# Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size

**Ignacy Misztal**[1]
Animal and Dairy Science, University of Georgia, Athens, Georgia 30602
ORCID ID: 0000-0002-0382-1897 (I.M.)

**ABSTRACT** Many computations with SNP data including genomic evaluation, parameter estimation, and genome-wide association studies use an inverse of the genomic relationship matrix. The cost of a regular inversion is cubic and is prohibitively expensive for large matrices. Recent studies in cattle demonstrated that the inverse can be computed in almost linear time by recursion on any subset of ~10,000 individuals. The purpose of this study is to present a theory of why such a recursion works and its implication for other populations. Assume that, because of a small effective population size, the additive information in a genotyped population has a small dimensionality, even with a very large number of SNP markers. That dimensionality is visible as a limited number of effective SNP effects, independent chromosome segments, or the rank of the genomic relationship matrix. Decompose a population arbitrarily into core and noncore individuals, with the number of core individuals equal to that dimensionality. Then, breeding values of noncore individuals can be derived by recursions on breeding values of core individuals, with coefficients of the recursion computed from the genomic relationship matrix. A resulting algorithm for the inversion called "algorithm for proven and young" (APY) has a linear computing and memory cost for noncore animals. Noninfinitesimal genetic architecture can be accommodated through a trait-specific genomic relationship matrix, possibly derived from Bayesian regressions. For populations with small effective population size, the inverse of the genomic relationship matrix can be computed inexpensively for a very large number of genotyped individuals.

**KEYWORDS** genomic relationship matrix; genomic selection; inversion; single-step GBLUP; recursion

FOR animals and plants, many genomic analyses with SNP data use one of two approaches. Either effects of SNP markers are estimated with best linear unbiased prediction (SNP-BLUP) (Meuwissen *et al.* 2001; VanRaden 2008; Gianola *et al.* 2009; Piepho 2009) or a genomic relationship matrix (GRM) is used in genomic BLUP (GBLUP) (VanRaden 2008). Estimation of SNP effects makes SNP selection and estimation of SNP variance easy, leading to straightforward single-trait prediction and genome-wide association study (GWAS). GBLUP is easier to use in more complex models (*e.g.*, multiple traits) and for parameter estimation because existing BLUP including parameter estimation methodology can be used, although the use of GBLUP for GWAS is more complex (Zhang *et al.* 2010). For prediction, SNP-BLUP (possibly with SNP weighting) and GBLUP are equivalent models (VanRaden 2008) but they differ in computing cost. SNP estimation includes the same number of SNPs independent of the number of individuals. Adding extra individuals incurs linear computing costs and no additional storage. GBLUP usually requires an inverse of GRM, and explicit inversion requires quadratic memory and cubic computations.

When only a small fraction of the population is genotyped, GBLUP can be extended to single-step GBLUP (ssGBLUP) (Aguilar *et al.* 2010; Christensen and Lund 2010). In this method, a numerator relationship matrix (NRM) for all individuals and a GRM are combined and then applied to BLUP. Benefits of ssGBLUP include simplicity of application (another BLUP), avoidance of double counting, and accounting for preselection on Mendelian sampling (Legarra *et al.* 2014). However, ssGBLUP also requires an inverse of GRM.

The inverse of GRM can be computed with general algorithms only for up to perhaps 150,000 individuals because of


doi: 10.1534/genetics.115.182089
Manuscript received August 18, 2015; accepted for publication November 15, 2015; published Early Online November 18, 2015.
Available freely online through the author-supported open access option.
[1]Address for correspondence: Animal and Dairy Science, University of Georgia, Athens, GA 30602. E-mail: ignacy@uga.edu

memory and computing time limitations. However, the number of genotyped individuals across animal populations is expanding. In dairy cattle, >950,000 Holsteins have been genotyped in the United States as of November 2015 (Council on Dairy Cattle Breeding; https://www.cdcb.us/Genotype/cur_density.html). Several algorithms were proposed to lower the cost of ssGBLUP (Legarra and Ducrocq 2012; Fernando *et al.* 2014; Liu *et al.* 2014), but either they are computationally expensive or the algorithms do not converge with a large number of genotypes.

Past progress in animal breeding resulted to a large degree from a fast algorithm to invert the NRM (Henderson 1976). Although the cost of explicit inversion of the NRM is cubic with the number of animals, the cost of creating that inverse directly by recursion is very low (Henderson 1976; Quaas 1988). When animals are ordered from oldest to youngest, a recursion for each animal includes at most only two terms (one for each parent). Consequently, the inverse of the NRM can be created at a linear cost.

Faux *et al.* (2012) applied recursions on relatives of genotyped individuals to the GRM; however, the inverse was not accurate. Misztal *et al.* (2014) postulated recursions on "proven" animals (with their own phenotypes or their progeny phenotyped) and called the methodology an algorithm for proven and young (APY) animals. The APY was tested in a population of Holsteins with a total of 100,000 genotyped animals and different groups of animals in recursion (Fragomeni *et al.* 2015). When recursions were on proven bulls only, the correlation of genomic estimated breeding values (GEBVs) for selection candidates with APY $\mathbf{G}^{-1}$ with GEBVs from a complete inverse was >0.99. When only cows were in the recursion, the correlation remained >0.99. When the recursion included random subsets of 5000, 10,000, and 15,000 animals, the correlations were 0.97, 0.98, and 0.99, respectively, with minimal variability among replicates. Moreover, the convergence rates with random subsets were superior, indicating better numerical conditioning. The APY was also applied to a commercial population of Angus cattle (Lourenco *et al.* 2015). With recursions on 4000, 8000, and 33,000 animals, the APY accounted for 84%, 97%, and 100% of accuracy gains of ssGBLUP over BLUP, respectively. The APY computing and storage costs when applied to cattle are almost linear, which allows for inverting practically any GRM size. However, why the APY works and its possible internal limitations have not been addressed. The first purpose of this study was to develop a theory explaining why recursion on a limited number of individuals results in an accurate GRM inverse. The second purpose is to determine implications of that theory for populations other than cattle.

