View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Citation for published version:
Pocrnic, I, Lourenco, DAL, Masuda, Y, Legarra, A & Misztal, I 2016, 'The Dimensionality of Genomic Information and Its Effect on Genomic Prediction', *Genetics*. https://doi.org/10.1534/genetics.116.187013

Digital Object Identifier (DOI):
10.1534/genetics.116.187013

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genetics

Publisher Rights Statement:
Copyright © 2016 by the Genetics Society of America
Available freely online through the author-supported open access option.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. May. 2020

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Ivan Pocrnic,*,[1] Daniela A. L. Lourenco,* Yutaka Masuda,* Andres Legarra,[†] and Ignacy Misztal*
*Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and [†]Institut National de la Recherche Agronomique, Génétique, Physiologie et Systèmes d'Elevage, F-31326 Castanet-Tolosan, France

ABSTRACT The genomic relationship matrix (GRM) can be inverted by the algorithm for proven and young (APY) based on recursion on a random subset of animals. While a regular inverse has a cubic cost, the cost of the APY inverse can be close to linear. Theory for the APY assumes that the optimal size of the subset (maximizing accuracy of genomic predictions) is due to a limited dimensionality of the GRM, which is a function of the effective population size (N_e). The objective of this study was to evaluate these assumptions by simulation. Six populations were simulated with approximate effective population size (N_e) from 20 to 200. Each population consisted of 10 nonoverlapping generations, with 25,000 animals per generation and phenotypes available for generations 1–9. The last 3 generations were fully genotyped assuming genome length L = 30. The GRM was constructed for each population and analyzed for distribution of eigenvalues. Genomic estimated breeding values (GEBV) were computed by single-step GBLUP, using either a direct or an APY inverse of GRM. The sizes of the subset in APY were set to the number of the largest eigenvalues explaining x% of variation (EIGx, x = 90, 95, 98, 99) in GRM. Accuracies of GEBV for the last generation with the APY inverse peaked at EIG98 and were slightly lower with EIG95, EIG99, or the direct inverse. Most information in the GRM is contained in $\sim N_eL$ largest eigenvalues, with no information beyond $4N_eL$. Genomic predictions with the APY inverse of the GRM are more accurate than by the regular inverse.

KEYWORDS GenPred; shared data resource; genomic selection; genomic relationship matrix; inversion; recursion; effective population size; single-step GBLUP

WHEN SNP information is available, genomic predictions most commonly use SNP-BLUP (and derivatives) or genomic BLUP (GBLUP) models (Meuwissen *et al.* 2001; VanRaden 2008). In the first model SNP effects are fitted directly, and the second model uses SNPs indirectly via a genomic relationship matrix. While both models are equivalent theoretically, analyses with complex models (multiple traits, several genetic effects, genotype-by-environment interactions) are simpler with GBLUP. For populations where only a small fraction of phenotyped animals are genotyped, there is a modification of GBLUP called single-step GBLUP (ssGBLUP) based on combining genomic and pedigree relationships (Aguilar *et al.* 2010; Christensen and Lund 2010). The ssGBLUP is becoming popular for commercial genetic evaluations because of simplicity of use, as existing models can be reused, and high accuracy due to joint modeling of phenotypes, pedigrees, and genotypes (Legarra *et al.* 2014).

GBLUP-based methods require an inverse of the genomic relationship matrix (GRM). A direct inverse has a cubic cost and can be computed efficiently for perhaps up to 150,000 individuals. Due to the popularity of commercial genotyping, some populations have >1 million genotyped animals (*e.g.*, U.S. Holstein cattle), and computing an inverse would be prohibitively expensive. Additionally, the GRM is not positive definite for larger dimensions and additional steps (*e.g.*, blending with a pedigree-based relationship matrix) are required to make the GRM positive definite (VanRaden 2008).

Misztal *et al.* (2014) postulated that the inverse can be computed efficiently using recursion on a small subset of animals (initially labeled as high-accuracy or "proven" in earlier studies) and named the method the algorithm for proven and young (APY). In this article, we refer to the inverse calculated with this algorithm as the APY inverse and to animals in the subset as core animals. While computing costs of APY are cubic for the subset, they are only linear for animals outside the subset. Fragomeni *et al.* (2015) analyzed Holstein data with 100,000 genotyped animals and found that any

Copyright © 2016 by the Genetics Society of America
doi: 10.1534/genetics.116.187013
Manuscript received January 19, 2016; accepted for publication February 29, 2016; published Early Online March 4, 2016.
Available freely online through the author-supported open access option.
[1]Corresponding author: Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602. E-mail: ipocrnic@uga.edu

subset of animals containing at least 10,000 animals resulted in an accurate inverse. The optimal subset size was estimated as slightly >8000 for Angus cattle (Lourenco *et al.* 2015b). The APY inverse was successfully computed for ~570,000 genotyped Holsteins in <2 hr of computing time on an average server (Masuda *et al.* 2016). More than 10,000 animals in the recursion did not improve genetic predictions. A regular inverse for 570,000 individuals would require weeks of computing and require memory available only in the largest computing clusters.

