Animal 17 (2023) 100766

Contents lists available at ScienceDirect

Animal
The international journal of animal biosciences

Strategies for choosing core animals in the algorithm for proven and young and their impact on the accuracy of single-step genomic predictions in cattle

A. Cesarani [a,b,*], M. Bermann [b], C. Dimauro [a], L. Degano [c], D. Vicario [c], D. Lourenco [b], N.P.P. Macciotta [a]

[a] Dipartimento di Agraria, Università di Sassari, 07100 Sassari, Italy
[b] Department of Animal and Dairy Science, University of Georgia, 30602 Athens, GA, USA
[c] Associazione Nazionale Allevatori Pezzata Rossa Italiana (ANAPRI), 33100 Udine, Italy

ARTICLE INFO

Article history:
Received 24 October 2022
Revised 24 February 2023
Accepted 28 February 2023
Available online 15 March 2023

Keywords:
Genomic selection
Key individuals
Prediction accuracy
Principal component analysis
Relationship matrix

ABSTRACT

Nowadays, in some populations, the number of genotyped animals is too large to obtain the inverse of the genomic relationship matrix. The algorithm for proven and young animals (**APY**) can be used to overcome this problem. In the present work, different strategies for defining core animals in APY were tested using either simulated or real data. In particular, core definitions based on random choice or on the contribution to the genomic relationship matrix ($\mathbf{G_{CONTR}}$) calculated using Principal Component Analysis were tested. Core sizes able to explain 90, 95, 98, and 99% of the total variance of the genomic relationship matrix (**G**) were used. Analyzed phenotypes were three simulated traits for 3 000 individuals, and milkability records for 136 406 Italian Simmental cows. The number of genotypes was 4 100 for the simulated dataset, and 11 636 for the Simmental data, respectively. The G_{CONTR} values in Simmental dataset were moderately correlated with the analyzed phenotype, and they showed a decreasing trend according to the year of birth of genotyped animals. The accuracy increased as the size of the core increased in both datasets. The inclusion in the core of animals with largest G_{CONTR} values led to the lowest accuracies (0.50 and 0.71 for the simulated and Simmental datasets, respectively; average across traits and core sizes). On the contrary, the selection of animals with the lowest rank according to their contribution to the **G** provided slightly higher accuracies, especially in the simulated dataset (0.68 for the simulated dataset, and 0.76 for the Simmental data; average across traits and core sizes). In real data, particularly for larger sizes of core animals, the criteria of choice appear less important, confirming the results of earlier studies. Anyway, the inclusion in the core of animals with the lowest values of G_{CONTR} led to increases in accuracy. These are preliminary results based on a small sample size that need to be confirmed on a larger number of genotypes.

Implications

The number of genotypes for some species and breeds is now too large to obtain the inverse of the genomic relationship matrix. The algorithm of proven and young animals, which identifies core and non-core animals, has been proposed to solve this problem. This study aimed to test different criteria to select core animals. The accuracy increased as the size of the core increased, and the accuracies decreased with the inclusion of top animals in the core.

Introduction

The single-step genomic BLUP (**ssGBLUP**) has enabled genomic selection use in livestock breeds where only a portion of the animals in the pedigree is genotyped. The blending of the genomic relationship matrix (**G**) with the pedigree matrix into the **H** matrix (Aguilar et al., 2010) has been the evolution of the early two-step approach of genomic selection (Meuwissen et al., 2001), providing an easier tool to be implemented in routine genetic evaluations. However, the standard ssGBLUP requires the inversion of **G**, which is computationally feasible for up to 150 000 genotyped individuals (Bradford et al., 2017). After more than ten years of genomic selection, the number of genotyped animals in some populations has exceeded one million (https://queries.uscdcb.com/Genotype/counts.html) thus reiterating the "curse of dimensionality"

* Corresponding author at: Dipartimento di Agraria, Università di Sassari, 07100 Sassari, Italy.
E-mail address: acesarani@uniss.it (A. Cesarani).

https://doi.org/10.1016/j.animal.2023.100766
1751-7311/

issue of genomics. The Algorithm for Proven and Young (**APY**) has been developed to solve this problem (Misztal et al., 2014). In APY, genotyped animals are partitioned into two groups, core and non-core. Only the portion of **G** corresponding to core animals is directly inverted in solving mixed model equations. The application of APY allowed for using large datasets in genomic selection, up to 4 M of genotyped individuals in the ssGBLUP model (Cesarani et al., 2022).

