scientific reports

Check for updates

OPEN

Probabilities of two alleles being identity by state at unobserved loci predicted by observed loci in cattle populations

Rintaro Nagai[1], Takeshi Honda[2], Masahiro Satoh[1] & Yoshinobu Uemoto[1]✉

This study aimed to investigate the prediction accuracy of the probability that alleles at unobserved loci are identity-by-state (IBS) using genome-based measures based on observed single nucleotide polymorphisms (SNPs). We performed a simulation analysis assumed to represent a cattle population with simulated and real SNP genotypes. The genome-based measures were based on the inbreeding coefficients in an individual and the additive relationship coefficients between two individuals. Reference values were defined as the probability that the alleles at unobserved SNPs were IBS. Reference values were predicted using both pedigree-based and genome-based measures with tens of thousands of SNPs. Prediction accuracy was calculated as the correlation coefficient between reference and predicted values. Our results showed that the inbreeding and additive relationship coefficients based on SNP-by-SNP with an allele frequency fixed at 0.5 and the coefficients based on the homozygous-segment with homozygous by descent and with run of homozygosity > 4 Mbp long demonstrated consistent high prediction accuracy in both simulated and real cattle populations. Our results also showed that the correlation coefficients of these measures were higher than those of pedigree-based measures. Our results indicate that genome-based measures utilizing observed SNPs can offer a more accurate prediction of IBS relationships at unobserved loci than pedigree-based measures in cattle populations.

Keywords Additive relationship coefficient, Genetic diversity, Identity-by-descent, Inbreeding coefficient, Simulated and real cattle populations

Maintaining genetic diversity in cattle breeds has been a long-standing issue, as evidenced by reports of the loss of genetic diversity in many cattle breeds[1–5]. Genetic diversity allows for selection flexibility to enhance the productivity of both existing and new economically important traits in the current population, as well as for adaptation to ensure food security and resilience in the context of future environmental changes in the future population[6,7]. Furthermore, reduced genetic diversity results in an increase of the homozygosity in the population. This may lead to an elevated frequency of recessive harmful alleles, which could subsequently impair individual performance (i.e., inbreeding depression[8]), with regard to fitness traits such as production, health, and reproductive traits in cattle[9–12].

Genetic diversity in cattle breeds has traditionally been managed by controlling inbreeding rates using inbreeding coefficients and additive relationship coefficients derived from pedigree information[8]. The inbreeding coefficient is defined as the probability that two homologous alleles at neutral loci that are unlinked to any locus under selection are identity-by-descent (IBD) in an individual[8,13]. The additive relationship coefficient, also referred to as the numerator relationship coefficient, is based on the concept of coancestry coefficients, which are defined as the probability that two homologous alleles at neutral loci are IBD between two individuals with a common ancestor. The additive relationship coefficient between two individuals is twice the coancestry coefficient[8]. This approach was adopted to avoid mating between animals with a common ancestor. However, pedigree-based measures may not accurately measure IBD because they may be influenced by biases associated with the definition of the base population[14]. If pedigree information contains errors in its recording, or if the depth of the pedigree is insufficient, it will not be possible to ascertain the precise base population[15].

[1]Graduate School of Agricultural Science, Tohoku University, Sendai 980-8572, Miyagi, Japan. [2]Food Resources Education and Research Center, Kobe University, Kasai 675-2103, Hyogo, Japan. ✉email: yoshinobu.uemoto.e7@tohoku.ac.jp

Genomic information offers new possibilities for monitoring the level of inbreeding. Genome-based measures have the potential to provide more accurate estimation of realized genomic inbreeding, and could be used in breeding strategies to control inbreeding more effectively than pedigree information[14,16,17]. Many types of genome-based inbreeding coefficients and additive relationship coefficients have been proposed to predict the probability of IBD at the genome level without pedigree information by utilizing high- or middle-density single nucleotide polymorphism (SNP) arrays[18]. Genome-based measures reflect the actual proportion of the genome shared by individuals and offer a more accurate representation than pedigree-based measures, which are limited to the expected fractions of genomic IBD. The use of these measures has facilitated the monitoring of genetic diversity in cattle breeds.

Genetic diversity monitoring using genome-based measures is focused on two points: preserving the genetic variability of causal variants at unknown loci that are relevant to future breeding goals and controlling the increased homozygosity of recessive harmful alleles at unknown loci. Consequently, the use of genome-based measures based on SNPs at observed loci involves predicting the probability that alleles at unobserved loci are identity-by-state (IBS) in an individual and between two individuals[19,20]. IBS is that two homologous alleles at a single locus are identical merely by state, irrespective of whether they were inherited from a recent common ancestor[19]. Although several studies have compared genome-based measures with pedigree-based coefficients or with simulated IBD/IBS values[17,21,22], few have explicitly evaluated their predictive accuracy for IBS at unobserved loci, particularly when combining simulated and real cattle populations. Therefore, the objective of this study was to investigate the accuracy of genome-based measures with observed SNPs for predicting the probability that alleles at unobserved loci are IBS in both simulated and real cattle populations. The genome-based measures were based on the inbreeding coefficients in an individual and the additive relationship coefficients between two individuals. We performed a simulation analysis based on genotypes assumed to represent a cattle population, and real genotypes reflecting the extent of linkage disequilibrium (LD) in Japanese Black cattle. In the simulated cattle population, the differences in prediction accuracy between genome- and pedigree-based measures were compared.

Materials and Methods

Simulated cattle population

Populations were simulated based on the forward-in-time process[23] using the QMSim software[24]. A schematic of the simulation process is shown in Fig. 1. First, a historical population was simulated to create mutation drift equilibrium and LD, which were the same as those described by Takeda et al.[25]. Briefly, the size of the historical population began with 1,000 individuals and the historical population was generated as generation zero (G0) to 1,000 (G1000), with a constant size of 1,000. Two different populations with different effective population size (N_e) were generated with a N_e of 20 and 100[26], which mimicked the recent N_e of Japanese Black cattle population[1] and other cattle breed populations such as Holstein cattle[3,27] and Japanese Shorthorn cattle[28], respectively. Thus, the number of individuals gradually decreased from 1000 to 20 (for N_e=20) or 100 (for N_e=100) from generations 1001 to 1030. This was then expanded to 10,000 individuals after three generations, resulting in 5,000 males and 5000 females in the final generation. In each population, 50 males and 500 females were randomly selected from the last historical generation as founders of the recent population. Different patterns of LD decay between N_e=20 and N_e=100 were obtained (Supplementary Fig. S1), similar to those described by Takeda et al.[25]. The simulated population with N_e=20 had higher LD coefficient (r^2) values[29] than those with N_e=100 for all distances between the two loci.

Second, four different types of recent populations were simulated with varying N_e (N_e=20 or N_e=100) and selection criteria (selection at random or selection based on estimated breeding values (EBVs)) over 100 overlapping generations with a constant progeny size per generation (n=1000). Each mating produced two progenies with equal changes in being male or female, and 1,000 progenies were obtained in each generation. Next, 50 male and 500 female progenies were then selected with replacement rate of 0.5 for males and 0.2 for females. The generated recent population was then used to evaluate the accuracy of the genome-based measures. The main parameters of the recent population are summarized in Table 1.

For the genomic structure, the simulated genome consisted of 29 autosomal chromosomes and had a total length of 24.71 Morgans with a length similar to the bovine autosomes based on ARS-UCD 1.2 reference sequence assembly (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002263795.1/). The details of the simulated genomes are presented in Supplementary Table S1. In the historical population, each chromosome contained twice as many SNPs as the Illumina BovineSNP50v2 BeadChip (Illumina Inc., San Diego, CA, USA) and 70 quantitative trait loci (QTL). The SNPs and QTL were biallelic and uniformly distributed along the chromosomes. The initial minor allele frequency (MAF) of SNPs and QTL was set to 0.5, and the mutation rates of SNPs and QTL were set to 2.5×10^{-5}. In the recent population, a total of 60,000 SNPs and 500 QTL with MAF ≥ 0.05 were randomly selected from all chromosomes in the founders. The QTL effects were sampled from a gamma distribution with a shape parameter of 0.4 and the scale parameter was determined internally for the simulated genetic variance. Simulated phenotypes with a set value of phenotypic variance (1.0) were generated with heritability (h^2 = 0.30) explained only by the simulated QTL. All individuals had phenotypes that were used for selection based on EBVs predicted by the best linear unbiased prediction (BLUP). One hundred replicates of historical and recent populations were simulated based on the genomic structure for each condition.

In the simulated population, the target population was defined as generations 10 (G10), 20 (G20), 50 (G50), and 100 (G100) of the recent population, and the accuracy of pedigree- and genome-based measures was evaluated for the target population. The progenies (n = 1000) from each generation of the target population were extracted, and their pedigree and SNP genotyping data were used for the evaluation. In each replicate, the diagonal element of an additive relationship matrix[30] (**A**) minus 1 was defined as the pedigree-based inbreeding coefficient, and the off-diagonal elements of **A** were defined as the pedigree-based additive relationship coefficient. Pedigree-

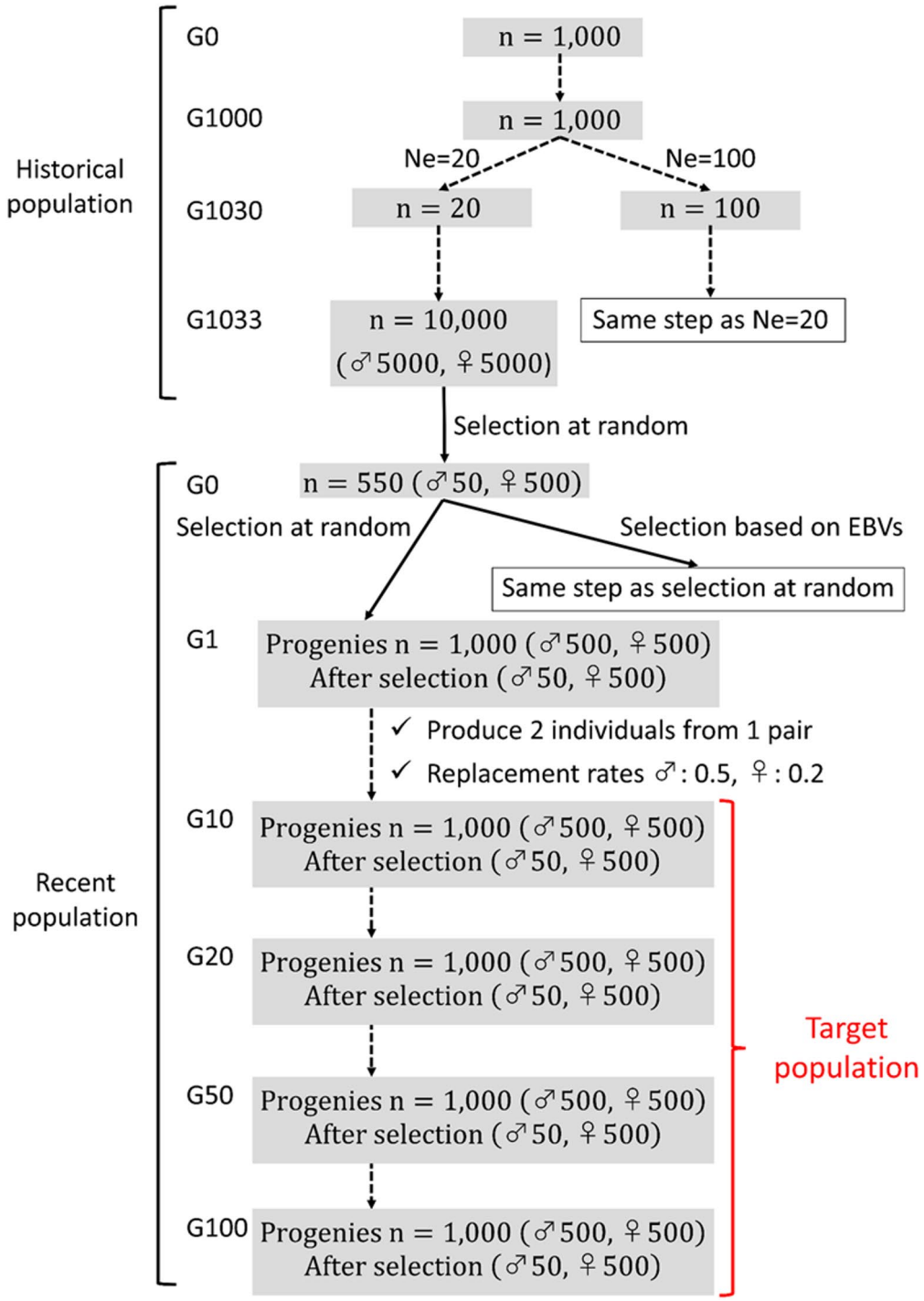


Fig. 1. Schematic illustration of the simulation.

