Mixed Model Methods for Genomic Prediction and Variance Component Estimation of Additive and Dominance Effects Using SNP Markers

Yang Da*, Chunkao Wang, Shengwen Wang, Guo Hu

Department of Animal Science, University of Minnesota, Saint Paul, Minnesota, United States of America

Abstract

We established a genomic model of quantitative trait with genomic additive and dominance relationships that parallels the traditional quantitative genetics model, which partitions a genotypic value as breeding value plus dominance deviation and calculates additive and dominance relationships using pedigree information. Based on this genomic model, two sets of computationally complementary but mathematically identical mixed model methods were developed for genomic best linear unbiased prediction (GBLUP) and genomic restricted maximum likelihood estimation (GREML) of additive and dominance effects using SNP markers. These two sets are referred to as the CE and QM sets, where the CE set was designed for large numbers of markers and the QM set was designed for large numbers of individuals. GBLUP and associated accuracy formulations for individuals in training and validation data sets were derived for breeding values, dominance deviations and genotypic values. Simulation study showed that GREML and GBLUP generally were able to capture small additive and dominance effects that each accounted for 0.00005–0.0003 of the phenotypic variance and GREML was able to differentiate true additive and dominance heritability levels. GBLUP of the total genetic value as the summation of additive and dominance effects had higher prediction accuracy than either additive or dominance GBLUP, causal variants had the highest accuracy of GREML and GBLUP, and predicted accuracies were in agreement with observed accuracies. Genomic additive and dominance relationship matrices using SNP markers were consistent with theoretical expectations. The GREML and GBLUP methods can be an effective tool for assessing the type and magnitude of genetic effects affecting a phenotype and for predicting the total genetic value at the whole genome level.

Citation: Da Y, Wang C, Wang S, Hu G (2014) Mixed Model Methods for Genomic Prediction and Variance Component Estimation of Additive and Dominance Effects Using SNP Markers. PLoS ONE 9(1): e87666. doi:10.1371/journal.pone.0087666

Editor: Yi-Hsiang Hsu, Harvard Medical School, United States of America

Received July 7, 2013; **Accepted** December 29, 2013; **Published** January 30, 2014

Funding: This research was supported by USDA National Institute of Food and Agriculture Grant no. 2011-67015-30333 and by project MN-16-043 of the Agricultural Experiment Station at the University of Minnesota. The Holstein SNP data were supported by National Research Initiative Competitive Grant no. 2008-35205-18846 from the USDA National Institute of Food and Agriculture and by a financial contribution from Holstein Association USA. Supercomputer computing time was provided by the Minnesota Supercomputer Institute at the University of Minnesota. Guo Hu was supported by Heilongjiang River Fishery Research Institute of Chinese Academy of Fishery Sciences, Harbin, China. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yda@umn.edu

Introduction

Genomic prediction using genome-wide single nucleotide polymorphism (SNP) markers has been shown to be a powerful tool to capture small genetic effects dispersed over the genome for predicting an individual's genetic potential of a phenotype [1–5]. Current large scale genomic prediction focused on additive effects [2,4,5]. Two SNP models for genomic prediction of additive effects were described: a traditional quantitative genetics model and a model with (-1)-0–1 SNP coding [2]. The traditional quantitative genetics model is attractive because it is equivalent to a conventional animal model with the relationship matrix calculated from the SNP genotypes [5] and it directly predicts genomic breeding values [2,4,5]. Method and computing tool are available for estimating genomic heritability using genome-wide SNP markers [6]. This method uses a standardization of the 0–1–2 additive coding and the subtraction step of this standardization leads to additive effects that are breeding values under the traditional quantitative genetics model assuming Hardy-Weinberg equilibrium [2,6,7]. The mixed model implementation of this method is ideal for a large number of markers but is not ideal for a large number of individuals because the size of the matrix that needs to be inverted increases as the number of individuals increases.

From the point of view of missing heritability [8–10], the ability to estimate genome-wide dominance contribution will help determine the total genetic contribution to a phenotype. Similarly, methods of genomic prediction taking into account of dominance can predict an individual's total genetic potential for phenotypes affected by additive and dominance effects. Substantial dominance effect should justify the inclusion of dominance in genomic prediction and the design of mating systems to maximize dominance effect. In dairy cattle, dominance variances estimated from pedigree data were reported to be 11–16% of the phenotypic variance of stature [11], and the increased availability of cows with phenotypes and genotypes provides an opportunity to estimate dominance effects and include those in mating programs [12]. However, only limited methodology studies on genomic prediction and variance component estimation of dominance were available [13–16].

