Multifactorial Methods Integrating Haplotype and Epistasis Effects for Genomic Estimation and Prediction of Quantitative Traits

Yang Da*, Zuoxiang Liang, Dzianis Prakapenka
Department of Animal Science, University of Minnesota, Saint Paul, MN 55108, USA

* Corresponding author

E-mail addresses:

YD: yda@umn.edu
ZL: zliang@umn.edu
DP: praka032@umn.edu

ABSTRACT

The rapid growth in genomic selection data provides unprecedented opportunities to discover and utilize complex genetic effects for improving phenotypes but methodology is lacking. Epistasis effects are interaction effects and haplotype effects may contain local high-order epistasis effects. Multifactorial methods with SNP, haplotype and epistasis effects up to the third-order are developed to investigate the contributions of global low-order and local high-order epistasis effects to the phenotypic variance and the accuracy of gnomic prediction of quantitative traits. These methods include genomic best linear unbiased prediction (GBLUP) with associated reliability for individuals with and without phenotypic observations including a computationally efficient GBLUP method for large validation populations, and genomic restricted maximum estimation (GREML) of the variance and associated heritability using a combination of EM-REML and AI-REML iterative algorithms. These methods were developed for two models, Model-I with 10 effect types, and Model-II with 13 effect types including intra- and inter-chromosome pairwise epistasis effects that replace the pairwise epistasis effects of Model-I. GREML heritability estimate and GBLUP effect estimate for each effect of an effect type are derived except for third-order epistasis effects. The multifactorial models evaluate each effect type based on the phenotypic values adjusted for the remaining effect types and can use more effect types than separate models of SNP, haplotype and epistasis effects; and provide a methodology capability to evaluate the contributions of complex genetic effects to the phenotypic variance and prediction accuracy, and to discover and utilize complex genetic effects for improving the phenotypes of quantitative traits.

INTRODUCTION

Genomic estimation and prediction of quantitative traits using single nucleotide polymorphism (SNP) markers and mixed models have become a widely approach for genetic improvement in livestock and crop species. The rapid growth in genomic selection data provides unprecedented opportunities to discover and utilize complex genetic mechanism but methodology and computing tools are lacking for investigating complex genetic mechanisms using the approach of genomic estimation and prediction. The integration of global low-order epistasis effects and local high-order epistasis effects contained in haplotypes for genomic estimation and prediction is a step forward for the discovery and application of complex genetic mechanisms to improve the phenotypes of quantitative traits. The integrated model with multiple types of genetic effects can use more effect types than separate models SNP, haplotype and epistasis effects, and may provide more accurate understanding of each effect type than the separate models due to the use of phenotypic values adjusted for the genetic values of the remaining effect types in the model.

The theory of genetic partition of two-locus genotypic values defines four types of epistasis values, additive × additive (A×A), additive × dominance (A×D), dominance × additive (D×A), and dominance × dominance (D×D) epistasis values by Cockerham and Kempthorne [1, 2]. The Cockerham method defines each epistasis coefficient as the product of the coefficients of the two interacting effects that each can be additive or dominance [1]. This definition of epistasis coefficient is the basis for defining epistasis model matrices in terms of the model matrices of additive and dominance effects. Cockerham also defines a pedigree epistasis relationship as the product between the pedigree additive and dominance relationships [1], and this definition is the theoretical basis for Henderson's approach to express epistasis relationship matrices as the Hadamard products of the additive and dominance relationship matrices [3].

The Henderson approach of Hadamard products for epistasis relationship matrices was suggested for genomic prediction using epistasis effects by replacing the pedigree additive and dominance relationship matrices with the genomic additive and dominance relationship matrices calculated from SNP markers [4-6]. This genomic version of the Henderson's Hadamard products calculates genomic epistasis relationship matrices based on the model matrices of SNP additive and dominance effects without creating large epistasis model matrices that can be difficult or impossible to compute. For m SNPs, each pairwise (second-order) epistasis model matrix is $(m-1)/2$ times as large as the SNP additive or dominance model matrix. For 50,000 SNPs, an epistasis model matrix is nearly 25,000 times as large as the SNP additive or dominance matrix. The calculation of such a large model matrix for pairwise epistasis effects is difficult and the calculation of the model matrices for epistasis effects higher than the second-order is practically impossible for 50,000 or more SNPs. Since genomic additive and dominance relationship matrices are calculated from the SNP additive and dominance model matrices [7-10], the Hadamard products between SNP additive and dominance relationship matrices removes the computing difficulty associated with the large epistasis model matrices for calculating genomic epistasis relationship matrices. However, this genomic version of the pedigree-based epistasis relationship matrices contains intra-locus epistasis effects that is not present in the epistasis model [11]. For this reason, the genomic version of Henderson's Hadamard product could be described as approximate genomic epistasis relationship matrices (AGERM). Formulations have been developed to obtain the exact genomic epistasis relationship matrices (EGERM) by removing the intra-locus epistasis effects contained in AGERM by modifying Henderson's Hadamard products without creating the epistasis model matrices [11-14]. The difference between AGERM and EGERM tends to diminish as the number of SNPs increases [13]. A Holstein dataset with 60,671 SNPs showed AGERM and EGERM had the same heritability estimates and the same accuracy of

69 predicting the phenotypic values [15] , and the swine dataset with 52,842 SNPs in this manuscript
70 showed the two methods had similar results. However, EGERM required many times of computing
71 time as required by AGERM. The methods in this article allow the use of either AGERM or
72 EGERM, and our computing package of EPIHAP implements both AGERM and EGERM.
73 Henderson's Hadamard products [3] and hence AGERM are applicable to any order of epistasis
74 effects, and EGERM also has a general formula for any order of epistasis effects [13]. However,
75 limited tests showed that fourth-order global epistasis virtually contributed nothing to the
76 phenotypic variance but generated considerable computing difficulty [16], raising question about
77 the value for global epistasis effects beyond the third-order. Methods of genomic estimation and
78 prediction of global epistasis effects up to the third-order should have a wide-range applications,
79 given that the number of reported epistasis effects lag far behind the number of single-point effects
80 [17-19] even though epistasis effects are important genetic effects [20-22]. In contrast to the
81 computing difficulty and uncertain impact of global high-order epistasis effects beyond the third-
82 order, local high-order epistasis effects in haplotypes with many SNPs were responsible for the
83 increased accuracy of predicting phenotypic values of certain traits. For examples, a haplotype
84 model with 12 SNPs per haplotype block had the best prediction accuracy for low density
85 lipoproteins in a human population [23], a haplotype model with 500 Kb haplotype blocks that on
86 average had 105 SNPs per block had the best prediction accuracy for average daily gain in a swine
87 population [24], and a haplotype model with 15 SNPs per haplotype block had the best prediction
88 accuracy in a wheat study [25]. The integration of haplotype and epistasis effects provides an
89 approach to investigate the contributions of global low-order epistasis effects and local high-order
90 epistasis effects to the phenotypic variance and the accuracy of genomic prediction under the same
91 model.

An epistasis GWAS in Holstein cattle showed that intra- and inter-chromosome epistasis effects affected different traits differently, e.g., daughter pregnancy rate was mostly affected by inter-chromosome epistasis effects whereas milk production traits were mostly affected by intra-chromosome epistasis effects [26]. Genomic heritability estimates of intra- and inter-chromosome heritabilities for daughter pregnancy rate using methods in this article showed that intra-chromosome A×A heritability was 0.031, and inter-chromosome A×A heritability 0.178 [15], consistent with the GWAS results of 21% intra-chromosome and 79% inter-chromosome A×A effects among the top 33,552 pairs of A×A effects in the GWAS study [26]. Therefore, dividing pairwise epistasis effects into intra- and inter-chromosome epistasis effects for genomic prediction and estimation allows the investigation of the contributions of intra- and inter-chromosome pairwise epistasis effects to the phenotypic variance and prediction accuracy.