The optimum size of APY core animals and the criteria for their definition are still debated (Abdollahi-Arpanahi et al., 2022). Pocrnic et al. (2016), working on simulated data, found that APY could provide the same accuracy for genomic estimated breeding values (**GEBV**) of the standard ssGBLUP with the full **G**, when the core size is equal to the number of independent chromosome segments segregating in the population. The number of independent chromosome segments is defined as four times the effective population size times the genome length in Morgans (Stam, 1980), and they can be approximated by the number of eigenvalues of **G** able to explain the 98% of the variance (Pocrnic et al., 2016). If core size is lower than the number of independent chromosome segments, then the core definition becomes important. Different strategies to select core animals have been tested (Ostersen et al., 2016; Bradford et al., 2017). The random choice seems to be the best option (Fragomeni et al., 2015; Bradford et al., 2017), and it is currently used in APY ssGBLUP evaluations. Abdollahi-Arpanahi et al. (2022) found the best prediction accuracy when core animals are well distributed across generations. Recently, Pocrnic et al. (2022) proposed an iterative algorithm for optimizing core composition based on their distance in a covariance sense.

A possible option for developing an analytic method for core definition should further exploit results of the **G** eigen decomposition, in order to identify key components of its structure (McVean, 2009). The coefficients of the eigenvectors of **G** have been used to calculate the genetic contributions of animals to genetic population covariance for identifying key individuals for sequencing and imputation purposes (Neuditschko et al., 2017). In the present work, the effectiveness of this approach for choosing core animals in APY ssGBLUP has been tested both on real and on simulated data. Moreover, the values computed using this approach were investigated to decipher their meaning.

Material and methods

Data

The first analyzed dataset (Table 1) was simulated for the XVII QTL MAS workshop (Usai et al., 2014). It consisted of five non-overlapping generations with a total population size of 4 100 individuals (100 males). Five chromosomes, each one with 4 000 equally spaced SNPs spread into 100 Mb and with a total of 50 quantitative trait loci, were simulated. Three phenotypes – mimicking milk yield, fat yield, and fat content – were also available for females with heritabilities of 0.35, 0.35, and 0.50, respectively. The phenotypes were not simulated for the females of the last generation. True breeding values were available for all animals.

The real dairy dataset (Table 1) included 136 406 milkability records of Italian Simmental cattle (one record per cow). This trait is scored at farm level using a scale from 1 to 3 (where 1 means slow milking speed, whereas 3 means fast milking speed). The heritability of the trait was 0.12. The pedigree file contained a total of 1 182 789 animals, and it was tracked back for three generations from animals with phenotypes or genotypes (Lourenco et al., 2014; Cesarani et al., 2021a). A total of 11 636 individuals genotyped at 42 141 SNPs were retained for the analyses. This is an updated dataset of the one used in Cesarani et al. (2021b), where other details about the investigated data are available.

Table 1
Data structure for the simulated and cattle datasets.

Dataset	Phenotypes	Genotypes	Animals in the analysis
Simulated	3 000	4 100	4 100
Simmental	136 406	11 636	279 032

Mixed model analysis

The breeding values for the simulated dataset were estimated using the following animal model:

$$\mathbf{y} = 1\mu + \mathbf{Za} + \mathbf{e} \quad (1)$$

where **y** is the vector of the phenotype (traits 1, 2, or 3), **1** is a vector of ones and μ is the general mean, **a** is the vector of animal random additive genetic effect, and **e** is the vector of random residuals.

Breeding values for milkability records of Simmental cattle were estimated with the following single-trait animal model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \quad (2)$$

where **y** is the vector of milkability records; **b** is the vector of fixed effects of herd x year subclass (29 623 levels), age at parity (63 levels), calving season (six levels), and days in milk (linear and quadratic covariates) fixed effects; **a** is the vector of random additive genetic effects, and **e** is the vector of random residuals. The **X** and **Z** are the incidence matrices relating milkability records to effects in **b** and **a**, respectively. The vector **e** was distributed as $N(0, \mathbf{I}\sigma_e^2)$ where **I** is an identity matrix and σ_e^2 is the variance associated with the residual error.

The ssGBLUP was used for the analyses, where **a** was distributed as $N(0, \mathbf{H}\sigma_a^2)$, with **H** the realized relationship matrix, σ_a^2 the additive genetic variance, and **Vg** is genetic (co)variance matrix for an individual. The inverse of **H** was obtained using (i) the direct $\mathbf{G}^{-1}$ or (ii) $\mathbf{G}^{-1}$ from APY ($\mathbf{G}_{APY}^{-1}$), with different core definitions. In all cases, GEBV were estimated using iteration on data with the software BLUP90IOD2OMP1 (ver. 3.122; Tsuruta et al., 2001; Tsuruta and Misztal, 2008).

Core definition

Different strategies were used for the core definition. First, four sizes of core were fixed at the number of eigenvalues able to explain 90 (V90), 95 (V95), 98 (V98) and 99% (V99) of the total variance of the G matrix, respectively. Within each size, animals were chosen according to two criteria: (1) RANDOM; (2) TOP or BOTTOM animals ranked according to their contribution to the **G** (G_{CONTR}). The latter was derived by using the principal component analysis applied to the genomic relationship matrix. In particular, G_{CONTR} was calculated as:

$$G_{CONTR} = \sum_{i=1}^{i=k} \left(\frac{a_{ij}}{\sqrt{\lambda_i}} \right)^2$$

where *a* is the coefficient of the i-th eigenvector for the j-th animal; λ is the i-th eigenvalue; *k* is the number of retained non-null eigenvectors (i.e., equal to the number of animals-1).