based inbreeding coefficients and additive relationship coefficients were calculated using pedigree information traced back to the full generations (named as F_{APED} and f_{APED}, respectively) and four generations (named as F_{PED} and f_{PED}, respectively) in the recent population. These coefficients were calculated using the tabular method[31] with our own program coded in R software. For genome-based measures, assuming middle-density SNP arrays, 50,000 SNPs were randomly extracted from the initial set of 60,000 founder SNPs and defined as observed SNPs. The remaining 10,000 SNPs were defined as the unobserved SNPs, because our preliminary study demonstrated that the correlation coefficients between the IBS relationships with 10,000 SNPs and 50,000 SNPs exceeded 0.90 for both inbreeding coefficients and additive relationship coefficients (Supplementary Fig. S2). The unobserved SNPs were used to calculate the reference values of the inbreeding coefficients and additive relationship coefficients. The observed SNPs were used to calculate the predicted values of the inbreeding and

Parameter	Value
Effective population size	20, 100
Number of generations	100 overlapping generations
Progeny size per generations	1,000
Number of breeding males	50
Number of breeding females	500
Number of females per male	10
Number of progeny per female	2
Replacement rate of selection	0.5 in males, 0.2 in females
Selection criteria	Random, higher EBV
Heritability	0.3
Number of chromosomes	29
Number of markers	60,000
Number of QTL	500
Number of replicates	100

Table 1. Main parameters for the recent population using QMSim software.

additive relationship coefficients. The means and SD of 100 replicates of the correlation coefficients between the reference and predicted values were then calculated.

To evaluate changes in genetic diversity over generations within the simulated population, the expected heterozygosity (E[Het]) was calculated using the unobserved SNPs as follows:

$$E\left[\mathrm{Het}\right] = \frac{1}{\mathrm{m}} \sum {}^{\mathrm{m}}_{\mathrm{j}=1} 2\mathrm{p}_{\mathrm{j}} \left(1 - \mathrm{p}_{\mathrm{j}}\right),$$

where m is the number of SNPs and p_{j} is the frequency of the second allele of j-th SNP in the target population. Unobserved SNPs are resampled per replicate, and E[Het] was calculated in each replicate with different combinations of N_e and selection criteria.

Real cattle population

We conducted a simulation analysis based on a real cattle SNP dataset to verify our results in simulated populations. A real SNP genotype dataset was used to determine the actual extent of LD in Japanese Black cattle. Details of the experimental population and SNP information have been previously reported by Uemoto et al.[32,33]. Overall, 1368 Japanese Black cattle with SNP genotypes on the Illumina BovineHD BeadsChip (Illumina Inc., San Diego, CA, USA) were used, and the animals had no pedigree information. The dataset excluded animals that were very close relatives by excluding animals with large off-diagonal elements in the genomic relationship matrix (GRM) (cut-off value of 0.4) in the GCTA software[34]. All SNP positions were updated according to the SNPchiMp v.3 database[35] and the ARS-UCD 1.2 reference sequence assembly was downloaded from Ensembl (release 97) (http://ftp.ensembl.org/pub/release-97/variation/vcf/bos_taurus/). The raw genotypes were phased, and the missing genotypes were imputed using Beagle5 software[36]. SNP quality control was assessed using PLINK 1.9 software[37] with the exclusion criteria of the Hardy-Weinberg equilibrium (HWE) test with a P-value < 0.0001. After quality control, 718,223 SNPs on autosomal chromosomes were available, and these genotyped animals were regarded as the target population of the real cattle population.

In this study, 43,421 SNPs located on the Illumina BovineSNP50v2 BeadChip (Illumina Inc., San Diego, CA, USA) were extracted from all SNPs and defined as observed SNPs. The observed SNPs were used to calculate predicted values. Then, a total of 10,000 SNPs with MAF ≥ 0.01 were randomly extracted from the remaining SNPs and were defined as the unobserved SNPs. A total of 100 replicates of the extraction of unobserved SNPs were simulated, and the unobserved SNPs were used to calculate the reference values. The means and SD of the 100 replicates were calculated for the correlation coefficients between the reference and predicted values.

Identity by state at unobserved loci

In this study, the reference values of inbreeding coefficients and additive relationship coefficients were defined as the probabilities that alleles at unobserved loci were IBS in an individual and between two individuals, respectively. Reference values were calculated as average similarity scores based on coancestry coefficients[38–40] using 10,000 unobserved SNPs. The reference values were derived from the elements of the GRM ($\mathbf{G}_{\mathrm{t}}$), which were calculated as follows:

$$\mathbf{G}_{\mathrm{t}} = 2\mathbf{K},$$

where $\mathbf{K}$ is a coancestry coefficient matrix and elements of $\mathbf{K}$ are obtained using the following steps: First, the similarity score between two individuals x and y at the j-th SNP ($\mathrm{k}_{\mathrm{xy,\,j}}$) was calculated as follows:

$$k_{xy,j} = \frac{1}{4}\left(I_{11} + I_{12} + I_{21} + I_{22}\right),$$

where I_{ab} is an indicator variable equal to 1 if allele *a* on the *j*-th SNP in individual *x* and allele *b* on the same SNP in individual *y* are identical; otherwise, it is 0. Second, the average similarity score between two individuals *x* and *y* over *m* SNPs (k_{xy}) was calculated for all individual pairs as follows:

$$k_{xy} = \frac{\sum_{j=1}^{m} k_{xy,j}}{m}.$$

The diagonal element of $\mathbf{G}_t$ minus 1 was defined as the reference value of the inbreeding coefficient (F_{IBS}), and the off-diagonal elements of $\mathbf{G}_t$ were defined as the reference values of the additive relationship coefficient (f_{IBS}). The F_{IBS} and f_{IBS} were calculated for the target populations of the simulated and real cattle populations in each replicate.

Genome-based inbreeding coefficients and additive relationship coefficients

Using the observed SNPs, the genome-based inbreeding coefficients and additive relationship coefficients were calculated for the target populations of the simulated and real cattle populations in each replicate. We defined ten different genome-based inbreeding coefficients derived from SNP-by-SNP (F_{GRMV1}, F_{GRMV2}, F_{GRMY}, and F_{HOM}), haplotype (F_{GHAP}), and homozygous segments (F_{ROH4}, F_{ROH16}, $F_{ROH4all}$, $F_{ROH16all}$, and F_{HBD}). In addition, we defined seven genome-based additive relationship coefficients derived from SNP-by-SNP (f_{GRMV1}, f_{GRMV2}, and f_{GRMY}), haplotype (f_{GHAP}), and homozygous segments (f_{GROH}, f_{SEG4}, and f_{SEG16}). The details of each genome-based inbreeding coefficient and additive relationship coefficient used are described in subsequent sections below. For the simulated and real cattle populations, the observed SNPs were assessed using the exclusion criterion of MAF < 0.01 in the target population with the except for $F_{ROH4all}$ and $F_{ROH16all}$. The number of SNPs with MAF ≥ 0.01 was 34,503 in the real cattle populations. Using the observed SNPs, genome-based measures were calculated as follows:

F_{GRMV1}, F_{GRMV2}, f_{GRMV1}, and f_{GRMV2} were derived from VanRaden's GRM[41] ($\mathbf{G}_V$), which was calculated as follows:

$$\mathbf{G}_V = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum_{j=1}^{m} p_j\left(1 - p_j\right)}, \quad (1)$$

where *m* is the number of SNPs, p_j is the frequency of the second allele of *j*-th SNP, and the elements of **Z** are calculated as $z_{ij} = x_{ij} - 2p_j$, where x_{ij} is the number of second alleles of the *i*-th individual at the *j*-th SNP. When p_j is calculated in the target population, the diagonal elements of $\mathbf{G}_V$ minus 1 are defined as F_{GRMV1} and the off-diagonal elements of $\mathbf{G}_V$ are defined as f_{GRMV1}. When p_j is set to 0.5, which is assumed to be the frequency of the founders, the diagonal element of $\mathbf{G}_V$ minus one is defined as F_{GRMV2} and the off-diagonal elements of $\mathbf{G}_V$ are defined as f_{GRMV2}. These coefficients were based on the assumption that rare homozygous genotypes contribute more to the inbreeding measure than common homozygous genotypes[42].

F_{GRMY} and f_{GRMY} were derived from Yang's GRM[43] ($\mathbf{G}_Y$), which was calculated as follows:

$$\mathbf{G}_Y = \frac{\mathbf{W}\mathbf{W}'}{m},$$

where the elements of **W** are calculated as $w_{ij} = \left(x_{ij} - 2p_j\right) / \sqrt{2p_j\left(1 - p_j\right)}$. When p_j was calculated in the target population, the diagonal elements of $\mathbf{G}_Y$ minus 1 were defined as F_{GRMY} and the off-diagonal elements of $\mathbf{G}_Y$ were defined as f_{GRMY}. These coefficients are based on the correlation of uniting gametes[43,44] and also give more weight to homozygotes for the minor allele than to homozygotes for the major allele[45].

F_{HOM} is based on excess homozygosity[46] and is calculated as follows:

$$F_{HOM} = \frac{O\,[\mathrm{hom}] - E\,[\mathrm{hom}]}{m - E\,[\mathrm{hom}]},$$

where *O*[hom] and *E*[hom] are the observed and expected numbers of homozygous SNPs in an individual under HWE, respectively. F_{HOM} was calculated using the PLINK software[37], and *E*[hom] was calculated from the allele frequency of the target population.