Genomic best linear unbiased prediction (GBLUP) and various Bayesian methods are available for genomic prediction, and GBLUP generally had good performance in real data [17]. Restricted maximum likelihood estimation (REML) [18] has been a widely accepted method for estimating variance components.

Objectives of this study were to develop mixed model methods for the joint genomic prediction of and variance component estimation of additive and dominance effects based on the traditional quantitative genetics model that partitions a genotypic value into breeding value and dominance deviation. The methodology will have two complimentary computing strategies for large numbers of individuals and markers, and the genomic prediction methods for have GBLUP and associated reliability for both training and validation data sets. Accuracies of the new methods will be evaluated using simulation data based a true dairy cattle SNP structure.

Methods

Genetic Model of SNP Markers and Mixed Model of Phenotypic Observations

The genetic model of SNP markers is an expansion of the additive model used in genomic evaluations [2,4,5] by adding a dominance component to the additive model. Using the traditional quantitative genetics model that partitions a genetic value into breeding value and dominance deviation under the assumption of Hardy-Weinberg equilibrium [7], the genetic value of each SNP marker can be expressed as:

$$g_{ij}=\mu+a_{ij}+d_{ij}=\mu+w_{\alpha ij}\alpha+w_{\delta ij}\delta \qquad (1)$$

where $g_{ij}=$genotypic value of SNP genotype $A_iA_j(i,j=1,2)$, $\mu=$common mean, $\alpha=$average effect of gene substitution, $\delta=$dominance effect, $a_{ij}=w_{\alpha ij}\alpha=$breeding value, $d_{ij}=w_{\delta ij}\delta=$ dominance deviation, $w_{\alpha 11}=2p_2$, $w_{\alpha 12}=p_2-p_1$, $w_{\alpha 22}=-2p_1$, $w_{\delta 11}=-2p_2^2$, $w_{\delta 12}=2p_1p_2$, $w_{\delta 22}=-2p_1^2$, and where $p_1=$fre$=$frequency of A_1 allele and $p_2=1-p_1=$frequency of A_2. Note that gene substitution effect (α) is a contrast of breeding values or a contrast of allelic effects, and dominance effect (δ) is a contrast of dominance deviations or a contrast of genotypic values (Text S1: Part A). In matrix notations, the genetic model of Equation 1 can be expressed as:

$$\begin{aligned}\begin{pmatrix} g_{11}\\ g_{12}\\ g_{22}\end{pmatrix}&=\begin{pmatrix}1\\1\\1\end{pmatrix}\mu+\begin{pmatrix} w_{\alpha 11}\\ w_{\alpha 12}\\ w_{\alpha 22}\end{pmatrix}\alpha+\begin{pmatrix} w_{\delta 11}\\ w_{\delta 12}\\ w_{\delta 22}\end{pmatrix}\delta\\ &=\begin{pmatrix}1\\1\\1\end{pmatrix}\mu+\begin{pmatrix} 2p_2\\ p_2-p_1\\ -2p_1\end{pmatrix}\alpha+\begin{pmatrix} -2p_2^2\\ 2p_1p_2\\ -2p_1^2\end{pmatrix}\delta\end{aligned} \qquad (2)$$

The quantitative genetics model of Equation 2 has the interpretation of 'breeding value' for additive effects. Assuming equal allele frequency and using a reparameterized μ, Equation 2 can achieve the (-1)-0-1 coding or 0-1-2 coding for additive effects and the 0-1-0 coding for dominance effects, but additive effects in those equal frequency models do not have the interpretation of 'breeding value' when the actual allele frequencies are unequal (Text S1: Part A). For each SNP marker, the variance of α and the variance of δ are assumed to be $Var(\alpha)=\sigma_\alpha^2$ and $Var(\delta)=\sigma_\delta^2$, and the covariance between α and δ is assumed null. Let $N=$the number of phenotypic observations, $q=$the number of individuals, $m=$the number of SNP markers, and $c=$the number of fixed effects. Based on Equation 2, the mixed model with SNP breeding values and dominance deviations can be expressed as:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{ZW}_\alpha\alpha+\mathbf{ZW}_\delta\delta+\mathbf{e} \qquad (3)$$

