BMC Genomics

RESEARCH

Open Access

# Expanding the application of haplotype-based genomic predictions to the wild: A case of antibody response against *Teladorsagia circumcincta* in Soay sheep

Seyed Milad Vahedi[1], Siavash Salek Ardetani[2*], Luiz F. Brito[3], Karim Karimi[4], Kian Pahlavan Afshari[5] and Mohammad Hossein Banabazi[6,7*]

## Abstract

**Background** Genomic prediction of breeding values (GP) has been adopted in evolutionary genomic studies to uncover microevolutionary processes of wild populations or improve captive breeding strategies. While recent evolutionary studies applied GP with individual single nucleotide polymorphism (SNP), haplotype-based GP could outperform individual SNP predictions through better capturing the linkage disequilibrium (LD) between the SNP and quantitative trait loci (QTL). This study aimed to evaluate the accuracy and bias of haplotype-based GP of immunoglobulin (Ig) A (IgA), IgE, and IgG against *Teladorsagia circumcincta* in lambs of an unmanaged sheep population (Soay breed) based on Genomic Best Linear Unbiased Prediction (GBLUP) and five Bayesian [BayesA, BayesB, BayesCπ, Bayesian Lasso (BayesL), and BayesR] methods.

**Results** The accuracy and bias of GPs using SNP, haplotypic pseudo-SNP from blocks with different LD thresholds (0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.00), or the combinations of pseudo-SNPs and non-LD clustered SNPs were obtained. Across methods and marker sets, higher ranges of genomic estimated breeding values (GEBV) accuracies were observed for IgA (0.20 to 0.49), followed by IgE (0.08 to 0.20) and IgG (0.05 to 0.14). Considering the methods evaluated, up to 8% gains in GP accuracy of IgG were achieved using pseudo-SNPs compared to SNPs. Up to 3% gain in GP accuracy for IgA was also obtained using the combinations of the pseudo-SNPs with non-clustered SNPs in comparison to fitting individual SNP. No improvement in GP accuracy of IgE was observed using haplotypic pseudo-SNPs or their combination with non-clustered SNPs compared to individual SNP. Bayesian methods outperformed GBLUP for all traits. Most scenarios yielded lower accuracies for all traits with an increased LD threshold. GP models using haplotypic pseudo-SNPs predicted less-biased GEBVs mainly for IgG. For this trait, lower bias was observed with higher LD thresholds, whereas no distinct trend was observed for other traits with changes in LD.

**Conclusions** Haplotype information improves GP performance of anti-helminthic antibody traits of IgA and IgG compared to fitting individual SNP. The observed gains in the predictive performances indicate that haplotype-based methods could benefit GP of some traits in wild animal populations.

*Correspondence:
Siavash Salek Ardetani
siavash.salek@znu.ac.ir
Mohammad Hossein Banabazi
mohammad.hossein.banabazi@slu.se
Full list of author information is available at the end of the article

**Keywords** Genomic prediction, Haplotype-based models, Pseudo-SNP, Individual SNP models, GBLUP, Bayesian, Soay sheep

## Background

Genomic prediction of breeding values (GP) was described over 20 years ago with the primary goal of accurately identifying the breeding candidates with the highest genetic merit using genome-wide single nucleotide polymorphism (SNP) markers [1]. The development of GP models and the affordable cost of genotyping have revolutionized animal breeding by increasing selection accuracies, shortening the generation interval, and increasing genetic progress of economically important traits [2, 3]. GP has also received considerable attention in other fields with similar practical needs. Genomic selection has been implemented in plant breeding programs for genetic improvement of quantitative traits [4, 5]. Moreover, GP has been applied to human genetics for identifying patient risk for particular diseases, referred to as "polygenic risk score" or selecting the best treatment option based on the individual's genotype [6, 7]. Despite the application of GP in livestock, plant, and human genetics, a limited number of studies have applied this method in wild or unmanaged animal populations [8–10].

The choice of the statistical model to be used for GP is a critical step in the success of genomic analyses. Statistical models commonly used for GP can be classified into two categories: (i) linear parametric methods referred to as "Best Linear Unbiased Prediction (BLUP) methods", such as genomic BLUP or GBLUP [11] and Single-step GBLUP [12], and (ii) non-linear parametric methods, such as BayesA [1], BayesB [1], BayesCπ [13], Bayesian Lasso (BayesL) [14], and BayesR [15]. These methods mainly differ in the assumptions used for the prior distribution of genetic effects. In GBLUP, a normal prior distribution is assumed for the marker effects, which means that a large number of quantitative trait loci (QTL) influence the trait, with most markers exhibiting a small effect [1]. The two models of BayesA and BayesB, described by Meuwissen et al. [1], assume SNP-specific variances; BayesA fits all SNPs, while BayesB fits approximately 1-π proportion of SNPs, where π is the percentage of markers which have no influence on the trait (zero effect). Therefore, when $\pi = 0$, BayesB is equivalent to BayesA. BayesCπ is similar to BayesB but treats π as an unknown parameter with a uniform (0, 1) prior distribution, and it assumes all SNP effects have a common variance [13]. The BayesL method assumes that the variance of the SNP marker effects follows a double exponential or Laplace distribution [16]. BayesR provides high flexibility by using a mixture of normal distributions as the prior for SNP effects [15]. In this method, four classes of SNP effect size (null, small, medium, and large) can be defined, for instance, and SNP effects would be modeled using a four-component normal mixture model [15]. In general, Bayesian approaches tend to be more accurate than GBLUP when the number of QTL explaining the trait variance is small [17].

Practical applications of GP have focused on single-SNP models fitting individual SNP as a locus in the mixed models without any information about the marker location. Instead, haplotype models have the potential to include structural genomic information to improve the accuracy of genomic evaluation [18–20]. Haplotypes are more informative than SNPs in describing recent identical-by-descent (IBD) relationships, and they may also be more effective in capturing linkage disequilibrium (LD) with multiallelic QTL than individual SNP, which are often biallelic [19]. In practice, the performance of GP based on haplotypes varies across traits and species, ranging from negligible to substantial increases in accuracy compared to SNP-based models [19, 21–24]. Three methods have been applied to define haploblocks, including (i) a fixed number of SNPs per haplotype block [25], (ii) fixed block length [26], and (iii) LD blocks [27]. The latter method is an efficient method that can decrease the number of explanatory variables without losing the information provided by the SNP markers [28]. By setting a minimum LD between SNP markers, they can be grouped into haploblocks that do not have a fixed length or a fixed number of SNPs. Due to the presence of relatively strong LD among particular markers, the number of variants per haploblock is reduced considerably compared to when haploblocks are defined by a fixed number of close SNPs [27]. Haplotypic information could be then integrated as pseudo-SNPs into BLUP [21, 22] or Bayesian [26] GP models, or based on a recent method applied by Araujo et al. [23], the pseudo-SNPs can contribute to a genomic relationship matrix (GRM) construction in combination with non-LD clustered SNPs, i.e., those located out of haploblocks.

Recently, GP methods have been adopted by researchers interested in quantitative evolution of wild animals [8, 9]. With the rise of wildlife infectious diseases, e.g., sea-star wasting disease [29], bats' white-nose syndrome [30], chytrid fungus in amphibians [31], and the

emergence of zoonotic infections such as SARS-CoV-2 in captive [32, 33] and unmanaged populations [34], GP models could be used to improve captive breeding and conservation strategies to select resistant individuals against pathogens. Moreover, GP models can be used in wild populations to investigate the microevolutionary trends of traits, and more accurate models can better demonstrate these changes. Then, we can better understand the "cryptic microevolution" process, which refers to traits being heritable and under directional selection, but they do not constantly evolve in response to that selection in the expected way [35]. Thus, it is of interest to know the accuracy of different GP models in prediction of the individuals' genomic merit for different traits in wild populations. Additionally, it is unclear if the GP of immune response traits in wild populations would potentially benefit from haplotypic information.

The goal of the present study was to investigate the GP accuracy of antibody response against a gastrointestinal strongyle nematode, *Teladorsagia circumcincta*, in Soay sheep lambs, as an example to explore the performance of haplotype-based GP models for their potential future applications in captive breeding or conservation strategies. Therefore, three different analyses were performed: (i) SNP markers were fitted, (ii) haplotypes constructed based on different LD threshold were fitted as pseudo-SNPs, and (iii) the pseudo-SNPs combined with the non-LD clustered SNPs were fitted. The accuracy and bias of GP from GBLUP and five Bayesian approaches, including BayesA, BayesB, BayesCπ, BayesL, and BayesR, were then compared. This study used a publicly-available dataset from a 25-year study that quantified antibody levels in unmanaged Soay breed lambs [36].

## Results

The descriptive statistics of the phenotypic records and adjusted phenotypes are presented in Table 1. The average ± standard error (SE) of inter-marker distance was 68.50 ± 2.19 Kb, and the minimum and maximum distances between SNPs were 3.17 Kb and 423.72 Kb, respectively.

Haploblock construction was performed based on ten thresholds of LD (measured as $r^2$) ranging from 0.15 to 1.00 (Table 2). As the LD threshold increased, the number of haploblocks and pseudo-SNPs decreased, ranging from 1,432 to 8,442 and 2,897 to 28,265, respectively. With an increase in $r^2$, the average number of SNPs per block decreased, ranging from 2.10 to 2.22 SNPs. Moreover, with stricter LD thresholds, the total number of SNPs applied to haploblocks and the total length of the autosome covered by the haplotypes decreased from 18,705 to 2,991 and 249.7 Mb to 32.2 Mb, respectively (Table 2).

Non-LD cluster SNPs were combined with pseudo-SNPs from haplotype blocks with different LD thresholds (Table 3). With an increase in $r^2$, the number of non-clustered SNPs decreased from 34,038 to 18,324, i.e., fewer SNP markers contributed to the haploblock construction, and more markers remained as individual SNP out of haploblocks. Notably, the proportion of pseudo-SNPs in the total variants decreased from 61 to 8% with an increase in the LD levels, and the total number of variants (i.e., the overall number of SNP and pseudo-SNP markers) decreased from 46,589 to 36,935.

### Genomic relationship matrices

The GRMs were constructed using the SNPs and pseudo-SNPs and based on the VanRaden method [11]. To investigate the presence of family structure in the studied population, the principal component analysis (PCA) of SNP markers was depicted (Fig. 1A). Furthermore, the distribution of the diagonal elements of the genomic relationship matrix based on individual SNP ($G_{SNP}$) was plotted (Fig. 1B and Fig. 1C). The bar and QQ plots of $G_{SNP}$ diagonal elements show that the distribution was close to normal. Moreover, no distinct cluster was observed in the PCA (Fig. 1A). These results confirm that there was minimal familial structure among the genotyped animals.