In order to better understand the G_{CONTR} values, their correlation with the involved traits (in terms of true breeding values for the simulated dataset or GEBV for the Simmental dataset) was computed. Moreover, since G_{CONTR} reflects the **G** matrix, the correlations with pedigree parameters were also estimated. In particular, we computed the number of offspring in the pedigree and the pedigree-based inbreeding for each genotyped animal. Moreover, the G_{CONTR} trend according to the year of birth was computed

for Simmental animals (the only dataset for which the years of birth were available).

Model validation

Two different validation strategies were tested: i) use of true breeding values for all animals in the last available generation (simulated data); ii) validation of candidate genotyped cows (Simmental data).

In particular, the accuracy of model used on the simulated dataset was evaluated by calculating the correlation between the GEBV estimated for the 1 000 animals in the last generation (without own phenotypes) and their true breeding values.

For Simmental data, two analyses were carried out: (i) complete dataset, with phenotypes of cows born from 1985 to 2019; (ii) reduced dataset, with phenotypes of cows born from 1985 to 2016. The focus group of animals consisted of the youngest genotyped cows (n = 321, born from 2017 to 2019) with no available phenotypes in the reduced dataset. The accuracy for Simmental data was based on the correlation between GEBV estimated for the focus group in the complete (with available phenotypes) and reduced (without phenotypes) datasets. This parameter indicates the stability of the model, i.e., how much the consecutive evaluations are consistent with the addition of new data. For all correlations used to evaluate the accuracy of the different scenarios, the 95% confidence intervals were computed using Fisher's z-transformation (Fisher, 1921).

Moreover, GEBV from all core definitions were compared against the model that used the direct $\mathbf{G}^{-1}$ in order to quantify the impact of the APY algorithm.

Results

Investigation of the contribution to the genomic relationship matrix

The number of eigenvalues (and thus of animals) able to explain a fixed amount of variance (from 90 to 99%) of **G** in the two datasets are reported in Table 2. The simulated dataset showed the lowest number of eigenvalues for all considered thresholds. These numbers were then used as core size in the different four considered scenarios.

Table 3 shows the correlations between the G_{CONTR} values and the true or estimated breeding values in the two investigated datasets. As far as the simulated data, correlations were very low, with the one computed for fat percentage not significantly different from zero. The G_{CONTR} values were almost independent from the number of offspring in Simmental animals (R = −0.07), whereas a moderate negative correlation was estimated between G_{CONTR} and the number of offspring in the simulated dataset (Table 3). Moreover, G_{CONTR} showed a moderate and negative correlation with the pedigree-based inbreeding in both simulated and Simmental datasets. G_{CONTR} estimated for the genotyped animals in the Simmental populations were weakly and negatively correlated with the milkability GEBV. The distribution of the G_{CONTR} in the simulated dataset set exhibited a bimodal shape, whereas it was skewed on the right of the real dataset (Fig. 1). Fig. 2 shows the analysis of G_{CONTR} computed for the Simmental data: a significant difference was observed between G_{CONTR} of males and females, with the latter group showing lower values (Fig. 2A). Moreover, a correlation of −0.23 was computed between the G_{CONTR} and the year of birth, as demonstrated also by the decreasing trend observed for G_{CONTR} according to the year of birth of the Simmental genotyped animals (Fig. 2B). In this breed, old animals with genotypes are only bulls (the first genotyped female was born in 2000), confirming the larger G_{CONTR} computed for males (Fig. 2A). Of interest are the correlations of the G_{CONTR} with the investigated traits. Whereas basically no relationships were detected in the simulated dataset, moderate negative correlation was obtained on Simmental real data.

Table 2
Number of eigenvalues (and core sizes) to explain 90, 95, 98, and 99% of variance of the **G** matrix in the simulated and cattle datasets.

Scenario	Variance explained (%)	Core size	
		Simulated	Simmental
V90	90	108	354
V95	95	195	786
V98	98	376	1 593
V99	99	568	2 358

Table 3
Correlations between the contribution to the genomic relationship matrix and the breeding values or the number of offspring in the simulated and cattle datasets.

Scenario	Type	Trait	Correlation with G_{CONTR}
Simulated	True breeding value	Milk yield	0.03
		Fat yield	0.04
		Fat percentage	0.01^{NS}
	Number of offspring		−0.28
	Pedigree inbreeding		−0.20
Simmental	GEBV	Milkability	−0.17
	Number of offspring		−0.07
	Pedigree inbreeding		−0.23

Abbreviations: G_{CONTR} = contribution to the genomic relationship matrix; GEBV = Genomic Estimated Breeding Value; NS = not significant.