The purpose of the multifactorial model in this article is to integrate haplotype effects and epistasis effects up to the third-order for genomic estimation and prediction of quantitative traits, to provide a general and flexible methodology framework for genomic prediction and estimation using complex genetic mechanisms, and to provide methodology details of the EPIHAP computer package that implements the integration of haplotype and epistasis effects [15, 16]. The multifactorial model has the advantage of using more effect types and assessing each effect type based on the phenotypic values adjusted for all remaining effect types over separate SNP, haplotype and epistasis models. We hypothesize that some traits involve only a small number of the effect types, some traits are more complex and involve more effect types, global low-order epistasis are more important than local high-order epistasis effects of haplotypes for some traits whereas the reverse is true for some other traits, and some traits may be affected by both global low-order and local high-order epistasis effects. The methodology in this article will provide an approach to evaluate these hypotheses, facilitate the discovery and utilization of global low-order

and local high-order epistasis effects relevant to the phenotypic variance and prediction accuracy of each trait, and obtain new knowledge of complex genetic mechanisms underlying quantitative traits.

METHODS

Quantitative genetics (QG) model with SNP, haplotype and epistasis effects and values

The mixed model with single-SNP additive and dominance effects, haplotype additive effects and pairwise SNP epistasis effects in this article is based on the quantitative genetics (QG) model resulting from the genetic partition of single-SNP genotypic values [9, 10], haplotype genotypic values [27], and pairwise genotypic values [1]. An advantage of this QG model is the readily available quantitative genetics interpretations of SNP additive and dominance effects, values and variances; haplotype additive effects, values and variances; epistasis effects, values and variances; and the corresponding SNP, haplotype and epistasis heritability estimates. Two QG models are developed: Model-I with 10 effect types including SNP additive and dominance effects, haplotype additive effects, and epistasis effects up to the third-order; and Model-II with 13 effect types resulting from replacing the pairwise epistasis effects of Model-I with intra- and inter-chromosome epistasis effects. Detailed descriptions of the effects, values, model matrices and the coding of the model matrices as well as the precise definition of each term in the two QG models are provided in **Supplementary Text S1 and Table S1.** With these precise definitions of genetic effects, values and model matrices in the QG models, a concise multifactorial QG model covering both Model-I and Model-II is established, i.e.:

$$\mathbf{g} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{W}_i \boldsymbol{\tau}_{io} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{u}_i \quad (1)$$

$$\mathbf{u}_i = \mathbf{W}_i \boldsymbol{\tau}_{io} \quad (2)$$

where $\boldsymbol{\tau}_{io}$ = genetic effects of the i^{th} effect type from the original QG model based on genetic partition, $\mathbf{W}_i$ = model matrix of $\boldsymbol{\tau}_{io}$, $\mathbf{u}_i$ = genetic values of the i^{th} effect type from the original QG model, and f = number of effect types. For Model-I, subscripts i=1,...,10 represent SNP additive (A), SNP dominance (D), haplotype additive, A×A, A×D, D×D, A×A×A, A×A×D, A×D×D and D×D×D effects sequentially. For Model-II, subscripts i=1,...,13 represent SNP additive, SNP dominance, haplotype additive, intra-chromosome A×A, intra-chromosome A×D, intra-chromosome D×D, inter-chromosome A×A, inter-chromosome A×D, inter-chromosome D×D, A×A×A, A×A×D, A×D×D and D×D×D effects sequentially. The variance-covariance matrix of the genetic values of Equations 1 and 2 is:

$$\mathbf{G} = \text{var}(\sum_{i=1}^{f} \mathbf{u}_i) = \sum_{i=1}^{f} \text{Var}(\mathbf{u}_i) = \sum_{i=1}^{f} \mathbf{G}_i = \sum_{i=1}^{f} \sigma_{io}^2 \mathbf{W}_i \mathbf{W}_i' \quad (3)$$

$$\text{Var}(\boldsymbol{\tau}_{io}) = \sigma_{io}^2 \mathbf{I} \quad (4)$$

$$\mathbf{G}_i = \text{Var}(\mathbf{u}_i) = \mathbf{W}_i \text{Var}(\boldsymbol{\tau}_{io}) \mathbf{W}_i' = \sigma_{io}^2 \mathbf{W}_i \mathbf{W}_i' \quad (5)$$

where $\sigma_{io}^2 = \text{Var}(\tau_{ijo})$ = genetic variance of the i^{th} effect type under the original QG model common to all individuals (all j values). Note that $\mathbf{W}_i \mathbf{W}_i'$ is not a genomic relationship matrix but is the primary information for calculating each genomic relationship matrix. The structure of the **G** matrix of Equation 3 assumes independence between the genetic values of different effect types. However, the GBLUP values of different effect types using the **G** matrix of Equation 3 could be correlated. Under the Hardy-Weinberg equilibrium (HWE) and LE assumptions, additive, dominance and epistasis effects are independent of each other [1, 2]. For genome-wide SNPs, the LE assumption generally does not hold for closely linked loci, and nonzero Hardy-Weinberg disequilibrium (HWD) may exist numerically. These and other unknown factors in real data may result in the existence of correlations between different effect types. Haplotype additive values are

161 correlated with SNP additive effects because a haplotype additive value is the sum of all SNP
162 additive values and an epistasis value within the haplotype block plus a potential haplotype loss
163 [28]. In two recent haplotype studies for genomic prediction, the integration of SNP and haplotype
164 effects increased the prediction accuracy for four of the seven traits in the human study [23] and
165 for three of the eight traits in the swine study [24], showing that SNP and haplotype additive values
166 compensated each other for prediction accuracy and that the correlation between SNP and
167 haplotype additive values were incomplete for those traits. The correlation between haplotype and
168 epistasis values can be complex: the correlation should be nonexistent if the A×A values are inter-
169 chromosome A×A values or intra-chromosome A×A values involving distal SNPs not covered by
170 the haplotypes, but the correlation could be strong if the A×A values are intra-chromosome A×A
171 values involving proximal SNPs covered by the haplotypes.

172 **The reparametrized and equivalent QG model for genomic estimation and prediction**

173 Genomic relationship matrices will be used for genomic estimation and prediction, and the use of
174 genomic relationship matrices results in a reparametrized and equivalent model of the original QG
175 model for genetic values, to be referred to as the RE-QG model, where "reparametrized" refers to
176 the reparameterization of the genetic effects, model matrix and genetic variance of each effect type;
177 and "equivalent" refers to the requirement of the same first and second moments for the original
178 QG model (Equations 1-5) and the RE-QG model described below. This RE-QG model of genetic
179 values can be expressed as:

180 $$\mathbf{g} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{T}_i \boldsymbol{\tau}_i = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{u}_i \quad (6)$$

181 $$\mathbf{G} = \mathrm{var}(\sum_{i=1}^{f} \mathbf{u}_i) = \sum_{i=1}^{f} \mathbf{G}_i = \sum_{i=1}^{f} \sigma_i^2 \mathbf{T}_i \mathbf{T}_i' = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_i = \sum_{i=1}^{f} \sigma_{io}^2 \mathbf{W}_i \mathbf{W}_i' \quad (7)$$

182 where

183 $$\boldsymbol{\tau}_i = \sqrt{k_i}\,\boldsymbol{\tau}_{io} = \text{genetic effects of the } i^{th} \text{ effect type} \quad (8)$$

$$\mathbf{T}_i = \mathbf{W}_i / \sqrt{k_i} = \text{model matrix of } \boldsymbol{\tau}_i \quad (9)$$

$$\sigma_i^2 = \text{Var}(\tau_{ij}) = \text{tr}(\mathbf{G}_i)/n = \sum_{j=1}^{n} G_i^{jj}/n = k_i \sigma_{io}^2 \quad (10)$$

= variance of the genetic effects of the i^{th} effect type common to all individuals

= average variance of all individuals for the genetic values of the i^{th} effect type

$$\mathbf{u}_i = \mathbf{T}_i \boldsymbol{\tau}_i = \mathbf{W}_i \boldsymbol{\tau}_{io} = \text{genetic values of the } i^{th} \text{ effect type} \quad (11)$$

$$\mathbf{G}_i = \text{Var}(\mathbf{u}_i) = \sigma_i^2 \mathbf{T}_i \mathbf{T}_i' = \sigma_i^2 \mathbf{S}_i = \sigma_{io}^2 \mathbf{W}_i \mathbf{W}_i' \quad (12)$$