Misztal (2016) proposed a theory for the APY inverse. Assume that the additive information in a population is contained in a limited number (say, n) of independent chromosome segments (M_e) or effective SNP markers (ESM). If M_e or ESM completely explain the additive variation, breeding values of n animals are linear functions of M_e or ESM and contain nearly all the information included in M_e or ESM. Treating any subset of n animals as core animals, a recursion on any n animals is sufficient, because there is a high redundancy in genomic information. Whereas the number of M_e is a function of effective population size, the number of ESM could be computed as the number of eigenvalues explaining nearly all the variation in **G**. Assuming that M_e and ESM describe the same concept, the optimal subset size must be a function of effective population size (N_e) and can be derived from eigenvalue analysis of the GRM.

The purpose of this article was to test the theory of the APY with simulated data. In particular, we wanted to find whether (1) the optimal size of the recursion is related to effective population size, (2) the optimal size can be derived from eigenvalue analysis of the GRM, and (3) genetic predictions obtained with APY $\mathbf{G}^{-1}$ are superior to those with a regular inverse.

Materials and Methods

Data simulation

Data for this study were simulated using the QMSim software (Sargolzaei and Schenkel 2009). The historical population consisted of 1000 generations with a gradual increase in size from 1000 to 100,000 breeding individuals, with equal sex ratio, nonoverlapping generations, random mating, no selection, and no migration to create initial linkage disequilibrium (LD) and establish mutation–drift balance in the population. Six populations with different effective population size were created by selecting different numbers of breeding animals from the last generation of the historical population. Whereas the number of breeding females per generation was kept constant at 12,500, the number of males varied from 5 to 50 (5, 10, 20, 30, 40, and 50), aiming for approximate effective population sizes from 20 to 200 (data sets P20, P40, P80, P120, P160, and P200). In each generation randomly selected male offspring were used as sires for the next generation, while all the females were used as dams for the next generation. Ten recent generations were simulated for each population by random mating and with litter size of 2. All 75,000 individuals in generations 8–10 had genotypic information available. The simulated genome was assumed to have 30 chromosomes of equal length of 100 cM each, with 49,980 evenly allocated biallelic SNP markers and equal allele frequencies in the first generation of the historical population. A total of 4980 biallelic and randomly distributed QTL affected the trait, with allelic effects sampled from a gamma distribution with a shape parameter of 0.4. The recurrent mutation rate of the markers and QTL was assumed to be 2.5×10^{-5} per locus per generation (Solberg *et al.* 2008). Phenotypes were simulated with an overall mean as the only fixed effect and assuming heritability of 0.3. All animals in the recent generations had phenotypes available, except for animals in the last generation. The simulation was replicated five times.

Matrices and models

The raw genomic relationship matrix was constructed as in VanRaden (2008),

$$\mathbf{G}_0 = \frac{\mathbf{Z}\mathbf{Z}'}{2\Sigma p_j(1-p_j)},$$

where **Z** is the centered matrix of gene content adjusted for gene frequencies, and p_j is the gene frequency for SNP j. The observed allele frequencies were used and calculated from the genotyped animals. The eigenvalues of this matrix were computed using subroutine DSYEV in LAPACK. Because $\mathbf{G}_0$ was not full rank, estimation of breeding values was based on the inverse of a blended **G** defined as

$$\mathbf{G} = 0.95\mathbf{G}_0 + 0.05\mathbf{A}_{22}$$

(VanRaden 2008), where $\mathbf{A}_{22}$ is a pedigree-based numerator relationship matrix for genotyped animals. In preliminary tests, the blending had very little impact on realized accuracies (Aguilar *et al.* 2010).

The APY for inversion of **G** is based on a recursion on a subset of animals (Misztal *et al.* 2014; Misztal 2016). Split animals arbitrarily into core (c) and noncore (n) such that the number of core animals is close to the dimensionality of **G** or the number of effective SNPs. Also denote

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix}.$$

Assume that the breeding values (BV) **u** for noncore animals are linear functions of those for core animals,

$$\mathbf{u}_n = \mathbf{P}_{nc}\mathbf{u}_c + \mathbf{\Phi}_n,$$

where $\mathbf{P}_{nc}$ is a matrix relating BV of noncore to core animals and $\mathbf{\Phi}_n$ is the error term. Then

$$\begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u}_c \\ \mathbf{\Phi}_n \end{bmatrix}$$