Table 4 shows the correlations between GEBV estimated for the candidate animals with the direct inversion of the **G** matrix and those estimated using the different APY scenarios. Values for the Simmental dataset ranged from 0.71 (TOP animals in the V90 scenario) to 0.95 (RANDOM animals in the V99 scenario). Correlations for the simulated dataset ranged from 0.54 (average across traits for TOP animals in the V90 scenario) to 0.95 (average across traits for BOTTOM animals in the V90 scenario). Correlations between GEBV estimated using or not the APY algorithm for all animals in the analyses were larger and close to the unity in the V98 and V99 scenarios (data not shown).

Validation accuracy

Table 5 shows the results about the validation in the different scenarios. Prediction accuracies of the simulated dataset were larger for the second trait in all the considered scenarios. The standard ssGBLUP provided larger accuracies compared to ssGBLUP_APY, especially for the smaller core sizes (Table 5). The accuracies were affected by core size and definition. As the number of core animals increases, an increase in accuracies and a reduction of differences among criteria for core definition can be observed. In ssGBLUP_APY, the choice of bottom-ranked G_{CONTR} animals yielded higher accuracies for almost all considered scenarios, with the largest differences observed in the smallest core sizes (V90 and V95). The lowest accuracies were obtained when only TOP animals for G_{CONTR} were included in the core.

Also for the Italian Simmental, an improvement of accuracy can be observed as the size of the core increased, together with a reduction among differences between core size definition criteria (Table 5). Moreover, the inclusion of BOTTOM G_{CONTR} ranked animals provided larger accuracies for all core scenarios, whereas the lowest accuracy was found when only TOP animals were included in the core.

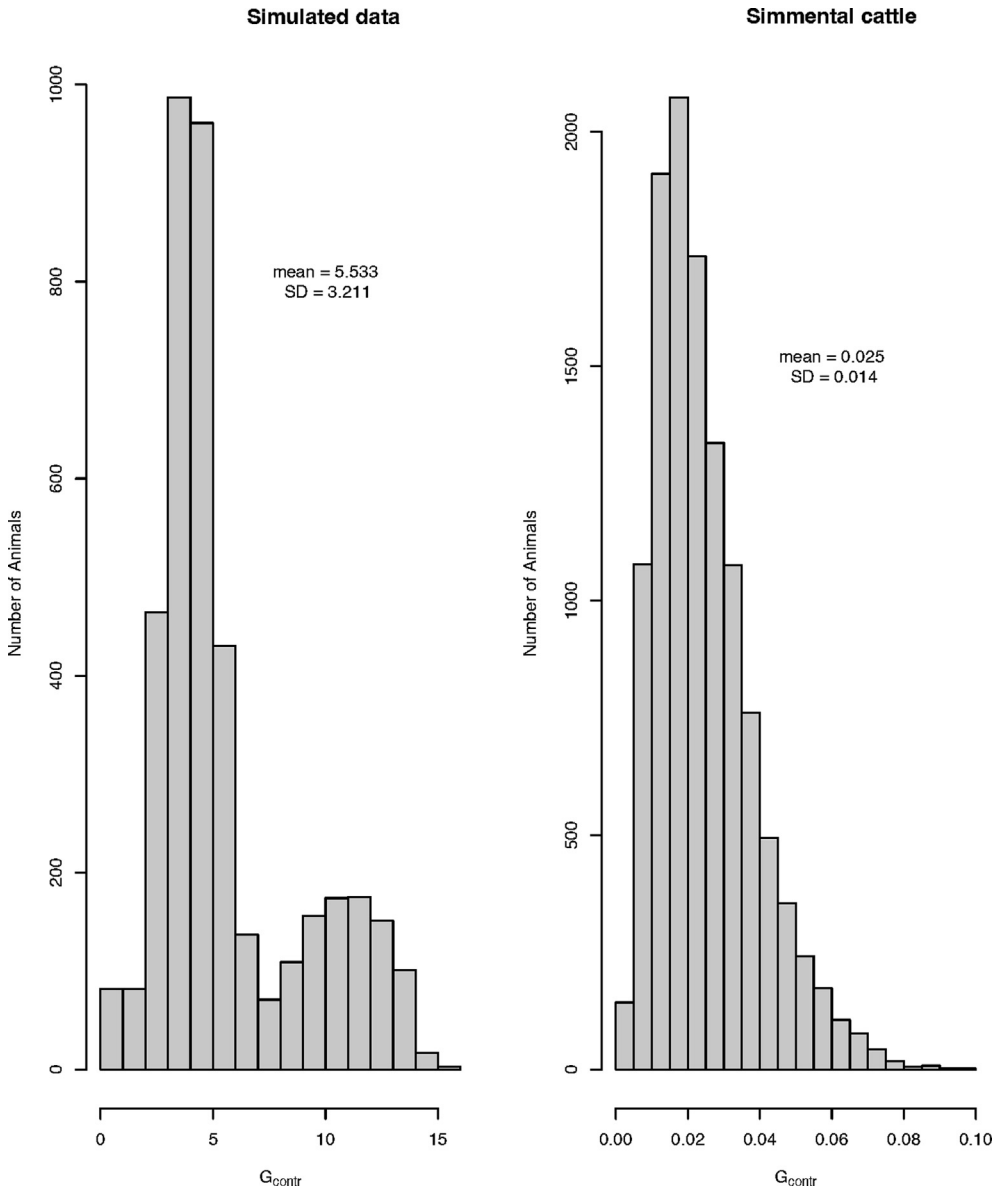

Fig. 1. Histograms of the contribution to the genomic relationship matrix (G_{CONTR}) computed in the simulated and cattle datasets.