## Methods

### *Recursions and the inverse of the numerator relationship matrix*

Henderson's (1976) inverse of the relationship matrix can be derived by recursion. Let $\mathbf{u}$ be a vector of additive effects or breeding values (BVs) distributed as $\mathbf{u} \sim N(0, \mathbf{A})$, where $\mathbf{A}$ is a numerator relationship matrix and, for simplicity, the additive variance is set to 1. Following developments in Misztal *et al.* (2014), the joint distribution of $u_1, ..., u_n$ can be written as

$$p(u_1, .., u_n) = p(u_1)\, p(u_2|u_1)\, p(u_3|u_2, u_1) \ldots \\ p(u_n|u_1, u_2, \ldots, u_{n-1}).$$

A notation below $n{:}m$ denotes an index from $n$ to $m$. Assuming normality, the conditional distributions are

$$p(u_i|u_1, u_2, \ldots, u_{i-1}) \sim N\Big[\mathbf{a}_{i,1:i-1}(\mathbf{A}_{1:i-1,1:i-1})^{-1}\mathbf{u}_{1:i-1}, a_{i,i} \\ - \mathbf{a}_{i,1:i-1}(\mathbf{A}_{1:i-1,1:i-1})^{-1}\mathbf{a}_{i,1:i-1}\Big]$$

with $\mathbf{a}_{i,1:i-1}$ part of the $i$th row of $\mathbf{A}$, and with the recursion equation

$$u_i|u_1 \ldots u_{i-1} = \sum_{j=1}^{i-1} p_{ij}u_j + \varphi_i,$$

where $\varphi_i$ is independent of $u_1 \ldots u_{i-1}$. In matrix notation

$$\mathbf{p}_{i,1:i-1} = \mathbf{a}_{i,1:i-1}(\mathbf{A}_{1:i-1,1:i-1})^{-1}, \\ \mathbf{M}_{i,i} = m_i = \text{var}(\varphi_i) = a_{i,i} - \mathbf{p}_{i,1:i-1}\mathbf{a'}_{i,1:i-1}.$$

In matrix notation, the recursions can be written as

$$\mathbf{u} = \mathbf{P}\mathbf{u} + \mathbf{\Phi}, \quad \text{var}(\mathbf{\Phi}) = \mathbf{M},$$

where $\mathbf{M}$ is a diagonal matrix. Because $\mathbf{u} = (\mathbf{I}-\mathbf{P})^{-1}\mathbf{\Phi}$, the inverse of $\mathbf{A}$ can be computed as

$$\mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P}),$$

which is a form of a Cholesky decomposition with diagonal $\mathbf{M} = \text{diag}\{m_i\}$ and lower diagonal $\mathbf{p} = \{p_{ij}\}$ with entries defined for elements with indexes $j \le i$. Each different ordering of individuals will lead to a different matrix $\mathbf{P}$ but identical $\mathbf{A}^{-1}$. For random ordering $\mathbf{P}$ is likely dense, and the cost of the Cholesky decomposition with dense coefficients is cubic. Additional costs of the inverse will include calculation of $\mathbf{A}$ and $\mathbf{P}$.

Henderson (1976) discovered rules to create $\mathbf{A}^{-1}$ at a low cost. Indirectly, the rules are based on a recursion (Quaas 1988),

$$u_i = 0.5(u_{s_i} + u_{d_i}) + \varphi_i,$$

where $s_i$ and $d_i$ refer to the sire and dam of animal $i$, and $\phi_i$ is the Mendelian sampling. Subsequently, when animals are ordered from the oldest to the youngest,

$$u_i|u_1 \ldots u_{i-1} = u_i|u_{s_i}, u_{d_i},$$

$\mathbf{P}$ has at most only two nonzero elements per row corresponding to parents and equal to 0.5, and the inverse can be

calculated at a low and linear cost. Additionally, computing **A** is no longer needed as the nonzero elements of **P** are known and diagonal elements of **M** are easy to compute. Even though the inverse by Henderson (1976) is very simple, it is not an approximation. In practice, it may be more accurate than an inverse derived from inverting **A** explicitly because of fewer computations and thus lower rounding errors.

### *Recursions and the inverse of the genomic relationship matrix*

Let $\mathbf{u}$ be distributed as $\mathbf{u} \sim N(0, \mathbf{G})$, where $\mathbf{G}$ is a genomic relationship matrix. The recursion equations are the same as previously,

$$u_i | u_1 \ldots u_{i-1} = \sum_{j=1}^{i-1} p_{ij} u_j + \varphi_i,$$

but with different coefficients,

$$\begin{aligned} \mathbf{p}_{i,1:i-1} &= \mathbf{g}_{i,1:i-1} (\mathbf{G}_{1:i-1,1:i-1})^{-1}, \\ \mathbf{M}_{i,i} &= m_i = \mathrm{var}(\varphi_i) = g_{i,i} - \mathbf{p}_{i,1:i-1} \mathbf{g'}_{i,1:i-1}, \end{aligned} \tag{1}$$

and a similar inverse

$$\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{P})' \mathbf{M}^{-1} (\mathbf{I} - \mathbf{P}). \tag{2}$$

This inverse can be computed at a low cost only when **P** is sparse and only a small fraction of **G** needs to be computed. The next sections show that both conditions can be met for populations with limited effective population size.