Euclidean distances between different pairs of GRMs are depicted in Fig. 2. Our results confirmed that GRMs

**Table 1** Total number (N), minimum (Min), maximum (Max), mean, standard deviation (SD) of phenotypes and adjusted phenotypes, and number of individuals in the training[a] and testing[b] sets

| Trait | N | Min | Max | Mean | SD | Training (N)[a] | Testing (N)[b] |
|---|---|---|---|---|---|---|---|
| IgA | 2,034 | 0 | 2.73 | 0.74 | 0.50 | 1,848 | 186 |
| IgA_adj[c] | | -0.84 | 2.02 | 0 | 0.50 | | |
| IgE | 2,034 | 0 | 1.08 | 0.09 | 0.12 | 1,848 | 186 |
| IgE_adj[a] | | -0.14 | 1.00 | 0 | 0.12 | | |
| IgG | 2,034 | 0 | 1.60 | 0.24 | 0.19 | 1,848 | 186 |
| IgG_adj[a] | | -0.32 | 1.36 | 0 | 0.19 | | |

[c] IgA_adj = adjusted phenotype of IgA; IgE_adj = adjusted phenotype of IgE; IgG_adj = adjusted phenotype of Ig

**Table 2** Statistics of haploblocks defined based on linkage disequilibrium (LD) levels of $r^2$

| | $r^2$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0.15** | **0.20** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** | **1.00** |
| **Total number of pseudo-SNP** | 28,265 | 26,673 | 22,971 | 19,562 | 16,384 | 13,354 | 10,663 | 8,031 | 5,816 | 2,897 |
| **Number of blocks** | 8,442 | 7,976 | 7,021 | 6,110 | 5,248 | 4,399 | 3,654 | 2,906 | 2,281 | 1,432 |
| **Min SNPs per block** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **Max SNPs per block** | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 5 |
| **Average number of SNPs per block** | 2.22 | 2.19 | 2.17 | 2.16 | 2.15 | 2.15 | 2.15 | 2.15 | 2.14 | 2.10 |
| **Total number of included SNPs (%[a])** | 18,705 (50.5%) | 17,505 (47.3%) | 15,232 (41.1%) | 13,186 (35.6%) | 11,301 (30.5%) | 9,476 (25.6%) | 7,868 (21.2%) | 6,256 (16.9%) | 4,886 (13.2%) | 2,991 (8.1%) |
| **Length of covered genome (Mb)** | 249.7 | 230.6 | 196.8 | 168.2 | 142.3 | 117.6 | 96.0 | 75.3 | 57.1 | 32.2 |

[a] Number of included SNPs divided by total number of SNPs after quality control

**Table 3** Number and proportion (%) of pseudo-SNPs and non-LD clustered SNPs combined for genomic prediction analyses, based on linkage disequilibrium (LD) level of $r^2$

| $r^2$ | Analysis | Pseudo-SNP (%) | SNP (%) | Total |
|---|---|---|---|---|
| 0.15 | $A_{COM0.15}$ | 28,265 (61%) | 18,324 (39%) | 46,589 |
| 0.20 | $A_{COM0.20}$ | 26,673 (58%) | 19,527 (42%) | 46,200 |
| 0.30 | $A_{COM0.30}$ | 22,971 (51%) | 21,797 (49%) | 44,768 |
| 0.40 | $A_{COM0.40}$ | 19,562 (45%) | 23,843 (55%) | 43,405 |
| 0.50 | $A_{COM0.50}$ | 16,384 (39%) | 25,728 (61%) | 42,112 |
| 0.60 | $A_{COM0.60}$ | 13,354 (33%) | 27,553 (67%) | 40,907 |
| 0.70 | $A_{COM0.70}$ | 10,663 (27%) | 29,161 (73%) | 39,824 |
| 0.80 | $A_{COM0.80}$ | 8,031 (21%) | 30,773 (79%) | 38,804 |
| 0.90 | $A_{COM0.90}$ | 5,816 (15%) | 32,142 (85%) | 37,958 |
| 1.00 | $A_{COM1.00}$ | 2,897 (8%) | 34,038 (91%) | 36,935 |

constructed based on SNPs, haplotypes, or their combinations were different from each other. Regarding GRMs based on haplotypes, with stricter LD thresholds, more SNP markers were eliminated from the analyses; fewer of them contributed to the GRM construction. Therefore, with the increase in the LD threshold, higher distances were observed from $G_{SNP}$. Concerning GRMs based on the combination of pseudo-SNPs and non-LD clustered SNPs, by increasing the LD threshold, more SNPs were removed from the haploblock construction; instead, they were used as individual SNP in the GRM construction. In contrast, with lower LD thresholds, more SNP markers would contribute to the haploblocks, and fewer non-LD clustered SNP markers were involved in the GRM construction. Consequently, with stricter LD thresholds, more similarity was observed between these GRMs and $G_{SNP}$. Interestingly, the magnitude of differentiation from $G_{SNP}$ was more apparent for GRMs based on the pseudo-SNPs than those observed for GRMs based on the combination of pseudo-SNPs and non-clustered SNPs at different levels of LD.

### Heritability estimates

Heritability estimates for different scenarios are shown in Fig. 3. We observed higher ranges of heritability for IgA (0.20 to 0.49) compared to IgG (0.15 to 0.30) and IgE (0.08 to 0.24). The Bayesian methods resulted in higher heritability estimates for all traits than restricted maximum likelihood (REML). Irrespective of the applied methods, the heritability estimates were close when analyses were based on the combination of SNPs and pseudo-SNPs ($A_{COM0.15}$-$A_{COM1.00}$). However, for the analyses based on haplotypes ($A_{HAP0.15}$-$A_{HAP1.00}$), with more relaxed LD thresholds, from 1.00 to 0.15, the heritability estimates increased by 14–23%, 6–12%, and 3–11% for IgA, IgE, and IgG, respectively.

For each trait, the highest and lowest heritabilities obtained based on individual SNP, pseudo-SNP, and the combinations of pseudo-SNPs and non-LD clustered SNPs are listed in Supplementary Table 1. The highest heritability (0.49) was obtained for IgA when the BayesL method was applied to the $A_{COM0.20}$ and $A_{COM0.50}$. On the contrary, the lowest heritability (0.20) was observed when $A_{HAP1.00}$ was applied to BayesB. Regarding IgE, the highest estimate (0.24) was obtained for BayesA based on $A_{SNP}$, $A_{HAP0.15}$, $A_{HAP0.20}$, $A_{HAP0.30}$, $A_{HAP0.40}$, and based on all the applied combinations of SNPs and pseudo-SNPs. However, the lowest estimate of 0.08 was observed by BayesR with $A_{HAP1.00}$. The highest heritability of IgG (0.30) was observed by BayesL based on $A_{COM0.20}$, $A_{COM0.50}$, and $A_{COM0.60}$. In contrast, the lowest heritability of 0.15 was found by BayesB based on $A_{HAP1.00}$.

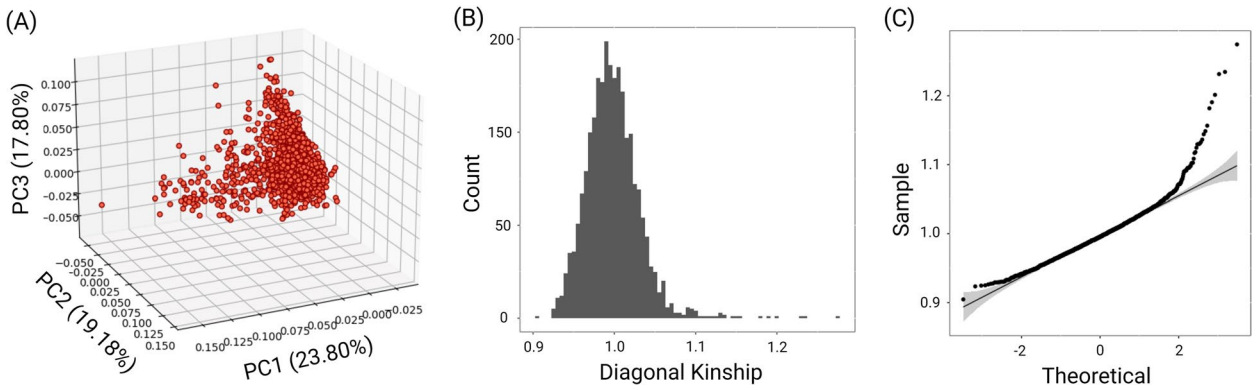


**Fig. 1** Plots of the principal components and the distribution of the genomic relationship matrix based on the genome-wide SNP markers. **A** Scatter plot of the first three principal components of the genomic dataset; no distinct clusters are observed in the studied dataset. **B** Histogram of the diagonal elements of the genomic relationship matrix. Despite the fact that a small portion of the population has high kinship values (> 1.1), no distinct peaks were observed for the studied dataset. (C) Quantile–quantile plot of the diagonal elements of the genomic relationship matrix. Overall, it is inferred that the genotypic dataset of Soay sheep was mainly generated from the animals with minimum family structure

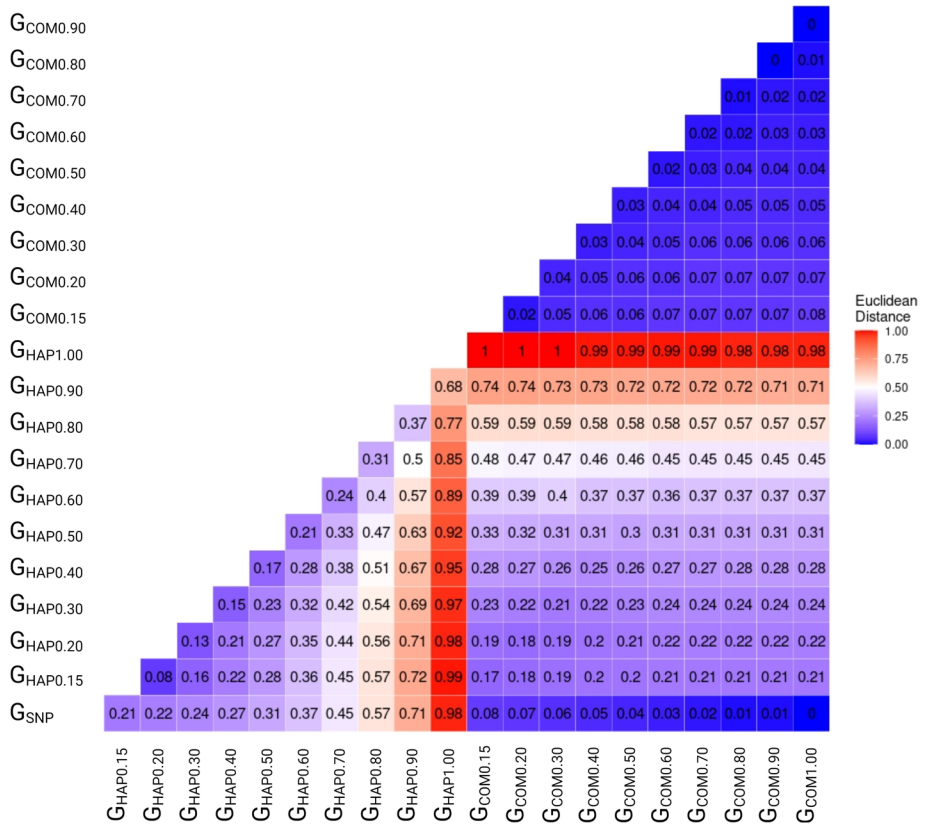


**Fig. 2** Heatmap of Euclidean distance between different genomic relationship matrices (GRM). Euclidean distance was used to compare a total of 21 GRMs, including $G_{SNP,}$ which refers to the GRM defined based on SNPs as markers, and $G_{HAP0.15}$, $G_{HAP0.20}$, $G_{HAP0.30}$, $G_{HAP0.40}$, $G_{HAP0.50}$, $G_{HAP0.60}$, $G_{HAP0.70}$, $G_{HAP0.80}$, $G_{HAP0.90}$, and $G_{HAP1.00}$ defined based on haplotypes constructed by LD thresholds of 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00 as markers, respectively. Values were scaled between 0 and 1

