, TOP10, TOP50, TOP100, TOP200, TOP400, and ‘SIG’ for Q200 and 50k, TOP10, TOP100, TOP500, TOP1000, TOP2000, TOP4000, and ‘SIG’ for Q2000 scenarios. Standard deviations were less than 0.02 for all scenarios. Fig. 6 shows the prediction accuracy depending on the number of QTN (Q2000 or Q200), trait heritability (H30, H90, and H99), and training data scenarios (EIGx and ALL). In general, as the training set size increased, prediction accuracy also increased. Different patterns were observed between $N_e$20 and $N_e$200. In $N_e$200, the prediction accuracy increased consistently as the training set size increased; however, the increase in prediction accuracy was not constant in $N_e$20. For example, in $N_e$20 Q200 H30, while the training data set went up from the EIG50 to EIG95, prediction accuracy increased only by about 0.03. A similar pattern was observed for $N_e$20 Q2000 H30, which showed a gain of about 0.02. In general, $N_e$20 showed greater prediction accuracy than $N_e$200. For example, when the smallest sample size (EIG50) was used, the average accuracy in $N_e$20 was $0.82 \pm 0.03$ and $N_e$200 was $0.74 \pm 0.09$, a difference of about 0.08. This difference became smaller with the largest sample size (ALL), which was 0.04. Therefore, no substantial differences in prediction accuracy between the $N_e$20 and $N_e$200 were found when the largest training set was used. Overall, the difference in prediction accuracy between EIG98 or EIG99 and ALL was smaller in $N_e$200 (0.88, 0.90, and 0.92 on average for EIG98, EIG99, and ALL) than in $N_e$20 (0.90, 0.92, and 0.96 on average for EIG98, EIG99, and ALL).
ALL), indicating a sample for training with the size of EIG98 or EIG99 would suffice.

The number of QTN marginally affected the prediction accuracy for \( N_e20 \) and \( N_e200 \) scenarios. For both \( N_e \) scenarios, Q200 showed greater prediction accuracy, especially for the low heritability with the smaller training sets. For example, in \( N_e20 \) H30, the prediction accuracy from EIG50 to EIG70 was 0.81 with Q200 and 0.79 with Q2000. With larger training set sizes and higher heritability, the difference between Q200 and Q2000 decreased. Prediction accuracies were highly influenced by the heritability of the trait, particularly with larger effective population size. For instance, with \( N_e200 \) Q200, when EIG50 was used, the prediction accuracy was 0.64, 0.81, and 0.81 for H30, H90, and H99, respectively. Even with the most extensive training set (ALL), prediction accuracy was 0.84, 0.96, and 0.98, following the previous order. Similar patterns were observed for \( N_e200 \) Q2000. Thus, using low to high-reliability genotyped animals for training could affect the prediction accuracy more in populations with large \( N_e \). Expanding the training set from EIG95 to EIG98 in \( N_e200 \) H30 increased prediction accuracy by 5.5% and 6.82% for Q200 and Q2000, respectively. However, the maximum increase for \( N_e200 \) H90 and H99 was observed when the training set size expanded from EIG80 to EIG90 (3.8% ~ 4.61%). Unlike \( N_e200 \), great improvements in accuracy were observed when moving from EIG99 to ALL in H30 and H90 for both Q200 and Q2000.

Regression coefficients \((b_1)\) used as indicators of inflation/deflation of GEBV are shown in Fig. 7. When the training set size was small, less inflation was observed in \( N_e200 \) than in \( N_e20 \). In both \( N_e \) scenarios, using a large training set alleviated the inflation, so when ALL was used for training, all scenarios had \( b_1 \) close to 1, especially for \( N_e20 \). Interestingly, more variation between the models was observed in \( N_e20 \) than in \( N_e200 \).