= variance-covariance matrix of the genetic values of the i^{th} effect type

$$\mathbf{S}_i = \mathbf{T}_i \mathbf{T}_i' = \mathbf{W}_i \mathbf{W}_i' / k_i = \text{genomic relationship matrix of the } i^{th} \text{ effect type} \quad (13)$$

$$k_i = \text{tr}(\mathbf{W}_i \mathbf{W}_i')/n = \text{average of the diagonal elements of } \mathbf{W}_i \mathbf{W}_i' \quad (14)$$

Equations 8-10 are the reparametrization of the genetic effects, model matrices and genetic variances of the original QG model, whereas Equations 11 and 12 show the genetic values and the variance-covariance matrix of the genetic values are the same under the RE-QG and QG models. In Equation 10, G_i^{jj} = the genetic variance of the j^{th} individual for the i^{th} effect type = the j^{th} diagonal element of the $\mathbf{G}_i$ matrix defined by Equation 12. The k_i formula of Equation 14 as the average of the diagonal elements of $\mathbf{W}_i \mathbf{W}_i'$ was originally proposed for genomic additive relationships [8], and was used for genomic dominance relationships [9, 10], haplotype additive genomic relationships [27], and pairwise epistasis genomic relationships [6]. The need of this RE-QG model is due to the use of the genomic relationship matrices (e.g., Equation 13), because the QG model does not contain genomic relationship matrices (Equation 3). Detailed notations of the QG model of Equations 1-5 in reference to the RE-QG model described by Equations 6-14 are summarized in **Supplementary Table S1**.

The formula of genomic relationship matrix ($\mathbf{S}_i$ of Equation 13) is based on the model matrix of each effect type and can be difficult or impossible to compute if epistasis model matrices are used. This computing difficulty of epistasis model matrices is removed by calculating $\mathbf{S}_i$ based on the model matrices of SNP additive and dominance effects without creating the epistasis model matrices using either AGERM or EGERM. AGERM refers the genomic version of Henderson's Hadamard products between pedigree additive and dominance relationship matrices [3] with the pedigree additive and dominance relationship matrices replaced by the genomic additive and dominance relationship matrices [4-6]. AGERM contains intra-locus epistasis that should not exist [11] and EGERM removes intra-locus epistasis from AGERM based on products between genomic additive and dominance relationship matrices [11, 13].

The QG and RE-QG models have the same prediction accuracy due to the equivalence between these two models. The genetic values ($\mathbf{u}_i$, Equations 1 and 6) and the variance-covariance matrix of the genetic values ($\mathbf{G}_i$, Equations 5 and 12) under the QG and RE-QG models are identical, although these two equations have different expressions for the genetic effects and model matrices. Consequently, the QG model without using genomic relationship matrices and the RE-QG model using genomic relationship matrices have identical accuracy of genomic prediction. The choice of the k_i formula for defining the genomic relationship matrix does not affect the accuracy of genomic prediction but affects the interpretation and application of the genetic variance and genomic relationships for each effect type. Since the interpretation of each genetic variance is a focus whereas the interpretation of the genomic relationships is not a focus in this study, the interpretation of the genetic variance and associated heritability is the consideration in choosing the k_i formula of Equation 14.

The RE-QG model using genomic relationships (Equations 6-14) has two major advantages over the QG model without using genomic relationship matrices (Equations 1-5) although the two models have the same prediction accuracy. First, the use of genomic relationships, originally proposed for genomic additive relationships [7], provides a genomic version of the traditional theory and methods of best linear unbiased prediction (BLUP) that uses pedigree relationships, and this genomic version can utilize a wealth of BLUP-based theory, methods and computing strategies. Second, the genetic variance of the genetic effects of each effect type under the RE-QG model can be used for estimating genomic heritability whereas the genetic variance of the genetic effects under the QG model cannot be used for estimating genomic heritability. With the k_i value defined by Equation 14, The variance of the genetic effects of the i^{th} effect type, $\sigma_i^2 = k_i\sigma_{io}^2$ (Equation 10), has the unique interpretation as the average variance of the genotypic values of all individuals and is a common variance to all individuals. Moreover, $\sigma_i^2 = k_i\sigma_{io}^2$ is unaffected by the number of levels for each effect type, unless the number of levels such as the number of SNPs is too small to provide sufficient coverage of the genome [9, 23, 29]. In contrast, the QG model does not have a method to estimate genetic variance components for calculating genomic heritabilities, because σ_{io}^2 is an inverse function of the number of effect levels. As the number of effect levels such as the number of SNPs increases or decreases, the value of each element in $\mathbf{W}_i\mathbf{W}_i'$ changes in the same direction and the σ_{io}^2 estimate changes in the opposite direction, i.e., as the number of effect levels increases or decreases, σ_{io}^2 decreases or increases. Consequently, the σ_{io}^2 estimate does not have a unique interpretation and cannot be used for estimating genomic heritability [9]. Moreover, the variance of the genetic value of an individual ($\sigma_{io}^2\ (\mathbf{W}_i\mathbf{W}_i')^{jj}$) cannot be used for

calculating genomic heritability because of the individual specificity of the $(\mathbf{W}_i\mathbf{W}_i')^{jj}$ values, as shown as follows.

The exact relationship between the genetic variance for the i^{th} effect type of the j^{th} individual under the RE-QG model and the QG model can be described based on the $\mathbf{G}_i$ matrix defined by Equation 12, i.e.:

$$G_i^{jj} = \mathrm{Var}(u_{ij}) = \sigma_i^2(\mathbf{S}_i)^{jj} = \sigma_{io}^2(\mathbf{W}_i\mathbf{W}_i')^{jj} \quad (15)$$

where G_i^{jj} = the j^{th} diagonal element of the $\mathbf{G}_i$ matrix defined by Equation 12 = the genetic variance of the j^{th} individual for the genotypic value of the i^{th} effect type, and u_{ij} = the j^{th} element of $\mathbf{u}_i$ defined by Equation 11. Equation 15 shows that different individuals do not have a common variance of the genetic values (G_i^{jj}) unless all diagonal elements of $\mathbf{S}_i$ or $\mathbf{W}_i\mathbf{W}_i'$ are identical, which could not happen with genome-wide SNP data in the absence of identical twins because genome-wide SNPs have a high degree of individual specificity. Consequently, G_i^{jj} is not a common variance to all individuals and cannot be used for calculating the genomic heritability of the i^{th} effect type. In contrast, σ_i^2 of Equation 10 under the RE-QG model as the average variance of the genotypic values of all individuals is common to all individuals and can be used for calculating the heritability of each effect type. For the example of Model-I, the exact genetic interpretation of G_i^{jj} is: $G_i^{jj} = \sigma_{aj}^2$ = the variance of the genomic additive (breeding) value of the j^{th} individual for $i=1$, $G_i^{jj} = \sigma_{dj}^2$ = the variance of the genomic dominance value of the j^{th} individual for $i=2$, $G_i^{jj} = \sigma_{ahj}^2$ = the variance of the genomic haplotype additive value of the j^{th} individual for $i=3$, $G_i^{jj} = \sigma_{aaj}^2$ = the variance of the A×A value of the j^{th} individual for $i=4$, $G_i^{jj} = \sigma_{adj}^2$ = the variance of the A×D value of the j^{th} individual for $i=5$, $G_i^{jj} = \sigma_{ddj}^2$ = the variance of the D×D value

of the j^{th} individual for $i=6$, $G_i^{jj}=\sigma^2_{aaaj}$ = the variance of the A×A×A value of the j^{th} individual for $i=7$, $G_i^{jj}=\sigma^2_{aadj}$ = the variance of the A×A×D value of the j^{th} individual for $i=8$, $G_i^{jj}=\sigma^2_{addj}$ = the variance of the A×D×D value of the j^{th} individual for $i=9$, and $G_i^{jj}=\sigma^2_{dddj}$ = the variance of the D×D×D value of the j^{th} individual for $i=10$. These genetic interpretations along with those for intra- and inter-chromosome pairwise epistasis effects of Model-II under the QG and RE-QG models are summarized in **Supplementary Table S1**.