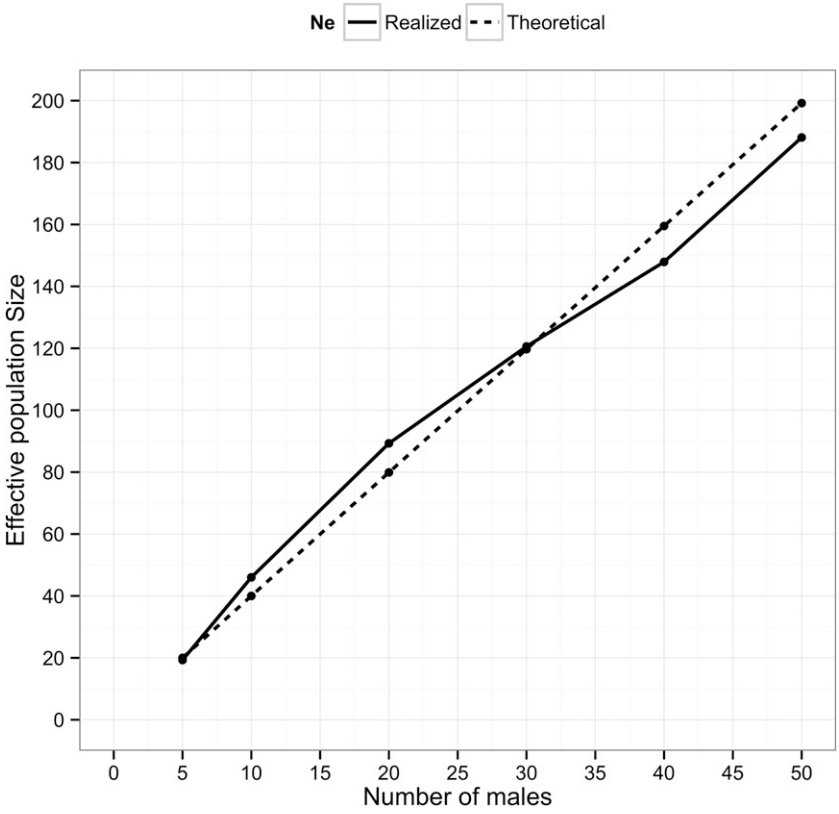

Figure 1 Theoretical and realized effective population size (N_e) as a function of breeding males per generation when the number of breeding females was 12,500 per generation.

and

$$\text{var}(\mathbf{u}) = \mathbf{G}_{\text{APY}} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

where $\mathbf{M}_{nn} = \text{var}(\mathbf{\Phi}_n)$. Subsequently,

$$\mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{P}_{cn} & \mathbf{I} \end{bmatrix}.$$

Using conditional distributions, $\mathbf{P}_{nc} = \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}$, $\mathbf{M}_{nn} = \text{diag}\{m_{nn,i}\} = \text{diag}\{g_{ii} - \mathbf{g}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci}\}$, and the final formula is as originally defined in Misztal *et al.* (2014):

$$\mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix}.$$

In this algorithm, the direct inversion is only for $\mathbf{G}_{cc}$ and computing $\mathbf{G}_{nn}$ is not needed.

Phenotypes were analyzed using the ssGBLUP model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{S}\mathbf{u} + \mathbf{e},$$

in which $\mathbf{y}$ is the observation vector for the first nine of the recent generations, μ is an overall mean, $\mathbf{u}$ is the vector of additive animal effects, $\mathbf{S}$ is the incidence matrix relating observations in $\mathbf{y}$ to additive genetic effects in $\mathbf{u}$, and $\mathbf{e}$ is the vector of random residuals. We assumed that the variances were

$$\text{var}(\mathbf{u}) = 0.3\mathbf{H} \text{ and } \text{var}(\mathbf{e}) = 0.7\mathbf{I},$$

where $\mathbf{H}$ is a matrix combining pedigree and genomic relationships, with its inverse as in Aguilar *et al.* (2010); *i.e.*,

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}^{-1}$ is the inverse of a numerator relationship matrix for all animals included in the analysis. The partition in blocks refers to animals with/without genotypes.

Computations

Effective population size was calculated using two formulas. Theoretical effective population size (N_{e_T}) was calculated using the formula

$$N_{e_T} = \frac{4N_m N_f}{N_m + N_f}$$

(Wright 1931), where N_m and N_f are the numbers of breeding males and females per generation, respectively. Inbreeding (or realized) effective population size (N_{e_F}) was calculated from the realized increase in inbreeding by generation, using the formula by Falconer and Mackay (1996),