Fig. 8 shows the prediction accuracy with 50k compared to 50k plus SIG, TOP400
(Q200), and TOP4000 (Q2000), together with the percentage of gain by adding possible causative variants. As the increase was not major across analyses combining the 50k and the top SNP, only the scenarios with the largest changes are shown in Fig. 8; all the other scenarios are in Additional file 2. Overall, the percentage of gain was greater in \( N_e200 \) (3.29% ~ 9.01%) than in \( N_e20 \) (0.86% ~ 1.98%). In addition, Q200 showed a higher percentage of gain than Q2000 in both \( N_e \) scenarios. Interestingly, the maximum accuracy gain was usually observed when the largest number of top SNP (TOP400 for Q200 and TOP4000 for Q2000) was added to 50k chip data, representing twice the number of simulated QTN. The only exceptions were Q200 H90 and H99 for \( N_e20 \) and \( N_e200 \), which had the highest accuracy gain with 50k plus SIG. This is probably because identifying significant QTN was easier in a less polygenic trait (Q200) with H90 and H99. In contrast, finding QTN in Q2000 or a low heritability trait was harder.

Assuming we know all the actual QTN, using them for GP returned the best predictive accuracy (Additional file 1: Fig. 2). However, it is hard to identify the real QTN using real data; therefore, the current study compared the prediction accuracy using TOP200 and TOP2000, QTN200 and QTN2000, and added those variants to the 50k panel to investigate the losses from having variants selected based on GWA. This was done for H99 when the selected variants were from using ALL (30,000 animals) as discovery and training sets. When only TOP200 (2000) and QTN200 (2000) were used, the difference in prediction accuracies was quite large (0.05 ~ 0.07). However, prediction accuracies were almost the same when each TOP200 (2000) and QTN200 (2000) were combined with the 50k chip. Based on our findings, if QTN are in the data, using the top variants from GWA in GP can provide reasonable accuracy.
Discussion

In this study, we comprehensively investigated the impact of using different sample sizes in GWA based on the dimensionality of the genomic information, the implications of using genotyped animals having low to high reliability of EBV in GWA, and the inclusion of preselected variants into a typical 50k SNP panel using ssGBLUP and WssGBLUP. These investigations brought insights into how different data structures can affect the performance of GWA and GP under the ssGBLUP framework. We used the concept of limited dimensionality of the genomic information [4]. Our results showed that this concept could be a helpful indicator of the number of genotyped animals required for GWA, depending on \( N_e \), \( M_e \), the number of QTN, and the reliability of EBV. Our results showed that having a sample size with the number of genotyped animals corresponding to that of EIG98 was appropriate for variant discovery, particularly in the population with large \( N_e \). Additionally, using genotyped animals with high EBV reliability could help better identify significant QTN regardless of \( N_e \) and the genetic architecture of the traits. Incorporating selected variants obtained from GWA to the 50k SNP chip could improve prediction accuracy when a training set with proper size was used; however, the gain could be limited in some scenarios.

GWA – preselection of variants

The most prevalent workflow for GP with sequence data is 1) pre-selection of significant variants, 2) incorporation of selected variants to the commercial chip data (i.e., 50k) or fitting separate genomic matrices in the model [9, 24, 25], 3) comparison of the GP performance with a benchmark SNP chip. Several studies have been conducted to improve GP using sequence data with either simulated or real data. However, conclusions about the advantage of using...
sequence data have not been very consistent in the literature, and they seem to be dependent on several factors such as the species, the genetic architecture of the trait, the size of data, and statistical methods [9, 24, 26, 27].

Among those factors, the most critical is the size of data for discovery, training, and test sets. Specifically, the sample size for the variant discovery set is essential as it is the first step and, thus, predominantly affects the results of the entire study. Current results indicated that using a small number of genotyped animals could not identify the significant SNP or QTN.

Lourenco et al. [22] used two different numbers of genotyped animals (N = 2,000 and 25,000) for GWA and reported that the best resolution was observed when more genotyped animals were used. In the same line, de Las Heras-Saldana et al. [28] outlined that using a larger dataset for GWA allowed to better identify quantitative trait loci (QTL) regions for carcass traits in Hanwoo cattle.