RESULTS AND DISCUSSION

The multifactorial model of phenotypic values

Based on the RE-QG model of Equations 6-14, the multifactorial model for phenotypic values is:

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb}+\mathbf{Zg}+\mathbf{e}=\mathbf{Xb}+\mathbf{Z}\textstyle\sum_{i=1}^{f}\mathbf{T}_i\boldsymbol{\tau}_i+\mathbf{e}\\ &=\mathbf{Xb}+\mathbf{Z}\textstyle\sum_{i=1}^{f}\mathbf{u}_i+\mathbf{e}\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbf{V} &= \mathbf{ZGZ}'+\sigma_e^2\mathbf{I}_N=\mathbf{Z}(\textstyle\sum_{i=1}^{f}\mathbf{G}_i)\mathbf{Z}'+\sigma_e^2\mathbf{I}_N\\ &=\mathbf{Z}(\textstyle\sum_{i=1}^{f}\sigma_i^2\mathbf{T}_i\mathbf{T}_i')\mathbf{Z}'+\sigma_e^2\mathbf{I}_N=\mathbf{Z}(\textstyle\sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i)\mathbf{Z}'+\sigma_e^2\mathbf{I}_N\end{aligned} \tag{17}$$

where $\mathbf{y}$ = N×1 column vector of phenotypic observations, $\mathbf{Z}$ = N×n incidence matrix allocating phenotypic observations to each individual = identity matrix for one observation per individual (N = n), N = number of observations, n = number of individuals, $\mathbf{b}$ = c×1 column vector of fixed effects such as heard-year-season in dairy cattle, c = number of fixed effects, $\mathbf{X}$ = N×c model matrix of $\mathbf{b}$, $\mathbf{e}$ = N×1 column vector of random residuals, σ_e^2 = residual variance, and $\mathbf{G}=\sum_{i=1}^{f}\mathbf{G}_i$ (Equation 7). The phenotypic values ($\mathbf{y}$) are assumed to follow a normal distribution with mean $\mathbf{Xb}$ and variance-covariance matrix of $\mathbf{V}$. The methods described below for genomic estimation and prediction are based on the conditional expectation (CE) method, which is more efficient

289 computationally than the methods based on mixed model equations (MME) when the number of
290 genetic effects is greater than the number of individuals [9, 27].

291 For Model-I with 10 effect types, the genomic epistasis relationship matrices can be calculated
292 using either EGERM or AGERM. However, EGERM or AGERM did not consider intra- and inter-
293 chromosome genomic epistasis relationship matrices that are required by Model-II with 13 effect
294 types. This research derives intra- and inter-chromosome genomic epistasis relationship matrices
295 for both EGERM and AGERM.

296 **Intra- and inter-chromosome genomic epistasis relationship matrices**

297 The main derivation of the intra- and inter-chromosome genomic epistasis relationship matrices is
298 the partition of the numerator of a genomic epistasis relationship matrix into intra- and inter-
299 chromosome numerators. The first step is to derive the intra-chromosome numerator, and the
300 second step is to derive the inter-chromosome numerator as the difference between whole-genome
301 numerator and the intra-chromosome numerator. The last step is to divide the intra-chromosome
302 numerator by the average of the diagonal elements of the intra-chromosome numerator, and to
303 divide the inter-chromosome numerator by the average of the diagonal elements of the inter-
304 chromosome numerator. Using this procedure, intra- and inter-chromosome epistasis relationship
305 matrices were derived for both EGERM and AGERM (**Supplementary Text 1**).

306 **Genomic best linear unbiased prediction (GBLUP) and reliability**

307 Based on the multifactorial genetic model of Equations 16 and 17, the GBLUP of the genetic
308 values of the i^{th} effect type ($\hat{\mathbf{u}}_i$) and the best linear unbiased estimator (BLUE) or generalized
309 least squares (GLS) estimator of fixed effect ($\hat{\mathbf{b}}$) are:

310 $$\hat{\mathbf{u}}_i = \sigma_i^2 \mathbf{S}_i \mathbf{Z}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_i^2 \mathbf{S}_i \mathbf{Z}' \mathbf{P}\mathbf{y}, \quad i = 1,...,f \tag{18}$$

311 $$\hat{\mathbf{b}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \tag{19}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$. The GBLUP of total genetic values of the n individuals is the summation of all types of genetic values:

$$\hat{\mathbf{g}} = \sum_{i=1}^{f} \hat{\mathbf{u}}_i \qquad (20)$$

Reliability of GBLUP is the squared correlation between the GBLUP of a type of genetic values and the unobservable true genetic values being predicted by the GBLUP. The expected accuracy of predicting the genetic values by the GBLUP is the square root of reliability, or the correlation between the GBLUP of a type of genetic effects and the unobservable true genetic effects being predicted by the GBLUP. In the absence of validation studies for observed prediction accuracy, reliability or the expected prediction accuracy is the measure of prediction accuracy of the GBLUP. The reliability of the GBLUP of the total genetic value (Equation 20) of the j^{th} individual is:

$$R^2_{gj} = [\mathbf{G}(\mathbf{Z}'\mathbf{P}\mathbf{Z})\mathbf{G}]^{jj} / \mathbf{G}^{jj} \qquad (21)$$

where $\mathbf{G} = \sum_{i=1}^{f} \mathbf{G_i} = \sum_{i=1}^{f} \sigma_i^2 \mathbf{T}_i \mathbf{T}_i' = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_i$ (Equation (36)), $\mathbf{G}^{jj} = \sum_{i=1}^{f} \mathbf{G}_i^{jj} = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_i^{jj}$, and subscript or superscript jj denotes the j^{th} diagonal element. The reliability formula for any or a combination of genetic values can be readily derived from Equation 21, e.g., the reliability of $\hat{\mathbf{u}}_3$ (GBLUP of haplotype additive values) is obtained from Equation 21 by deleting all terms except $\mathbf{G}_3(\mathbf{Z}'\mathbf{P}\mathbf{Z})\mathbf{G}_3$ in the numerator and $\sigma_3^2 \mathbf{S}_3^{jj}$ in the denominator, with changes in the $\mathbf{V}$ and $\mathbf{P}$ matrices accordingly.

Calculation of GBLUP and reliability for individuals with and without phenotypic observations separately

Two strategies are available for calculating GBLUP and reliability of Equations 20 and 21. Strategy-1 is a one-step strategy that include all individuals with and without phenotypic observations in the same system of equations so that GBLUP and reliability are calculated

simultaneously for all individuals. This strategy essentially augments the mixed model for individuals with phenotypic observations with a set of null equations consists of '0's but uses each genomic relationship matrix for all individuals, and these null equations and the use of the relationship matrix for all individuals do not affect the GBLUP, reliability and heritability of individuals with phenotypic observations. The advantage of this one-step strategy is the simplicity of data preparation. For example, for a k-fold cross validation study, the phenotypic input file only needs to have k columns of the trait observations, with one column for each validation where the phenotypic observations for the validation individuals are set as 'missing' and the **X** and **Z** model matrices for the 'missing' observations are set to zero. With this strategy, the genotypic data needs to be processed only once. As the number of traits increases for validation studies, this one-step strategy becomes more appealing due to the savings in data preparation work. This strategy has been implemented in our computing tools of GVCBLUP [30], GVCHAP [31] and EPIHAP [15, 16]. However, when the number of validation individuals or individuals without phenotypic values is large, each genomic relationship matrix ($\mathbf{S}_i$ matrix) is large and the one-step strategy becomes more difficult as the number of individuals increases.