F_{GHAP} and f_{GHAP} were based on haplotype information. A haplotype is defined as a group of closely linked SNPs on the same homologous chromosomes that are frequently inherited. These measures were calculated using the following steps: First, the phased SNP genotypes for all individuals were obtained from the outputs of QMSim software[24] in the simulated cattle population and those of Beagle5 software[36] in the real cattle population and were then used for the analysis. The LD coefficient (r^2) values[29] were calculated to evaluate the LD between SNPs and to construct haploblocks with a threshold of 0.25 (the default value) based on the big-LD approach using the *GPART* package in R software[47]. In this study, a haploblock was defined as a genomic region spanning at least two SNPs. Second, the haplotype alleles were converted to pseudo-SNPs[48], and each pseudo-SNP allele corresponded to one of the unique haplotype alleles present within a haploblock. The number of copies of a specific pseudo-SNP allele was counted and coded as 0, 1, or 2 for each individual, similar to the code

used for SNP. Pseudo-SNPs were assessed using the exclusion criterion of MAF < 0.01. Third, pseudo-SNPs were used to construct a haplotype-based GRM ($\mathbf{G}_{\mathrm{H}}$) based on Eq. (1), with $\mathrm{p_j}$ calculated for the target population. The diagonal element of $\mathbf{G}_{\mathrm{H}}$ minus one is defined as $\mathrm{F_{GHAP}}$ and the off-diagonal element of $\mathbf{G}_{\mathrm{H}}$ is defined as $\mathrm{f_{GHAP}}$.

$\mathrm{F_{ROH4}}$, $\mathrm{F_{ROH4all}}$, $\mathrm{F_{ROH16}}$, and $\mathrm{F_{ROH16all}}$ are based on run of homozygosity (ROH). ROH are defined as contiguous homozygous stretches in an individual genome, resulting from the transmission of identical haplotypes from parents to offspring[49,50]. The identification and characterization of ROH in a population can provide insights into how the population structure has evolved over time. In this study, ROH was estimated using PLINK software[37], which uses a sliding window approach to define ROH as a stretch that includes a minimum specified number of homozygous SNPs within a specified distance. We applied the following parameters and thresholds to define an ROH: (i) a sliding window size of 50 SNPs, (ii) the number of heterozygous SNPs allowed in the ROH was 1, (iii) the maximum allowed distance between consecutive SNPs was 1 Mbp, (iv) the minimum density of SNP in a sliding window was 1 SNP every 100 kbp, (v) the minimum length of an ROH was set to 4 Mbp and 16 Mbp, and (vi) the minimum number of consecutive homozygous SNP included in the ROH (*L*), which was calculated in each target population as follows:

$$\mathrm{L} = \frac{ln\frac{\alpha}{\mathrm{n_s n_i}}}{ln\left(1-\bar{\mathrm{het}}\right)},$$

where α is the false positive probability set to 0.05, $\mathrm{n_s}$ is the number of SNPs per individual, $\mathrm{n_i}$ is the number of SNP genotyped individual, and $\bar{\mathrm{het}}$ is the mean of heterozygosity across all SNPs[51,52]. The inbreeding coefficient based on ROH was defined as the total length of the ROH divided by the overall length of the autosomal genome[50]. The overall length of the autosomal genome was defined as 2,489,385 kbp based on the ARS-UCD 1.2 reference sequence assembly (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002263795.1/). The scale of the inter-marker distance (cM) in the simulated cattle population was assumed to be in Mbp. In this study, four different types of ROH-based inbreeding coefficients ($\mathrm{F_{ROH4}}$, $\mathrm{F_{ROH4all}}$, $\mathrm{F_{ROH16}}$and $\mathrm{F_{ROH16all}}$) were defined based on the size of the ROH and MAF conditions. The size of the ROH is inversely correlated with its age: longer ROH originate from recent common ancestors while shorter ROH come from ancient ancestors because they have been broken down by recombination over many generations[53]. Two thresholds for the size of the ROH from ancient and recent ancestors were defined based on the minimum length of an ROH (4 Mbp or 16 Mbp, respectively). The $\mathrm{F_{ROH4}}$ and $\mathrm{F_{ROH4all}}$ used SNPs with MAF ≥ 0.01 and all SNPs, respectively, to estimate the ROH with a minimum length of 4 Mbp. The $\mathrm{F_{ROH16}}$ and $\mathrm{F_{ROH16all}}$ used SNPs with MAF ≥ 0.01 and all SNPs, respectively, to estimate the ROH with a minimum length of 16 Mbp.

$\mathrm{f_{GROH}}$ was derived from the ROH-based GRM[54] ($\mathbf{G}_{\mathrm{ROH}}$), and indicates whether animals are inbred at the same genomic positions[54]. To quantify the effect of an SNP in an ROH, the SNP genotype in an ROH and not being an ROH was set to 1 and 0, respectively, for each individual. Then, $\mathbf{G}_{\mathrm{ROH}}$ was calculated as follows:

$$\mathbf{G}_{\mathrm{ROH}} = \frac{\mathbf{YY}'}{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{q_j}\left(1-\mathrm{q_j}\right)},$$

where the elements of $\mathbf{Y}$ ($\mathrm{y_{ij}}$) are $1-\mathrm{q_j}$ for being in an ROH or $0-\mathrm{q_j}$ for not being in an ROH of the *i*-th individual at the *j*-th SNP, and $\mathrm{q_j}$ is the frequency of *j*-th SNP being in a ROH. In this study, ROH was estimated using PLINK software[37] with the same parameters and thresholds as $\mathrm{F_{ROH4}}$ to define ROH. After calculating $\mathbf{G}_{\mathrm{ROH}}$, the off-diagonal elements of $\mathbf{G}_{\mathrm{ROH}}$ were defined as $\mathrm{f_{GROH}}$.

$\mathrm{F_{HBD}}$ is based on homozygous-by-descent[55] (HBD). HBD segments are chromosomal segments inherited twice from a common ancestor without recombination. HBD segments result in long stretches of homozygous genotypes (i.e., ROH), and the ROH segment is autozygous. The classification of HBD segments was estimated using the *RZooRoH* package in R software[56], which is a model-based approach based on a Hidden Markov Model (HMM) to identify HBD segments[57]. In this model, the genomic region is described as a mosaic of HBD and non-HBD segments with HMM. The age of inbreeding is estimated for HBD classes based on the transition probability between different HBD and non-HBD segments and is conditional on class specificity. The probability of staying in a particular state is calculated as e^{-R_k}, where R_k is the rate specific to the *k*-th class. This implies that the length of an HBD segment of any class follows an exponential distribution with rate R_k. In this study, 10 HBD classes (the default settings) were used, and the proportion of HBD loci to all loci was calculated and defined as $\mathrm{F_{HBD}}$.

$\mathrm{f_{SEG4}}$ and $\mathrm{f_{SEG16}}$ are additive relationship coefficients based on shared homozygous segments between two individuals ($\mathrm{f_{SEG}}$), which are interpreted as the proportion of ROH sharing between two individuals[58,59]. The $\mathrm{f_{SEG}}$ for individuals *x* and *y* ($\mathrm{f_{SEG}}\ (\mathrm{x},\ \mathrm{y})$) is defined as follows:

$$\mathrm{f_{SEG}}\ (\mathrm{x},\ \mathrm{y}) = \frac{\sum_{\mathrm{k}}\sum_{\mathrm{a_x=1}}^{2}\sum_{\mathrm{b_y=1}}^{2}\left\{\mathrm{L_{seg_k}}\ (\mathrm{a_x},\ \mathrm{b_y})\right\}}{4\mathrm{L}},$$

where *L* is the total length of the genome, $\mathrm{L_{seg_k}}\ (\mathrm{a_x},\ \mathrm{b_y})$ is the *k*-th ROH region between chromosome *a* of individual *x* and chromosome *b* of individual *y*. The $\mathrm{f_{SEG}}$ was predicted using the *optisel* package in R software[59]. We calculated $\mathrm{f_{SEG4}}$ and $\mathrm{f_{SEG16}}$ for all pairs of individuals in the target population using the

following parameters: the minimum number of SNP in $L_{seg_k}(a_x, b_y)$ was 50, and the minimum length of $L_{seg_k}(a_x, b_y)$ was 4 and 16 Mbp, respectively.

Results

Simulated cattle population

To evaluate the effect of N_e and selection criteria on the genomic structure of the simulated cattle population, Fig. 2 shows the trends in expected heterozygosity derived from 10,000 unobserved SNPs for populations with four different combinations of N_e (N_e =20 or N_e =100) and selection criteria (selection at random or selection based on EBVs). The results of this study indicated that expected heterozygosity declined over generations across all simulated populations. The magnitude of this decline was more pronounced in populations subjected to selection based on EBVs than in those subjected to random selection. Additionally, the difference of the expected heterozygosity with N_e =20 and N_e =100 was modest in later generations under selection. The mean ± SD of E[Het] in G100 were 0.175 ± 0.009 and 0.180 ± 0.010 in N_e =20_sel and N_e =100_sel, respectively. It is generally accepted that a smaller N_e should accelerate genetic drift and reduce heterozygosity more rapidly. This result may be indicative of the simulation settings or a limited sensitivity of the heterozygosity metric used. In the simulation settings, the progeny size per generation is moderate (n = 1,000) in the recent population, and the selection intensity is low (from 500 to 50 in male and no selection intensity in female). Selection is also based on overlapping generations with low replacement rate (0.5 for males and 0.2 for females). Our setting values of overlapping generations and selection/replacement conditions masked differences in N_e within the time frame, which may have contributed to the modest differences in the expected heterozygosity observed in our results.

The correlation coefficients between the reference and predicted values of the inbreeding coefficients in the target population of the simulated cattle are presented in Fig. 3 and Supplementary Table S2. There were no notable differences in the correlation coefficients between the N_e =20_ran and N_e =100_ran scenarios and between the N_e =20_sel and N_e =100_sel scenarios. In the population subjected to selection at random, the correlation coefficients ranged from 0.56 to 0.64 in pedigree-based measures and from 0.53 to 0.94 in genome-based measures. In particular, F_{GRMV2}, F_{HOM}, and F_{HBD} exhibited correlation coefficients exceeding 0.90 in all generations. There were no large differences in the correlation coefficients between generations for any of the measures. In the population subjected to selection based on EBVs, the correlation coefficients were ranged from 0.44 to 0.68 in pedigree-based measures and from 0.02 to 0.97 in genome-based measures. Notably,

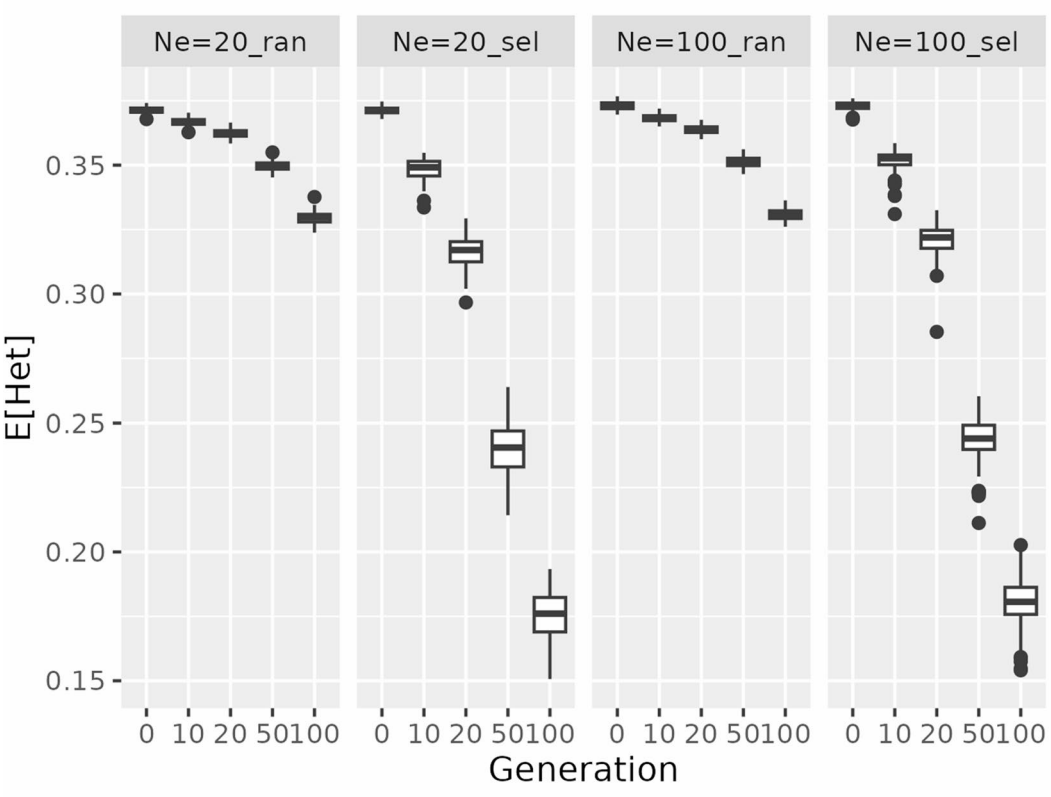


Fig. 2. Trends of expected heterozygosity in the simulated cattle population. Expected heterozygosity (E[Het]) derived from 10,000 unobserved single nucleotide polymorphisms (SNPs) was calculated at generations 10, 20, 50, and 100 in a simulated cattle population with four combinations of effective population size (N_e=20 and N_e=100) and selection criteria (ran: selection at random and sel: selection based on estimated breeding values) in each replicate.