$$N_{e_F} = \frac{1}{2\Delta F},$$

where

$$\Delta F = \frac{F_n - F_{n-1}}{1 - F_{n-1}}$$

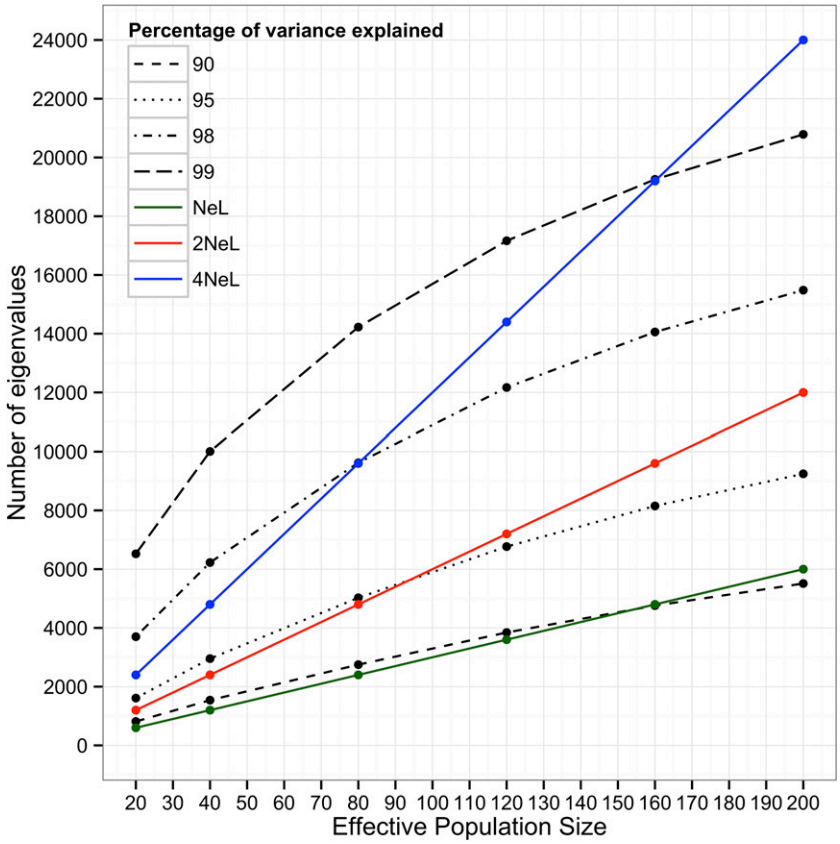

Figure 2 Number of largest eigenvalues that explain 90%, 95%, 98%, and 99% of variation in the genomic relationship matrix for populations with different effective population sizes (N_e). Solid lines show N_eL, $2N_eL$, and $4N_eL$, where L = 30 M.

and F_n is the average inbreeding in the nth generation. The observed average inbreeding coefficients per generation were obtained from QMSim software (Sargolzaei and Schenkel 2009).

Genomic estimated breeding values (GEBV) were calculated using either an explicit inverse of **G** or the APY inverse. Core animals in the APY were selected randomly and their number corresponded to the number of the largest eigenvalues in $\mathbf{G}_0$ that explained 90% (EIG90), 95% (EIG95), 98% (EIG98), and 99% (EIG99) of the retained variance. Validation accuracies were computed only for animals in the 10th generation (without phenotypes) and defined as correlations between simulated breeding values and GEBV computed with either the regular inversion ($GEBV_{REG}$) or the APY ($GEBV_{APY}$) and a different number of core animals. All computations were applied to each of the six data sets.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results and Discussion

Figure 1 shows N_{e_T} and N_{e_F} with the different number of breeding males per generation. Both N_{e_T} and N_{e_F} were very similar and increased with the number of breeding males, from 20 and 19.3 to 199.2 and 188.1 for P20 and P200, respectively. If we take into account family size and variation in family size (Laporte and Charlesworth 2002), the N_e values would be in the same range (20–200) as in the simulation. Thus, the simulation scheme was effective in creating populations with close to the desired N_e. For simplicity, all graphs and discussions use rounded N_{e_T}.

The number of eigenvalues of $\mathbf{G}_0$ that accounted for 90%, 95%, 98%, and 99% of the original variation is shown in Figure 2 and *Appendix* Table A1. Accounting for 90% of the original variation (EIG90) required 814 ± 14 eigenvalues in population P20 and 5512 ± 19 eigenvalues in population P200. Accounting for 99% of the original variation (EIG99) required 6523 ± 68 eigenvalues in population P20 and 20,786 ± 29 eigenvalues in population P200. Thus, increasing N_e ~10 times increased the number of selected eigenvalues by 6.8 for EIG90 and by 3.2 for EIG99. While the number of eigenvalues increased with N_e, the increase was less than proportional especially for higher N_e. Graphically (Figure 2), the increases in the number of eigenvalues corresponding to 90% and 95% were close to linear, but less so corresponding to 98% and especially for 99% past N_e = 120. The total number of positive eigenvalues in **G** is bounded by the number of SNPs (49,980 in this study) and the number of genotyped individuals (75,000 in this study). Subsequently, steeper declines from a linear trend when the number of eigenvalues is very high could be due to a limited number of SNPs and individuals used in the simulation.