### *Recursion and SNP model*

Assume availability of an optimal number of SNP markers (called effective SNP) such that increasing that number would not increase the accuracy of prediction. Let $\mathbf{y} = \mu + \mathbf{Za} + \mathbf{e}$ with $\mathrm{var}(\mathbf{a}) = \mathbf{I}\sigma_a^2$ and $\mathrm{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$ be a SNP BLUP, where $\mathbf{y}$ is a vector of phenotypes (or phenotype equivalents), $\mu$ is population mean, $\mathbf{a}$ is a vector of SNP marker effects, $\mathbf{Z}$ is a centered matrix of gene content, $\mathbf{e}$ is a vector of residual effects, $\mathbf{I}$ is an identity matrix, $\sigma_a^2$ is SNP marker variance, and $\sigma_e^2$ is residual variance. If $\mathbf{u}$ is a vector of BVs, then $\mathbf{u} = \mathbf{Za} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the fraction of BVs unexplained by SNP effects with $\mathrm{var}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma_\varepsilon^2$. Applications to farm animals using medium-size SNP chips usually assume $\varepsilon \approx 0$ (VanRaden 2008; Goddard *et al.* 2011).

Divide individuals arbitrarily into two groups: core individuals denoted as *c* and other (noncore) individuals denoted as *n*. Then $\mathbf{u}_c = \mathbf{Z}_c\mathbf{a} + \boldsymbol{\varepsilon}_c$ and $\mathbf{u}_n = \mathbf{Z}_n\mathbf{a} + \boldsymbol{\varepsilon}_n$. The conditional expectation of $\mathbf{a} | \mathbf{u}_c$ is BLUP prediction of SNP effects calculated from BVs of core animals

$$\hat{\mathbf{a}} = (\mathbf{Z'}_c\mathbf{Z}_c + \mathbf{I}\alpha)^{-1} \mathbf{Z'}_c\mathbf{u}_c,$$

where $\alpha = \sigma_\varepsilon^2 / \sigma_a^2$. Let $\mathbf{a} = \hat{\mathbf{a}} + \boldsymbol{\varepsilon}_a$, where $\boldsymbol{\varepsilon}_a$ is prediction error, $\mathrm{var}(\boldsymbol{\varepsilon}_a)$ is prediction error variance (PEV), and $\hat{\mathbf{a}}$ and $\boldsymbol{\varepsilon}_a$ are independent. Then

$$\begin{aligned} \mathbf{u}_n &= \mathbf{Z}_n \left[ (\mathbf{Z'}_c\mathbf{Z}_c + \mathbf{I}\alpha)^{-1} \mathbf{Z'}_c\mathbf{u}_c + \boldsymbol{\varepsilon}_a \right] + \boldsymbol{\varepsilon}_n \\ &= \mathbf{Z}_n (\mathbf{Z'}_c\mathbf{Z}_c + \mathbf{I}\alpha)^{-1} \mathbf{Z'}_c\mathbf{u}_c + \boldsymbol{\varepsilon}_n + \mathbf{Z}_n\boldsymbol{\varepsilon}_a. \end{aligned}$$

If SNP effects nearly fully explain BVs ($\boldsymbol{\varepsilon}_c \approx \mathbf{0}$, $\boldsymbol{\varepsilon}_n \approx \mathbf{0}$),

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

where $\mathbf{P}_{nc} = \mathbf{Z}_n (\mathbf{Z'}_c\mathbf{Z}_c + \mathbf{I}\alpha)^{-1} \mathbf{Z'}_c$ is a matrix that relates breeding values of noncore to core individuals and $\boldsymbol{\Phi}_n = \boldsymbol{\varepsilon}_n + \mathbf{Z}_n\boldsymbol{\varepsilon}_a$. Note that the combined error term $\boldsymbol{\Phi}_n$ has a nondiagonal variance but can be approximated as diagonal especially when the number of core individuals is equal to or greater than the number of SNPs ($\boldsymbol{\varepsilon}_a \approx \mathbf{0}$).

In the formula that relates noncore to core animals, using fewer core animals than the number of SNPs necessary would lead to increased PEV, and using more core animals than the number of SNPs should not affect PEV. Note that when the number of core animals is the same as the number of effective SNPs and $\mathbf{Z}_c$ is invertible, BVs of core animals contain almost the same information as these SNP effects or

$$\mathbf{a} \approx \mathbf{Z}_c^{-1}\mathbf{u}_c.$$

Consequently, BVs of core individuals act as linear combinations of effective SNP effects and BVs of noncore individuals depend approximately only on BVs of core individuals. The above formula is useful only for presentation as in practice the differences in GEBVs with slightly different numbers of core animals are minuscule (see Figure 1).