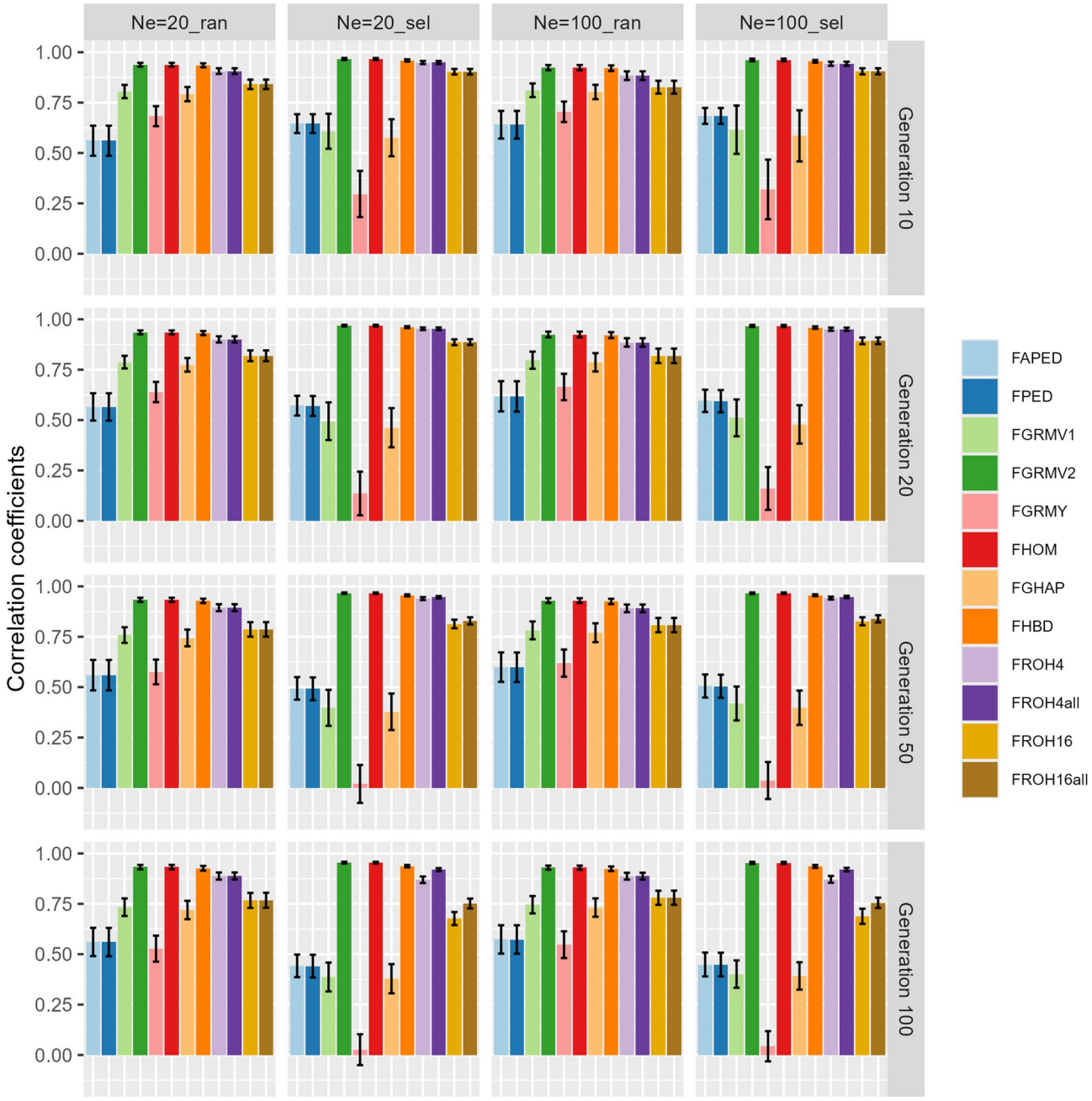


Fig. 3. Correlation coefficients between reference and predicted values of inbreeding coefficients in the simulated cattle population. Correlation coefficients were calculated at generations 10, 20, 50, and 100 in the simulated cattle population with four combinations of effective population size (Ne = 20 and Ne = 100) and selection criteria (ran: selection at random and sel: selection based on estimated breeding values) in each replicate. Mean and SD of 100 replicates were calculated and plotted. The calculations of the reference and predicted values of inbreeding coefficients are explained in the main text. The abbreviations of the pedigree- and genome-based measures are also explained in the main text.

while the correlation coefficients for many of the measures declined over the generations, those for F_{GRMV2}, F_{HOM}, F_{HBD}, and $F_{ROH4all}$ remained constant, exceeding 0.90 for all generations. In contrast, the correlation coefficients for F_{GRMY} were the lowest, ranging from 0.02 to 0.32. The correlation coefficients for F_{ROH4}, $F_{ROH4all}$, F_{ROH16} and $F_{ROH16all}$, the correlation coefficients for F_{ROH4} and $F_{ROH4all}$ were higher than those for F_{ROH16} and $F_{ROH16all}$, respectively, in all scenarios. Moreover, the correlation coefficients for $F_{ROH4all}$ and $F_{ROH16all}$ were higher than those for F_{ROH4} and F_{ROH16}, respectively, in the later generation in the population with selection based on EBVs, while there were no differences in the coefficients between SNPs with MAF ≥ 0.01 and all SNPs in the other populations.

Table 2 shows the mean values of inbreeding coefficients for the target simulated cattle populations. There were no notable differences in the inbreeding coefficients between the N_e =20 and N_e =100 scenarios. The F_{IBS}, F_{APED}, F_{GRMV2}, F_{HBD}, and $F_{ROH4all}$, and $F_{ROH16all}$ increased over generations in all scenarios, particularly in the population selected based on EBVs. In contrast, F_{GRMV1}, F_{GRMY}, F_{HOM}, and F_{GHAP} were constantly around zero over generations in all scenarios, and the F_{PED}, F_{ROH4}, and F_{ROH16} decreased from the later generations in the population with selection based on EBVs.

The correlation coefficients between the reference and predicted values of the additive relationship coefficients in the target population of simulated cattle are shown in Fig. 4 and Supplementary Table S3. There were no notable differences in the correlation coefficients between the N_e =20 and N_e =100 scenarios. In the population subjected to selection at random, the correlation coefficients ranged from 0.77 to 0.83 in pedigree-based measures and from 0.08 to 0.96 in genome-based measures. In particular, all genome-based measures, except for f_{GROH} and f_{SEG16} had correlation coefficients exceeding 0.90 in all generations, and those for f_{GROH} exhibited the lowest values, ranging from 0.08 to 0.17. There were no large differences in the correlation coefficients between generations for any of the measures. In the population subjected to selection based on EBVs, the correlation coefficients ranged from 0.58 to 0.80 in pedigree-based measures and from 0.11 to 0.98 in genome-based measures. Notably, while the correlation coefficients for many of the measures exhibited a decline over generations, the correlation coefficients for f_{GRMV2} and f_{SEG4} exceeded 0.90 for all generations. Conversely, the correlation coefficients for f_{GROH} were the lowest, ranging from 0.11 to 0.27.

Real cattle population

The correlation coefficients between the reference and predicted values of the inbreeding coefficients in a real cattle population are shown in Fig. 5 and Supplementary Table S4. The correlation coefficients ranged from

	Ne=20_ran				Ne=20_sel			
Measures	G10	G20	G50	G100	G10	G20	G50	G100
F_{IBS}	0.63	0.64	0.65	0.67	0.65	0.68	0.76	0.82
F_{APED}	0.01	0.02	0.06	0.11	0.05	0.11	0.24	0.37
F_{PED}	0.01	0.02	0.02	0.02	0.05	0.06	0.05	0.04
F_{GRMV1}	0.00	0.00	0.00	0.00	−0.01	−0.01	0.00	0.00
F_{GRMV2}	0.26	0.27	0.28	0.30	0.29	0.32	0.36	0.37
F_{GRMY}	0.00	0.00	0.00	0.00	−0.01	−0.01	0.00	0.00
F_{HOM}	0.00	0.00	0.00	0.00	−0.01	−0.01	0.00	0.00
F_{GHAP}	0.00	0.00	0.00	0.00	−0.01	−0.01	0.00	0.00
F_{HBD}	0.23	0.23	0.24	0.27	0.25	0.29	0.39	0.46
F_{ROH4}	0.09	0.10	0.12	0.14	0.13	0.20	0.33	0.29
$F_{ROH4all}$	0.09	0.10	0.12	0.14	0.13	0.21	0.37	0.51
F_{ROH16}	0.04	0.04	0.03	0.03	0.07	0.11	0.15	0.08
$F_{ROH16all}$	0.04	0.04	0.03	0.03	0.07	0.11	0.19	0.24
	Ne=100_ran				Ne=100_sel			
Measures	G10	G20	G50	G100	G10	G20	G50	G100
F_{IBS}	0.63	0.63	0.65	0.67	0.63	0.68	0.76	0.82
F_{APED}	0.01	0.02	0.06	0.11	0.01	0.11	0.24	0.37
F_{PED}	0.01	0.02	0.02	0.02	0.01	0.06	0.05	0.04
F_{GRMV1}	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00
F_{GRMV2}	0.26	0.27	0.28	0.30	0.26	0.32	0.36	0.37
F_{GRMY}	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00
F_{HOM}	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00
F_{GHAP}	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00
F_{HBD}	0.21	0.21	0.23	0.25	0.21	0.27	0.37	0.44
F_{ROH4}	0.06	0.07	0.09	0.12	0.06	0.17	0.32	0.29
$F_{ROH4all}$	0.06	0.07	0.09	0.12	0.06	0.17	0.35	0.49
F_{ROH16}	0.02	0.03	0.03	0.03	0.02	0.10	0.15	0.09
$F_{ROH16all}$	0.02	0.03	0.03	0.03	0.02	0.10	0.18	0.22

Table 2. Mean values of inbreeding coefficients in the simulated cattle population. Inbreeding coefficients were calculated at generation 10 (G10), 20 (G20), 50 (G50), and 100 (G100) in the simulated cattle population with four combinations of effective population size (Ne=20 and Ne=100) and selection criteria (ran: selection at random and sel: selection based on estimated breeding values) in each replicate. Mean values of 100 replicates were calculated. The calculations of the reference and predicted values of inbreeding coefficients are explained in the main text. The abbreviations of the pedigree- and genome-based measures are also explained in the main text.