The number of eigenvalues can also be expressed in terms of N_e and genome length L (L = 30). The value of EIG90

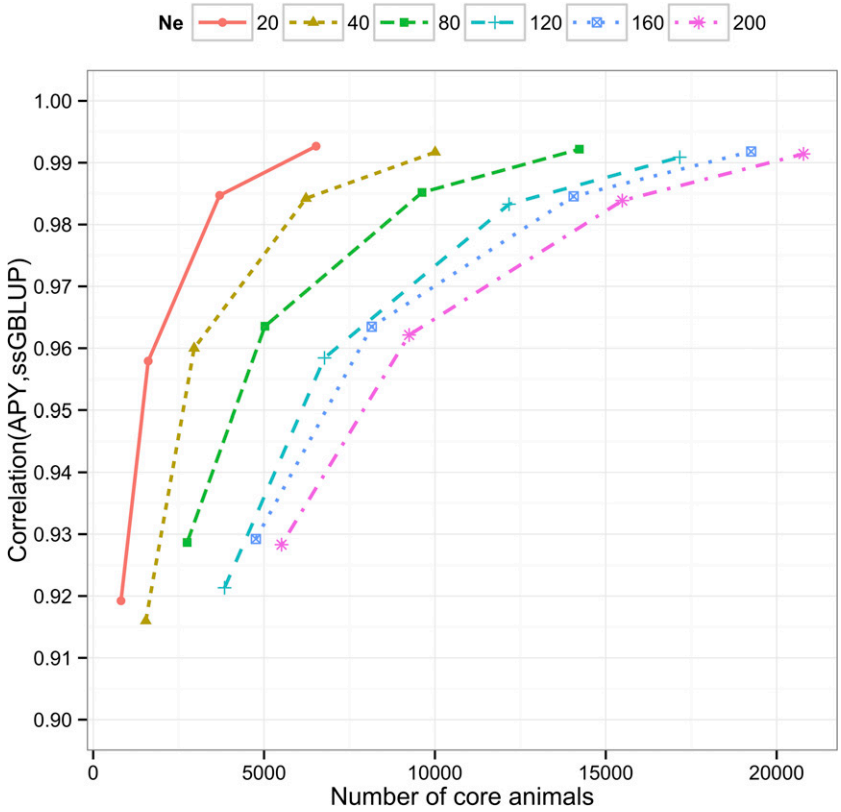

Figure 3 Correlations between genomic estimated breeding values obtained with the direct inverse (ssGBLUP) and the inverse with the algorithm for proven and young (APY) of the genomic relationship matrix for six simulated populations as a function of the number of core animals.

varies from $\sim 40N_e$ (P20) to $27N_e$ (P200), the value of EIG95 varies from $80N_e$ (P20) to $46N_e$ (P200), and the value of EIG98 varies from $\sim 185N_e$ (P20) to $75N_e$ (P200). Assuming that the increase in the number of eigenvalues is indeed linear with N_e but affected by limited number of SNPs and genotyped individuals, the approximate values would be EIG90 $\approx N_eL$, EIG95 $\approx 2N_eL$, and EIG98 $\approx 4N_eL$.

Figure 3 shows the correlation between $GEBV_{REG}$ and $GEBV_{APY}$ for validation animals with variable numbers of core animals (from EIG90 to EIG99). Populations with greater N_e required a larger number of core animals to reach equivalent correlations. For all populations, the correlations were >0.99 with the number of core animals equal to EIG99 and >0.98 with the number of core animals equal to EIG98. Figure 4 shows the results as in Figure 3 but with the percentage of explained variance on the abscissa. The curves are linear and nearly identical. This means that the correlations between $GEBV_{REG}$ and $GEBV_{APY}$ are nearly a linear function of the percentage of the explained variance, regardless of N_e. The correlations are slightly higher than the percentage of explained variance, probably because GEBV contain not only contributions due to genomics but also some due to parent average (VanRaden 2008; Lourenco *et al.* 2015a).

Figure 5 shows true accuracies (defined as the correlation between simulated breeding value and GEBV) across the six simulated populations as a function of the number of eigenvalues explaining the given amounts of variance. All SDs of accuracies across replicates were ≤0.01. The accuracy is inversely related to N_e as it was highest for population P20 (0.89 ± 0.01) and lowest for P200 (0.77 ± 0.01). In simulated populations, Muir (2007) and Goddard (2009) showed that accuracy of GEBV decreases as N_e increases. Smaller N_e means fewer M_e or ESM to estimate and subsequently smaller prediction error variance of these effects. The accuracies were only ~0.03 below the peak level with the number of core animals corresponding to EIG90; the accuracies increase by ~0.02 at EIG95, peaking at EIG98; and they are slightly lower at EIG99. The accuracies with the regular inverse (noted as 100% in the graph) were slightly lower than with EIG99. The results indicate that the majority of the information for GEBV is provided by EIG90 largest eigenvalues. The accuracy provided by eigenvalues present beyond EIG95 in EIG98 was small but required almost doubling the number of core animals. Eigenvalues corresponding to the last 2% variation do not provide any information and in fact slightly reduce the accuracies, which shows that the genomic information may be redundant and in fact overfitted the data. Subsequently we can conclude that the dimensionality of the genomic information (defined as the number of the larger, informative eigenvalues) in this study does not exceed EIG98. For genomic prediction, using the number of core animals corresponding to EIG98 is sufficient, with reduction to EIG95 when computing is expensive.