### *Recursions and independent chromosome segments*

For populations with limited effective population size ($N_e$), the genome is broken into a small number called homogenic or independent chromosome segments (ICS), with the number of segments inversely proportional to $N_e$ (Stam 1980; Daetwyler *et al.* 2010). If the number of ICS in a population is $M_e$ and each segment has an additive effect, the BV of each individual is a sum of effects of chromosome segments present in that individual. The following derivations use the previous derivations while substituting effective SNP effects by ICS effects.

Let $\mathbf{s}$ be a vector of additive effects of ICS. Assume that these effects explain nearly all the additive variance. Let $t_{ij}$ be a fraction of segment *j* in individual *i*, and assume that the value of $t_{ij}$ is $t_{ij}s_j$. Assuming a gametic model, $t_{ij}$ would take values of 0, 1, or 2, although this assumption is not critical here. Then $\mathbf{u} = \mathbf{Ts} + \boldsymbol{\varepsilon}$, where $\mathbf{T}$ is a matrix that relates $\mathbf{u}$ to chromosome segments, $\mathrm{var}(\mathbf{s}) = \mathbf{I}\sigma_t^2$, $\sigma_t^2$ is segment variance, and $\boldsymbol{\varepsilon}$ is the fraction of BVs unexplained by ICS effects.

Following the previous derivations, divide the individuals into two groups: core and noncore, $\mathbf{u}_c = \mathbf{T}_c\mathbf{s} + \boldsymbol{\varepsilon}_c$ and $\mathbf{u}_n = \mathbf{T}_n\mathbf{s} + \boldsymbol{\varepsilon}_n$. If the number of core animals is equal to $M_e$, $\mathbf{T}$ is full rank, and $\boldsymbol{\varepsilon}_c \approx \mathbf{0}$,

$$\mathbf{s} \approx \mathbf{T}_c^{-1}\mathbf{u}_c$$

or nearly all the additive information present in ICS is also present in BVs of core animals. Then,

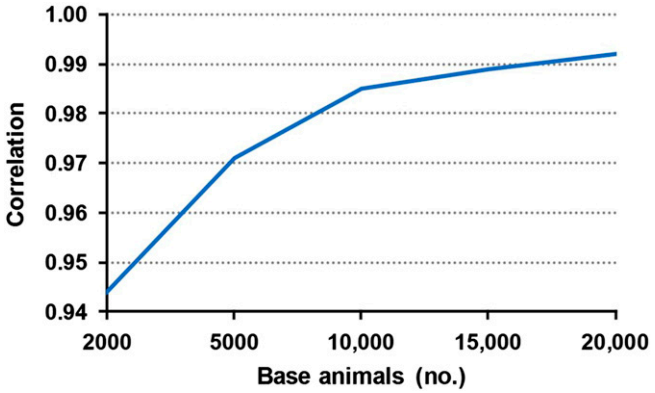


**Figure 1** Correlations between genomic estimated breeding values (GEBVs) for selection candidates using regular and the APY inverse of the genomic relationship matrix (GRM) with various numbers of base individuals (Fragomeni *et al.* 2015). Correlations are based on analysis of 10,102,702 final scores on 6,930,618 Holstein cows, with genotypes available on 100,000 animals; and correlations are based on GEBVs for 49,611 selection candidates.

$$\mathbf{u}_n \approx \mathbf{T}_n\mathbf{T}_c^{-1}\mathbf{u}_c + \boldsymbol{\varepsilon}_n = \mathbf{P}_{nc}\mathbf{u}_c + \boldsymbol{\varepsilon}_n$$

and like before, BVs of noncore individuals depend approximately only on BVs of core individuals.

### *Recursions and the APY formula*

Based on previous derivations, when the number of core animals is sufficiently large, BVs of noncore animals depend only on BVs of core animals. However, obtaining the values of $\mathbf{P}_{nc}$ and corresponding errors from the formula with effective SNPs or ICS is hard. If the GRM is available, the inverse can be obtained indirectly, by applying Equations 1 and 2. In (1), decompose the component matrices $\mathbf{P}$ and $\mathbf{M}$ into sections due to core and noncore animals

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{cc} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\Phi} = \{\varphi_i\} = \begin{bmatrix} \boldsymbol{\Phi}_c \\ \boldsymbol{\Phi}_n \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_{cc} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn} \end{bmatrix}.$$

Subsequently, the complete recursion is

$$\begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{cc} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi}_c \\ \boldsymbol{\Phi}_n \end{bmatrix},$$

as the term relating BVs of noncore animals to BVs of noncore animals ($\mathbf{P}_{nn}$) is null. Following Equation 2, the inverse of $\mathbf{G}$ is

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

If the inverse corresponding to core animals is available, $\mathbf{G}_{cc}^{-1} = (\mathbf{I} - \mathbf{P}'_{cc})\mathbf{M}_{cc}^{-1}(\mathbf{I} - \mathbf{P}_{cc})$, and the complete inverse can be simplified to

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{P}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} [\, -\mathbf{P}_{nc} \quad \mathbf{I}\,].$$

Computing $\mathbf{P}_{nc}$ directly is impossible especially when the effective SNPs or ICS are not known; however, $\mathbf{P}_{nc}$ can be computed indirectly from the GRM. Denote

$$\text{var}\begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} = \mathbf{G} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix} \sigma_u^2,$$

where $\mathbf{G}$ is a GRM and $\sigma_u^2$ is the additive variance. Using conditional distributions, $\mathbf{P}_{nc} = \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}$, $\mathbf{M}_{nn} = \text{diag}\{g_{i,i} - \mathbf{p}_{i,1:i-1}\mathbf{g}'_{i,1:i-1}\}$ for individual $i$ in the noncore group, and the inverse of $\mathbf{G}$ can be calculated as

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

The above equation is the same as that reported by Misztal *et al.* (2014) for the APY with proven animals as the core group and young animals as the noncore group. The formula above requires that only $\mathbf{G}_{cc}$ be full rank. When $\mathbf{G}$ is not of full rank, the APY inverse may in fact be a generalized inverse.