Discussion

After about twelve years of genomic selection, calculation problems due to the huge increase of genotyped animals in many livestock species (Cesarani et al., 2022) have raised again the issue of influent individual detection. In particular, with the introduction of single-step methodology, a main point is the proper definition of core animals for optimizing genomic predictions using the APY algorithm. Results of the present study basically confirm the importance of core definition in APY, both in terms of size and of choice criteria, although some interesting aspects can be pinpointed. Theoretically, to maximize the reliability of genomic predictions genotyped animals should represent all the genetic variability of the population, usually identified in the number of segregating independent chromosome segments (Misztal, 2016). Previous studies identified the value of 98% of the variance explained as the threshold for a good approximation of independent chromosome segments of the population (Pocrnic et al., 2022): a core size equal to the number of eigenvalues explaining this amount of variance should lead to the same results of using the full **G**. The same justification could be valid also for the correlations between GEBV estimated using or not the APY algorithm for the candidate animals (Table 4). As expected, the correlations increased as the core size increased. The inclusion of only TOP animals in the core resulted in the lowest correlations in both datasets. Looking at the core size V90, correlations for TOP were on average 0.13 (simulated dataset) and 0.07 (Simmental dataset) points lower than BOTTOM or RANDOM.

Most of the studies aimed at investigating various strategies for defining core individuals have concluded that, when the size is no more a limiting factor, selection criteria are less relevant and, in any case, the random choice of animals is the best option (Bradford et al., 2017; Abdollahi-Arpanahi et al., 2022). In the present paper, an analytical approach for choosing core animals is presented and it has been compared with random selection. Core animals were selected based on their contribution to the genomic relationship matrix, obtained by squaring the sum of the correlations of the individual with each **G** eigenvector (Neuditschko

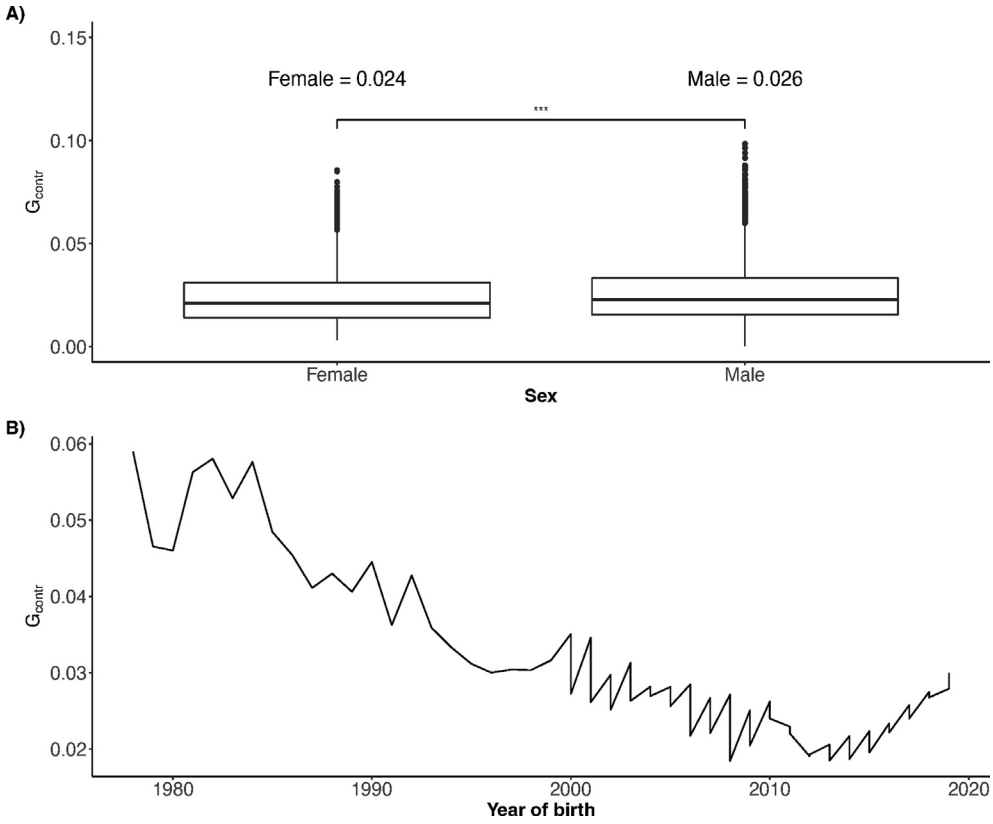

Fig. 2. (A) Differences between the contribution to the genomic relationship matrix (G_{CONTR}) estimated for males (average value of 0.026) and females (average value of 0.024) in the Italian Simmental cattle breed. (B) Trend of the contribution to the genomic relationship matrix according to the year of birth of the genotyped animals.