For large numbers of calculating GBLUP for individuals with and without phenotypic values separately is more efficient computationally than calculating GBLUP for all individuals in the same system of equations by applying Henderson's BLUP for animals without phenotypic observations [32] to GBLUP. Let n_1 = number of individuals with phenotypic observations, n_0 = number of individuals without phenotypic observations, $n = n_1 + n_0$, and let the $\mathbf{S}_i$ matrix be partitioned as:

$$\mathbf{S}_i = \begin{bmatrix} \mathbf{S}_{i11} & \mathbf{S}_{i10} \\ \mathbf{S}_{i01} & \mathbf{S}_{i00} \end{bmatrix}, \quad i=1,...,f \tag{22}$$

where $\mathbf{S}_{i11} = n_1 \times n_1$ genomic relationship matrix of the genetic values of the i^{th} effect type for individuals with phenotypic observations, $\mathbf{S}_{i01} = n_0 \times n_1 =$ genomic relationship matrix of the genetic values of the i^{th} effect type between individuals without phenotypic observations and individuals with phenotypic observations, $\mathbf{S}_{i10} = \mathbf{S}_{i01}' = n_1 \times n_0 =$ genomic relationship matrix between individuals with phenotypic observations and individuals without phenotypic observations, and $\mathbf{S}_{i00} = n_0 \times n_0$ genomic relationship matrix of the genetic values of the i^{th} effect type for individuals without phenotypic observations. In Equations 16 and 17, $\mathbf{y} = \mathbf{y}_1$, and the $\mathbf{Z}$ matrix needs to be changed to $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{0}]$, the $\mathbf{u}_i$ vector partitioned as $\mathbf{u}_i = [\mathbf{u}_{i1}' \quad \mathbf{u}_{i0}']'$, and the $\mathbf{g}$ vector partitioned as $\mathbf{g} = [\mathbf{g}_1' \quad \mathbf{g}_0']$, where $\mathbf{Z}_1 = N \times n_1$ incidence matrix allocating phenotypic observations to individuals with phenotypic observations, $\mathbf{0} = N \times n_0$ incidence matrix with elements '0' connecting phenotypic observations to individuals without phenotypic observations. With these changes and Equation 22, the $\mathbf{V}$ matrix of Equation 17 can be re-written as:

$$\mathbf{V} = \mathbf{Z}_1(\sum_{i=1}^{f}\mathbf{G}_i)\mathbf{Z}_1' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}_1(\sum_{i=1}^{f}\sigma_i^2\mathbf{S}_{i11})\mathbf{Z}_1' + \sigma_e^2\mathbf{I}_N \quad (23)$$

and the GBLUP and reliability for individuals with and without phenotypic observations can be calculated as:

$$\hat{\mathbf{u}}_{i1} = \sigma_i^2\mathbf{S}_{i11}\mathbf{Z}_1'\mathbf{V}^{-1}(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{b}}) = \sigma_i^2\mathbf{S}_{i11}\mathbf{Z}_1'\mathbf{P}\mathbf{y}_1, \quad i = 1,...,f \quad (24)$$

$$\hat{\mathbf{g}}_1 = \sum_{i=1}^{f}\hat{\mathbf{u}}_{i1} \quad (25)$$

$$R_{g1j}^2 = [\mathbf{G}_{11}(\mathbf{Z}_1'\mathbf{P}\mathbf{Z}_1')\mathbf{G}_{11}]^{jj} / \mathbf{G}_{11}^{jj} \quad (26)$$

$$\hat{\mathbf{u}}_{i0} = \sigma_i^2\mathbf{S}_{i01}\mathbf{Z}_1'\mathbf{V}^{-1}(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{b}}) = \sigma_i^2\mathbf{S}_{i01}\mathbf{Z}_1'\mathbf{P}\mathbf{y}_1, \quad i = 1,...,f \quad (27)$$

$$= \sigma_i^2\mathbf{S}_{i01}\mathbf{S}_{i11}^{-1}\mathbf{S}_{i11}\mathbf{Z}_1'\mathbf{P}\mathbf{y}_1 = \mathbf{G}_{i01}\mathbf{G}_{i11}^{-1}\mathbf{G}_{i11}\mathbf{Z}_1'\mathbf{P}\mathbf{y}_1 = \mathbf{G}_{i01}\mathbf{G}_{i11}^{-1}\hat{\mathbf{u}}_{i1}, \quad i = 1,...,f \quad (28)$$

$$\hat{\mathbf{g}}_0 = \sum_{i=1}^{f}\hat{\mathbf{u}}_{i0} \quad (29)$$

$$R^2_{g0j} = [\mathbf{G}_{01}(\mathbf{Z}_1'\mathbf{P}\mathbf{Z}_1')\mathbf{G}_{10}]^{jj} / \mathbf{G}^{jj}_{00} \quad (30)$$

where $\hat{\mathbf{u}}_{i1}$ = $n_1 \times 1$ column vector of the GBLUP of the genetic values of the i^{th} effect type for individuals with phenotypic observations, $\hat{\mathbf{g}}_1$ = $n_1 \times 1$ column vector of the GBLUP of the total genetic values for individuals with phenotypic observations, R^2_{g1j} = reliability for the j^{th} individuals with phenotypic observations, $\hat{\mathbf{u}}_{i0}$ = $n_0 \times 1$ column vector of the GBLUP of the genetic values of the i^{th} effect type for individuals without phenotypic observations $\hat{\mathbf{g}}_0$ = $n_0 \times 1$ column vector of the GBLUP of the total genetic values for individuals without phenotypic observations, R^2_{g0j} = reliability for the j^{th} individuals without phenotypic observations, $\mathbf{G}_{11} = \sum_{i=1}^{f} \mathbf{G}_{i11} = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_{i11}$, $\mathbf{G}_{01} = \sum_{i=1}^{f} \mathbf{G}_{i01} = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_{i01}$, $\mathbf{G}_{10} = \sum_{i=1}^{f} \mathbf{G}_{i10} = \sum_{i=1}^{f} \sigma_i^2 \mathbf{S}_{i10}$, $\mathbf{G}^{jj}_{11} = \sum_{i=1}^{f} \mathbf{S}^{jj}_{i11} \sigma_i^2$, and $\mathbf{G}^{jj}_{00} = \sum_{i=1}^{f} \mathbf{S}^{jj}_{i00} \sigma_i^2$.

Equations 27 and 28 yield identical results if $\mathbf{S}^{-1}_{i11}$ exists. However, when the number of individuals is greater than the number of effect levels such as the number of SNPs, $\mathbf{S}^{-1}_{i11}$ in Equation 28 does not exist and Equation 27 still can calculate the GBLUP. The usefulness of Equation 28 is showing the GBLUP of individuals without phenotypic observations is the regression of the genetic values of individuals without phenotypic observations on the genetic values of individuals with phenotypic observations. The advantage of Equation 27 is that it does not calculate $\mathbf{S}^{-1}_{i11}$ and hence is unaffected by the singularity of $\mathbf{S}_{i11}$. Therefore, Equation 27 is recommended for calculating GBLUP for individuals without phenotypic observation when the number of such individuals is large. The GBLUP calculations of Equations 24, 27 and 28 do not involve the genomic relationship matrix among individuals without phenotypic observations $\mathbf{S}_{i00}$, which is much larger than $\mathbf{S}_{i11}$ when n_1 is much larger than n_0. The reliability calculation for individuals

without phenotypic observations (Equation 30) only uses the diagonal elements of $\mathbf{S}_{i00}$, not the entire $\mathbf{S}_{i00}$.

Advantage of integrated model over separate models

The multifactorial model of Equations 16 and 17 integrating SNP, haplotype and epistasis effects have the advantage of using more effect types and assessing each effect type based on the phenotypic values adjusted for all remaining effect types over separate models for SNP, haplotype and epistasis effects that do not have a mechanism to adjust for effect types not in the model and each uses a smaller number of genetic effects in the model.

This advantage of the multifactorial model assessing each effect type based on the phenotypic values adjusted for all remaining effect types can be shown using the MME version of the GBLUP for the i^{th} effect type, i.e.,

$$\begin{aligned}\hat{\mathbf{u}}_i &= (\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{G}_i^{-1})^{-1}[\mathbf{Z}_i'\mathbf{y} - (\mathbf{Z}_i'\mathbf{X}\hat{\mathbf{b}} + \sum_{\substack{j=1 \\ j'i}}^{f}\mathbf{Z}_i'\mathbf{Z}_j\hat{\mathbf{u}}_j)] \\ &= (\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{G}_i^{-1})^{-1}\mathbf{Z}_i'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_{\substack{j=1 \\ j'i}}^{f}\mathbf{Z}_j\hat{\mathbf{u}}_j) = (\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{G}_i^{-1})^{-1}\mathbf{Z}_i'\mathbf{y}_{bu}^*\end{aligned} \quad (31)$$

$$\begin{aligned}\hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i) \\ &= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{y} - \sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}_u^*\end{aligned} \quad (32)$$

where $\mathbf{y}_{bu}^* = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_{\substack{j=1 \\ j'i}}^{f}\mathbf{Z}_j\hat{\mathbf{u}}_j$ = phenotypic observations adjusted for the fixed effects and all random genetic values except those of $\hat{\mathbf{u}}_i$, $\mathbf{y}_u^* = \mathbf{y} - \sum_{i=1}^{f}\mathbf{Z}_i\hat{\mathbf{u}}_i$ = phenotypic observations adjusted for all random genetic values, and $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. Equation 31 shows the MME version of $\hat{\mathbf{u}}_i$ uses the phenotypic values adjusted for the GBLUP of all other effect types in the model. Since the MME version of $\hat{\mathbf{u}}_i$ (Equation 31) and $\hat{\mathbf{b}}$ (Equation 32) are identical to the CE version of $\hat{\mathbf{u}}_i$ (Equation 18) and $\hat{\mathbf{b}}$ (Equation 19), the CE version of $\hat{\mathbf{u}}_i$ (Equation 18)

417 uses the phenotypic values adjusted for the GBLUP of all other effect types in the model even
418 though the CE version does not do such adjustments explicitly.