As the number of genotyped animals has currently increased in many species, for instance, about 5 million U.S. Holsteins (https://queries.uscdb.com/Genotype/cur_freq.html), and about 1 million American Angus (K. Retallick, American Angus Association, Saint Joseph, MO, personal communication) have been genotyped as of March 2022, it is important to know how many genotyped animals are effectively required to detect the significant variants. Current results showed that using at least the number of genotyped animals equivalent to EIG98 could identify the most informative QTN. Using EIG99 or all available genotyped animals little improved the ability to identify significant QTN in N_e200 for both Q200 and Q2000 scenarios. This result could be helpful for both small and large genotyped populations with large N_e. For breeding populations with fewer resources, the number of animals to genotype may be limited; therefore, accessing what would be the effective sample size could benefit cost-effective genotyping or sequencing. For large populations, our study showed that not all animals are
needed for variant discovery, and a balanced data set should be constructed for discovery, training, and testing to avoid biases and maximize the power to detect the significant variants. When $N_e$ was small and the trait was highly polygenic, the small sample size could not identify any significant QTN until it reached ALL and EIG98 for H30 and H90. With more information on genotyped animals (i.e., H99), a sample size equivalent to EIG95 helped identify a few QTN. Chicken and pigs had smaller $N_e$ (32 ~ 48) than cattle among the livestock species [29]. Therefore, using a sample size corresponding to less than ALL would not be enough to detect significant signals for those species. Gozalo-Marcilla et al. [30] carried out large-scale GWA for backfat thickness in pigs using around 15k to 55k genotyped animals. They found 264 significant SNP across 8 different lines for traits with moderate to high heritability (0.30 ~ 0.58). As backfat thickness has been known for its polygenic architecture (more than 1400 QTL associated backfat thickness is reported in https://www.animalgenome.org/QTLdb), their discovery is supported by our findings in populations with small $N_e$ and moderate heritability.

Pocrnic et al. [4] described the number of largest eigenvalues explaining a certain proportion of $G$ as a function of $N_e$ and genome length in Morgans, such that $EIG90 \approx N_e L$, $EIG95 \approx 2N_e L$, and $EIG98 \approx 4N_e L$. Stam [3] expressed the expected number of independent chromosome segments as $4N_e L$. In this study, $N_e$ was 20 or 200 and $L$ was 23, so $M_e$ was approximated as 1,840 and 18,400, between EIG95 and EIG98 for $N_e 20$, and EIG98 and EIG99 for $N_e 200$ in Table 2. As $N_e$ and $M_e$ are proportional, smaller $N_e$ denotes fewer $M_e$, indicating fewer blocks are existed in the genome with a strong LD between variants because of the close relationship among individuals. $N_e$ also has been reported as a factor affecting the performance of GWA [22, 31]. In the current study, we showed that when the same number of genotyped animals were used (ALL), $N_e 200$ could better identify the significant QTN explaining more genetic variance than $N_e 20$ for all heritability and QTN scenarios. This might be because of
smaller chromosome segments and weaker LD between the QTN and SNP in $N_e$200 than in $N_e$20. Pinpointing QTN is harder in $N_e$20 because many SNP may capture the QTN signal. The noise in GWA resolution is possibly due to the strong relationships between the SNP and QTN, established by a highly structured population over the generations in $N_e$20; therefore, identifying the true causative variant is not trivial in smaller populations. In general, $N_e$ of farm animals such as chickens, pigs, dairy, and beef cattle is less than 200 and could range from 40 to 150 [29]; thus, the current findings would be helpful information for the future GWA in those species. However, identifying all significant variants is not assured due to the polygenic nature of most traits in livestock animals, and most of the causal variants have a small effect. For example, even with the largest number of genotyped animals for GWA in our study (ALL, $N = 30k$), identifying QTN with very small effects was not possible due to limited statistical power. Misztal et al. [12] showed that identifying all simulated QTN was impossible even when all had the same effect, $N_e$ was 600, and the sample size was 6000. For a population with $N_e$ of 60, a sample size three times larger resulted in more true signals in GWA, but not as many as with $N_e$ of 600. The same authors argued that with smaller $N_e$, more data is required to overcome the noise stage and capture the actual signals.

Different heritability scenarios were compared to investigate the performance of GWA when genotyped animals with low to high reliability of EBV were used for the variant discovery stage. Our findings highlighted that regardless of the number of QTN, $N_e$, $M_e$, and sample size, high heritability scenarios could capture more significant QTN explaining a larger portion of the variance. However, Takeda et al. [32] observed no differences in power to detect QTL when heritability 0.2 and 0.5 were simulated but outlined that QTL detection was better with the increasing number of phenotyped progenies ($N = 1,500, 4,500, 9,000$). As the use of more phenotyped progeny data indicated higher reliability of EBV of the parents, which is the
case of genotyped animals in the higher heritability scenarios in our study, those findings agreed with the current results. Besides, van den Berg et al. [33] reported that the number of false positives in QTL detection decreased with increasing heritability and number of records. Thus, using genotyped animals with high EBV reliability could sufficiently detect the QTN although few animals were used.