Table 4
Correlations between the genomic estimated breeding values assessed using or not the algorithm for proven and young animals in the simulated and cattle datasets.

		Simulated			Simmental
Scenario	Core definition	Trait 1	Trait 2	Trait 3	Milkability
V90	Bottom	0.77	0.81	0.78	0.81
	Top	0.52	0.55	0.56	0.71
	Random	0.69	0.70	0.72	0.76
V95	Bottom	0.86	0.88	0.83	0.86
	Top	0.63	0.64	0.64	0.81
	Random	0.75	0.78	0.81	0.86
V98	Bottom	0.93	0.94	0.91	0.91
	Top	0.77	0.78	0.72	0.90
	Random	0.88	0.89	0.88	0.92
V99	Bottom	0.95	0.96	0.95	0.95
	Top	0.83	0.84	0.80	0.95
	Random	0.93	0.94	0.93	0.95

Abbreviations: Bottom = animals with lowest values of contribution to the genomic relationship matrix; Top = animals with highest values of contribution to the genomic relationship matrix; Random = animals randomly selected.

et al., 2017). This total R^2 coefficient provides the amount of the total **G** variance explained by each individual and, thus, we expect that animals with a larger impact on **G** (i.e., the most important animals in the population) have larger contributions. The larger G_{CONTR} values of the most important animals were confirmed by the negative correlation between these values and the pedigree inbreeding coefficients: very old animals that contributed mostly to the recent population have lower inbreeding. On the contrary, recent animals that have a smaller impact on the population have larger inbreeding and lower G_{CONTR}. This pattern was also confirmed by the G_{CONTR} trend according to the year of birth highlighted in the Simmental population. However, the negative correlation between G_{CONTR} and the number of offspring in the pedigree seems to disagree with the importance of the animals in the population. However, it should be considered that this correlation in the real dataset was very low (−0.07). As reported in Table 3, the G_{CONTR} values seem to be population specific. For example, these values were basically not related with phenotypic traits in the simulated dataset, whereas they showed moderate and negative correlations with milkability in the Simmental data. The negative correlation observed between GEBV and G_{CONTR} in the real dataset seems to confirm the larger contributions of old animals. In fact, GEBVs are expected to increase during the time: recent animals (that contribute less to the population) have usually larger GEBV and, thus, lower G_{CONTR}. The moderate relationship of the contributions to the **G** matrix with the investigated phenotypes suggests that this metrics is able to capture, at least to a certain extent, a portion of the variability of **G** that is related with the considered production traits.

As far as the GEBV accuracies, two strategies were tested to represent different breeding programs. In the simulated dataset, all animals in the last generation were considered as potential candidates, whereas in the Simmental population, the attention was focused only on young genotyped cows. The adoption of the G_{CONTR} as a criterion for core definition has resulted in differences, in some cases moderate, in genomic prediction accuracies. Considering the meaning of this metrics, it was reasonable to hypothesize that animals that mostly contribute to the variance of **G** (i.e., TOP animals) were those that better represent the genetic composition of the population and therefore, the most suitable for being considered as core individuals in APY. On the contrary, in the present study, best accuracies were provided when animals with smallest G_{CONTR} were included in the APY core. This result could be due to the relationship between the impact of genotyped animals in the popula-

Table 5
Accuracies of predicted breeding values estimated in the different scenarios for the simulated and cattle datasets in terms of correlation (and 95% confidence interval).