### Genomic prediction accuracies

The accuracies of GPs based on different methods and analyses are presented in Fig. 4. Higher ranges of accuracies were observed for IgA (0.20 to 0.49), followed by IgE (0.08 to 0.20), and IgG (0.05 to 0.14). Through comparing the accuracy of different marker sets in each applied method, we found up to 3% improvement in GP accuracy of IgA using the combination of the pseudo-SNPs with non-LD clustered SNPs. In contrast, for IgE, comparable accuracies to GBLUP using individual SNP were obtained by the combination of the pseudo-SNPs with non-LD clustered SNPs and haplotype-based GPs. For IgG, up to 8% gains in GP accuracy were observed in analyses based on the haplotypic pseudo-SNPs (Fig. 4).

Bayesian methods outperformed GBLUP in all three studied traits. By comparing the accuracy of different methods in each applied marker set, we found that BayesB outperformed all other methods for IgA. In most cases of IgE and IgG, Bayesian approaches obtained higher accuracy than GBLUP (Fig. 4). For each trait, the highest and lowest accuracies achieved based on individual SNP, pseudo-SNP, and the combinations of pseudo-SNPs and non-LD clustered SNPs were listed in Supplementary Table 2. Regarding IgA, the highest GP accuracy (0.49) was obtained with the BayesB method based on $A_{COM0.15}$ or $A_{COM0.20}$ compared to GBLUP using $A_{SNP}$ (0.31). In contrast, the lowest accuracy (0.20) was estimated based on BayesL and GBLUP with $A_{HAP0.90}$. Concerning IgE, the highest accuracy (0.20) was given by BayesL with $A_{SNP}$ and BayesR with $A_{COM0.60}$ or $A_{COM0.70}$. On the contrary, the lowest accuracy (0.08) was obtained for BayesB using $A_{HAP0.90}$. With regards to IgG, the highest accuracy (0.14) was estimated for BayesB based on $A_{HAP0.70}$. Conversely, the weakest performance (0.04) was given by BayesL based on $A_{COM0.40}$.

Regardless of the applied methods, a general trend was observed for IgA GPs based on haplotypic pseudo-SNPs and the combination of pseudo-SNPs and non-clustered SNPs, indicating lower accuracies with the increase in the LD threshold. Meanwhile, higher accuracies were obtained for IgG with higher LD threshold when the combinations of pseudo-SNPs and non-clustered SNPs were used; however, a slight reduction in the accuracies was observed with stringent LD levels (> 0.70).

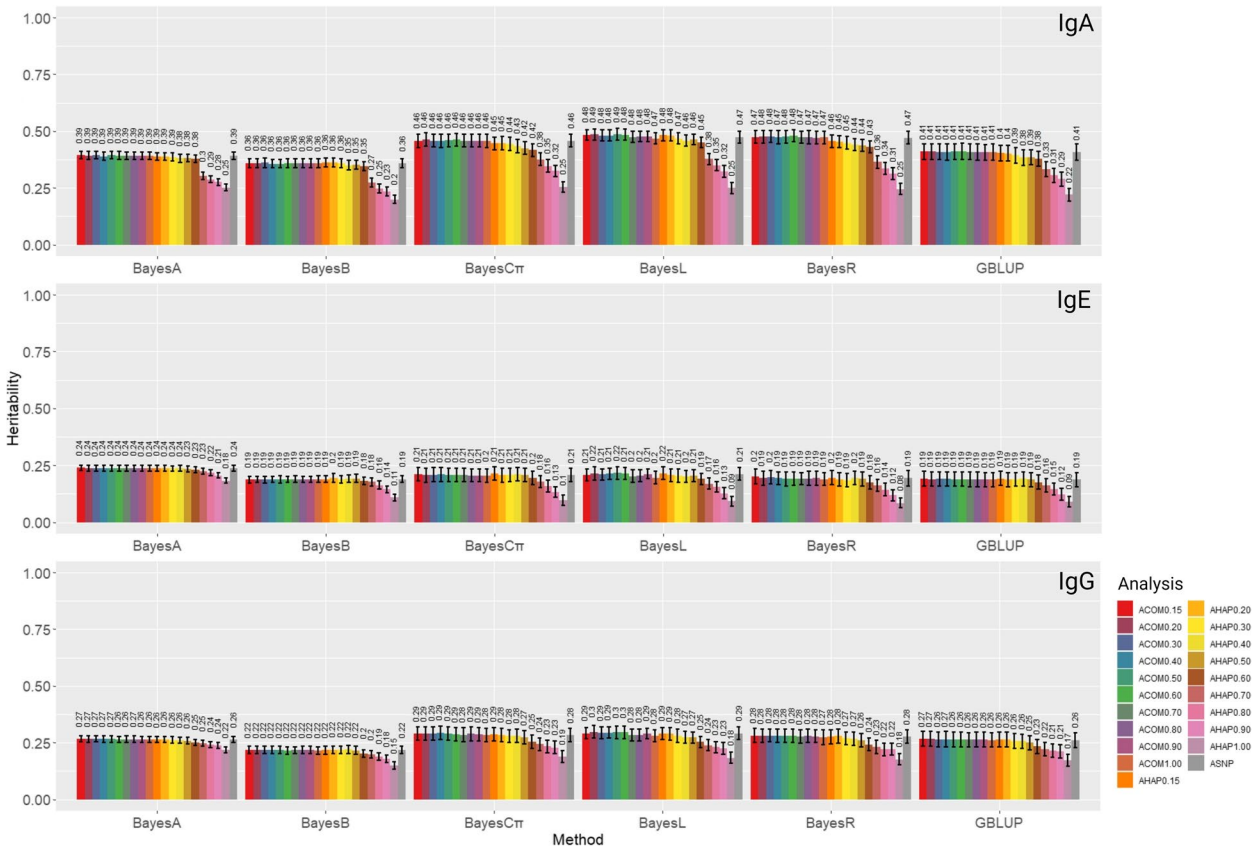


**Fig. 3** Heritability estimates for IgA, IgE, and IgG applying different methods and analyses. Heritability was computed as the ratio of the additive genetic variance to the phenotypic variance. The methods evaluated are GBLUB, BayesA, BayesB, BatesCπ, BayesLasso, and BayesR based on different analyses, including $A_{SNP}$, $A_{COM0.15}$, $A_{COM0.20}$, $A_{COM0.30}$, $A_{COM0.40}$, $A_{COM0.50}$, $A_{COM0.60}$, $A_{COM0.70}$, $A_{COM0.80}$, $A_{COM0.90}$, $A_{COM1.00}$, $A_{HAP0.15}$, $A_{HAP0.20}$, $A_{HAP0.30}$, $A_{HAP0.40}$, $A_{HAP0.50}$, $A_{HAP0.60}$, $A_{HAP0.70}$, $A_{HAP0.80}$, $A_{HAP0.90}$, and $A_{HAP1.00}$. Definitions of the analyses are given in Tables 3 and 4

In 13 out of the 18 model-trait combinations, i.e., all IgG scenarios along with five out of six IgA's, at least one of the models based on haplotypes or the combination of haplotypes and non-clustered SNPs achieved a higher accuracy than the model fitting SNPs (Supplementary Table 3). Only in one scenario (BayesL for IgE), a higher accuracy was obtained by the model fitting individual SNPs. In four scenarios, comparable performances were observed between models fitting individual SNP and those using haplotypic information (Supplementray Table 3).

We revealed that the magnitude of differences between the highest and the lowest GP accuracies was higher for haplotype-based approaches than those based on the models fitting the combinations of pseudo-SNPs and SNP markers. These differences were 0.27, 0.11, and 0.09 for haplotype-based GPs of IgA, IgE, and IgG, respectively. In contrast, we obtained lower differences of 0.18, 0.05, and 0.02 for GPs of IgA, IgE, and IgG based on the combination of pseudo-SNPs and non-clustered SNP markers, respectively.

## Genomic prediction biases

The bias in Genomic Estimated Breeding Value (GEBV) predictions for all scenarios is presented in Fig. 5 as deviations from 1 (bias – 1). Considering all methods and analyses, the bias deviation values ranged from -0.44 to 0.30, -0.63 to 0.19, and -0.80 to -0.30 for IgA, IgE, and IgG, respectively. In most scenarios (17 out of 18 model-trait combinations), at least one of the models based on haplotypes or the combination of haplotypes and non-clustered SNPs achieved a lower bias than the model fitting SNPs individually (Supplementary Table 4). Only in one scenario a comparable bias was observed between models fitting individual SNPs and those using haplotypic information (Supplementary table 4).