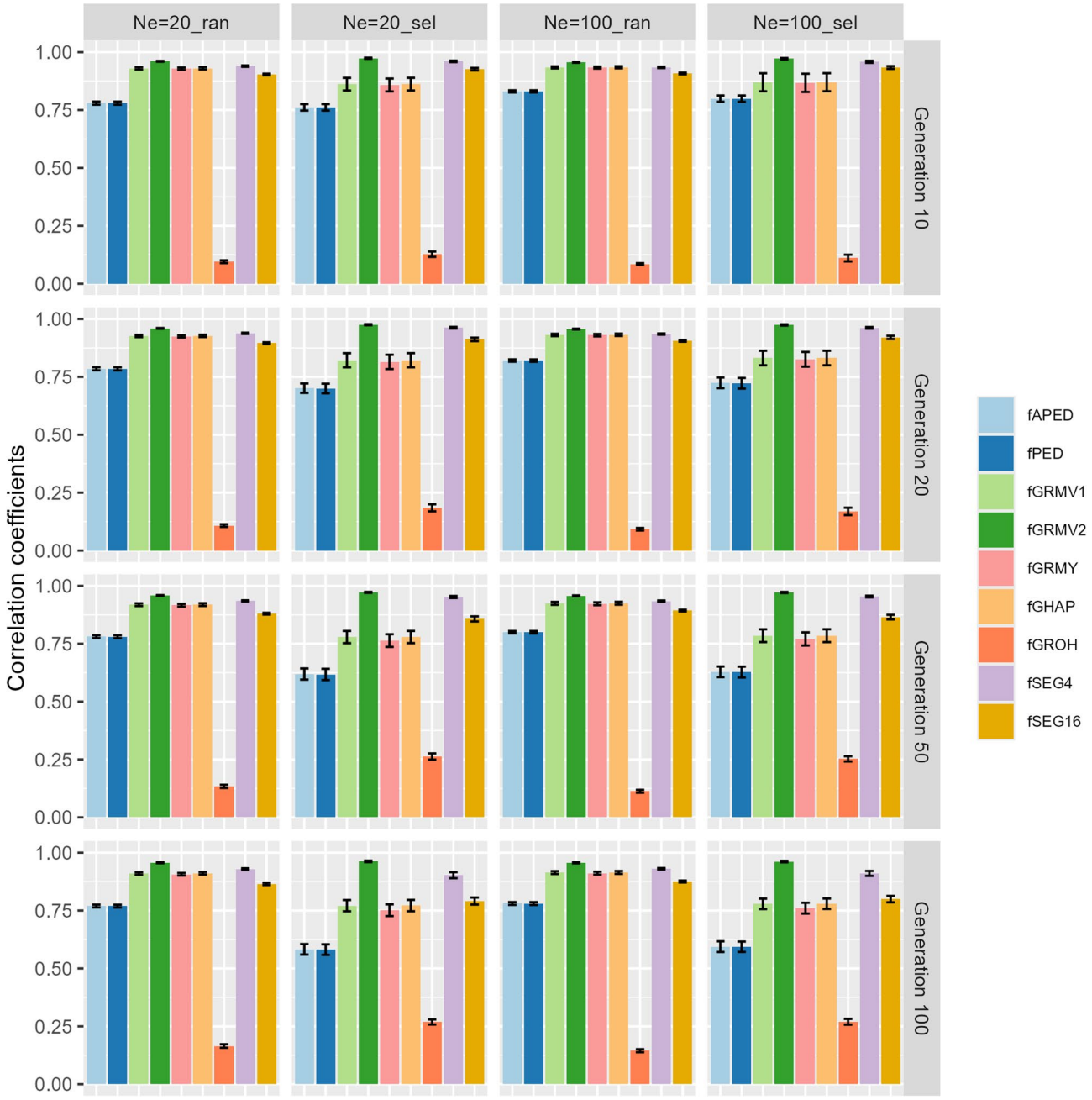


Fig. 4. Correlation coefficients between reference and predicted values of additive relationship coefficients in the simulated cattle population. Correlation coefficients were calculated at generation 10, 20, 50, and 100 in the simulated cattle population with four combinations of effective population size (N_e=20 and N_e=100) and selection criteria (ran: selection at random and sel: selection based on estimated breeding values) in each replicate. Mean and SD of 100 replicates were calculated and plotted. The calculations of the reference and predicted values of additive relationship coefficients are explained in the main text. The abbreviations of the pedigree- and genome-based measures are also explained in the main text.

−0.11 to 0.96, with F_{GRMV2} and F_{HOM} exhibiting the highest coefficients (0.96). F_{HBD}, F_{ROH4}, and $F_{ROH4all}$ exhibited high coefficients (ranging from 0.85 to 0.90), F_{ROH16}, and $F_{ROH16all}$ exhibited moderate coefficients (0.77), and F_{GRMV1}, F_{GRMY}, and F_{GHAP} displayed coefficients below 0.20. As for the difference in correlation coefficients among F_{ROH4}, $F_{ROH4all}$, F_{ROH16}, and $F_{ROH16all}$, the correlation coefficients for F_{ROH4} and $F_{ROH4all}$ (0.85 and 0.86, respectively) were higher than those for F_{ROH16} and $F_{ROH16all}$ (0.77), and there were no differences in the coefficients between SNPs with MAF ≥ 0.01 and all SNPs. Table 3 shows the descriptive statistics of the inbreeding coefficients in real cattle populations. The mean values of the inbreeding coefficients showed a similar trend among the genome-based measures compared with those in G10 of the simulated population.

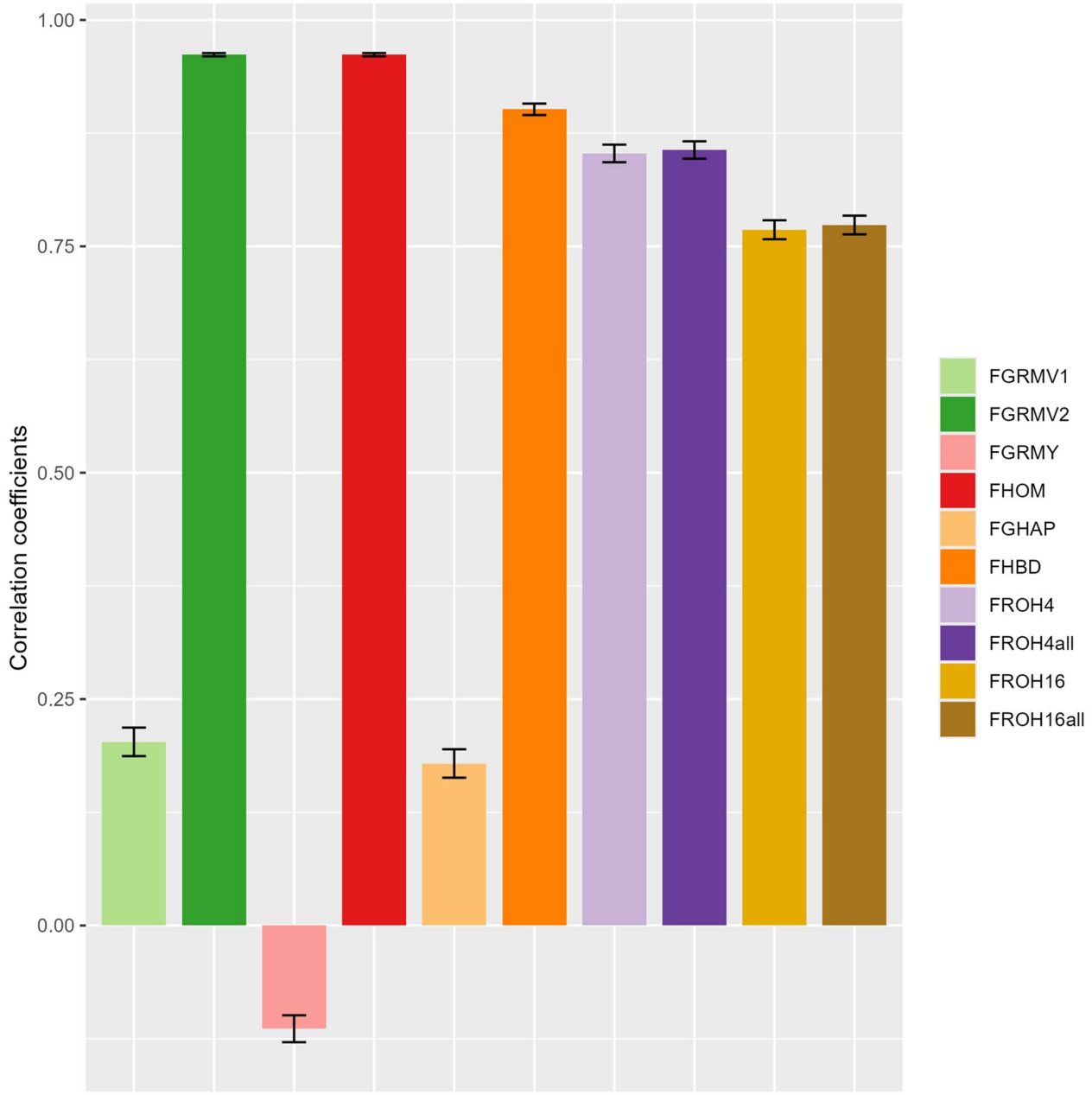


Fig. 5. Correlation coefficients between reference and predicted values of inbreeding coefficients in the real cattle population. Correlation coefficients were calculated in the real cattle population, which was simulated by a total of 100 replicates of random extraction for the unobserved single nucleotide polymorphisms (SNPs). Mean and SD of 100 replicates were calculated and plotted. The calculations of the reference and predicted values of inbreeding coefficients are explained in the main text. The abbreviations of the genome-based measures are also explained in the main text.

The correlation coefficients between the reference and predicted values of the additive relationship coefficients in the real cattle population are shown in Fig. 6 and Supplementary Table S5. The correlation coefficients ranged from 0.05 to 0.97, with f_{GRMV2} and f_{SEG4} exhibiting the two highest correlation coefficients (0.97 and 0.86, respectively) and f_{GROH} displaying a coefficient below 0.10.

Discussion

Management of genetic diversity in cattle breeds is important for maintaining genetic resources and avoiding inbreeding depression, and an appropriate indicator for monitoring the levels of inbreeding is necessary. There are several established genome-based measures for evaluating the levels of genomic inbreeding[41,43,46,47,50,54,55,59]. Among these genome-based measures based on SNPs at observed loci, it is crucial to identify reliable indicators

Measures[a]	Mean	SD	Min	Max
F_{GRMV1}	0.00	0.00	−0.18	0.34
F_{GRMV2}	0.37	0.00	0.29	0.55
F_{GRMY}	0.00	0.01	−0.26	0.46
F_{HOM}	0.00	0.00	−0.13	0.29
F_{GHAP}	0.00	0.00	−0.18	0.35
F_{HBD}	0.12	0.00	0.03	0.34
F_{ROH4}	0.06	0.00	0.00	0.26
$F_{ROH4all}$	0.06	0.00	0.00	0.26
F_{ROH16}	0.03	0.00	0.00	0.23
$F_{ROH16all}$	0.03	0.00	0.00	0.24

Table 3. Summary statistics for the genome-based inbreeding coefficients in the real cattle population. [a]The calculations and the abbreviations of genome-based measures are shown in the main text.

that can accurately predict IBS relationships at unobserved loci. Accordingly, this study aimed to conduct a comprehensive investigation of the most appropriate indicators of inbreeding. This was achieved by comparing the predictive performance of existing indicators with the observed SNPs, under the assumption that the IBS relationship at unobserved loci is regarded as a reference value, and the IBS relationships represent IBS-based similarity scores rather than biologically defined IBD probability. A comprehensive analysis of the efficacy of genome-based measures in both simulated and real cattle populations is presented.