419 **Genomic restricted maximum estimation (GREML) of variances and heritabilities**

420 The estimation of variance components uses GREML and a combination of EM-REML and AI-
421 REML algorithms of iterative solutions. EM-REML is slow but converges whereas AI-REML is
422 fast but fails for zero heritability estimates. In our GVCBLUP and GVCHAP computing packages
423 that implement these two algorithms [30, 31], EM-REML is used automatically when AI-REML
424 fails. The EM-REML iterative algorithm for the multifactorial model of Equations 16 and 17 is:

425 $$\sigma_i^{2(j+1)} = \sigma_i^{2(j)}\mathbf{y}\mathbf{P}^{(j)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\mathbf{P}^{(j)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(j)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'),\ i=1,...,f \quad (33)$$

426 $$\sigma_e^{2(j+1)} = \sigma_e^{2(j)}\mathbf{y}\mathbf{P}^{(k)}\mathbf{P}^{(j)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(j)}) \quad (34)$$

427 where j = iteration number. The AI-REML iterative algorithm is an extension of the early
428 formulations [33, 34] to the multifactorial model of Equations 16 and 17:

429 $$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \left(\mathbf{AI}^{(j)}\right)^{-1}\boldsymbol{\Delta}^{(j)} \quad (35)$$

430 where $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, ..., \sigma_f^2, \sigma_{f+1}^2)'$ = $(f+1)\times 1$ column vector of variance-covariance components,
431 $\sigma_{f+1}^2 = \sigma_e^2$ = residual variance, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, ..., \Delta_f, \Delta_{f+1})'$ = $(f+1)\times 1$ column vector of the partial
432 derivatives of the log residual likelihood function with respect to each variance component, and j
433 = iteration number. A typical term in $\boldsymbol{\Delta}$ (Δ_i) and a typical term in **AI** (AI_{ik}) are:

434 $$\begin{aligned}\Delta_i &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\right) + \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\mathbf{y} \\ &= -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{Z}\mathbf{S}_i\mathbf{Z}') + \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{y}, \quad i=1,...,f+1\end{aligned} \quad (36)$$

435 $$\begin{aligned}\mathrm{AI}_{ik} &= \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_k^2}\mathbf{P}\mathbf{y} \\ &= \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}'\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{S}_k\mathbf{Z}'\mathbf{P}\mathbf{y}, \quad i, k=1,...,f+1\end{aligned} \quad (37)$$

where $\mathbf{S}_{f+1} = \mathbf{I}_N$. For the full Model-I or Model-II, some effect types inevitably may have zero variances. In those cases, AI-REML (Equations 35-37) fails, and EM-REML (Equations 33 and 34) still converges although slow convergence rate can be expected for the full Model-I or Model-II. Once the effect types with zero variances are removed from the model, AI-REML converges, and fast convergence rate can be expected. The estimate of the genomic heritability for each type of genetic effects (h_i^2) and the total heritability of all types of genetic effects (H^2) are:

$$h_i^2 = \sigma_i^2 / \sigma_y^2 \qquad i = 1,...,f \tag{38}$$

$$H^2 = \sum_{i=1}^{f} h_i^2 \tag{39}$$

where $\sigma_y^2 = \sum_{i=1}^{f} \sigma_i^2 + \sigma_e^2$ = phenotypic variance.

The heritability estimates of Equation 38 can be used for model selection by removing effect types with heritability estimates below a user determined threshold value from the prediction model. Since different traits may have different genetic architectures, we hypothesize that some traits may involve only a small number of the effect types and some traits are more complex and involve more effect types, global epistasis may be more important than local high-order epistasis effects of haplotypes for some traits whereas the reverse may be true for other traits, and some traits may be affected by both global high-order and local high-order epistasis effects. The heritability estimates from Equation 37 provide an approach to evaluate these hypotheses and identify effect types relevant to the phenotypic variance whereas the total heritability of Equation 38 provides an estimate of the total genetic contribution to the phenotypic variance. In addition to the use of heritability estimates, prediction accuracy based on GBLUP can be used for model selection by requiring a threshold accuracy level for the effect type to be included in the prediction model, e.g., we identified the A + A×A model to have the same accuracy of predicting the phenotypic values of daughter pregnancy rate as the full Model-I in U.S. Holstein cows [15].

Estimation of pairwise epistasis effect and heritability

The heritability of a SNP, haplotype block or pairwise epistasis effect is the contribution of the genetic effect to the phenotypic variance and is also the contribution to the heritability of the effect type, and is estimated through the GBLUP of the corresponding genetic effects. These heritability estimates can be used to identify genome locations with large contributions to the phenotypic variance. The estimation of pairwise epistasis effects and heritability is the most demanding computing because the pairwise epistasis model matrices must be creased and are no longer avoidable. Estimating the effects and heritabilities for third-order epistasis effects is computationally unfeasible and is not considered. GBLUP of SNP, haplotype and pairwise epistasis effects of Model-I (**Supplementary Table S1**) are calculated as:

$$\hat{\boldsymbol{\tau}}_i = \sigma_i^2 \mathbf{T}_i' \mathbf{Z}' \mathbf{P} \mathbf{y} = \mathbf{T}_i' \mathbf{S}_i^{-1} \hat{\mathbf{u}}_i, \quad i = 1\text{-}6 \tag{40}$$

where $\hat{\boldsymbol{\tau}}_i$ is the $m \times 1$ column vector of SNP additive effects for $i = 1$, or SNP dominance effects for $i = 2$; or $b \times 1$ column vector of haplotype additive effects for $i = 3$; or $\binom{m}{2} \times 1$ column vector of A×A epistasis effects for $i = 4$, or $2\binom{m}{2} \times 1$ column vector of A×D epistasis effects for $i = 5$, or $\binom{m}{2} \times 1$ column vector of D×D epistasis effects for $i = 6$. For $i = 5$, the order of A×D and D×A effects is determined by the order of the model matrices of those effects, i.e., $\hat{\boldsymbol{\tau}}_5 = (\hat{\boldsymbol{\tau}}_{\alpha\delta}', \hat{\boldsymbol{\tau}}_{\delta\alpha}')'$ if $\mathbf{T}_5 = (\mathbf{T}_{\alpha\delta}, \mathbf{T}_{\delta\alpha})$, or $\hat{\boldsymbol{\tau}}_5 = (\hat{\boldsymbol{\tau}}_{\delta\alpha}', \hat{\boldsymbol{\tau}}_{\alpha\delta}')'$ if $\mathbf{T}_5 = (\mathbf{T}_{\delta\alpha}, \mathbf{T}_{\alpha\delta})$. The heritability of the j^{th} effect of the i^{th} effect type ($\hat{h}_{ij}^2$) is estimated as a faction of the genomic heritability of the i^{th} effect type ($\hat{h}_i^2$):

$$\hat{h}_{ij}^2 = (\hat{\tau}_{ij}^2 / \sum_{i=1}^{m} \hat{\tau}_{ij}^2) \hat{h}_i^2 = (\hat{\tau}_{ij}^2 / \hat{\boldsymbol{\tau}}_i' \hat{\boldsymbol{\tau}}_i) \hat{h}_i^2 = \hat{\sigma}_{ij}^2 / \hat{\sigma}_y^2 \tag{41}$$