The above derivations can be simplified if the recursion includes only the noncore animals. Then

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

Subsequently

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{0} \\ 0 & \mathbf{M}_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

and

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

which leads to the same formula as derived previously.

### *Computing costs*

Inversion of $\mathbf{G}$ by APY has a cubic cost (and quadratic memory) for core individuals and a linear cost (and linear memory) for noncore individuals. Savings in memory and computing are due to ignoring storage and computations for the blocks of noncore × noncore animals (except diagonal elements) for both $\mathbf{G}$ and APY $\mathbf{G}^{-1}$ (see Figure 2). Assume $n$ core and $p$ noncore individuals. Although regular $\mathbf{G}^{-1}$ requires $\sim(n+p)^2$ memory and $(n+p)^3$ computations, APY $\mathbf{G}^{-1}$ requires only $\sim 2np$ memory and $n^3 + 2n^2p$ computations. If $\beta = n/(n+p)$ is the fraction of individuals that compose the core group and $n << p$, APY $\mathbf{G}^{-1}$ would require only $\sim 2\beta$ memory and $2\beta^2$ computations of the regular algorithm. If $n$ = 10,000 and $n + P$ = 600,000, this is equivalent to 3% of the storage and 0.05% of the computations of the regular algorithm. When the number of core animals is limited, the APY effectively removes limits from computing $\mathbf{G}^{-1}$.

### *Determination of effective number of SNP markers*

Assume an SNP model with a very large number of possibly redundant SNP markers. The real dimensionality of the SNP information and subsequently the number of core animals required to account for all information in these markers can be

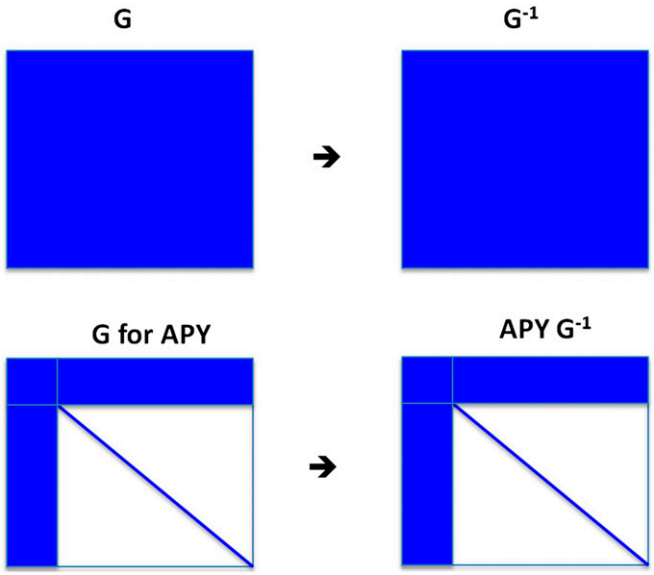


**Figure 2** Sparsity pattern of a regular genomic relationship matrix (**G**) and its inverse ($\mathbf{G}^{-1}$) and elements of the genomic relationship matrix needed for construction of the APY (**G** for the APY) and the APY inverse (APY $\mathbf{G}^{-1}$).

determined by singular value decomposition (*e.g.*, Wall *et al.* 2003). In a formula that relates BVs to SNP effects $\mathbf{u} = \mathbf{Za}$ (ignoring the error term), apply singular value decomposition $\mathbf{Z} = \mathbf{U\Delta V}$, where **U** and **V** are unitarian matrices ($\mathbf{UU}' = \mathbf{I}$ and $\mathbf{VV}' = \mathbf{I}$) and $\mathbf{\Delta} = \{\delta_{ij}\}$ is a diagonal matrix with singular values on the diagonal. Then

$$\mathbf{u} = \mathbf{U\Delta Va}.$$

Let $\mathbf{\Delta_s}$ be a matrix of rows of $\mathbf{\Delta}$ where small singular values (say $< \varphi$) are zeroed. A well-chosen $\varphi$ would retain the accuracy of BVs

$$\mathbf{u} = \mathbf{Za} = \mathbf{U\Delta Va} \approx \mathbf{U\Delta}_s\mathbf{Va}$$

while reducing the dimensionality of the SNP information; the number of nonzero values in $\mathbf{\Delta}_s$ can be called the effective number of SNPs and the nonzero fraction of $\mathbf{\Delta}_s\mathbf{Va}$ could be called effective SNP markers. The parameter $\varphi$ and the effective number of SNPs can be obtained from eigenvalues of the GRM. On the variance scale

$$\mathbf{G} = \mathrm{var}(\mathbf{U\Delta Va}) \sim \mathbf{U\Delta\Delta}'\mathbf{U}',$$

where the formula above shows eigendecomposition of **G** with $\mathbf{\Delta\Delta}'$ as a diagonal matrix of eigenvalues. Following Janss *et al.* (2012), the average fraction of retained variance over all individuals with small eigenvalues set to 0 is proportional to the sum of retained eigenvalues. Subsequently $\phi^2$ can be chosen to retain a high fraction of the variance (say 0.98) by

$$\varphi : \frac{\sum_{\delta_i > \varphi} \delta_i^2}{\sum \delta_i^2} = 0.98.$$

Summarizing, the number of effective independent SNPs and subsequently the minimum number of core animals can be derived from eigenvalue analysis of the GRM.