Measures for monitoring genetic diversity

Management of genetic diversity plays a crucial role in maintaining genetic resources and encompasses the preservation of the genetic variability of causal variants at unknown loci for a trait that may prove valuable in the future despite currently being of no interest. The genetic selection of cattle with low enteric methane emission is a typical example[60]. Furthermore, the management of genetic variability contributes to alleviating the risk of inbreeding depression through increased homozygosity. Some recessive alleles with deleterious effects on reproductive performance, including bull conception rate and semen production traits, have been identified in genome-wide association studies (GWAS) of cattle populations[61–63]. Nevertheless, there could be still unknown loci that negatively affect reproductive traits. Consequently, it is necessary to investigate the accuracy of genome-based measures using observed SNPs to predict IBS relationships at unobserved loci as reference values for monitoring the levels of genomic inbreeding[19,20]. However, little is known about the predictive accuracy of IBS relationships in cattle. Some studies have assessed the accuracy of genome-based measures in predicting the IBD relationships in cattle[21]. If IBD relationships are used as reference values for monitoring genomic inbreeding, problems arise such as differences in results depending on which generation is assumed to be the base population. In addition, some of studies have assessed the accuracy of genome-based measures in predicting the pedigree-based inbreeding coefficient as reference values in beef[64,65] and dairy cattle populations[66,67], despite the potential biases of the method. Therefore, we investigated the accuracy of genome-based inbreeding coefficients and additive relationship coefficients with observed SNPs in predicting the IBS relationship with unobserved SNPs, which were assumed to be reference values.

In this study, SNP-by-SNP[41,43,46], haplotype[48] and homozygous segments[49,50,54,57,58] measures were employed, which were derived from different genomic features. To obtain genome-based measures based on SNP-by-SNP and haplotype, it is necessary to have knowledge of the marker allele frequencies in the base population[41,43,48]. However, homozygous-segment-based measures do not require such information. Homozygous-segment-based measures can provide insight into historical and recent selection events by estimating short and long ROHs, respectively, which can be inferred from the extent and frequency of the ROH regions[52]. Haplotypes can also be used to describe recent IBD relationships that capture the LD between SNPs and QTL[68]. This is possible because long-shared haplotypes are likely inherited from more recent common ancestors. In contrast, SNP-by-SNP measures do not distinguish between these two time events. As these different genomic features may result in different accuracies in IBS relationship prediction, this study examined the impact of these different methods on the accuracy of IBS relationship prediction. It should be noted that our study does have certain limitations. We predict IBS relationships at unobserved loci using observed SNPs. However, this is based on indirect estimation via LD and genomic structure, rather than actual observed or biologically defined IBD probabilities.

Inbreeding coefficients

In this study, the correlation coefficients between the reference and predicted values of the inbreeding coefficients were calculated for simulated cattle populations generated using different N_e values and selection criteria. The results showed that many genome-based measures had higher accuracy than pedigree-based measures and that the accuracy of pedigree-based measures decreased over generations in the selected population. This suggests that caution should be exercised when evaluating inbreeding levels by using pedigree-based measures as reference values. The decrease of the accuracy over generations in pedigree-based measures under EBV-based selection may reflect limitations in pedigree depth and resolution. Pedigree-based measures, which represent expected values for each individual, may become similar between individuals over generations. This suggests that they may deviate from the actual values of genomic inbreeding. In contrast, genome-based measures, such as

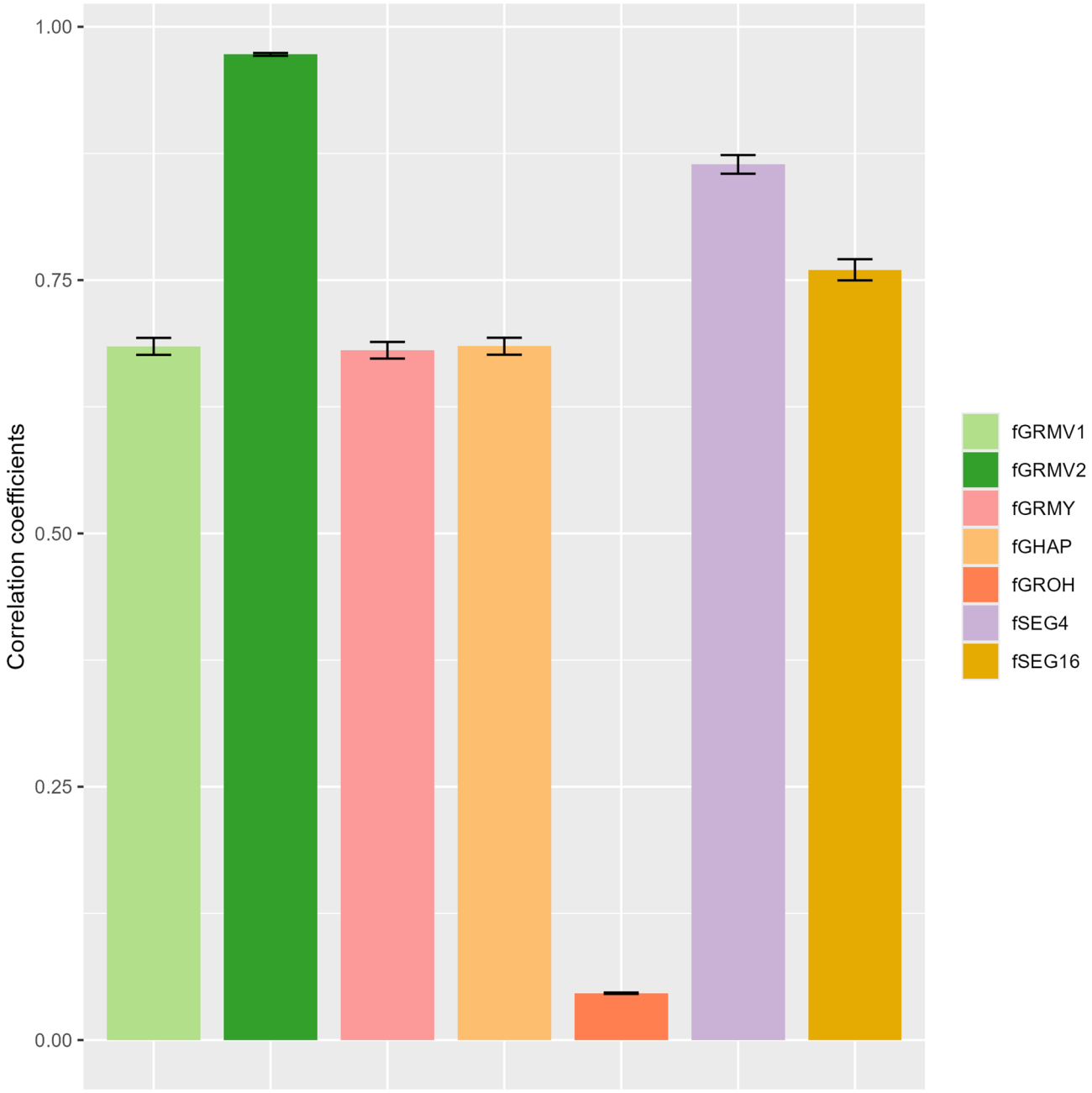


Fig. 6. Correlation coefficients between reference and predicted values of additive relationship coefficients in the real cattle population. Correlation coefficients were calculated in the real cattle population, which was simulated by a total of 100 replicates of random extraction for the unobserved single nucleotide polymorphisms (SNPs). Mean and SD of 100 replicates were calculated and plotted. The calculations of the reference and predicted values of additive relationship coefficients are explained in the main text. The abbreviations of the genome-based measures are also explained in the main text.

F_{GRMV2}, F_{HOM}, and F_{HBD} demonstrated consistently high prediction accuracies across all generations and scenarios. In addition, $F_{ROH4all}$ exhibited a high prediction accuracy, especially in the simulated population with selection based on EBVs.

When comparing the predicted values among different genome-based measures, it should be noted that these measures exhibit different ranges due to their definitions. For example, F_{GRMV2} assumes an allele frequency of 0.5 in the base population and thus takes values ranging from −1 to 1, while F_{GRMV1} and F_{GRMY} take values ranging from −1 to ∞[69]. In addition, homozygous segments measures such as $F_{ROH4all}$ take values ranging from 0 to 1. Consequently, comparing predicted values among genome-based measures is difficult, and the trends in the inbreeding coefficients for each generation were investigated. The results showed that the F_{GRMV2}, F_{HBD}, and $F_{ROH4all}$ increased over generations in all scenarios as F_{IBS} increased, whereas

the F_{HOM} was constantly around zero over generations in all scenarios. The results of F_{HOM} was due to the assumption that the expected numbers of homozygous SNPs were used to calculate F_{HOM}. When the expected numbers of homozygous SNPs are calculated in the current population, the mean is zero. This was the reason why the F_{HOM} was consistently around zero over generations in all scenarios, and this result indicated that the F_{HOM} has a potential to reflect the masking of true inbreeding trends.

The correlation coefficients between the reference and predicted values of the inbreeding coefficients were evaluated in a real cattle population, which indicated that the genome-based measures with high correlation coefficients (>0.90) were F_{GRMV2}, F_{HOM}, and F_{HBD}. The real cattle population used in this study was considered similar to a non-selected population, as animals that were very close relatives were excluded, and the results were similar to those observed in the simulated cattle population with random selection. Consequently, the results indicate that F_{GRMV2} and F_{HBD} could be useful indicators with the highest prediction accuracy in both selected and non-selected populations. Additionally, $F_{ROH4all}$ may be a useful indicator of high prediction accuracy in selected populations. These indicators also allow for the evaluation of changes in the levels of inbreeding over generations due to selection. Because most cattle populations are under selection for economically important traits, these three indicators are considered more effective than pedigree-based measures in evaluating the levels of inbreeding.

For SNP-by-SNP and haplotype-based measures, F_{GRMV2} was identified as the most accurate measure and could evaluate changes in the levels of inbreeding over generations due to selection in the present study. This is contingent on the assumption that the allele frequency utilized in the calculation was fixed at 0.5, which is the frequency observed in the base population. In contrast, this approach differs from other approaches (F_{GRMV1}, F_{GRMY}, F_{HOM}, and F_{GHAP}) that assume that the allele frequencies utilized in the calculation are obtained from the target population. Prediction of the inbreeding coefficient using allele frequency has been reported to be influenced by the allele frequency of the target population[69]. In this study, the accuracy of the measures with allele frequencies of the target population decreased, especially in the later generations of the selected population. Furthermore, these measures did not account for changes in the levels of inbreeding over generations due to selection. Conversely, F_{GRMV2} exhibits high accuracy in both simulated and real cattle populations, which is thought to be due to its fixed allele frequency. F_{GRMV2} is highly correlated with F_{PED} in cattle populations[70] and has been employed to monitor genetic diversity in dairy cattle at the Animal Genomics and Improvement Laboratory of the USDA[71]. Obtaining allele frequencies for a base population is challenging because the base population is often unknown. This study indicates that even the assumption of fixing the allele frequency in a base population of 0.5, provides high prediction accuracy. Additionally, caution should be exercised when utilizing SNP-by-SNP and haplotype-based measures with allele frequencies of the target population as indicators, particularly in selected populations.