For IgA, we observed unbiased GP by haplotype-based scenarios of BayesR using $A_{SNP}$, $A_{COM0.70}$, and $A_{COM0.90}$, which were comparable to the bias obtained by GBLUP using $A_{SNP}$. In contrast, the most biased scenario was observed when BayesL was applied with $A_{HAP0.90}$ ($-0.44 \pm 0.20$). Regarding IgE, GBLUP based on $A_{COM0.30}$ and $A_{COM0.40}$ provided unbiased GEBV prediction. On

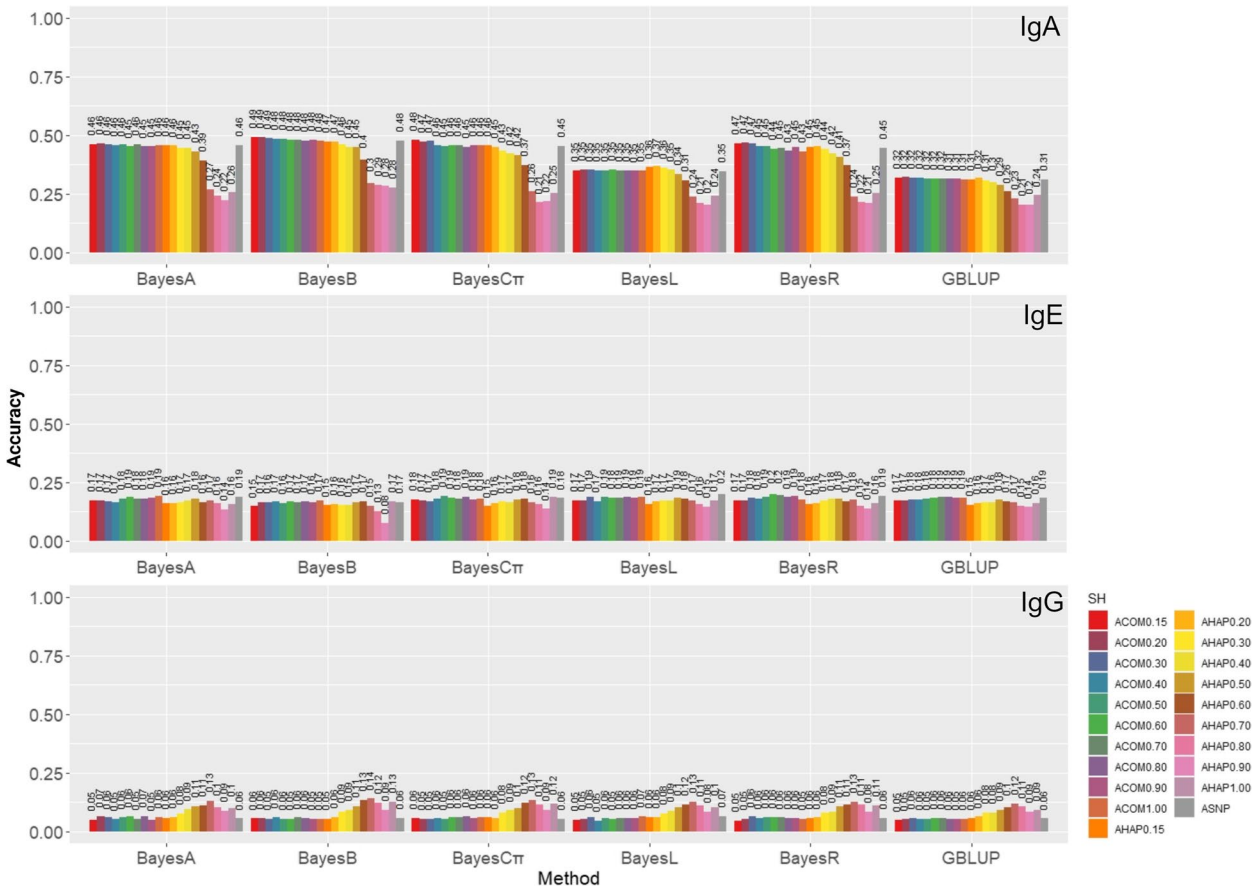


**Fig. 4** The estimates of genomic prediction accuracy of IgA, IgE, and IgG applying different methods and analyses. The genomic prediction accuracy was measured by the correlation between adjusted phenotypes ($y_c$) and GEBV for the validation subset. Methods under evaluation were GBLUB, BayesA, BayesB, BatesCπ, BayesLasso, and BayesR based on different analyses, including $A_{SNP}$, $A_{COM0.15}$, $A_{COM0.20}$, $A_{COM0.30}$, $A_{COM0.40}$, $A_{COM0.50}$, $A_{COM0.60}$, $A_{COM0.70}$, $A_{COM0.80}$, $A_{COM0.90}$, $A_{COM1.00}$, $A_{HAP0.15}$, $A_{HAP0.20}$, $A_{HAP0.30}$, $A_{HAP0.40}$, $A_{HAP0.50}$, $A_{HAP0.60}$, $A_{HAP0.70}$, $A_{HAP0.80}$, $A_{HAP0.90}$, and $A_{HAP1.00}$. Definitions of the analyses are given in Tables 3 and 4

the contrary, the highest level of bias was obtained by BayesB based on $A_{HAP0.90}$ (-0.63 ± 0.36). Concerning IgG, the least biased haplotype-based GP was observed by GBLUP based on $A_{HAP0.70}$ (-0.30 ± 0.70). However, the BayesL method using $A_{COM0.40}$ showed the most biased GP (-0.80 ± 0.31). Haplotype-based GPs predicted less-biased GEBVs in most IgG scenarios with high LD thresholds compared with SNP-based models, whereas no improvement in bias was observed for other traits with an increase in LD level. Moreover, in most scenarios of IgE and all IgG scenarios, GEBV inflation (bias < 0) was observed.

## Discussion

GP has been widely used to enhance the genetic gain of complex traits in livestock and plant breeding and predict polygenic risk scores of particular human diseases [37–39]. Recently, more attention has been given to this tool in the evolutionary genetics topic to improve captive breeding strategies and understand the microevolution of breeding values [8, 9]. In this study, we applied a haplotypic GP approach on three helminth-specific immune response traits of IgA, IgE, and IgG against *T. circumcincta* in the unmanaged population of Soay sheep. Haplotype-based GP has achieved little to substantial improvements in prediction accuracies compared with SNP models in domesticated species [21, 22, 40], including sheep [41]. However, to our knowledge, the application of haplotype-based GP and assessment of haplotypic GP performance have not been reported in wild or unmanaged populations.

Haploblocks were constructed based on different LD thresholds (Table 2). The Big-LD method applied to our study constructs the LD blocks using the weights estimated based on the number of SNPs from all possible overlapping intervals [42]. We observed the construction of haploblocks with LD = 1, which is not very common in commercial populations of domesticated species,

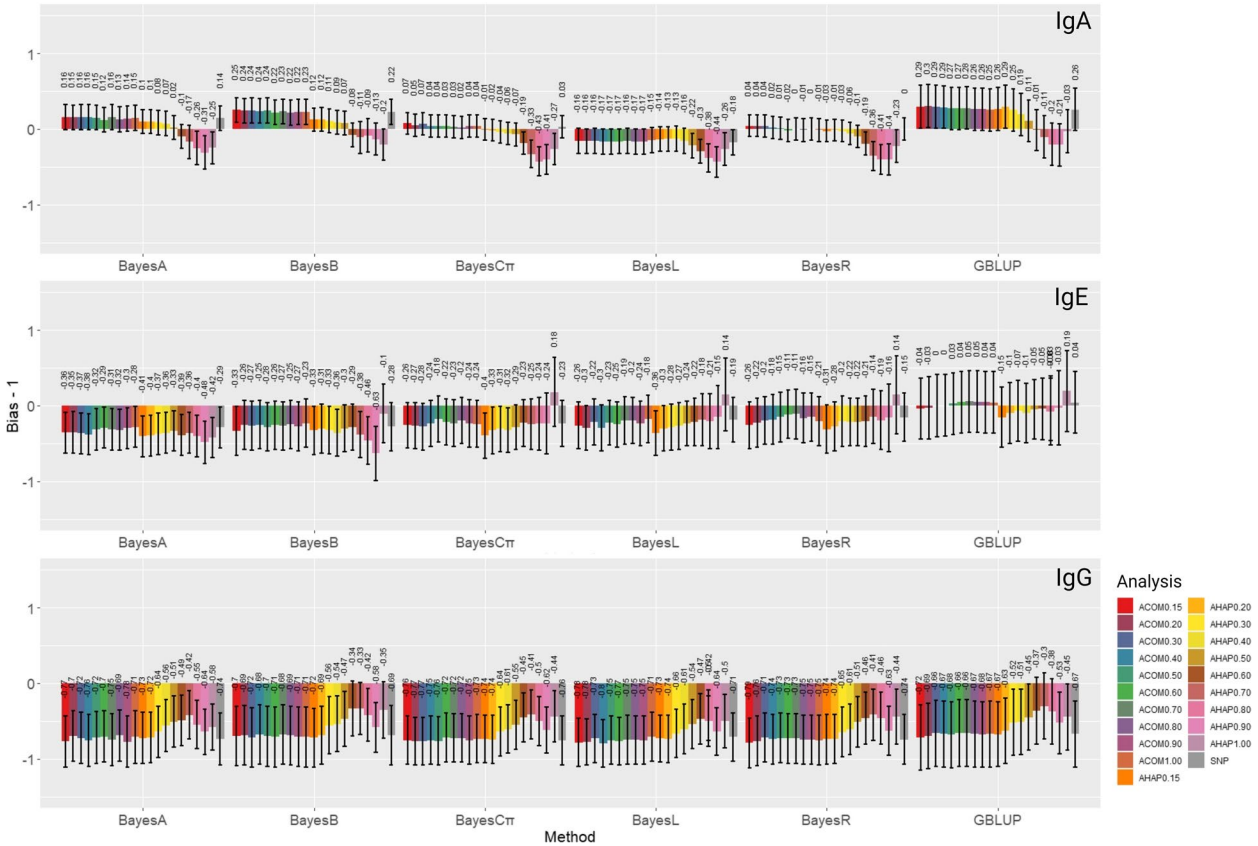


**Fig. 5** The bias estimates of genomic estimated breeding values (GEBV) of IgA, IgE, and IgG applying different methods and analyses. The genomic prediction bias was measured as the regression coefficients obtained by regressing the adjusted phenotypes ($y_c$) upon the precited direct genomic values GEBV in the validation subset. Methods under evaluation were GBLUB, BayesA, BayesB, BatesCπ, BayesLasso, and BayesR based on different analyses, including $A_{SNP}$, $A_{COM0.15}$, $A_{COM0.20}$, $A_{COM0.30}$, $A_{COM0.40}$, $A_{COM0.50}$, $A_{COM0.60}$, $A_{COM0.70}$, $A_{COM0.80}$, $A_{COM0.90}$, $A_{COM1.00}$, $A_{HAP0.15}$, $A_{HAP0.20}$, $A_{HAP0.30}$, $A_{HAP0.40}$, $A_{HAP0.50}$, $A_{HAP0.60}$, $A_{HAP0.70}$, $A_{HAP0.80}$, $A_{HAP0.90}$, and $A_{HAP1.00}$. Definitions of the analyses are given in Tables 3 and 4

as markers in such a high LD are typically eliminated in the process of designing SNP panels. When setting low LD thresholds to construct the LD blocks, more intervals of linked SNPs are obtained, the number of blocks is increased, fewer SNPs are excluded, and a higher portion of the genome is covered by haploblocks (and vice versa). Consequently, a greater number of blocks are expected with lower LD thresholds, as observed when comparing the numbers of blocks across LD thresholds from 0.15 to 1.00 (Table 2). The average number of SNPs per block showed that most of the haploblocks were constructed by two SNPs; however, the proportion of two-SNP-blocks increased with stricter LD thresholds. The reason could be that we used genotypes obtained from a medium-density SNP chip (50 K) for haploblock construction. This could result in the less presence of haploblocks with > 2 SNP markers with high LD thresholds since one essential criterion by which SNP markers are selected for SNP chips in commercial species is the gaps between markers, more importantly, the distance between two adjacent SNPs. Moreover, genotype imputation based on higher-density reference panels can increase SNP density and, therefore, haplotype construction. However, higher-density SNP data was unavailable for the current study, and we suggest genotype imputation in future studies. The number of total variants increased for analyses based on haplotypic pseudo-SNPs and individual SNP with stricter LD thresholds (Table 3). The reason is that with higher LD, fewer individual SNP were blocked in haplotypes, and more SNP markers remained non-clustered.