		Simulated			Simmental
Scenario	Core definition	Trait 1	Trait 2	Trait 3	Milkability
ssGBLUP		0.73 (0.70–0.76)	0.77 (0.74–0.79)	0.76 (0.73–0.79)	0.83 (0.79–0.86)
ssGBLUP APY					
V90	Bottom	0.62 (0.58–0.66)	0.66 (0.62–0.69)	0.59 (0.55–0.63)	0.71 (0.65.0.76)
	Top	0.38 (0.33–0.43)	0.42 (0.37–0.47)	0.41 (0.36–0.46)	0.60 (0.53–0.67)
	Random	0.50 (0.45–0.55)	0.60 (0.56–0.64)	0.49 (0.44–0.54)	0.63 (0.56–0.69)
V95	Bottom	0.68 (0.65–0.71)	0.69 (0.66–0.72)	0.61 (0.57–0.65)	0.73 (0.67–0.78)
	Top	0.48 (0.43–0.53)	0.50 (0.45–0.55)	0.44 (0.39–0.49)	0.69 (0.63–0.74)
	Random	0.53 (0.48–0.57)	0.59 (0.55–0.63)	0.62 (0.58–0.66)	0.71 (0.65.0.76)
V98	Bottom	0.71 (0.68–0.74)	0.74 (0.71–0.77)	0.70 (0.67–0.73)	0.79 (0.74–0.83)
	Top	0.57 (0.53–0.61)	0.58 (0.54–0.62)	0.49 (0.44–0.54)	0.77 (0.72–0.81)
	Random	0.67 (0.63–0.70)	0.71 (0.68–0.74)	0.67 (0.63–0.70)	0.78 (0.73–0.82)
V99	Bottom	0.71 (0.68–0.74)	0.75 (0.72–0.78)	0.70 (0.67–0.73)	0.81 (0.77–0.84)
	Top	0.60 (0.56–0.64)	0.63 (0.59–0.67)	0.55 (0.51–0.59)	0.78 (0.73–0.82)
	Random	0.70 (0.67–0.73)	0.74 (0.71–0.77)	0.72 (0.69–0.75)	0.80 (0.76–0.84)

Abbreviations: ssGBLUP = single-step genomic BLUP; APY = algorithm for proven and young animals; Bottom = animals with lowest values of contribution to the genomic relationship matrix; Top = animals with highest values of contribution to the genomic relationship matrix; Random = animals randomly selected.

tion and their G_{CONTR}. As pointed out before, the older animals with a larger contribution in the population have high G_{CONTR}: these animals are largely considered through the pedigree, and they are already reflected in the phenotypes of their offspring. Thus, there are no improvements by adding those animals in the core set. On the contrary, young animals are usually less represented in pedigree and data; when these young animals with low G_{CONTR} are included in the core, they increase the amount of information considered for the GEBV estimation.

Recently, Pocrnic et al. (2022) developed an iterative algorithm based on the conditional covariance for the core definition that aims at choosing animals that are distant in covariance sense. Their approach provided more stable prediction results in ssGBLUP APY compared to the random choice of animals, especially when the core size is under the optimum. The results of Pocrnic et al. (2022) can be considered in agreement with those of the present study, where slightly better accuracies were obtained when animals with smaller G_{CONTR} were considered in the core. In any case, observed differences among the various tested scenarios suggest that an optimization of core definition criteria could represent a way for improving genomic prediction accuracy, especially when the size of core animals is limited. Maybe there are more criteria, not only one, that could help in an optimum core definition. Some papers have highlighted that is important that core animals are uniformly distributed across different generations (Bradford et al., 2017; Abdollahi-Arpanahi et al., 2022) and also across breeds in case of crossbred populations or multibreed analysis (Mäntysaari et al., 2017; Vandenplas et al., 2018; Cesarani et al., 2022). The proposed method here presented offers one analytical approach that could be integrated with others. The peculiarity of the use of G_{CONTR} is that selected animals (i.e., the BOTTOM ranked) are not related to a particular classification criterion in all the considered datasets, but they changed according to the data structure (i.e., the oldest in the simulated dataset, and the youngest in the Simmental population).

Genotyped populations considered in the present study were of small size, representative of different genetic structures. However, the calculation of G_{CONTR} in large datasets could be hampered by the huge computing requirements due to the need for the eigenvalue analysis of the **G**. To address this issue, the G_{CONTR}-based approach for core definition could be improved by implementing algorithms for the optimization of memory use in PCA (Rachakonda et al., 2016). On the other hand, in the present form, it could be efficient in medium size populations that have already started their genomic programs.

Conclusion

Results of the present study confirmed that the choice of animals to be included in the core set of the APY is more important when the core size is not well representing the number of independent chromosome segments. The proposed approach of core definition, based on the individual contribution to the **G** matrix, provided an improvement of GEBV accuracy for the smallest core size considered. However, even if the number of core animals was large enough, the inclusion of BOTTOM animals selected according to their contribution to the **G** matrix led to small increases in accuracy.

Ethics approval

Animal Care and Use Committee approval was not needed as data were obtained from preexisting databases.

Data and model availability statement

None of the data were deposited in an official repository. The simulated dataset could be available upon reasonable request from M. Usai (gmusai@agrisricerca.it), whereas restrictions apply for the Simmental data that was used under agreement for this manuscript.

Author ORCIDs

A. Cesarani: https://orcid.org/0000-0003-4637-8669.
M. Bermann: https://orcid.org/0000-0002-5374-0710.
C. Dimauro: https://orcid.org/0000-0002-6588-923X.
D. Lourenco: https://orcid.org/0000-0003-3140-1002.
N.P.P. Macciotta: https://orcid.org/0000-0001-5504-9459.