### *Genetic architecture and the GRM*

While the derivations for the APY included SNP effects or effects of ICS, these effects are absent from final APY derivations, which depend on GRM only. Therefore, any information on specific architecture of a trait, if present, needs to be included in the GRM. In the GBLUP case (same variance of each SNP effect) the GRM can be derived from SNP BLUP as $\mathbf{G} = \mathbf{ZZ}'/q$, where $q$ is a scaling factor (VanRaden 2008). For weighted SNP BLUP, where $\mathrm{var}(\mathbf{a}) = \mathbf{D}\sigma_a^2$ and **D** is a diagonal matrix of weights, the GRM becomes $\mathbf{G} = \mathbf{ZDZ}'/q$. Weights can be computed with the SNP model (*e.g.*, Gianola *et al.* 2009), the GBLUP model (Zhang *et al.* 2010, 2015; Sun *et al.* 2012), or ssGBLUP (Wang *et al.* 2012).

## Discussion

### *Derivations using SNP effects*

Because adjacent SNP markers carry limited information due to linkage disequilibrium, the required number of core individuals in the APY may be equal to some number of "effective independent" SNP markers, as discussed; such a number would be almost independent of the actual number of markers if the number of actual markers is large enough. In Holsteins, predictions using SNP-BLUP were considerably more accurate with 40,000 than with 10,000 SNPs (VanRaden *et al.* 2009). However, improvements in accuracy using SNP-BLUP with higher-density SNP chips (VanRaden *et al.* 2013) or sequencing (Druet *et al.* 2014) are very small. Predictions with APY $\mathbf{G}^{-1}$ for the national evaluation of ~7 million Holsteins were accurate when the number of core animals was >10,000 (Figure 1), with little improvement beyond that number (Fragomeni *et al.* 2015). This suggests the effective number of SNPs for Holstein cattle to be ~10,000. For GWAS, Li *et al.* (2012) presented methods for estimating the effective number of independent markers based on eigenvalues of a SNP correlation matrix. For a human HapMap CEU population, they estimated the number of independent SNP markers at <620,000. The effective population size for that population was estimated at ~3100 (Tenesa *et al.* 2007). Assuming that the number of such markers is proportional to an effective population size, a simple extrapolation for Holsteins (effective population size ~100) leads to <20,000 effective independent SNP markers. Pintus *et al.* (2013) found that ~15,000 eigenvalues extracted from a matrix similar to the correlation matrix based on 40,000 SNPs explained 99% of variance for various traits of Holstein bulls. Marginally higher realized accuracies with higher SNP density could partially be due to lower sampling errors in the GRM (VanRaden 2008; Goddard *et al.* 2011).

There is a question of whether the number of core animals is trait dependent. In this study, APY $\mathbf{G}^{-1}$ was derived as an

extension of the theory for the inverse of the numerator relationship matrix, which is not trait specific. On the other hand, when all QTL are SNPs and identified, the number of effective SNPs is equal to the number of QTL.

### ICS

The main advantage of derivations with ICS is linking the number of core animals to an effective population size. Many definitions exist for the average number of ICS ($M_e$) (Daetwyler *et al.* 2010), with the upper bound $M_e = 4N_eL$, where $N_e$ is effective population size and $L$ is the length of the genome in morgans (Stam 1980). Assuming that most commercial populations of animals have $N_e \leq 100$ and $L \leq 30$, $M_e \leq 12{,}000$.

The concept of ICS is abstract, with segments not directly identifiable. Because ICS are associated with linkage-disequilibrium blocks (Cuppen 2005), the number of ICS and the number of effective independent SNPs may be similar under the polygenic model. Estimates of the number of ICS in commercial livestock populations vary greatly when estimated from realized accuracies (Brard and Ricard 2015), probably because those accuracies are dependent on selection intensity, with smaller accuracies under stronger selection (Bijma 2012; Lourenco *et al.* 2015). Another reason could be an implicit assumption of a constant size of each chromosome segment. In fact, Stam (1980) gave formulas for a distribution of segment size. Assuming that the proportion of variance explained is approximately proportional to segment size, a relatively small number of the largest segments may explain a large fraction of the variance while the remaining majority of segments may explain a small fraction of the variance. This is analogous to eigenvalue distribution. In a study by Fragomeni *et al.* (2015), the correlations of GEBVs obtained with regular and APY $\mathbf{G}^{-1}$ were very high, 0.94 with only 2000 core animals, increasing to 0.99 with 15,000 animals. In a study by Lourenco *et al.* (2015), 84% of accuracy gains due to genomic information were obtained using 4000 animals, with an increase to 97% with 8000 animals and 100% with 33,000 animals. As the reason for both a small number of ICS and reduced dimensionality of SNP information is linkage disequilibrium, the number of the largest effective SNPs and the number of largest ICS explaining the same amount of variance may be similar.

If the number of effective SNPs and the number of ICS are similar, these numbers can be estimated from eigenvalue decomposition of the GRM, as presented. This requires availability of a genotyped population with the size a few times larger than the number of ICS. For very large populations, computations of eigenvalues may be a limiting factor.