**Sample size approximation**

Approximated sample sizes were obtained through polynomial regression and a formula relating $N_e, M_e$, and percentage of variance to be explained by significant QTN. The latter was useful to investigate the sample size given the average reliability of EBV in the set of animals available for GWA. Overall, the comparison showed a better approximation with Q2000 scenarios but a very different approximation scale with Q200 scenarios (Additional file 1: Table 2). One possible reason for the inaccurate approximated sample size would be an unbalanced simulation design for $N_e$ and heritability. As we only simulated two scenarios of $N_e$: 20 and 200 for the smallest and largest $N_e$ in livestock species with three heritability scenarios: 0.3, 0.9, and 0.99 representing low, high, and very high reliability of EBV, there is a large gap between $N_e$ 20 and $N_e$ 200, and an irregular pattern of heritability scenarios. However, the proposed equation could be applied for more polygenic traits. For example, given the number of genotyped or sequenced animals available, reliability of those animals, reliability of target animals, and proportion of variance explained by identified QTN, i.e., identified SNP in real data, with $N_e$ and $M_e$, we can approximate the sample size for GWA.

**Genomic prediction**
In general, the accuracy of GP was improved as the size of training data increased, and combining selected variants to a 50k SNP panel could improve accuracy when the GP was performed with the proper size of training sets. It was demonstrated that increasing the number of animals in training sets improved the accuracy of GP [34-36]. Our findings support those results, although only a tiny improvement (< 1.0%) was reported when using training sets with the number of genotyped animals equal to EIG50 to EIG70 in Ne20. Moser et al. [37] observed no improvements in prediction accuracy when the training size was enlarged from 1,239 to 1,880 in Australian dairy cattle. Therefore, adding a substantial number of genotyped animals to the training set is necessary to improve prediction accuracy. As the current study suggested, the training set size can be based on the number of eigenvalues explaining a certain percentage of the variance in G. Those improvement patterns were very similar in both Ne20 and Ne200; however, the prediction accuracies were generally smaller for the Ne200 when the same number of genotyped animals were used. Daetwyler et al. [38] investigated the impact of the genomic structure of the population (Ne and Me) on the accuracy of GBLUP. In their study, with the same number of individuals in the training sets, smaller Ne showed better accuracy than larger Ne regardless of the number of QTL.

Daetwyler et al. [34] proposed the following equation for prediction accuracy: 

\[ r_{\hat{g}g} = \sqrt{N_p h^2 / N_p h^2 + M_e} \], where \( N_p \) is a training set size, \( h^2 \) is the trait heritability, and \( M_e \) is the number of independent chromosome segments. Equation to approximate the \( M_e \) proposed by Stam [3] was \( 4N_e L \), where \( M_e \) is proportional to \( N_e \), thus current results that showed greater accuracy with smaller \( N_e \) theoretically sounded when the \( N_p \) and \( h^2 \) are equivalent. In addition, the small size of \( N_e \) means that fewer \( M_e \) to estimate, thus a smaller prediction error variance would be estimated [4].
We selected variants based on a p-value of 0.05 with a Bonferroni correction for multiple testing and the order of the significance level; however, the Bonferroni correction might generate a stringent threshold, increasing the number of false negatives. Therefore, we tested GP by combining selected variants based on sample size (TOPv) and significant variants. We demonstrated that when a large training set incorporated a relatively large number of variants (i.e., twice the number of simulated QTN), prediction accuracy improved by up to 9%.

Several studies used selected variants from imputed sequence data to improve GP in single-breed populations. Veerkamp et al. [11] reported that when selected variants were used for GP, accuracy decreased, and bias increased. However, VanRaden et al. [26] observed an improvement in accuracy up to 5% when 16k selected variants were added to 60k chip data. In single-breed populations, an improvement in prediction accuracy using selected variants from sequence data could be limited due to long-range LD; thus, precise identification of variants is much harder than multi-breed or across-breed [11].