For homozygous-segment-based measures, F_{HBD}, which assumes that two ROH segments are IBD, was identified as the most accurate indicator of homozygous-segment-based measures in both simulated and real cattle populations. In addition, $F_{ROH4all}$, which assumes that the two ROH segments are IBS, exhibits high prediction accuracy, especially in the selected population. In the present study, both measures were also effective in evaluating changes in the levels of inbreeding over generations due to selection. The information of ROH has been employed in numerous inbreeding coefficient predictions in cattle populations, with the utility of this approach having been reported[64–67]. However, appropriate parameter settings are required to estimate ROH. In this study, F_{HBD} was calculated under the default setting, and the correlation coefficients for F_{HBD} exceeded 0.90 in both simulated and real cattle populations. Although F_{HBD} is a computationally time-consuming method, our results indicated that F_{HBD} is a useful indicator for predicting IBS relationships at unobserved loci.

In this study, we calculated the inbreeding coefficients (F_{ROH4}, $F_{ROH4all}$, F_{ROH16}, and F_{ROH16}), which required less computation time than F_{HBD}, using the estimated ROH under two different conditions: the length of the ROHs and the MAF condition for SNPs. Regarding the length of the ROH in the ROH estimation, the minimum length of the ROH was set to 4 and 16 Mbp in this study. This was done based on the assumption that ROH > 4 Mbp corresponds to homozygous segments of both past and recent origins, and that ROH > 16 Mbp corresponds to homozygous segments of only recent origins. Genome-based inbreeding coefficients calculated from long ROH are comparable to those derived from pedigree data in cattle population[66]. In this study, F_{ROH4} and $F_{ROH4all}$ had more accurate predictions than F_{ROH16} and F_{ROH16}, respectively, suggesting that an ROH > 4 Mbp is more desirable for predicting IBS relationships at unobserved loci. However, caution should be exercised in evaluating inbreeding depression. Inbreeding depression may be negatively associated with recent inbreeding, and not only because of the total level of inbreeding[67]. Therefore, it may be better to use a long ROH when evaluating the negative effects of inbreeding on phenotypic values. Furthermore, this study did not assess the prediction accuracy under the condition that the ROHs were finely divided into shorter segments, such as 1–2 Mbp and 2–4 Mbp of ROH length. A finer classification of ROHs may be useful for predicting IBS relationships at unobserved loci. Further investigations are warranted based on these considerations.

As for the MAF condition for SNPs in the estimation of ROH, Meyermans et al.[72] reported that extracting SNPs by the MAF condition would have a negative impact on the estimation of ROH due to the possibility of losing informative SNPs. Therefore, all SNPs and SNPs with MAF ≥ 0.01 were used to estimate ROH in this study. The prediction accuracy of homozygous-segment-based inbreeding coefficients using all SNPs ($F_{ROH4all}$ and $F_{ROH16all}$) was higher than those using SNPs with MAF ≥ 0.01 (F_{ROH4} and F_{ROH16}, respectively) in simulated population. Moreover, the inbreeding coefficients using SNPs with MAF ≥ 0.01 decreased from the later generation in the simulated population with selection, while those using all SNPs increased over generations. Information on homozygous alleles, which are contained within the segment and fixed within the target population, can be considered important in the estimation of ROH. Therefore, this may be an important factor in the prediction of IBS relationships at unobserved loci. These findings indicate that the estimation of ROH should consider both the length of ROH and the MAF condition for the SNP.

Additive relationship coefficients

In this study, many genome-based additive relationship coefficients had higher prediction accuracy than pedigree-based coefficients in the simulated population, and f_{GRMV2} and f_{SEG4} demonstrated consistently high prediction accuracy in both simulated and real cattle populations. f_{GRMV2} and f_{SEG4} can be classified using SNP-by-SNP and homozygous-segment-based methods, respectively. f_{GRMV2} are the off-diagonal elements of $\mathbf{G}_V$ under $p_j = 0.5$ baseline and do not discriminate alleles that are IBD or IBS at the observed loci. f_{SEG4} are the predicted values of the IBD relationships of the chromosomal segments based on SNPs at the observed loci. Both genome-based measures showed high correlations with the IBS relationships at unobserved loci, but observed differences across scenarios advise cautious interpretation; we do not claim equivalence. Their prediction accuracy exhibited a trend similar to that of the inbreeding coefficients in both the simulated and real cattle populations. This study demonstrates that IBS relationships at unobserved loci between two individuals can be accurately predicted using genome-based additive relationship coefficients such as f_{GRMV2} and f_{SEG4}.

For the practical application in breeding program, additive relationship coefficients serve as indicators of genetic relatedness between individuals and play an important role, particularly in genetic conservation programs aimed at minimizing the loss of genetic variability[73]. Consequently, studies on the application of additive relationship coefficients to conservation programs must be necessary. Optimal contribution selection (OCS) uses additive relationship coefficients to minimize the loss of genetic variability while improving genetic ability[74]. Some studies on OCS using genetic relationship matrices with additive relationship coefficients as elements have been conducted[75,76]. Morales-González et al.[75] employed VanRaden's GRM with the allele frequencies of the base population and conducted a selection simulation, which demonstrated that it was feasible to maintain allele frequencies close to those of the base population. The results of OCS simulation using f_{SEG} in a simulated pig population also demonstrated that genomic mating using f_{SEG} is more efficient than traditional mating schemes[76]. The f_{SEG} is an indicator that evaluates the proportion of ROH sharing between two individuals[59], and genomic mating using f_{SEG} can reduce the length of ROH segments within the offspring in the next generation. As this study focused on the prediction of IBS relationships at unobserved loci, suitable indicators for OCS were not evaluated. Future investigation should be conducted to determine the efficacy of OCS using f_{GRMV2} and f_{SEG4} in cattle.

Conclusion

We investigated whether genome-based inbreeding coefficients and additive relationship coefficients, using observed SNPs, could be used as indicators for predicting the probability of alleles at unobserved loci being IBS in an individual and between two individuals in both simulated and real cattle populations. The results demonstrated that the inbreeding coefficients and additive relationship coefficients based on SNP-by-SNP with an allele frequency fixed at 0.5 (F_{GRMV2} and f_{GRMV2}, respectively) exhibited high prediction accuracy in both simulated and real cattle populations. Additionally, homozygous-segment-based inbreeding coefficients and additive relationship coefficients, which include F_{HBD} and measures with short ROH ($F_{ROH4all}$ and f_{SEG4}, respectively), were highly accurate measures in both simulated and real cattle populations. Furthermore, F_{GRMV2}, F_{HBD}, and $F_{ROH4all}$ were effective in evaluating changes in inbreeding levels over generations due to selection. Our results indicate that these indicators are more reliable than pedigree-based measures specifically for predicting IBS relationships at unobserved loci. Our results show that it is possible to predict the IBS relationships at unobserved loci by indirect estimation via LD and genomic structure using observed SNPs in cattle populations.

Data availability

The datasets analyzed during the current study are not publicly available due to the fact that the SNP genotype data of Japanese Black cattle is owned by the Japan Livestock Technology Association but are available from the corresponding author on reasonable request.

Received: 16 February 2025; Accepted: 22 January 2026
Published online: 05 February 2026

References

1. Nomura, T., Honda, T. & Mukai, F. Inbreeding and effective population size of Japanese black cattle. *J. Anim. Sci.* **79** (2), 366–370. https://doi.org/10.2527/2001.792366x (2001).
2. Honda, T., Nomura, T., Yamaguchi, Y. & Mukai, F. Monitoring of genetic diversity in the Japanese black cattle population by the use of pedigree information. *J. Anim. Breed. Genet.* **121** (4), 242–252. https://doi.org/10.1111/j.1439-0388.2004.00452.x (2004).
3. Stachowicz, K., Sargolzaei, M., Miglior, F. & Schenkel, F. S. Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle. *J. Dairy. Sci.* **94** (10), 5160–5175. (2011).
4. Pryce, J. E. & Daetwyler, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim. Prod. Sci.* **52** (3), 107–114. (2012).
5. Makanjuola, B. O. et al. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. *J. Dairy. sci.* **103** (6), 5183–5199. (2020).
6. Boettcher, P. J. et al. Objectives, criteria and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. *Anim. Genet.* **41 Supplement 1**, 64–77. https://doi.org/10.1111/j.1365-2052.2010.02050.x (2010).
7. Boettcher, P. J. et al. Genetic resources and genomics for adaptation of livestock to climate change. *Front. Genet.* **5**, 461. https://doi.org/10.3389/fgene.2014.00461 (2014).
8. Falconer, D. S. & Mackay, T. F. C. Introduction to quantitative genetics (ed. 4th)Longman, (1996).
9. Howard, J. T., Pryce, J. E., Baes, C. & Maltecca, C. Invited review: inbreeding in the genomics era: inbreeding, inbreeding depression, and management of genomic variability. *J. Dairy. Sci.* **100** (8), 6009–6024. (2017).
10. Baes, C. F. et al. Symposium review: The genomic architecture of inbreeding: How homozygosity affects health and performance. *J. Dairy Sci.* **102** (3), 2807–2817. (2019).