We showed that with higher LD thresholds, GRMs constructed based on haplotypes are more differentiated than that based on individual SNP, whereas, for GRMs based on the combinations of pseudo-SNPs and non-clustered SNPs, the trend was reversed (Fig. 2). For GRMs based on haplotypes, with stricter LD thresholds, the relationship among individuals in the population is defined based on a shorter length of the genome and a lower number of SNPs; consequently, it could cause more differentiation among GRMs. However, for GRMs based

on the combinations of pseudo-SNPs and non-clustered SNPs with higher LD thresholds, greater similarity was observed with $G_{SNP}$ due to the lower number of blocked SNPs and more contribution of individual SNP to the GRM (Table 3).

#### Heritability estimates

We estimated a wider range of heritabilities (0.08 to 0.49) for the antibody traits against *T. circumcincta* compared to previous studies on Soay lambs, where the estimated heritability of the antibody traits ranged from 0.21 to 0.39 [36, 43]. The differences among our estimated heritabilities and the previous studies could be due to the different methods, SNP/haplotype information, and models. For IgA and IgG, higher heritability estimates were achieved when a combination of haplotype information with non-clustered SNPs was used, and for IgE, haplotypes achieved an equal heritability to fitting individual SNPs (Supplementary Table 1). In all combinations of SNPs and non-clustered SNPs, the total number of variants were more than individual SNP or haplotypic pseudo-SNPs with different LD thresholds (Tables 3). Therefore, more variants were available to explain the phenotypic variances of IgA, IgE, and IgG. In parallel to our results, Won et al. [44] obtained higher heritability estimates from haplotypes than individual SNP for carcass traits in pigs. Estimated heritabilities among haplotype-based GPs tended to decrease as the LD threshold increased, the length of haplotypes shortened, and the number of haplotypes declined. With higher LD thresholds, a smaller number of SNP markers and shorter genomic length contributed to the haplotype block construction (Table 2). Therefore, fewer haplotypic pseudo-SNPs were available to explain the phenotypic variance, and a lower proportion of total variance could be captured, resulting in lower heritability estimates.

#### Genomic prediction accuracy

Gains in the accuracy of IgG's GPs were observed using haplotype-based pseudo-SNPs. Our results are in concordance with the previous studies, revealing that significant improvements in haplotype-based GPs could be gained when oligogenic traits or those affected by major genes are evaluated [22, 44, 45]. For instance, Won et al. [44] reported an increase of 4.6% in GEBV accuracy with haplotypic GP for eye muscle area in Korean cattle. Moreover, a 9.8% improvement in the accuracy of carcass weight GEBV was documented by incorporating haplotype information based on SNP markers from functionally related genomic regions [45]. Additionally, up to 22% gain in accuracy was observed using haplotypes from fixed length or LD blocking strategies in genomic evaluation of milk production traits in French dairy goats [22].

One explanation for the higher performance of haplotype-based GPs, particularly for IgG, could be that $G_{SNP}$ is constructed based on marker alleles being IBS. As SNP chip markers typically represent old mutations, $G_{SNP}$ mainly traces old relationships among distant relatives and may not precisely account for changes due to recent selection [19]. Meanwhile, haplotype blocks can provide more information on recent mutations and better show close relationships [25]. Another explanation is that haplotype blocks are multi-allelic; consequently, they can better capture the LD with multi-allelic QTLs than biallelic individual SNP [19]. Moreover, haplotype blocks are derived from common ancestors; thus, GRMs based on haplotypes can differentiate between IBD and IBS, while $G_{SNP}$ lacks this ability [46]. Another advantage of haplotype information is that haplotypes include the local epistatic effects among QTLs located within the haplotype blocks [25]. We observed lower performances in GPs of IgA when haplotype-based pseudo-SNPs based on high LD thresholds were applied to analyses compared to individual SNP. This might be due to the reason that these haplotypes are not sufficient to capture the effects of all the important chromosomal regions controlling the trait.

In each applied method, gains in accuracy were observed for IgA when methods were applied based on the combination of pseudo-SNPs and non-clustered SNPs. Our results are in disagreement with the previous study in which the combinations of SNPs and non-clustered pseudo-SNPs were used for GP for the first time and showed no improvement in accuracy [23]. In their study, highly polygenic traits were simulated, and GP was performed based on the Single-step GBLUP method, while we used traits with different genetic architectures, and we conducted GP based on GBLUP and Bayesian methods. The difference in genetic architecture could result in capturing of a higher LD between the pseudo-SNPs and multiallelic QTLs when haplotypic pseudo-SNPs are added to individual SNP [47].

In all studied traits, Bayesian methods outperformed GBLUP. The genetic correlation between the studied traits was reported to be more than 69.5%, with the highest correlation of 82.4% between IgA and IgG [36]. Moreover, around 44% and 10% of additive genetic variance of IgA and IgG traits in Soay sheep were explained by three and two QTLs, with one overlapping genomic region on chromosome 20 [36]. The high genetic correlation between the studied traits, the overlapping QTLs between them, and the outperformance of Bayesian approaches for all the traits in the current study suggest that these traits are more likely to be less polygenic, with some similarities in their genetic architectures. In parallel to our results, several studies have shown that

Bayesian approaches could yield higher GP accuracies than GBLUP for oligogenic traits [13, 48, 49].

Generally, the heritability of traits is positively associated with GP accuracy [50]. While we observed this trend in IgA and IgE, with higher LD thresholds, lower heritabilities and higher GP accuracies were obtained for IgG up to LD = 0.70. This could be due to the overlaps between the IgG major genes and the haplotypes constructed based on high LD thresholds of 0.60 and 0.70. We could not find any overlapped regions between the IgG QTLs identified by Sparks et al. [36] and the genomic positions of haplotypes in $A_{HAP0.70}$; however, there might still be some unknown QTLs for this trait. Moreover, the obtained higher accuracies might be due to the better capture of LD between multiallelic QTLs, which could be missed in biallelic studies. For instance, the multiallelic polymorphism of the major histocompatibility complex region, which significantly contributes to antigen recognition and antibody production such as IgA, IgE, and IgG, has been well-documented in the Soay sheep population [51–53].

We achieved remarkably lower accuracies for IgG compared with IgA and IgE. While higher GP accuracies are more desirable, there are some explanations for why a trait may not evolve as expected, including: (a) there might be a genetic correlation between the trait of interest and fitness-related traits [54, 55], (b) considering the breeder's equation ($R = h^2S$, where $R$ is the response to selection, $h^2$ is the narrow-sense heritability, and $S$ is the strength of selection; [56]), the fluctuations in environmental conditions covarying with the heritability of the trait [57], the strength of selection [58], or both [59], can affect the response to selection, (c) in particular situations, the trait has responded to the selection, but a change in environmental conditions caused the phenotypic trend to mask the underlying genetic trend, which is referred to as "cryptic microevolution" [35].

Considering the higher accuracy achieved in this study with haplotype-based GPs for IgG and with the combination of pseudo-SNPs and non-clustered SNPs for IgA, there is an opportunity to apply these models in evolutionary and conservation genetics to improve captive breeding strategies. Wild animals could be genotyped, and haplotype-based GP models could be used to select the best individuals for the traits of interest. Considering the rise in wildlife infectious diseases and the emergence of zoonotic infections in wild animal populations, haplotype-based GP models could be used to improve captive breeding and conservation strategies to select pathogen-resistant individuals. Meanwhile, GP models have already been successfully applied for better immune responses to pathogens in livestock species breeding programs, such as tuberculosis resistance in dairy cattle [60], resistance against *Piscirickettsia salmonis* in Atlantic salmon [61], and higher immune response for porcine reproductive and respiratory syndrome in pigs [62].

#### Genomic prediction bias

The magnitude of the bias was lower for IgA among the three studied traits. The reason could be that bias in genomic evaluations was generally lower for the traits with higher heritability [63]. Less biased GPs were observed for some haplotypic scenarios in all traits (Supplementary Table 4). GP models using haplotypic pseudo-SNPs, which gained higher accuracy compared to those fitting individual SNP, predicted less-biased GEBVs for IgG. In contrast, the higher accuracy achieved by some combinations of pseuso-SNPs and non-clustered SNPs came with the cost of more-biased GEBVs for IgA. Karimi et al. [21] also reported a less biased GP based on haplotypic pseudo-SNPs for traits with moderate-to-high heritabilities in Holstein cattle. However, Feitosa et al. [24] observed a more biased prediction for beef fatty acid profile using the haplotype model compared with the SNP model. An explanation for the less biased genomic evaluation based on haplotypes in some haplotype-based scenarios could be that haploblocks account for local epistasis, i.e., the interaction between SNPs within haplotype block, which can reduce the bias of GPs [45, 64].

#### Future studies

Several opportunities exist for additional assessments of haplotype-based GPs in wild populations. We investigated the application of haplotype-based GP on three traits with similar genetic architectures. Therefore, we suggest evaluating the performance of haplotype approaches on other traits with different genetic architectures. Also, the benefits of the haplotype-based methods need to be investigated with larger populations. We used the forward validation method to estimate GPs accuracy, as also used in other sheep GP studies [65]. With larger sample sizes, the accuracy of GPs could be evaluated based on alternative validation methods, such as random and k-means cross-validation. Furthermore, several other methods can be used for fitting haplotypes in GP analyses (references) and future studies could compare alternative methods. A well-known approach is using haplotype information in a multi-allelic mixed model treating each haplotype block as a locus and each haplotype within the haplotype block as an allele [66]. We assessed the efficacy of GP using haplotypes constructed based on different LD thresholds. Haplotypic pseudo-SNP can also be produced based on a fixed number of SNPs per haplotype block or a fixed block length.