Fragomeni et al. [7] outlined that including causative QTN in the unweighted G through ssGBLUP increased accuracies by 0.04 when the number of QTN was 100 and 1000, which was similar to our results (0.02 ~ 0.06). Additionally, when those authors added weights derived from SNP effects to G, accuracies increased by 0.10 and 0.03 for 100 and 1000 QTN scenarios, respectively, meaning that weighting SNP was more important for the scenario with a smaller number of QTN (less polygenic). However, in the current study, WssGBLUP resulted in no improvement in accuracy, and more inflation of GEBV was observed compared to ssGBLUP. However, the inflation of GEBV was reduced when more genotyped animals were added to the training set. The major difference between ssGBLUP and WssGBLUP is that ssGBLUP assumes that all SNP explain the same amount of genetic variance, whereas WssGBLUP assigns different variances for each SNP [18]. In general, weighting G may not
increase the accuracy of GP but may improve the GWA resolution (Wang et al., 2012).

In our study, the resolution of variant detection was improved using at least the number of genotyped animals corresponding to the number of eigenvalues explaining 98% of the variation in $G$ for the $N_r200$ scenarios, meaning that when the number of genotyped animals for discovery is close to the approximated $M_e$, precise detection of significant variants is feasible. As the genomic information has limited dimensionality, it could be expressed as the number of non-redundant SNP, genotyped animals, and $M_e$ [39]. Therefore, investigating the dimensionality of the genomic information can help determine the sample size required for discovery and training. Since the performance of GWA and GP depends on several factors such as the genetic architecture of the trait, population structure, heritability, and sample size, more research is needed with real data to validate our results.

**Conclusion**

Accurately identifying causative variants from sequence data depends on the effective population size and, therefore, the dimensionality of genomic information. This dimensionality can help identify the suitable sample size for GWA and should be considered for variant selection. Assigning genotyped animals with high breeding value reliability to the discovery set helps better identify the significant QTN. As sequence data become available, preselecting variants, and adding them to regular chip data could improve prediction accuracy if the dimensionality of the genomic information is considered; however, the improvement is mostly limited.
List of abbreviations

A pedigree relationship matrix
EBV estimated breeding value
EIG eigenvalues of G
EMMAX efficient mixed-model association expedited
G genomic relationship matrix
GEBV genomic estimated breeding value
GP genomic prediction
GWA genome-wide association
H realized relationship matrix
I identity matrix
LD linkage disequilibrium
Me independent chromosome segments
Ne effective population size
QTL quantitative trait loci
QTN quantitative trait nucleotides
ssGBLUP single-step genomic best linear unbiased prediction
SE standard errors
SNP single nucleotide polymorphisms
TBV true breeding value
WssGBLUP weighted ssGBLUP

Declarations
Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
This study was partially funded by Agriculture and Food Research Initiative Competitive Grant no. 2020-67015-31030 from the US Department of Agriculture’s National Institute of Food and Agriculture.

Acknowledgments

Author’s information

Affiliations
Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA

Contributions
DL and SJ conceived and designed the study. SJ analyzed the data and wrote the first draft of the manuscript. DL helped with the computations and structure of the manuscript. NG, ST, IG, and DL provided critical insights and revised the manuscript. All authors read and approved the manuscript.
the final manuscript.  

**Corresponding author**  

Correspondence to Sungbong Jang

**Literature Cited**


12. Misztal I, Pocrnic I, Lourenco D, editors. Factors Influencing Accuracy of Genomic Selection with Sequence Information. JOURNAL OF ANIMAL SCIENCE; 2021: OXFORD UNIV PRESS INC.