Author contributions

A. Cesarani: Conceptualization, Methodology, Formal analysis, Validation, Investigation, Writing - Original Draft, Writing - Review and Editing, Funding acquisition. **M. Bermann**: Validation, Writing - Review and Editing. **C. Dimauro**: Writing - Review and Editing. **L. Degano**: Resources, Data Curation. **D. Vicario**: Supervision, Project administration. **D. Lourenco**: Writing - Review and Editing, Supervision, Project administration, Resources, Funding acquisition. **N.P.P. Macciotta**: Conceptualization, Methodology,

Investigation, Writing - Original Draft, Writing - Review and Editing, Supervision, Project administration, Funding acquisition.

Declaration of interest

None.

Acknowledgments

The authors would like to thank Mario Graziano Usai (AGRIS, Italy) for providing the simulated dataset, the Italian Simmental Breeders Association (ANAPRI) for providing Simmental data.

Financial support statement

This research was funded by the grant "Fondo di Ateneo per la Ricerca 2020 – Una tantum per la Ricerca – Macciotta" and the research contract between the University of Georgia and the University of Sassari. Moreover, financial support for this research was provided by the Italian Ministry of Agriculture (grant DUAL-BREEDING Fase 2 - CUP J22C21000670005).

References

Abdollahi-Arpanahi, R., Lourenco, D., Misztal, I., 2022. A comprehensive study on size and definition of the core group in the proven and young algorithm for single-step GBLUP. Genetics Selection Evolution 54, 1–14.

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Bradford, H.L., Pocrnic, I., Fragomeni, B.O., Lourenco, D.A.L., Misztal, I., 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. Journal of Animal Breeding and Genetics 134, 545–552.

Cesarani, A., Garcia, A., Hidalgo, J., Degano, L., Vicario, D., Macciotta, N.P.P., Lourenco, D., 2021a. Genomic information allows for more accurate breeding values for milkability in dual-purpose Italian Simmental cattle. Journal of Dairy Science 104, 5719–5727.

Cesarani, A., Masuda, Y., Tsuruta, S., Nicolazzi, E.L., VanRaden, P.M., Lourenco, D., Misztal, I., 2021b. Genomic predictions for yield traits in US Holsteins with unknown parent groups. Journal of Dairy Science 104, 5843–5853.

Cesarani, A., Lourenco, D., Tsuruta, S., Legarra, A., Nicolazzi, E.L., VanRaden, P.M., Misztal, I., 2022. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. Journal of Dairy Science 105, 5141–5152.

Fisher, R., 1921. On the probable error of a coefficient of correlation deduced from a small sample. Metron 1, 3–32.

Fragomeni, B.O., Lourenco, D.A.L., Tsuruta, S., Masuda, Y., Aguilar, I., Misztal, I., 2015. Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses – a simulation study. Journal of Animal Breeding and Genetics 132, 340–345.

Lourenco, D.A.L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T.J., Forni, S., Weller, J.I., 2014. Are evaluations on young genotyped animals benefiting from the past generations? Journal of Dairy Science 97, 3930–3942.

Mäntysaari, E.A., Evans, R.D., Strandén, I., 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. Journal of Animal Science 95, 4728–4737.

McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genetics 5, e1000686.

Meuwissen, T.H., Hayes, B.J., Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Misztal, I., 2016. Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics 202, 401–409.

Misztal, I., Legarra, A., Aguilar, I., 2014. Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science 97, 3943–3952.

Neuditschko, M., Raadsma, H.W., Khatkar, M.S., Jonas, E., Steinig, E.J., Flury, C., Signer-Hasler, H., Frischknecht, M., von Niederhäusern, R., Leeb, T., Rieder, S., 2017. Identification of key contributors in complex population structures. PLoS One 12, e0177638.

Ostersen, T., Christensen, O.F., Madsen, P., Henryon, M., 2016. Sparse single-step method for genomic evaluation in pigs. Genetics Selection Evolution 48, 1–10.

Pocrnic, I., Lourenco, D.A., Masuda, Y., Legarra, A., Misztal, I., 2016. The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics 203, 573–581.

Pocrnic, I., Lindgren, F., Tolhurst, D., Herring, W.O., Gorjanc, G., 2022. Optimisation of the core subset for the APY approximation of genomic relationships. Genetics Selection Evolution 54, 1–17.

Rachakonda, S., Silva Rogers, F., Liu, J., Calhoun, V.D., 2016. Memory Efficient PCA Methods for Large Group ICA. Frontiers in Neuroscience 10, 17.

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetics Research 35, 131–155.

Tsuruta, S., Misztal, I., 2008. Computing options for genetic evaluation with a large number of genetic markers. Journal of Animal Science 86, 1514–1518.

Tsuruta, S., Misztal, I., Stranden, I., 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed model equations in animal breeding applications. Journal of Animal Science 79, 1166–1172.

Usai, M.G., Gaspa, G., Macciotta, N.P.P., Carta, A., Casu, S., 2014. XVIth QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. BMC Proceedings 8, S1.

Vandenplas, J., Calus, M.P., Napel, J., 2018. Sparse single-step genomic BLUP in crossbreeding schemes. Journal of Animal Science 96, 2060–2073.