The theory for the APY was developed either based on the dimensionality of **G** as computed from eigenvalues or based on the independent chromosome segments (M_e) (Misztal 2016). Both concepts may be closely related. In particular, the number of M_e is similar to the number of core animals beyond which the accuracy of GEBV does not increase. In this study, such a

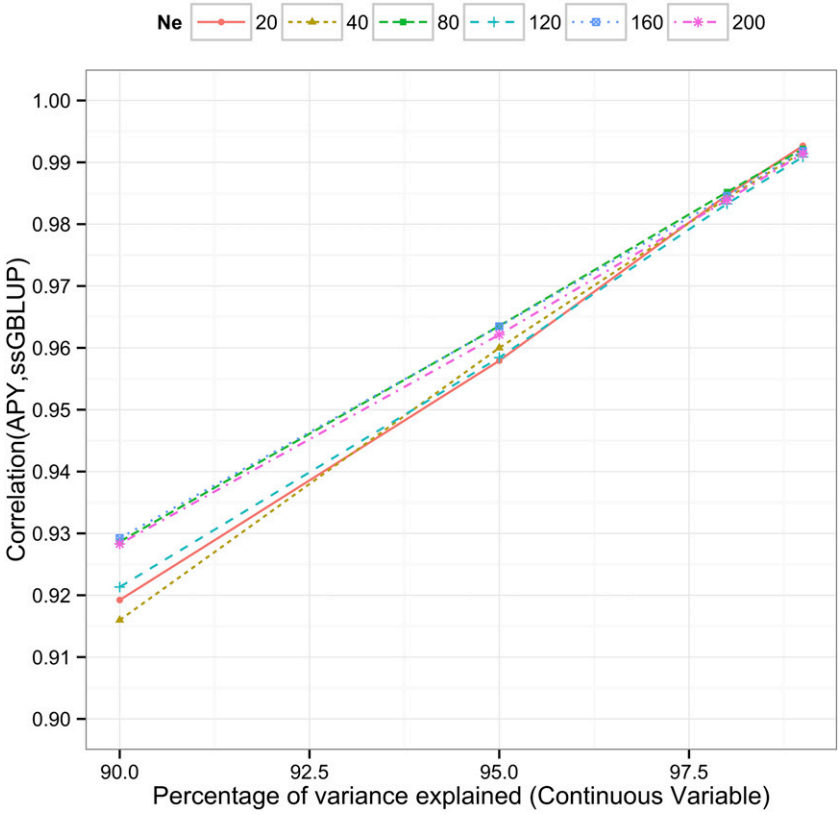

Figure 4 Correlations between genomic estimated breeding values obtained with the direct inverse (ssGBLUP) and the inverse with the algorithm for proven and young (APY) of the genomic relationship matrix for six simulated populations where the number of core animals is defined as the number of eigenvalues that explain 90%, 95%, 98%, and 99% of original variability.

number corresponded to EIG98. Stam (1980) derived a probability density function for a size of an independent chromosome segment, which leads to the expected number of segments $M_e = 4N_eL$, where L is the size of the genome in morgans. With $L = 30$ in this study, $M_e = 120N_e$, which is close to an estimate of $140N_e$ for EIG98. The approximate values in this study, EIG90 $\approx 35N_e$, EIG95 $\approx 70N_e$, and EIG98 $\approx 140N_e$, could be simplified to EIG90 $\approx N_eL$, EIG95 $\approx 2N_eL$, and EIG98 $\approx 4N_eL$, respectively. As the segments are of variable size, Goddard (2009) argued that a more relevant formula is $2N_eL/\log(4N_eL)$, which is equivalent to $M_e = 8N_e$ ($N_e = 20$) to $6N_e$ ($N_e = 200$). Both numbers are well below EIG90. Several formulas for M_e were compared in a meta-analysis by Brard and Ricard (2015), and none were found satisfactory. Such conclusions could be due to several factors. First, their study looked at realized accuracies, and these are strongly affected by selection (Bijma 2012; Lourenco *et al.* 2015a). Second, the implicit assumption was of segments of equal size, while segment sizes were variable (Stam 1980). Third, Brard and Ricard (2015) pointed out that part of the difficulty resided in getting good estimates of N_e, a parameter that is not always well defined and that changes over time. We can posit that in genomic selection we can estimate the effects of the largest chromosome segments well and those of smaller segments not as well but they are still useful for prediction, and the remaining smallest segments provide insufficient accuracy for prediction. Compared to methods reviewed by Brard and Ricard (2015), the possible definition of M_e by EIG98 does not depend on realized accuracies or trait definition, but does require genotype collection.