## Conclusions

Haplotypic information could improve the accuracy of genomic evaluations for antibody production of IgG and IgA traits. The gains in accuracy were more remarkable for IgG in most scenarios applied pseudo-SNPs. The improvement in accuracy was more significant for IgA using some combinations of pseudo-SNPs with individual SNP, particularly when lower LD thresholds applied. However, the slightly higher accuracy in IgA comes with the cost of more bias compared to the SNPs. In all studied traits, Bayesian approaches outperformed GBLUP. Although genomic evaluations based on haplotypes require additional steps, achieved improvements in GEBVs accuracy for some traits could be advantageous. We anticipate that this method could be applied to evolutionary and conservation quantitative genetics research to improve captive breeding and conservation strategies and better understand unmanaged populations' microevolution.

## Methods

### Study population

The St. Kilda archipelago (54o49′08o34'W) is located at 65 km west of the Outer Hebrides, Scotland, and consists of four islands: Hirta, Soay, Boreray, and Dun. Soay sheep are descendants of primitive European domestic sheep introduced to the island of Soay several millennia ago [67]. A population of unmanaged Soay sheep has inhabited the island of Hirta since 1932 [67]. The Hirta Soay sheep population is well-characterized by periods of growth followed by considerable declines due to cold winters, feed availability, and parasitic infections, leading to reduced body weight and increased mortality rates [68–70]. A longitudinal individual-based study on the Soay sheep population in the Village Bay area of Hirta began in 1985 [67]. Since that time, > 95% of the lambs have been captured during the lambing season in March–May to collect a variety of measures, including immunoglobulin (Ig) levels against *T. circumcincta* third larval stage [36, 71].

### Data preparation

A total of 2,061 IgA, IgE, and IgG records from 2,061 Soay sheep lambs against antigens of *T. circumcincta* were obtained from a publicly available dataset belonging to the study conducted by Sparks et al. [36]. In brief, antibody levels were measured as optical density (OD) values using direct ELISA tests on blood samples collected between 1990 and 2015. The procedure of capturing animals, sample collection, and ELISA methods were previously described in detail by Sparks et al. [36]. Samples belonging to the lambs within 50 days of birth were already removed from the dataset due to the potential presence of maternal antibodies [36]. Only animals with genotypic data were included in this study. Therefore, a total of 2,034 IgA, IgE, and IgG records belonging to 2,034 Soay lambs with an average ± SE, minimum and maximum age of 115.19 ± 0.17, 77, and 146 days remained for subsequent analyses.

A fixed-effects model was used to obtain the adjusted phenotypes for subsequent analyses. IgA, IgE, and IgG ODs were corrected for the systematic effects of animal age (in days), birth year, and sex. Two other "Plate ID" and "Run Date" effects were also present in the downloaded dataset. However, less than 5% of phenotypic variances of IgA, IgE, and IgG in lambs were explained by plate ID, and no variance was explained by Run Date in the Sparks et al. [36] study. Therefore, we did not use them in our analyses since they did not have remarkable effects. No information on other potentially significant effects was available in the dataset obtained. The fixed effect model was fitted using the AIREMLF90 package [72]. The residual effects were obtained and used as pseudo-phenotypes for the subsequent analyses.

Samples were already genotyped using the Illumina Ovine SNP50 BeadChip (Illumina; San Diego, CA, USA) by Sparks et al. [36]. While quality control was already conducted on the dataset by Sparks et al. [36], we reperformed quality control of 39,176 SNPs using PLINK 1.9 [73] on the lamb population ($N = 2{,}034$) with our criteria to ensure the quality of the data. Remaining markers with minor allele frequency (MAF) < 0.01, SNP calling rate < 0.90, extreme departure from Hardy–Weinberg equilibrium ($p$-value < $10^{-6}$), and SNPs located on non-autosomal chromosomes were removed. Moreover, samples with a genotype call rate of less than 90% were discarded from downstream analyses. A total of 37,031 SNPs from 2,034 lamb passed the quality control steps with the average genotype call rate > 99%. Then, the missing genotypes were imputed using the Beagle 5.2 software [74]. Subsequently, SNPs with MAF < 0.01 were filtered out. At the end, 37,029 SNPs from 2,034 lamb, all having the Ig records, remained for further analyses.

Principal component analysis was performed on SNP markers using the PLINK 1.9 software [73] to investigate the presence of family structure in the studied population. Furthermore, GRM was constructed by individual SNP ($G_{SNP}$) using the AGHmatrix 2.0.4 R package [75] based on the VanRaden method [11]. Then, the distribution of the diagonal elements of the GRM was evaluated using bar and Q-Q plots to investigate the presence of significant family structures. All plots were created using the ggplot2 R package [76].

### Haploblock construction

The Big-LD method [42], which has been linked with higher accuracy in estimating the recombination hotspots than other existing methods, was used to construct the haplotype blocks. This method is based on interval graph modeling of LD bins which are clusters of strong pairwise LD SNPs, not necessarily physically consecutive [42]. As described by previous studies [23, 77], the gpart 1.13.0 package [78] in R [79] was used to implement the Big-LD method for haploblocks construction, using the default settings; however, the MAFcut was set to zero since the data was already passed this quality control test. Moreover, the CLQcut was set based on the common pairwise LD measure of $r^2$, and ten LD thresholds were considered, including 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00. These LD thresholds were applied to capture different block structures from the biggest blocks with more SNPs in low LD ($r^2 = 0.15$) to the smallest blocks with a lower number of SNPs in high LD ($r^2 = 1.00$). Finally, the haplotype alleles were transformed to pseudo-SNPs, as described by Teissier et al. [22], using the GHap 2.0.0 R package [80]. Notably, many haploblocks can be multi-allelic, and several pseudo-SNPs can be created from the multiallelic haploblocks. The pseudo-SNPs were subjected to the same quality control criteria as the SNPs before their use for GP.

### Training and validation sets

We used the forward validation approach to evaluate the performance of the applied GPs models. The holdout approach has two main advantages over the common cross-validation approach, including: (i) in terms of breeding, holdout validation is generally preferred over cross-validation, as it provides a more realistic estimate of the accuracy of the model on new data. This is very important for breeding purposes, where the goal is to predict the genomic merit of future offspring based on the genotypes of their parents or individuals from previous generations. In contrast, in the cross-validation approach, animals from different generations can be denoted to the validation set, which is not realistic [65, 81, 82]; (ii) forward validation is computationally less intensive, as it requires training the model only once [13]. Forward validation approach has been widely used for evaluating GP models in different species, including sheep [83, 84], pigs [85], and cattle [86]. The validation set included lambs born in 2014 and 2015 ($N$=186), comprising 10% of the total population. Lambs born before 2014 ($N$=1,848) were classified as the training set and used to predict the GEBV of animals in the validation set. The size of the training and validation sets were comparable for all traits. Adjusted phenotypes calculated for animals born from 1995 to 2013 were used as pseudo-phenotype for the training, and those calculated for animals born in 2014 and 2015 were applied for validation.

### Genomic prediction of breeding values

Overall, six methods, including GBLUP, BayesA, BayesB, BayesCπ, BayesL, and BayesR were used in GP analyses. In each method, GP was computed based on three different analyses, including:

(i) analyses based on individual SNP ($A_{SNP}$) fitted in the models;

(ii) analyses using haplotypes constructed based on different LD thresholds fitted as pseudo-SNPs in the GP models. Therefore, $A_{HAP0.15}$, $A_{HAP0.20}$, $A_{HAP0.30}$, $A_{HAP0.40}$, $A_{HAP0.50}$, $A_{HAP0.60}$, $A_{HAP0.70}$, $A_{HAP0.80}$, $A_{HAP0.90}$, and $A_{HAP1.00}$ refer to the analyses using haplotypes constructed by LD thresholds of 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00 as pseudo-SNPs, respectively;

(iii) analyses using pseudo-SNPs in combination with non-LD clustered SNPs, located out of haploblocks, fitted in the model. After defining the haplotypic pseudo-SNPs based on the different LD thresholds, we combined them with individual SNP that were not blocked in haplotypes. Thus, $A_{COM0.15}$, $A_{COM0.20}$, $A_{COM0.30}$, $A_{COM0.40}$, $A_{COM0.50}$, $A_{COM0.60}$, $A_{COM0.70}$, $A_{COM0.80}$, $A_{COM0.90}$, and $A_{COM1.00}$ refer to the analyses in which the pseudo-SNPs with different LD thresholds were combined with non-LD clustered SNPs.

**GBLUP method:** The GBLUP model used was as follows:

$$\mathbf{y_c} = 1\mu + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y_c}$ is the vector of adjusted phenotype in the reference population, $\mu$ is the overall mean effect, $\mathbf{g}$ is the vector of additive genetic effects accounted for by all markers, i.e., SNPs in $A_{SNP}$, pseudo-SNPs in $A_{HAP0.15}$-$A_{HAP1.00}$, or a combination of SNPs and pseudo-SNPs in $A_{COM0.15}$-$A_{COM1.00}$, and $\mathbf{e}$ is a vector of random residual. $\mathbf{Z}$ is the incidence matrix relating GEBV to adjusted phenotypes of individual animals. It was assumed that $\mathbf{g} \sim N\left(0, \mathbf{G}\sigma_g^2\right)$ and $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$, where $\mathbf{G}$ is the GRM constructed based on only SNP markers, only haplotypes fitted as pseudo-SNPs, or a combination of pseudo-SNPs with non-clustered SNPs, $\mathbf{I}$ is an identity matrix, $\sigma_g^2$ is the additive genetic variance, and $\sigma_e^2$ is the residual variance. The GRM was constructed as follows [11]:

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{j=1}^{i} p_j(1-p_j)},$$

where, $\mathbf{Z}$ contains genotypes adjusted by the allele frequency and $p_j$ is the MAF of marker j. Variance components were estimated using the Average Information Restricted Maximum Likelihood (AIREML) algorithm. This process and the prediction of GEBVs with GBLUP models were performed using the GVCBLUP software [87].

### Bayesian methods

Five Bayesian GP models were fitted, including BayesA, BayesB, BayesCπ, BayesL, and BayesR. For these methods, the general statistical model was:

$$\mathbf{y_c} = 1\mu + \sum_{j=1}^{K} \mathbf{z_j}\beta_j + \mathbf{e},$$

where, $\mathbf{y_c}$ is the vector of adjusted phenotype in the reference population, μ is the overall mean effect, K is the number of markers fitted, including SNPs in $A_{SNP}$, pseudo-SNPs in $A_{HAP0.15}$- $A_{HAP1.00}$, or a combination of SNPs and pseudo-SNPs in $A_{COM0.15}$-$A_{COM1.00}$, $\mathbf{z_j}$ is a vector denoting the genotypes of the animals for marker j, $\beta_j$ is the effect of marker j, and $\mathbf{e}$ is a vector of random residuals. The vector of residuals $\mathbf{e}$ was assumed to be distributed as $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where $\sigma_e^2$ is the residual variance and $\mathbf{I}$ is an identity matrix. The hypothetical distribution of all markers' effects in each Bayes method and the formula of the effect distribution are shown in Table 4.