where $\hat{\boldsymbol{\tau}}_{ij}$ = the j^{th} effect of $\hat{\boldsymbol{\tau}}_i$ $i = 1\text{-}6$; $\hat{\sigma}_i^2$ = estimated variance of the i^{th} effect type; $\hat{\sigma}_{ij}^2$ = estimated variance of the j^{th} effect of the i^{th} effect type; $\hat{h}_i^2$ = the genomic heritability of the i^{th}

effect type defined by Equation (52). For proving Equation 57, $\hat{\sigma}_i^2$ and $\hat{\sigma}_{ij}^2$ can be formulated based on the method of mixed model equations (MME), i.e.,

$$\hat{\sigma}_i^2=\hat{\boldsymbol{\tau}}_i'\hat{\boldsymbol{\tau}}_i/[m_i-tr(\mathbf{C}^{ii})\lambda_i]=\sum_{j=1}^{m_i}\tau_{ij}^2/[m_i-tr(\mathbf{C}^{ii})\lambda_i]=\sum_{j=1}^{m_i}\hat{\sigma}_{ij}^2 \quad (42)$$

$$\hat{\sigma}_{ij}^2=\hat{\tau}_{ij}^2/[m_i-tr(C^{ii})\lambda_i] \quad (43)$$

where $\mathbf{C}^{ii}$ is the submatrix in the inverse or generalized inverse of the coefficient matrix of the MME corresponding to the i^{th} effect type, m_i = number of effects of the i^{th} effect type, and $\lambda_i=\hat{\sigma}_e^2/\hat{\sigma}_i^2$. Dividing Equation 43 by $\hat{\sigma}_y^2$ and multiplying by $\hat{\sigma}_i^2/\hat{\sigma}_i^2$ yield Equation 41, i.e.,

$$\hat{h}_{ij}^2=(\hat{\sigma}_{ij}^2/\hat{\sigma}_y^2)(\hat{\sigma}_i^2/\hat{\sigma}_i^2)=(\hat{\sigma}_{ij}^2/\hat{\sigma}_i^2)(\hat{\sigma}_i^2/\hat{\sigma}_y^2)=(\hat{\tau}_{ij}^2/\sum_{i=1}^{m}\hat{\tau}_{ij}^2)\hat{h}_i^2=(\hat{\tau}_{ij}^2/\hat{\boldsymbol{\tau}}_i'\hat{\boldsymbol{\tau}}_i)\hat{h}_i^2=(\hat{\sigma}_{ij}^2/\hat{\sigma}_y^2).$$

It is readily seen that the sum of all heritability estimates of the i^{th} effect type is the genomic heritability of the i^{th} effect type, i.e., $\sum_{i=1}^{m_i}\hat{h}_{ij}^2=\hat{h}_i^2$. Note that Equations 42 and 43 using MME are only for proving Equation 41. The MME method is computationally prohibitive for estimating genetic effects and their variances under the multifactorial model although the MME method yield identical results as the CE method, which is computationally feasible for genomic estimation and prediction under the multifactorial model.

Comparison between exact and approximate genomic epistasis relationship matrices

We evaluated the differences between AGERM and EGERM in genomic heritability estimates and prediction accuracies using a publicly available swine genomics data set that had 3534 animals from a single PIC nucleus pig line with five anonymous traits and 52,842 genotyped and imputed autosome SNPs after filtering by requiring minor allele frequency (MAF) > 0.001 and proportion of missing SNP genotypes < 0.100 [35]. The EGERM used the method of Jiang and Reif [13] and

the AGERM methods were described in **Supplementary Text 1**. The heritability results showed that EGERM had slightly higher heritability estimates than AGERM except the A×A heritability of T3 where AGERM had slightly high estimate than EGERM (0.280 vs. 0.278, **Table 1**). From **Table 1**, effect type with nonzero heritability estimates was included in the prediction model for evaluating the observed prediction accuracy as the correlation between the GBLUP of genotypic values and the phenotypic values in each validation population and then averaged over all 10 validation populations. The results showed that AGERM and EGERM had the same prediction accuracy for this swine sample (**Table 2**). A disadvantage of EGERM is the computing time for the construction of EGERM, about 9.51 times as much time for pairwise relationship matrices, 8.29 as much time for third-order and 9.44 times as much time for fourth-order as required for AGERM (**Table 3**). However, computing time is not the deciding factor for choosing between the exact and approximate methods, because the multi-node approach that calculate each genomic relationship matrix in pieces and adds those pieces together can reduce the computing time to an acceptable level when multiple threads/cores are available and the two-step strategy can be used so that each genomic relationship is calculated only once for different traits and validation populations [31]. Prediction accuracy is the ultimate deciding factor for choosing between different methods. We reported results of comparing AGERM and EGERM using 60,671 SNPs and 22,022 first-lactation Holstein cows with phenotypic observations of daughter pregnancy rate, showing that AGERM and EGERM had the same heritability estimates and prediction accuracy, but EGERM required 21 times as much computing time as required by AGERM, which required 1.32 times as much time for the genomic additive relationship matrix [15]. The combined results of the swine and Holstein samples indicated that EGERM and AGERM had similar results and that the computing difficulty of EGERM over AGERM increased rapidly as the sample size increased. Given the computing difficulty of EGERM and the negligible differences between EGERM and

AGERM in prediction accuracy, AGERM should be favored for its mathematical simplicity and computing efficiency at least for samples with 50,000 SNPs or more.

Numerical demonstration

The methods of genomic epistasis relationship matrices based on the additive and dominance model matrices, GREML, GBLUP and reliability, and estimation of effect heritability are demonstrated using a R program (DEMO.R) and a small artificial sample for the convenience of reading the numerical results (**Supplementary Text S2 and R program**). Because of the artificial nature and the extremely small sample size, this numerical demonstration does not have any genetic and methodology implications and is for showing calculations of the methods only. This R program is an extension of the R demo program of GVCHAP that integrates SNP and haplotype effects and has a computing pipeline for producing the input haplotype data from the SNP data [31].

CONCLUSION

The multifactorial methods with SNP, haplotype and epistasis effects up to the third-order provide an approach to investigate the contributions of global low-order and local high-order epistasis effects to the phenotypic variance and the accuracy of gnomic prediction. Genomic heritability of each effect type from GREML and prediction accuracy from validation studies using GBLUP can be used jointly to identify effect types contributing to the phenotypic variance and the accuracy of genomic prediction, and the GBLUP for the multifactorial model with selected effect type can be used for genomic evaluation. With many capabilities including the use of intra- and inter-chromosome separately, the multifactorial methods offer a significant methodology capability to investigate and utilize complex genetic mechanisms for genomic prediction and for understanding the complex genome-phenome relationships.

AUTHOR CONTRIBUTIONS

YD conceived this study and derived the formulations. ZL contributed to formulations of the epistasis genomic relationships, implemented the epistasis methods in EPIHAP, validated and evaluated the methods. DP contributed to the data processing for methodology evaluation. YD and ZL prepared the manuscript.

FUNDING

This research was supported by the National Institutes of Health's National Human Genome Research Institute, grant R01HG012425 as part of the NSF/NIH Enabling Discovery through GEnomics (EDGE) Program; grant 2020-67015-31133 from the USDA National Institute of Food and Agriculture; and project MIN-16-124 of the Agricultural Experiment Station at the University of Minnesota. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

The supercomputer of Minnesota Supercomputer Institute at the University of Minnesota and the Ceres and Atlas high performance computing system of USDA/ARS were used for the evaluation and testing of the methods and EPIHAP computing package. The use of the Ceres and Atlas computers by this research were supported by USDA-ARS projects 8042-31000-002-00-D and 8042-31000-001-00-D. The authors thank Paul VanRaden, Steven Schroeder and Ransom Baldwin for help with the use of the USDA-ARS computing facilities.

SUPPLEMENTARY MATERIAL

Text S1. Quantitative Genetics Models and Genomic Epistasis Relationship Matrices

Text S2. Numerical Demonstration

R program. DEMO.R for Numerical Demonstration

TABLE S1 | Notations of the quantitative genetics (QG) model, reparameterized and equivalent QG (RE-QG) model, and multifactorial (MF) model.