A relationship between $N_e$ and $M_e$ allows one to determine applicability of the APY for different populations. In general, the APY can lead to computational savings when the number of genotyped individuals is large and at least twice the number of ICS. Assuming (from Holsteins) $N_e \approx 100$, the APY is useful for ≥20,000 individuals. Extrapolating, with $N_e \approx 3000$ (a smaller estimate in some human populations), the APY would be useful for ≥600,000 individuals. Therefore, the APY appears to be more useful for animal populations with small $N_e$.

### Determining the number of core individuals

Too few core individuals would reduce accuracy of the inverse, whereas too many would increase cost and possibly decrease numerical stability because of unnecessary computations. The optimal number of core individuals could be defined such that increasing that number brings no notable improvement in accuracy of GEBVs calculated using APY $\mathbf{G}^{-1}$. This number of individuals can be determined by performing many analyses with different numbers of core individuals and comparing GEBVs or realized accuracies. Alternatively, the number of effective SNPs can be calculated from eigenvalue decomposition of the GRM as the number of the largest eigenvalues that explain 98–99% of the variation in the GRM, assuming that the remaining variation is due to sampling noise. The choice of the number of core animals is not critical.

### Choice of animals for recursion

Based on the provided theory, the choice of animals for recursion as core animals is not critical as long as appropriate matrices are full rank, and this excludes only multiple copies of clones. In practice, the choice may have some impact. The APY $\mathbf{G}^{-1}$ is sparse, with the location of dense blocks dependent on the definition of core animals. When such an inverse is used in mixed-model equations solved iteratively, the convergence rate may vary with the choice of core animals. In Holsteins, the best convergence rate was found with core animals selected randomly whereas the slowest was with cows as core animals (Fragomeni *et al.* 2015). Another factor is quality of genotypes. In commercial populations, animals may be genotyped with SNP chips of different density followed by, sometimes multiple, imputation to a standard SNP density. In genetic evaluation including ∼570,000 genotyped animals, use of popular sires as core resulted in slightly greater accuracy at the same size of recursion than random choices (Masuda *et al.* 2016). Therefore, the selection process may include the quality of genotypes. In particular, more "valuable" animals, *e.g.*, popular sires, are more likely to have more accurate genotypes.

### Which inverse is more accurate?

*A priori* is not clear which of regular and APY $\mathbf{G}^{-1}$ is superior. Masuda *et al.* (2016) looked at realized accuracies of Holstein bulls and found that those obtained with APY $\mathbf{G}^{-1}$were marginally (<0.01) greater. If the proposed theory for the APY is applicable, the APY $\mathbf{G}^{-1}$ is accurate if the number of core individuals reaches the number of ICS or effective number of independent SNPs. Therefore, for large populations most computations with regular $\mathbf{G}^{-1}$ are redundant and may lead to numerical problems. As argued before, the real rank of GRM (disregarding very small eigenvalues) is likely equal to the number of ICS (or independent SNPs). The GRM with a larger number of individuals is singular and standard

procedures to invert it include blending, a weighted mean of the original GRM with a NRM (VanRaden 2008; Aguilar *et al.* 2010); the primary purpose of blending is numerical stability of inversion as GEBVs obtained with blending 1–10% of the NRM were nearly identical (Misztal *et al.* 2010). Another reason for higher accuracy of APY $\mathbf{G}^{-1}$ could be less influence from sampling errors in the GRM, as the APY does not use the block of GRMs due to noncore individuals (except for diagonals); the standard error of an element of the GRM assuming 0.5 allele frequencies is $1/\sqrt{8m}$, where $m$ is the number of SNP markers (VanRaden 2008).

### *Concluding remarks*

The presented theory explains why and when recursions on a small subset of animals lead to an efficient computation of inverse of the GRM. Such an inverse is accurate when the subset is as large as the number of ICS or the rank of the GRM; for $N_e = 100$ the number of ICS is ~10,000. When the number of genotyped individuals is much larger than the number of ICS, the APY inverse is sparse and facilitates genomic evaluation, parameter estimation, and GWAS at greatly reduced cost and potentially higher accuracy than a regular inverse.

## Acknowledgments

Helpful comments by Ignacio Aguilar, Gustavo de los Campos, Andres Legarra, Daniela Lourenco, Yutaka Masuda, Bill Muir, Ivan Pocrnic, Paul VanRaden, and George Wiggans and editing by Suzanne Hubbard are gratefully acknowledged. The author also acknowledges the very helpful and meticulous comments by the two anonymous reviewers. This research was primarily supported by grants from Holstein Association USA (Brattleboro, VT) 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.

Cuppen, E., 2005 Haplotype-based genetics in mice and rats. Trends Genet. 21: 318–322.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

Druet, T., I. M. Macleod, and B. J. Hayes, 2014 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112: 39–47.

Faux, P., N. Gengler, and I. Misztal, 2012 A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J. Dairy Sci. 95: 6093–6102.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick, 2014 A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46: 50.

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.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183: 347–363.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409–421.

Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32: 69–83.

Janss, L., G. de los Campos, N. Sheehan, and D. Sorensen, 2012 Inferences from genomic models in stratified populations. Genetics 192: 693–704.

Legarra, A., and V. Ducrocq, 2012 Computational strategies for national integration of phenotypic, genomic, and pedigree data in single-step best linear unbiased prediction. J. Dairy Sci. 95: 4629–4645.