Figures

Figure 1. GWA results – EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne20 Q2000 H30

*The x-axis and y-axis are indicating the number of variants and -log10(p-value). The red horizontal dot line represents the Bonferroni correction threshold. Blue vertical lines are pointing out the QTN position which was significantly identified.
Figure 2. GWA results - EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne20 Q2000 H99

*The x-axis and y-axis are indicating the number of variants and -log10(p-value). The red horizontal dot line represents the Bonferroni correction threshold. Blue vertical lines are pointing out the QTN position which was significantly identified.
Figure 3. GWA results - EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne200 Q2000 H30

The x-axis and y-axis are indicating the number of variants and -log10(p-value). The red horizontal dot line represents the Bonferroni correction threshold. Blue vertical lines are pointing out the QTN position which was significantly identified.
Figure 4. GWA results - EIG50, EIG90, EIG95, EIG98, EIG99, All – Ne200 Q2000 H99

*The x-axis and y-axis are indicating the number of variants and -log10(p-value). The red horizontal dot line represents the Bonferroni correction threshold. Blue vertical lines are pointing out the QTN position which was significantly identified.
Figure 5. Total variance explained by significant QTN across the different sample sizes and heritabilities

(a) Ne20 Q200

(b) Ne20 Q2000

(c) Ne200 Q200

(d) Ne200 Q2000

*Each x-axis and y-axis indicate the number of genotyped animals (EIGx scenarios) and percentage of total variance explained by QTN for three heritability scenarios (H0.3, H0.9, and H0.99).
Figure 6. Prediction accuracy

![Graph showing prediction accuracy across different models and conditions.](image-url)
Figure 7. Regression coefficients (b1)
Figure 8. Prediction accuracy of 50k and TOPv scenario which showed maximum gain (%)

(1) Ne20

![Graph showing prediction accuracy for Ne20 scenarios.]

(2) Ne200

![Graph showing prediction accuracy for Ne200 scenarios.]

<table>
<thead>
<tr>
<th>Geno</th>
<th>50k</th>
<th>50k+SIG</th>
<th>50k+TOP400</th>
<th>50k+TOP4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ne200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Description of all GWA scenarios

<table>
<thead>
<tr>
<th>Scenario description</th>
<th>Ne</th>
<th>Number of QTN</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne20 Q200 H30</td>
<td>20</td>
<td>200</td>
<td>0.3</td>
</tr>
<tr>
<td>Ne20 Q200 H90</td>
<td>20</td>
<td>200</td>
<td>0.9</td>
</tr>
<tr>
<td>Ne20 Q200 H99</td>
<td>20</td>
<td>200</td>
<td>0.99</td>
</tr>
<tr>
<td>Ne20 Q2000 H30</td>
<td>20</td>
<td>2000</td>
<td>0.3</td>
</tr>
<tr>
<td>Ne20 Q2000 H90</td>
<td>20</td>
<td>2000</td>
<td>0.9</td>
</tr>
<tr>
<td>Ne20 Q2000 H99</td>
<td>20</td>
<td>2000</td>
<td>0.99</td>
</tr>
<tr>
<td>Ne200 Q200 H30</td>
<td>200</td>
<td>200</td>
<td>0.3</td>
</tr>
<tr>
<td>Ne200 Q200 H90</td>
<td>200</td>
<td>200</td>
<td>0.9</td>
</tr>
<tr>
<td>Ne200 Q200 H99</td>
<td>200</td>
<td>200</td>
<td>0.99</td>
</tr>
<tr>
<td>Ne200 Q2000 H30</td>
<td>200</td>
<td>2000</td>
<td>0.3</td>
</tr>
<tr>
<td>Ne200 Q2000 H90</td>
<td>200</td>
<td>2000</td>
<td>0.9</td>
</tr>
<tr>
<td>Ne200 Q2000 H99</td>
<td>200</td>
<td>2000</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Table 2. Number of genotyped animals for all scenarios in both discovery and training sets