11. Doekes, H. P., Bijma, P. & Windig, J. J. How depressing is inbreeding? A meta-analysis of 30 years of research on the effects of inbreeding in livestock. *Genes* **12** (6), 926. (2021).
12. Gutiérrez-Reinoso, M. A., Aponte, P. M. & García-Herreros, M. A review of inbreeding depression in dairy cattle: current status, emerging control strategies, and future prospects. *J. Dairy. Res.* **89** (1), 3–12. https://doi.org/10.1017/S0022029922000188 (2022).
13. Wright, S. Coefficients of inbreeding and relationship. *Am. Nat.* **56** (645), 330–338. https://doi.org/10.1086/279872 (1922).
14. Pryce, J. E., Hayes, B. J. & Goddard, M. E. Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *J. Dairy. Sci.* **95** (1), 377–388. (2012).
15. Oldenbroek, J. K. Genomic Management of Animal Genetic Diversity. (Wageningen Academy Pubishers, 2017).
16. Keller, M. C., Visscher, P. M. & Goddard, M. E. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* **189** (1), 237–249. https://doi.org/10.1534/genetics.111.130922 (2011).
17. Wang, J. Pedigrees or markers: which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.* **107**, 4–13. https://doi.org/10.1016/j.tpb.2015.08.006 (2016).
18. Meuwissen, T. H. E., Sonesson, A. K., Gebregiwergis, G. & Woolliams, J. A. Management of genetic diversity in the era of genomics. *Front. Genet.* **11**, 880. https://doi.org/10.3389/fgene.2020.00880 (2020).
19. Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* **11** (11), 800–805. https://doi.org/10.1038/nrg2865 (2010).
20. Toro, M. A., Villanueva, B. & Fernández, J. Genomics applied to management strategies in conservation programmes. *Livest. Sci.* **166**, 48–53. https://doi.org/10.1016/j.livsci.2014.04.020 (2014).
21. Forutan, M. et al. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genom.* **19** (1), 98. https://doi.org/10.1186/s12864-018-4453-z (2018).
22. Caballero, A., Fernández, A., Villanueva, B. & Toro, M. A. A comparison of marker-based estimators of inbreeding and inbreeding depression. *Genet. Sel. Evol.* **54** (1), 82. https://doi.org/10.1186/s12711-022-00772-0 (2022).
23. Carvajal-Rodríguez, A. & GENOMEPOP A program to simulate genomes in populations. *BMC Bioinform.* **9** https://doi.org/10.1186/1471-2105-9-223 (2008).
24. Sargolzaei, M. & Schenkel, F. S. QMSim: a large-scale genome simulator for livestock. *Bioinform* **25** (5), 680–681. https://doi.org/10.1093/bioinformatics/btp045 (2009).
25. Takeda, M., Uemoto, Y. & Satoh, M. Effect of genotyped bulls with different numbers of phenotyped progenies on quantitative trait loci detection and genomic evaluation in a simulated cattle population. *Anim. Sci. J.* **91** (1), e13432. https://doi.org/10.1111/asj.13432 (2020).
26. Lourenco, D. A. L. et al. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J. Anim. Breed. Genet.* **134** (6), 463–471. https://doi.org/10.1111/jbg.12288 (2017).
27. Sørensen, A. C., Sørensen, M. K. & Berg, P. Inbreeding in Danish dairy cattle breeds. *J. Dairy. Sci.* **88** (5), 1865–1872. https://doi.org/10.3168/jds.S0022-0302(05)72861-7 (2005).
28. Uemoto, Y., Suzuki, K., Yasuda, J., Roh, S. & Satoh, M. Evaluation of inbreeding and genetic diversity in Japanese shorthorn cattle by pedigree analysis. *Anim. Sci. J.* **92** (1), e13643. https://doi.org/10.1111/asj.13643 (2021).
29. Hill, W. G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60** (3), 615–628. https://doi.org/10.1093/genetics/60.3.615 (1968).
30. Henderson, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32** (1), 69–83. https://doi.org/10.2307/2529339 (1976).
31. Mrode, R. CABI International. *Linear Models for the Prediction of Animal Breeding Values*. 3rd (2014). https://doi.org/10.1079/9781780643915.0000
32. Uemoto, Y., Sasaki, S., Sugimoto, Y. & Watanabe, T. Accuracy of high-density genotype imputation in Japanese black cattle. *Anim. Genet.* **46** (4), 388–394. https://doi.org/10.1111/age.12314 (2015).
33. Uemoto, Y., Sasaki, S., Kojima, T., Sugimoto, Y. & Watanabe, T. Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese black cattle. *BMC Genet.* **16**, 134. https://doi.org/10.1186/s12863-015-0287-8 (2015).
34. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88** (1), 76–82 (2011).
35. Nicolazzi, E. L. et al. SNPchiMp v. 3: integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genom.* **16** (1), 283. https://doi.org/10.1186/s12864-015-1497-1 (2015).
36. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103** (3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015 (2018).
37. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** (3), 559–575. https://doi.org/10.1086/519795 (2007).
38. Nejati-Javaremi, A., Smith, C. & Gibson, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* **75** (7), 1738–1745. https://doi.org/10.2527/1997.7571738x (1997).
39. Eding, H. & Meuwissen, T. H. E. Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *J. Anim. Breed. Genet.* **118** (3), 141–159. https://doi.org/10.1046/j.1439-0388.2001.00290.x (2001).
40. Caballero, A. & Toro, M. A. Analysis of genetic diversity for the management of conserved subdivided populations. *Conserv. Genet.* **3** (3), 289–299. https://doi.org/10.1023/A:1019956205473 (2002).
41. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy. Sci.* **91** (11), 4414–4423. (2008).
42. Solé, M. et al. Age-based partitioning of individual genomic inbreeding levels in Belgian blue cattle. *Genet. Sel. Evol.* **49** (1), 92. (2017).
43. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** (7), 565–569. https://doi.org/10.1038/ng.608 (2010).
44. Caballero, A., Villanueva, B. & Druet, T. On the Estimation of inbreeding depression using different measures of inbreeding from molecular markers. *Evol. Appl.* **14** (2), 416–428. https://doi.org/10.1111/eva.13126 (2021).
45. Nietlisbach, P., Muff, S., Reid, J. M., Whitlock, M. C. & Keller, L. F. Nonequivalent lethal equivalents: models and inbreeding metrics for unbiased Estimation of inbreeding load. *Evol. Appl.* **12** (2), 266–279. https://doi.org/10.1111/eva.12713 (2019).
46. Li, C. C. & Horvitz, D. G. Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* **5** (2), 107–117 (1953).
47. Kim, S. A. et al. Gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinform* **35** (21), 4419–4421. https://doi.org/10.1093/bioinformatics/btz308 (2019).
48. Teissier, M. et al. Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. *J. Dairy. Sci.* **103** (12), 11559–11573. (2020).
49. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'etude du polymorphisme Humain. *Am. J. Hum. Genet.* **65** (6), 1493–1500. (1999).
50. McQuillan, R. et al. Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83** (3), 359–372. https://doi.org/10.1016/j.ajhg.2008.08.007 (2008).
51. Lencz, T. et al. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *PNAS* **104** (50), 19942–19947. https://doi.org/10.1073/pnas.0710021104 (2007).
52. Purfield, D. C., Berry, D. P., McParland, S. & Bradley, D. G. Runs of homozygosity and population history in cattle. *BMC Genet.* 13, 70; (2012). https://doi.org/10.1186/1471-2156-13-70

53. Ceballos, F. et al. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234. https://doi.org/10.1038/nrg.2017.109 (2018).
54. Doekes, H. P. et al. Inbreeding depression across the genome of Dutch Holstein Friesian dairy cattle. *Genet. Sel. Evol.* **52** (1), 64. https://doi.org/10.1186/s12711-020-00583-1 (2020).
55. Leutenegger, A. L. et al. Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.* **73** (3), 516–523. https://doi.org/10.1086/378207 (2003).
56. Bertrand, A. R., Kadri, N. K., Flori, L., Gautier, M. & Druet, T. RZooRoH: an R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods Ecol. Evol.* **10** (6), 860–866. https://doi.org/10.1111/2041-210X.13167 (2019).
57. Druet, T. & Gautier, M. A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Mol. Ecol.* **26** (20), 5820–5841. https://doi.org/10.1111/mec.14324 (2017).
58. de Cara, M. Á. R., Villanueva, B., Toro, M. Á. & Fernández, J. Using genomic tools to maintain diversity and fitness in conservation programmes. *Mol. Ecol.* **22** (24), 6091–6099. https://doi.org/10.1111/mec.12560 (2013).
59. Wellmann, R. Optimum contribution selection for animal breeding and conservation: the R package optisel. *BMC Bioinform.* **20** (1), 25. https://doi.org/10.1186/s12859-018-2450-5 (2019).
60. Knapp, J. R., Laur, G. L., Vadas, P. A., Weiss, W. P. & Tricarico, J. M. Invited review: enteric methane in dairy cattle production: quantifying the opportunities and impact of reducing emissions. *J. Dairy. Sci.* **97** (6), 3231–3261. (2014).
61. Hiltpold, M. et al. Activation of cryptic splicing in bovine WDR19 is associated with reduced semen quality and male fertility. *PLOS Genet.* **16** (5), e1008804. (2020).
62. Nagai, R. et al. Genome-wide detection of non-additive quantitative trait loci for semen production traits in beef and dairy bulls. *Animal* **16** (3), 100472. https://doi.org/10.1016/j.animal.2022.100472 (2022).
63. Kinukawa, M. et al. A potent allele marker related to low bull conception rate in Japanese black bulls. *Animal* **17** (5), 100804. https://doi.org/10.1016/j.animal.2023.100804 (2023).
64. Sumreddee, P. et al. Inbreeding depression in line 1 Hereford cattle population using pedigree and genomic information. *J. Anim. Sci.* **97** (1), 1–18. https://doi.org/10.1093/jas/sky385 (2019).
65. Nishio, M. et al. Comparing pedigree and genomic inbreeding coefficients, and inbreeding depression of reproductive traits in Japanese black cattle. *BMC Genom.* **24** (1), 376. https://doi.org/10.1186/s12864-023-09480-5 (2023).
66. Ferenčaković, M. et al. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. *J. Anim. Breed. Genet.* **130** (4), 286–293. https://doi.org/10.1111/jbg.12012 (2013).
67. Makanjuola, B. O., Maltecca, C., Miglior, F., Schenkel, F. S. & Baes, C. F. Effect of recent and ancient inbreeding on production and fertility traits in Canadian Holsteins. *BMC Genom.* **21** (1), 605. https://doi.org/10.1186/s12864-020-07031-w (2020).
68. Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I. & Grindflek, E. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46** (1), 49. https://doi.org/10.1186/1297-9686-46-49 (2014).
69. Villanueva, B. et al. The value of genomic relationship matrices to estimate levels of inbreeding. *Genet. Sel. Evol.* **53** (1), 42. (2021).
70. VanRaden, P. M., Olson, K. M., Wiggans, G. R., Cole, J. B. & Tooker, M. E. Genomic inbreeding and relationships among Holsteins, Jerseys, and brown Swiss. *J. Dairy. Sci.* **94** (11), 5673–5682. (2011).
71. Guinan, F. L. et al. Changes in genetic trends in US dairy cattle since the implementation of genomic selection. *J. Dairy. Sci.* **106** (2), 1110–1129. (2023).
72. Meyermans, R., Gorssen, W., Buys, N. & Janssens, S. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genom.* **21** (1), 94. (2020).
73. Wellmann, R. & Bennewitz, J. Key genetic parameters for population management. *Front. Genet.* **667** https://doi.org/10.3389/fgene.2019.00667 (2019).
74. Meuwissen, T. H. E. Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* **75** (4), 934–940. https://doi.org/10.2527/1997.754934x (1997).
75. Morales-González, E., Fernández, J., Pong-Wong, R., Toro, M. Á. & Villanueva, B. Changes in allele frequencies when different genomic coancestry matrices are used for maintaining genetic diversity. *Genes* **12** (5), 673. https://doi.org/10.3390/genes12050673 (2021).
76. Zhao, F. et al. Genetic gain and inbreeding from simulation of different genomic mating schemes for pig improvement. *J. Anim. Sci. Biotechnol.* **14** (1), 87. https://doi.org/10.1186/s40104-023-00872-x (2023).

Acknowledgements

This work was partially supported by the Livestock Promotional Subsidies of the Japan Racing Association (JRA). Rintaro Nagai was granted by JST, the establishment of university fellowships for the creation of science technology innovation (Grant Number JPMJFS2102). The authors thank the Japan Livestock Technology Association for providing the Japanese Black cattle dataset with the SNP genotypes on the Illumina Bovine HD BeadsChip. We thank Editage (www.editage.com) for the English language editing.

Author contributions

RN designed the experiment, performed the statistical analysis, and contributed to writing and improving the manuscript. TH and MS designed the experiment and contributed to writing and improving the manuscript. YU designed the experiment, collected SNP genotypes, performed the statistical analysis, managed the entire project, and contributed to writing and improving the manuscript. All authors have read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The Animal Care and Use Committee approval was not required for this study because the dataset was simulated and obtained from previously published data[32,33].

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-026-37530-x.

Correspondence and requests for materials should be addressed to Y.U.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026