In all the Bayesian methods, the marker effects were estimated using a total of 100,000 Markov chain Monte Carlo (MCMC) iterations, with the first 20,000 discarded as burn in, and a thinning interval of 100. All the Bayesian methods were implemented using the hibayes R package [89]. We diagnosed convergence using a criterion of the accuracy of estimation of a quantile using the R package coda [90].

### Comparison of genomic relationship matrices

Overall, 21 GRMs were constructed for GPs, which could be classified into three categories:

(i) $G_{SNP}$, which refers to the GRM defined based on SNPs as markers;

(ii) $G_{HAP0.15}$, $G_{HAP0.20}$, $G_{HAP0.30}$, $G_{HAP0.40}$, $G_{HAP0.50}$, $G_{HAP0.60}$, $G_{HAP0.70}$, $G_{HAP0.80}$, $G_{HAP0.90}$, and $G_{HAP1.00}$ defined based on haplotypes constructed by LD thresholds of 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00 as markers, respectively;

(iii) $G_{COM0.15}$, $G_{COM0.20}$, $G_{COM0.30}$, $G_{COM0.40}$, $G_{COM0.50}$, $G_{COM0.60}$, $G_{COM0.70}$, $G_{COM0.80}$, $G_{COM0.90}$, and $G_{COM1.00}$ constructed using pseudo-SNPs with different LD thresholds combined with non-LD clustered SNPs.

As previously applied by Karimi et al. [21], to investigate the differences between matrices, pairwise Euclidean distance was calculated by $d(\mathbf{C}, \mathbf{D}) = \sqrt{\sum_i \sum_j (c_{ij} - d_{ij})^2}$, where $c_{ij}$ and $d_{ij}$ are elements of two comparing GRMs of $\mathbf{C}$ and $\mathbf{D}$, respectively. Finally, the calculated values were scaled between 0 and 1.

### Heritability estimation

Variance components were estimated using the GVCBLUP software [87] and the hibayes package [89] for GBLUP and Bayesian approaches, respectively. The REML algorithm and MCMC method were applied to variance component estimation in GBLUP and Bayesian methods, respectively. In GBLUP, heritability was computed as $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where $\sigma_g^2$ and $\sigma_e^2$ are the additive genetic and residual variances, respectively. For the Bayesian methods, heritability was estimated by

**Table 4** Assumption of effect size distribution of markers for the Bayesian methods used in this study

| Method | Joint distribution of SNP/ pseudo-SNP effect | SNP/pseudo-SNP effect distribution[a] | Variance distribution of the SNP/pseudo-SNP effects | Reference |
|---|---|---|---|---|
| BayesA | t | $\beta_j \sim N(0, \sigma_g^2)$ | $\sigma_g^2 \sim \chi^{-2}(\nu, S)$ | [1] |
| BayesB | point-t | $\beta_j \sim 0.05\left(0, \sigma_g^2\right) + 0.95\delta_0$ | $\sigma_g^2 \sim \chi^{-2}(\nu, S)$ | [1] |
| BayesCπ | t mixture | $\beta_j \sim (1-\pi)N\left(0, \sigma_g^2\right) + \pi\delta_0$ | $\sigma_g^2 \sim \chi^{-2}(\nu, S)$ | [13] |
| BayesL | double exponential or Laplace | $\beta_j \sim N\left(0, \sigma_g^2\right)$ | $\sigma_g^2 \sim \text{Exp}(\frac{\lambda^2}{2})$ | [16] |
| BayesR | point-normal mixture | $\beta_j \sim$ $\pi_1\delta_0 + \pi_2 N\left(0, 10^{-4}\sigma_g^2\right) + \pi_3 N\left(0, 10^{-3}\sigma_g^2\right) + \pi_4 N\left(0, 10^{-2}\sigma_g^2\right)$ | $\sigma_g^2 \sim \chi^{-2}(\nu, S)$ | [88] |

[a] where $\beta_j$ is the effect of SNP/haplotype j, $\sigma_g^2$ is the additive genetic variance, ν and S are the degree freedom and scale parameter for inverse chi-square distribution, t represents student's t-distribution, $\delta_0$ is the effect size equals to zero, $\lambda^2$ is the rate parameter which is assigned a gamma prior, $\pi = (\pi_1 + \pi_2 + \pi_3 + \pi_4)$ is the mixing proportions such that $\sum_{i=1}^{4} \pi_i = 1$

$h^2 = \frac{V_A}{V_A+\sigma_e^2}$. In this equation, $V_A$ is the total additive genetic variance which was estimated by $V_A = \pi \times 2\widehat{\sigma}^2_{SNP}\sum_{j=1}^{m} p_j q_j$, where $\pi$ is the proportion of the markers with non-zero effect, $\widehat{\sigma}^2_{SNP}$ is the marker variance, and $p_j$ and $q_j$ are allele frequencies of $j^{th}$ SNP or pseudo-SNP.

### Performance of genomic prediction of breeding values

The accuracy of the GEBV was obtained by dividing the Pearson correlation between adjusted phenotypes ($y_c$) and GEBV for the validation subset. Bias was defined as the inflation or deflation of GEBV compared to adjusted phenotypes for the validation subset. The bias of the GEBV was calculated as the deviation from the unity of regression coefficient of adjusted phenotypes on GEBV for the validation subset [i.e., $y_c = b_0 + b_1 GEBV$ where $y_c$ is the adjusted phenotype in the validation set, GEBV corresponds to the predicted direct genomic values in the validation set, and $b_0$ and $b_1$ are the intercept and slope, respectively]. Therefore, the value of 0 indicates no bias in GEBV estimates, while bias $< 0$ and $> 0$ show inflation and deflation, respectively [3, 91].

#### Abbreviations

| | |
|---|---|
| GP | Genomic prediction of breeding values |
| SNP | Single nucleotide polymorphism |
| LD | Linkage disequilibrium |
| REML | Restricted maximum likelihood |
| QTL | Quantitative trait loci |
| GBLUP | Genomic best linear unbiased prediction |
| BayesL | Bayesian Lasso |
| Ig | Immunoglobulin |
| GEBV | Genomic estimated breeding values |
| IBD | Identical-by-descent |
| GRM | Genomic relationship matrix |
| OD | Optical density |
| SE | Standard error |
| PCA | Principal component analysis |
| MAF | Minor allele frequency |
| AIREML | Average information restricted maximum Llremlikelihood |
| MCMC | Markov chain Monte Carlo |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09407-0.

**Additional file 1.**

#### Acknowledgements

The authors are grateful to Sparks et al. [36] for publicly sharing the Soay sheep data for research purposes. We also acknowledge the valuable comments provided by the two reviewers, which greatly contributed to improve the quality of the paper.

#### Authors' contributions

SV, SS, LFB, and MB conceived the study. SV carried out the analyses and drafted the manuscript. SS, LFB, and KP helped in conducting analyses and writing the manuscript. SS, MB, LFB, and KK critically revised and edited the manuscript. All authors contributed to the paper and approved the submitted version. All authors have read and agreed to the published version of the manuscript.

#### Funding

Open access funding provided by Swedish University of Agricultural Sciences. We have received no specific funding for this study.

#### Availability of data and materials

The datasets analyzed during the current study are available in the Dryad database: https://datadryad.org/stash/dataset/doi:10.5061/dryad.zgmsbcc6f and described in Sparks et al. [36]. All generated pseudo-SNPs, GEBVs, and scripts for the analyses are available at: https://osf.io/njm8v/

## Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

#### Author details

[1]Department of Animal Science and Aquaculture, Dalhousie University, Truro, NS B2N5E3, Canada. [2]Department of Animal Science, University of Zanjan, Zanjan 4537138791, Iran. [3]Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA. [4]Molecular Diagnostics Program, Verspeeten Clinical Genome Centre, London Health Sciences Centre, London, ON N6A 5W9, Canada. [5]Department of Animal Sciences, Islamic Azad University, Varamin Varamin-Pishva Branch3381774895, Iran. [6]Department of Animal Breeding and Genetics (HGEN), Centre for Veterinary Medicine and Animal Science (VHC), Swedish University of Agricultural Sciences (SLU), 75007 Uppsala, Sweden. [7]Department of Biotechnology, Animal Science Research Institute of IRAN (ASRI), Agricultural Research, Education & Extension Organization (AREEO), Karaj 3146618361, Iran.

Received: 8 December 2022 Accepted: 24 May 2023
Published online: 17 June 2023

#### References

1. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. Goddard M, Hayes B. Genomic selection. J Anim Breed Genet. 2007;124:323–30.
3. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221–37.
4. Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 2007;47:1082–90.
5. Lorenzana RE, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet. 2009;120:151–61.
6. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9: e1003348.
7. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19:581–90.
8. Gienapp P, Calus MP, Laine VN, Visser ME. Genomic selection on breeding time in a wild bird population. Evol Lett. 2019;3:142–51.
9. Ashraf B, Hunter DC, Bérénos C, Ellis PA, Johnston SE, Pilkington JG, et al. Genomic prediction in the wild: a case study in Soay sheep. Mol Ecol. 2020;31:6541–55.
10. Hunter D, Ashraf B, Bérénos C, Ellis PA, Johnston S, Wilson A, et al. Using genomic prediction to detect microevolutionary change of a quantitative trait. Proc R Soc B. 2022;289:20220330.
11. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.

12. Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743–52.
13. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:1–12.
14. Park T, Casella G. The bayesian lasso. J Am Stat Assoc. 2008;103:681–6.
15. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.
16. Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008;179:1045–55.
17. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021–31.
18. Calus M, Meuwissen T, De Roos A, Veerkamp R. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 2008;178:553–61.
19. Meuwissen TH, 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. 2014;46:1–8.
20. Da Y. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. BMC Genet. 2015;16:1–12.
21. Karimi Z, Sargolzaei M, Robinson JAB, Schenkel FS. Assessing haplotype-based models for genomic evaluation in Holstein cattle. Can J Anim Sci. 2018;98:750–9.
22. Teissier M, Larroque H, Brito LF, Rupp R, Schenkel FS, Robert-Granié C. 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. 2020;103:11559–73.
23. Araujo AC, Carneiro PL, Oliveira HR, Schenkel FS, Veroneze R, Lourenco DA, et al. A comprehensive comparison of haplotype-based single-step genomic predictions in livestock populations with different genetic diversity levels: a simulation study. Front Genet. 2021;12:1843.
24. Feitosa FLB, Pereira ASC, Amorim ST, Peripolli E, Silva RM de O, Braz CU, et al. Comparison between haplotype-based and individual snp-based genomic predictions for beef fatty acid profile in Nelore cattle. J Anim Breed Genet. 2020;137:468–76.
25. Hickey J, Kinghorn B, Tier B, Clark SA, van der Werf J, Gorjanc G. Genomic evaluations using similarity between haplotypes. J Anim Breed Genet. 2013;130:259–69.
26. Hess M, Druet T, Hess A, Garrick D. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol. 2017;49:1–14.
27. Cuyabano BC, Su G, Lund MS. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics. 2014;15:1–11.
28. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296:2225–9.
29. Hewson I, Button JB, Gudenkauf BM, Miner B, Newton AL, Gaydos JK, et al. Densovirus associated with sea-star wasting disease and mass mortality. Proc Natl Acad Sci. 2014;111:17278–83.
30. Blehert DS, Hicks AC, Behr M, Meteyer CU, Berlowski-Zier BM, Buckles EL, et al. Bat white-nose syndrome: an emerging fungal pathogen? Science. 2009;323:227–227.
31. Garner TW, Schmidt BR, Martel A, Pasmans F, Muths E, Cunningham AA, et al. Mitigating amphibian chytridiomycoses in nature. Philos Trans R Soc B Biol Sci. 2016;371:20160207.
32. Karikalan M, Chander V, Mahajan S, Deol P, Agrawal R, Nandi S, et al. Natural infection of Delta mutant of SARS-CoV-2 in Asiatic lions of India. Transbound Emerg Dis. 2021;69:3047–55.
33. Grome HN, Meyer B, Read E, Buchanan M, Cushing A, Sawatzki K, et al. SARS-CoV-2 Outbreak among Malayan Tigers and Humans, Tennessee, USA, 2020. Emerg Infect Dis. 2022;28:833.
34. Hale VL, Dennis PM, McBride DS, Nolting JM, Madden C, Huey D, et al. SARS-CoV-2 infection in free-ranging white-tailed deer. Nature. 2022;602:481–6.
35. Merilä J, Kruuk L, Sheldon B. Cryptic evolution in a wild bird population. Nature. 2001;412:76–9.
36. Sparks AM, Watt K, Sinclair R, Pilkington JG, Pemberton JM, McNeilly TN, et al. The genetic architecture of helminth-specific immune responses in a wild population of Soay sheep (Ovis aries). PLoS Genet. 2019;15:e1008461.
37. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:1–11.
38. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics. 2019;211:1131–41.
39. Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Cerón-Rojas J, et al. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014;112:48–60.
40. Bian C, Prakapenka D, Tan C, Yang R, Zhu D, Guo X, et al. Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. Genet Sel Evol. 2021;53:1–19.
41. Araujo AC, Carneiro PL, Oliveira HR, Lewis RM, Brito LF. SNP-and haplotype-based single-step genomic predictions for body weight, wool, and reproductive traits in North American Rambouillet sheep. J Anim Breed Genet. 2022;140:216–34.
42. Kim SA, Cho C-S, Kim S-R, Bull SB, Yoo YJ. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. Bioinformatics. 2018;34:388–97.
43. Hayward AD, Garnier R, Watt KA, Pilkington JG, Grenfell BT, Matthews JB, et al. Heritable, heterogeneous, and costly resistance of sheep against nematodes and potential feedbacks to epidemiological dynamics. Am Nat. 2014;184:S58-76.
44. Won S, Park J-E, Son J-H, Lee S-H, Park BH, Park M, et al. Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. Front Genet. 2020;11:134.
45. Xu L, Gao N, Wang Z, Xu L, Liu Y, Chen Y, et al. Incorporating genome annotation into genomic prediction for carcass traits in Chinese Simmental beef cattle. Front Genet. 2020;11:481.
46. Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet. 1999;65:1493–500.
47. Liang Z, Tan C, Prakapenka D, Ma L, Da Y. Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. Front Genet. 2020;11: 588907.
48. de Los CG, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193:327–45.
49. Pryce J, Arias J, Bowman P, Davis S, Macdonald K, Waghorn G, et al. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J Dairy Sci. 2012;95:2108–19.
50. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91.
51. McDevitt HO. Discovering the role of the major histocompatibility complex in the immune response. Annu Rev Immunol. 2000;18:1.
52. Dicks KL, Pemberton JM, Ballingall KT. Characterisation of major histocompatibility complex class IIa haplotypes in an island sheep population. Immunogenetics. 2019;71:383–93.
53. Dicks KL. Unravelling major histocompatibility complex diversity in the Soay sheep of St Kilda. 2018.
54. Lehmann L, Keller L. Synergy, partner choice and frequency dependence: their integration into inclusive fitness theory and their interpretation in terms of direct and indirect fitness effects. J Evol Biol. 2006;19:1426–36.
55. Berenbaum M. Coumarins and caterpillars: a case for coevolution. Evolution. 1983;37:163–79.
56. Falconer D. Introduction to quantitative genetics. 3rd ed. New York: Longman; 1989.
57. Hoffmann AA, Merilä J. Heritable variation and evolution under favourable and unfavourable conditions. Trends Ecol Evol. 1999;14:96–101.
58. Grant BR, Grant PR. Evolution of Darwin's finches caused by a rare climatic event. Proc R Soc Lond B Biol Sci. 1993;251:111–7.

59. Wilson AJ, Pemberton JM, Pilkington J, Coltman DW, Mifsud D, Clutton-Brock TH, et al. Environmental coupling of selection and heritability limits evolution. PLoS Biol. 2006;4: e216.
60. Tsairidou S, Woolliams JA, Allen AR, Skuce RA, McBride SH, Wright DM, et al. Genomic prediction for tuberculosis resistance in dairy cattle. PLoS ONE. 2014;9: e96728.
61. Bangera R, Correa K, Lhorente JP, Figueroa R, Yáñez JM. Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar). BMC Genomics. 2017;18:1–12.
62. Boddicker NJ, Bjorkquist A, Rowland RR, Lunney JK, Reecy JM, Dekkers J. Genome-wide association and genomic prediction for host response to porcine reproductive and respiratory syndrome virus infection. Genet Sel Evol. 2014;46:1–14.
63. Meher PK, Rustgi S, Kumar A. Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. Heredity. 2022;128:1–12.
64. Jónás D, Ducrocq V, Fouilloux M-N, Croiseau P. Alternative haplotype construction methods for genomic evaluation. J Dairy Sci. 2016;99:4537–46.
65. Brito LF, Clarke SM, McEwan JC, Miller SP, Pickering NK, Bain WE, et al. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. BMC Genet. 2017;18:1–17.
66. Prakapenka D, Wang C, Liang Z, Bian C, Tan C, Da Y. GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. Front Genet. 2020;11:282.
67. Clutton-Brock TH, Pemberton JM. Soay sheep: dynamics and selection in an island population. 1st ed. Cambridge University Press; 2004.
68. Gulland F. The role of nematode parasites in Soay sheep (Ovis aries L.) mortality during a population crash. Parasitology. 1992;105:493–503.
69. Coulson T, Catchpole EA, Albon SD, Morgan BJ, Pemberton J, Clutton-Brock TH, et al. Age, sex, density, winter weather, and population crashes in Soay sheep. Science. 2001;292:1528–31.
70. Craig B, Pilkington J, Pemberton J. Gastrointestinal nematode species burdens and host mortality in a feral sheep population. Parasitology. 2006;133:485–96.
71. Leivesley JA, Bussière LF, Pemberton JM, Pilkington JG, Wilson K, Hayward AD. Survival costs of reproduction are mediated by parasite infection in wild Soay sheep. Ecol Lett. 2019;22:1203–13.
72. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D. BLUPF90 and related programs (BGF90). In: The 7th World Congress Genetics Application Livestock Production. Montpellier; p. 28.
73. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
74. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48.
75. Amadeu RR, Cellon C, Olmstead JW, Garcia AA, Resende Jr MF, Muñoz PR. AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome. 2016;9:1–10.
76. Wickham H, Chang W, Wickham MH. Package 'ggplot2.' Create Elegant Data Vis Using Gramm Graph Version. 2016;2:1–189.
77. Araujo AC, Carneiro PL, Alvarenga AB, Oliveira HR, Miller SP, Retallick K, et al. Haplotype-based Single-step GWAS for Yearling Temperament in American Angus Cattle. Genes. 2021;13:17.
78. Kim SA, Brossard M, Roshandel D, Paterson AD, Bull SB, Yoo YJ. gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. Bioinformatics. 2019;35:4419–21.
79. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020.
80. Utsunomiya YT, Milanesi M, Utsunomiya AT, Ajmone-Marsan P, Garcia JF. GHap: an R package for genome-wide haplotyping. Bioinformatics. 2016;32:2861–2.
81. Hayes BJ, Goddard M. Prediction of breeding values using marker-derived relationship matrices. J Anim Sci. 2008;86:2089–92.
82. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92:433–43.
83. Oget C, Teissier M, Astruc J-M, Tosser-Klopp G, Rupp R. Alternative methods improve the accuracy of genomic prediction using information from a causal point mutation in a dairy sheep model. BMC Genomics. 2019;20:1–14.
84. Pickering NK, Auvray B, Dodds KG, McEwan JC. Genomic prediction and genome-wide association study for dagginess and host internal parasite resistance in New Zealand sheep. BMC Genomics. 2015;16:1–11.
85. Salek Ardestani S, Jafarikia M, Sargolzaei M, Sullivan B, Miar Y. Genomic Prediction of Average Daily Gain, Back-Fat Thickness, and Loin Muscle Depth Using Different Genomic Tools in Canadian Swine Populations. Front Genet. 2021;12:735.
86. Guarini A, Lourenco D, Brito L, Sargolzaei M, Baes CF, Miglior F, et al. Comparison of genomic predictions for lowly heritable traits using multi-step and single-step genomic best linear unbiased predictor in Holstein cattle. J Dairy Sci. 2018;101:8076–86.
87. Wang C, Prakapenka D, Wang S, Pulugurta S, Runesha HB, Da Y. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. BMC Bioinformatics. 2014;15:1–9.
88. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11:e1004969.
89. Yin L, Zhang H, Li X, Zhao S, Liu X. hibayes: an R package to fit individual-level, summary-level and single-step Bayesian regression models for genomic prediction and genome-wide association studies. bioRxiv. 2022;1–37.
90. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. 2006;6:7–11.
91. Rolf MM, Garrick DJ, Fountain T, Ramey HR, Weaber RL, Decker JE, et al. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. Genet Sel Evol. 2015;47:1–14.


## Publisher's Note