<table>
<thead>
<tr>
<th>EIG50</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>80</th>
<th>510</th>
<th>520</th>
<th>550</th>
<th>530</th>
<th>500</th>
<th>510</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIG60</td>
<td>130</td>
<td>140</td>
<td>140</td>
<td>140</td>
<td>130</td>
<td>900</td>
<td>910</td>
<td>940</td>
<td>920</td>
<td>900</td>
<td>900</td>
</tr>
<tr>
<td>EIG70</td>
<td>220</td>
<td>240</td>
<td>230</td>
<td>220</td>
<td>220</td>
<td>1,500</td>
<td>1,550</td>
<td>1,600</td>
<td>1,540</td>
<td>1,500</td>
<td>1,500</td>
</tr>
<tr>
<td>EIG80</td>
<td>390</td>
<td>410</td>
<td>410</td>
<td>400</td>
<td>400</td>
<td>2,600</td>
<td>2,650</td>
<td>2,700</td>
<td>2,650</td>
<td>2,600</td>
<td>2,600</td>
</tr>
<tr>
<td>EIG90</td>
<td>860</td>
<td>900</td>
<td>890</td>
<td>890</td>
<td>880</td>
<td>5,100</td>
<td>5,250</td>
<td>5,300</td>
<td>5,300</td>
<td>5,100</td>
<td>5,100</td>
</tr>
<tr>
<td>EIG95</td>
<td>1,700</td>
<td>1,770</td>
<td>1,800</td>
<td>1,800</td>
<td>1,750</td>
<td>8,600</td>
<td>8,800</td>
<td>8,800</td>
<td>8,800</td>
<td>8,600</td>
<td>8,600</td>
</tr>
<tr>
<td>EIG98</td>
<td>4,000</td>
<td>4,100</td>
<td>4,100</td>
<td>4,100</td>
<td>4,000</td>
<td>15,000</td>
<td>15,200</td>
<td>15,200</td>
<td>15,200</td>
<td>15,000</td>
<td>15,000</td>
</tr>
<tr>
<td>EIG99</td>
<td>6,900</td>
<td>7,100</td>
<td>7,100</td>
<td>7,000</td>
<td>6,900</td>
<td>21,000</td>
<td>22,000</td>
<td>21,400</td>
<td>22,000</td>
<td>22,000</td>
<td>21,000</td>
</tr>
<tr>
<td>All</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
</tr>
</tbody>
</table>
Table 3. Approximated sample size based on local polynomial regression and proposed equation using ‘ALL’ as benchmark

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Heritability</th>
<th>%Var1</th>
<th>SSpol2</th>
<th>EIGxapp1</th>
<th>SSrel4</th>
<th>EIGxapp2</th>
<th>Diff6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne20 Q200</td>
<td>0.3</td>
<td>44.9</td>
<td>30000</td>
<td>EIG95~98</td>
<td>2698</td>
<td>EIG95~98</td>
<td>475</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>44.9</td>
<td>2223</td>
<td>EIG90~95</td>
<td>2453</td>
<td>EIG95~98</td>
<td>791</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>44.9</td>
<td>1662</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ne20 Q2000</td>
<td>0.3</td>
<td>3.9</td>
<td>30000</td>
<td>EIG95~98</td>
<td>3007</td>
<td>EIG95~98</td>
<td>-42</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>3.9</td>
<td>3049</td>
<td>EIG90~95</td>
<td>2734</td>
<td>EIG95~98</td>
<td>356</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>3.9</td>
<td>2378</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ne200 Q200</td>
<td>0.3</td>
<td>77.4</td>
<td>30000</td>
<td>EIG80~90</td>
<td>4324</td>
<td>EIG80~90</td>
<td>-881</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>77.4</td>
<td>5205</td>
<td>EIG80~90</td>
<td>3931</td>
<td>EIG80~90</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>77.4</td>
<td>3814</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ne200 Q2000</td>
<td>0.3</td>
<td>13.9</td>
<td>30000</td>
<td>EIG80~90</td>
<td>4719</td>
<td>EIG80~90</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>13.9</td>
<td>4524</td>
<td>EIG80~90</td>
<td>4290</td>
<td>EIG80~90</td>
<td>626</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>13.9</td>
<td>3664</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

%Var1: percentage of variance explained by significantly identified QTN
SSpol2: Approximated sample size using local polynomial regression
EIGxapp1: EIGx scenario range including Sampleapp1
SSrel4: Approximated sample size using proposed equation
EIGxapp2: EIGx scenario range including Sampleapp2
Diff6: Difference between Sampleapp2 and Sampleapp1

* Dotted arrows in Fig. 2 matched with each line of H90 and H99 scenario was approximated sample size using local polynomial regression and those were described above as Sampleapp1.

* EIGx scenario ranges including either Sampleapp1 (EIGxapp1) or Sampleapp2 (EIGxapp2) are described with the difference between Sampleapp1 and Sampleapp2 as Sampleapp2 − Sampleapp1 (Diff) in Table.

CC-BY-NC-ND 4.0 International license made available under a which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.