This study focused on dimensionality of the GRM. In fact, the eigenvalue distribution of a SNP BLUP matrix ($\mathbf{Z}'\mathbf{Z}$, where $\mathbf{Z}$ is gene content) is the same as both share the same singular values from singular value decomposition of $\mathbf{Z}$. Therefore, the dimensionality of the GRM can be defined as dimensionality of the SNP genomic information in general. In this study, the eigenvalues were computed from the GRM explicitly constructed. However, for large data sets, it is possible to compute them from the singular value decomposition of matrix $\mathbf{Z}$, with a cost quadratic in the number of markers and only linear in the number of individuals (*e.g.*, by subroutine DGESVD in LAPACK).

Some results of this study could be influenced by simulation parameters. In particular, a larger number of genotyped animals and the number of SNP markers could have increased the dimensionality especially for higher N_e. Genotypes by simulation are perfect while in real data they are affected by quality control and possibly imputation. In addition, the simulated population was not selected and the number of genotyped generations was small. Further studies will show applicability of the results of this article to real populations undergoing selection.

Although the largest population size simulated in this study had $N_e = 200$, the dimensionality of the GRM can be extrapolated for populations with a larger N_e. In general, the dimensionality of the genomic information ($\mathbf{G}$ or $\mathbf{Z}'\mathbf{Z}$) is $\leq \min(M_e, N_{snp}, N_{ind})$, where N_{snp} is the number of SNPs and N_{ind} is the number of genotyped individuals. In this study, N_{snp} and N_{ind} were several times larger than M_e although the dimensionality of the GRM was depressed by limited N_{snp} and N_{ind} especially for a large N_e. It appears that the dimensionality of

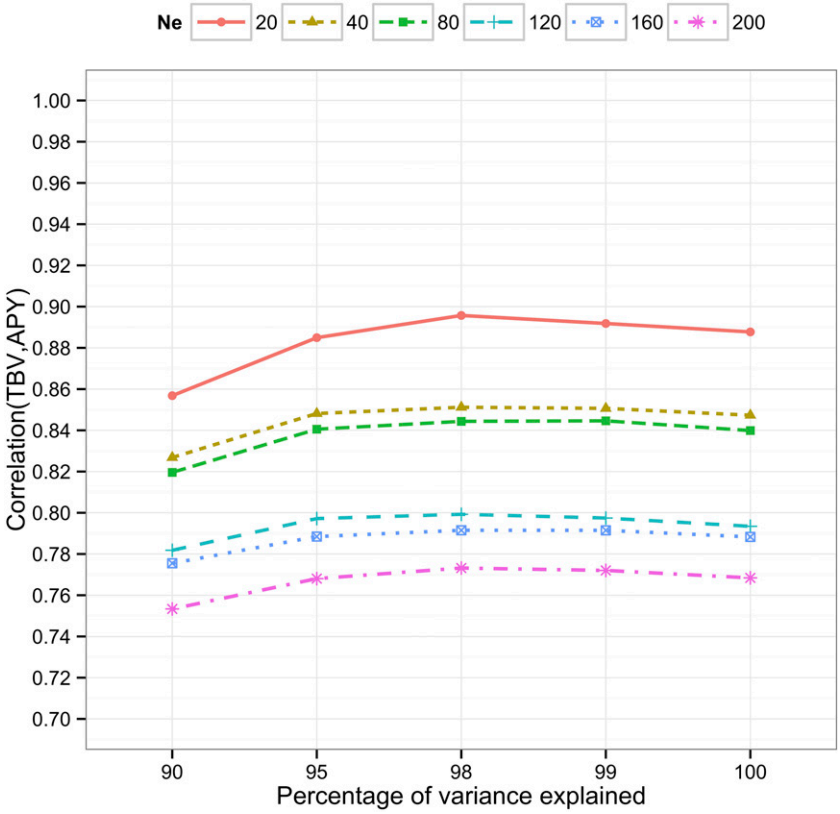

Figure 5 Accuracies of genomic estimated breeding values across six simulated populations where the number of core animals is defined as the number of eigenvalues that explain 90%, 95%, 98%, and 99% of original variability; values for 100% correspond to the regular inverse of the genomic relationship matrix. Accuracies are defined as correlations between true breeding values (TBV) and genomic estimated breeding values obtained with the algorithm for proven and young (APY) inverse.

the GRM is close to M_e when N_{snp} and N_{ind} are a few times larger than M_e. In fact, MacLeod *et al.* (2005) found that detection of 90% of junctions between independent chromosome segments required ~12 times as many markers as the number of junctions ($\approx M_e$). Assume N_e = 3000 and M_e = 360,000. If N_{ind} or N_{snp} is low in comparison to M_e, dimensionality will be close to min (N_{ind}, N_{snp}). The dimensionality will reach M_e only when both N_{ind} and N_{snp} are $>>M_e$. For polygenic traits, N_{snp} determines a fraction of the additive variance explained by the genomic information (Jensen *et al.* 2012). Hypothesizing that in fact it is the ratio of N_e to M_e that is important, even a large N_{snp} can create a "missing heritability" problem in humans where N_e and subsequently M_e are large (Yang *et al.* 2010).