REFERENCES

1. Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics. 1954;39(6):859.
2. Kempthorne O. The correlation between relatives in a random mating population. Proceedings of the Royal Society of London Series B-Biological Sciences. 1954;143(910):103-13.
3. Henderson C. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of Animal Science. 1985;60(1):111-7.
4. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS One. 2012;7(9):e45293.
5. Muñoz PR, Resende MF, Gezan SA, Resende MDV, de los Campos G, Kirst M, et al. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics. 2014;198(4):1759-68.
6. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics. 2017;206(3):1297-307.
7. VanRaden PM. Efficient methods to compute genomic predictions. Journal of Dairy Science. 2008;91(11):4414-23.
8. Hayes B, Goddard M. Genome-wide association and genomic selection in animal breeding. Genome. 2010;53(11):876-83.

9. Da Y, Wang C, Wang S, Hu G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS One. 2014;9(1):e87666.

10. Wang C, Da Y. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. PLoS One. 2014;9(12):e114484.

11. Martini JW, Toledo FH, Crossa J. On the approximation of interaction effect models by Hadamard powers of the additive genomic relationship. Theoretical population biology. 2020;132:16-23.

12. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201(2):759-68.

13. Jiang Y, Reif JC. Efficient algorithms for calculating epistatic genomic relationship matrices. Genetics. 2020;216(3):651-69.

14. Martini JW, Wimmer V, Erbe M, Simianer H. Epistasis and covariance: how gene interaction translates into genomic relationship. Theoretical and Applied Genetics. 2016;129(5):963-76.

15. Liang Z, Prakapenka D, Da Y. Comparison of two methods of genomic epistasis relationship matrices using daughter pregnancy rate in U.S. Holstein cattle. Abstract 2466V, page 409 of ADSA2022 Abstracts. https://www.adsa.org/Portals/0/SiteContent/Docs/Meetings/2022ADSA/Abstracts_BOOK_2022.pdf?v=20220613. 2022.

16. Liang Z, Prakapenka D, Da Y. EPIHAP: a computing tool for genomic estimation and prediction using global epistasis effects and haplotype effects. Abstract P167, page 223 of ADSA2021 Abstracts, ADSA 2021 Virtual Annual Meeting. https://www.adsa.org/Portals/0/SiteContent/Docs/Meetings/2021ADSA/ADSA2021_Abstracts.pdf. 2021.

17. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 2008;9(11):855-67.

18. Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics. 2004;5(8):618-25.

19. Ritchie MD, Van Steen K. The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. Annals of translational medicine. 2018;6(8):157.

20. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nature Reviews Genetics. 2014;15(1):22-33.

21. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human molecular genetics. 2002;11(20):2463-8.

22. Segre D, DeLuna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. Nature genetics. 2005;37(1):77-83.

23. Liang Z, Tan C, Prakapenka D, Ma L, Da Y. Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. Frontiers in Genetics. 2020;11(1461):588907. doi: 10.3389/fgene.2020.588907.

24. Bian C, Prakapenka D, Tan C, Yang R, Zhu D, Guo X, et al. Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. Genetics Selection Evolution. 2021;53(1):1-19.

25. Sallam AH, Conley E, Prakapenka D, Da Y, Anderson JA. Improving prediction acuracy uing mlti-allelic hplotype pediction and taining ppulation otimization in weat. G3: Genes, Genomes, Genetics. 2020;10(7):2265-73.

26. Prakapenka D, Liang Z, Jiang J, Ma L, Da Y. A Large-scale genome-wide association study of epistasis effects of production traits and daughter pregnancy rate in US Holstein cattle. Genes. 2021;12(7):1089.

27. Da Y. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. BMC Genetics. 2015;16(1):144.

28. Da Y, Tan C, Parakapenka D. Joint SNP-haplotype analysis for genomic selection based on the invariance property of GBLUP and GREML to duplicate SNPs. Journal of Animal Science. 2016;94(Suppl_5):161-2.

29. Tan C, Wu Z, Ren J, Huang Z, Liu D, He X, et al. Genome-wide association study and accuracy of genomic prediction for teat number in Duroc pigs using genotyping-by-sequencing. Genetics Selection Evolution. 2017;49(1):35.

30. Wang C, Prakapenka D, Wang S, Pulugurta S, Runesha HB, Da Y. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. BMC Bioinformatics. 2014;15(1):270.

31. Prakapenka D, Wang C, Liang Z, Bian C, Tan C, Da Y. GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. Frontiers in Genetics. 2020;11:282. doi: 10.3389/fgene.2020.00282.

32. Henderson C. Best linear unbiased prediction of breeding values not in the model for records. Journal of Dairy Science. 1977;60(5):783-7.

33. Johnson D, Thompson R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of dairy science. 1995;78(2):449-56.

34. Lee SH, van der Werf JH. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genetics Selection Evolution. 2006;38(1):1-19.

35. Cleveland MA, Hickey JM, Forni S. A common dataset for genomic analysis of livestock populations. G3: Genes| Genomes| Genetics. 2012;2(4):429-35.

Table 1 | Genomic heritability estimates of additive, dominance and epistasis effects up to the third-order for five traits in a swine population

	Trait				
	T1	**T2**	**T3**	**T4**	**T5**
Effect	Exact genomic epistasis relationship matrices (EGERM)				
A	0.023	0.217	0.131	0.336	0.366
D	0.000	0.013	0.000	0.000	0.052
A×A	0.046	0.186	0.278	0.017	0.054
A×D	0.000	0.000	0.091	0.000	0.000
D×D	0.000	0.000	0.091	0.000	0.000
A×A×A	0.000	0.000	0.000	0.000	0.000
A×A×D	0.000	0.000	0.079	0.000	0.000
A×D×D	0.000	0.000	0.102	0.000	0.000
D×D×D	0.000	0.000	0.117	0.000	0.000
Total heritability	0.069	0.416	0.889	0.354	0.471
Effect	Approximate genomic epistasis relationship matrices (AGERM)				
A	0.022	0.215	0.139	0.329	0.360
D	0.000	0.013	0.000	0.000	0.051
A×A	0.043	0.176	*0.280*	0.016	0.050
A×D	0.000	0.000	0.091	0.000	0.000
D×D	0.000	0.000	0.090	0.000	0.000
A×A×A	0.000	0.000	0.000	0.000	0.000
A×A×D	0.000	0.000	0.075	0.000	0.000
A×D×D	0.000	0.000	0.095	0.000	0.000
D×D×D	0.000	0.000	0.109	0.000	0.000
Total heritability	0.065	0.404	0.879	0.346	0.461

Table 2 | Observed prediction accuracy of epistasis models relative to the additive model for five traits in a swine population

	Trait				
	T1	**T2**	**T3**	**T4**	**T5**
Prediction accuracy of SNP model					
A	0.066	0.495	0.326	0.468	0.493
A+D	0.056	0.495	0.326	0.468	0.496
Epistasis model	A+AA	A+D+AA	A+AA+AD+DD+ AAD+ADD+DDD	A+AA	A+D+AA
EGERM					
Prediction accuracy	0.063	0.498	0.336	0.468	0.497
Accuracy increase (%)	−4.545	0.606	3.067	0.000	0.202
AGERM					
Prediction accuracy	0.063	0.498	0.336	0.468	0.497
Accuracy increase (%)	−4.545	0.606	3.067	0.000	0.202

'Prediction accuracy' is the observed prediction accuracy calculated as the correlation between the GBLUP of genotypic values and the phenotypic values in each validation population and then averaged over all 10 validation populations. 'Accuracy increase' is percentage increase of the observed prediction accuracy of the epistasis model over the observed prediction accuracy of the best SNP model, which was the additive model (A) for T1-T4 and the A+D model for T5. A = additive effects, D = dominance effects, AA = A×A effects, AD = A×D effects, DD = D×D effects, AAA = A×A×A effects, AAD = A×A×D effects, ADD = A×D×D dominance effects, DDD = D×D×D dominance effects.

Table 3 | Computing time (in seconds) for the construction of exact and approximate genomic epistasis relationship matrices for a swine population with 3534 pigs and 52,843 SNPs using 20 threads of the Mangi supercomputer of the Minnesota Supercomputer Institute at the University of Minnesota.

Genomic epistasis relationship matrices	Pairwise	Third_order	Fourth_order
EGERM	666	796	1256
AGERM	70	96	133
EGERM/AGERM	9.51	8.29	9.44