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

Li, M.-X., J. M. Y. Yeung, S. S. Cherny, and P. C. Sham, 2012 Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet. 131: 747–756.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents, 2014 A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97: 5833–5850.

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

Masuda, Y., I. Misztal, S. Tsuruta, and A. Legarra, I. Aguilar*et al.*, 2015 Implementation of genomic recursions in single-step genomic BLUP for US Holsteins with a large number of genotyped animals. J. Dairy Sci. 98: 4090–4094.

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., I. Aguilar, A. Legarra, and T. J. Lawlor, 2010 Choice of parameters for single-step genomic evaluation for type (Abstr. 616). J. Dairy Sci. 93(E-Suppl. 1): 533.

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.

Piepho, H. P., 2009 Ridge regression and extensions for genome-wide selection in maize. Crop Sci. 49: 1165–1176

Pintus, M. A., E. L. Nicolazzi, J. B. C. H. M. Van Kaam, S. Biffani, A. Stella *et al.*, 2013 Use of different statistical models to predict direct genomic values for productive and functional traits in Italian Holsteins. J. Anim. Breed. Genet. 130: 32–40.

Quaas, R. L., 1988 Additive genetic model with groups and relationships. J. Dairy Sci. 71: 1338–1345.

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

Sun, X., L. Qu, D. J. Garrick, J. C. M. Dekkers, and R. L. Fernando, 2012 A fast EM algorithm for BayesA-like prediction of genomic breeding values. PLoS One 7: e49157.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

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

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.*, 2009 Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16–24.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker *et al.*, 2013 Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96: 668–678.

Wall, M. E., A. Rechtsteiner, and L. M. Rocha, 2003 Singular value decomposition and principal component analysis, pp. 91–109 in *A Practical Approach to Microarray Data Analysis*, edited by D. P. Berrar, W. Dubitzky, and M. Granzow. Kluwer, Norwell, MA.

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.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D. J. de Koning *et al.*, 2010 Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5: e12648.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao *et al.*, 2015 Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 5: 615–627.

*Communicating editor: E. A. Stone*

## Appendix

### Numerical Example

Consider five individuals with two SNP genotypes aA/BB, AA/bB, aA/bB, AA/BB, and aa/BB; the third and fourth individuals could be progeny of the first two individuals while the fifth would be unrelated. Construct a GRM as in VanRaden (2008), assuming 0.5 gene frequencies, and add 0.01 to the diagonal as otherwise the GRM has a rank of 2:

$$\mathbf{Z} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 1 & 1 \\ -1 & -1 \end{bmatrix},$$

$$\mathbf{G} = \frac{ZZ'}{1.25} + \mathbf{I}0.01 = \begin{bmatrix} 0.81 & 0 & 0 & 0.80 & -0.80 \\ & 0.81 & 0 & 0.80 & -0.80 \\ & & 0.01 & 0 & 0 \\ & & & 1.61 & -1.60 \\ \text{symm.} & & & & 1.61 \end{bmatrix}.$$

Treat the first two individuals as core and the rest as noncore. Subsequently

$$\mathbf{G}_{cc}^{-1} = \begin{bmatrix} 1.235 & \mathbf{0} \\ \mathbf{0} & 1.235 \end{bmatrix},$$

$$\mathbf{P}_{cn} = \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} = \begin{bmatrix} 0.000 & 0.988 & -0.988 \\ 0.000 & 0.988 & -0.988 \end{bmatrix},$$

$$\mathbf{M}_{nn} = \begin{bmatrix} 0.010 & 0 & 0 \\ & 0.030 & 0 \\ 0 & 0 & 0.030 \end{bmatrix}, \mathbf{G}^{-1} = \begin{bmatrix} 66.8 & 65.5 & 0 & -33.1 & 33.1 \\ & 66.804 & 0 & -33.1 & 33.1 \\ & & 100.0 & 0 & 0 \\ & & & 33.6 & 0 \\ \text{symm.} & & & & 33.6 \end{bmatrix}.$$

For comparison, a regular inverse of the GRM ($\mathbf{G}_{\text{reg}}^{-1}$) is quite different,

$$\mathbf{G}_{\text{reg}}^{-1} = \begin{bmatrix} 40.6 & 39.4 & 0 & -19.9 & 19.9 \\ & 40.6 & 0 & -19.9 & 19.9 \\ & & 100.0 & 0 & 0 \\ & & & 60.0 & 39.9 \\ \text{symm.} & & & & 60.0 \end{bmatrix},$$

although the inverse of the APY inverse is almost identical to the original GRM

$$\left(\mathbf{G}^{-1}\right)^{-1} = \begin{bmatrix} 0.81 & 0 & 0 & 0.80 & -0.80 \\ & 0.81 & 0 & 0.80 & -0.80 \\ & & 0.01 & 0 & 0 \\ & & & 1.61 & -1.58 \\ \text{symm.} & & & & 1.61 \end{bmatrix}.$$

Large differences on the inverse but not the original scale can be explained by eigen-decomposition $\mathbf{G} = \mathbf{UDU}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues, and $\mathbf{G}^{-1} = \mathbf{U}'\mathbf{D}^{-1}\mathbf{U}$. The smaller the eigenvalue is, the smaller its impact on $\mathbf{G}$ but the larger the impact on $\mathbf{G}^{-1}$.