Simulations in this study assumed a polygenic model with an equal variance of each SNP. Heterogeneous SNP variance can be incorporated via weights for each trait separately as discussed in Misztal (2016). In particular, if positions of all causative SNPs were known, the rank of **G** would be equal to the number of causative SNPs (Misztal 2016); with 200 causative SNPs the rank of **G** would be 200. If only a few causative SNPs are known and their position/variance is not known precisely, we can expect the rank of **G** to be lower than that estimated from the effective population size but larger than the number of causative SNPs.

Results of this study may be applicable toward understanding the limits of genome-wide association studies (GWAS) resolution. Wang *et al.* (2012) found in a simulation study that the highest correlation of a simulated QTL value was not with the SNP effect closest to the QTL but with the average of 8–16 adjacent SNP markers. Su *et al.* (2014) investigated individual or block variances on 50,000 SNPs in Holstein cattle and found that slightly higher accuracy was obtained when the same variances were imposed on a block of 30 SNPs, which corresponds to 2 Mb or ~15N_e segments (assuming N_e = 100 for Holsteins). In a simulation study, Hassani *et al.* (2015) found that QTL effects were better predicted by averages of ±100 flanking markers than by an average of a smaller number of flanking markers. The resolution of GWAS may be limited to a size of an individual chromosome segment and subsequently by N_e.

In summary, when the number of SNP markers and genotyped animals is large, the dimensionality of the SNP genomic information defined by the eigenvalue of the GRM is approximately a linear function of effective population size. Subsequently, an inverse of the GRM based on limited recursion can be computed inexpensively for a large number of individuals. Such an inverse results in more accurate estimation of GEBV than a direct inverse.

Acknowledgments

The authors thank Paul VanRaden for helpful suggestions. Editing by Heather L. Bradford and Shogo Tsuruta is gratefully acknowledged. This research was primarily supported by grants from the Holstein Association USA (Brattleboro, VT), the American Angus Association, Zoetis, Cobb-Vantress, Smithfield Premium Genetics, the Pig Improvement Company, and the U.S. Department of Agriculture's National Institute of Food and Agriculture (Agriculture and Food Research Initiative competitive grant 2015-67015-22936).

Literature Cited

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

Bijma, P., 2012 Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J. Anim. Breed. Genet. 129: 345–358.

Brard, S., and A. Ricard, 2015 Is the use of formulae a reliable way to predict the accuracy of genomic selection? J. Anim. Breed. Genet. 132: 207–217.

Christensen, O. F., and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, Essex, UK.

Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar *et al.*, 2015 Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J. Dairy Sci. 98: 4090–4094.

Goddard, M., 2009 Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136: 245–257.

Hassani, S., M. Saatchi, R. L. Fernando, and D. J. Garrick, 2015 Accuracy of prediction of simulated polygenic phenotypes and their underlying quantitative trait loci genotypes using real or imputed whole-genome markers in cattle. Genet. Sel. Evol. 47: 99.

Jensen, J., G. Su, and P. Madsen, 2012 Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet. 13: 44.

Laporte, V., and B. Charlesworth, 2002 Effective population size and population subdivision in demographically structured populations. Genetics 162: 501–519

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal, 2014 Single Step, a general approach for genomic selection. Livest. Sci. 166: 54–65.

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach *et al.*, 2015a Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genet. Sel. Evol. 47: 56.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar *et al.*, 2015b Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93: 2653–2662.

MacLeod, A. K., C. S. Haley, J. A. Woolliams, and P. Stam, 2005 Marker densities and the mapping of ancestral junctions. Genet. Res. 85: 69–79.

Masuda, Y., I. Misztal, S. Tsuruta, A. Legarra, I. Aguilar *et al.*, 2016 Implementation of genomic recursions in single-step genomic BLUP for US Holsteins with a large number of genotyped animals. J. Dairy Sci. 99: 1968–1974.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 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., A. Legarra, and I. Aguilar, 2014 Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97: 3943–3952.

Muir, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124: 342–355.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680–681.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

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

Su, G., O. F. Christensen, L. Janss, and M. S. Lund, 2014 Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Sci. 97: 6547–6559.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir, 2012 Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73–83.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Communicating editor: D. J. de Koning

Appendix

Table A1 Number of largest eigenvalues (mean ± SE) explaining a given percentage of variation for populations with different effective population sizes N_e (P*x* means effective population size *x*)

N_e	90%	95%	98%	99%
P20	814 ± 14	1,611 ± 26	3,701 ± 48	6,523 ± 68
P40	1,540 ± 8	2,954 ± 16	6,226 ± 29	10,006 ± 37
P80	2,749 ± 14	5,026 ± 25	9,622 ± 40	14,226 ± 47
P120	3,844 ± 14	6,769 ± 22	12,169 ± 31	17,163 ± 34
P160	4,760 ± 7	8,151 ± 12	14,058 ± 15	19,253 ± 15
P200	5,512 ± 19	9,245 ± 25	15,483 ± 29	20,786 ± 29