where $\mathbf{Z}=N\times q$ model matrix allocating phenotypic observations to SNP marker genotypes of individuals, $\mathbf{W}_\alpha=q\times m$ model matrix for gene substitution effects of SNP markers, $\alpha=$column vector of gene substitution effects of SNP markers, $\mathbf{W}_\delta=q\times m$ model matrix for dominance effects of SNP markers, $\delta=$column vector of dominance effects of SNP markers, $\mathbf{X}=N\times c$ model matrix for fixed non-genetic effects such as herd-year-season in dairy cattle, and $\mathbf{b}=$vector of fixed effects. Assumptions for the first and second moments are: $E(\mathbf{y})=\mathbf{Xb}$, $Var(\alpha)=\mathbf{I}_m\sigma_\alpha^2$, $Var(\delta)=\mathbf{I}_m\sigma_\delta^2$, and $Var(\mathbf{e})=\mathbf{R}=\mathbf{I}_N\sigma_e^2$, where $\sigma_e^2=$residual variance, $\mathbf{I}_m=m\times m$ identity matrix, and $\mathbf{I}_N=N\times N$ identity matrix. With the model and assumptions of Equations 1–3, methods for GBLUP and genomic variance component estimation using restricted maximum likelihood estimation (GREML) can be developed.

Results and discussion

Genomic Additive and Dominance Relationship Matrices

As the number of SNP markers increases, the values of the diagonal elements of $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ increase. Two methods to normalize the $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ matrices can be used. The first method divides $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ by the expected variance of the diagonal elements of each matrix (Definition I, [2]). The second method divides $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ by the average of the diagonal elements of $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ (Definition II, [4]), and we apply this method to $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ for defining dominance relationship matrix. In addition, we use a transformation to transform $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ into correlation matrices so that off-diagonal elements are mathematically comparable, and we refer to this definition as Definition III and refer to the resulting correlation matrices as genomic additive and dominance correlation matrices. The additive correlations of Definition III are the genomic version of Wright's coefficient of relationship [19]. Each of these three definitions of additive and dominance relationship or correlation matrices can be represented by two transformation matrices, $\mathbf{Q}_\alpha$ or $\mathbf{Q}_\delta$. Let $\mathbf{Q}_\alpha=diag\{(k_{\alpha ii})^{1/2}\}$ an $q\times q$ diagonal matrix, and $\mathbf{Q}_\delta=diag\{(k_{\delta ii})^{1/2}\}$ an $q\times q$ diagonal matrix, where $k_{\alpha ii}$ is the expected variance of the diagonal elements of $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $k_{\delta ii}$ is the expected variance of the diagonal elements of $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ for Definition I ($2\sum_{i=1}^{m}\mathrm{p}_i(1-\mathrm{p}_i)$ for additive relationships [2], and $4\sum_{i=1}^{m}\mathrm{p}_i^2(1-\mathrm{p}_i)^2$ for dominance relationships, personal communication from P. VanRaden to Y. Da, March 3, 2013), $k_{\alpha ii}$ is the average of diagonal elements of $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $k_{\delta ii}$ is the average diagonal elements of $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ for Definition II, and $k_{\alpha ii}$ is the i^{th} diagonal element of $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $k_{\delta ii}$ is the i^{th} diagonal element of $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ for Definition III. Then,

$$\mathbf{T}_\alpha=\mathbf{Q}_\alpha^{-1}\mathbf{W}_\alpha \qquad (4)$$

$$\mathbf{T}_\delta = \mathbf{Q}_\delta^{-1}\mathbf{W}_\delta \tag{5}$$

The additive relationship or correlation matrix ($\mathbf{A}_g$) and dominance relationship or correlation matrix ($\mathbf{D}_g$) can be expressed as.

$$\mathbf{A}_g = \mathbf{T}_\alpha\mathbf{T}_\alpha' = \mathbf{Q}_\alpha^{-1}\mathbf{W}_\alpha\mathbf{W}_\alpha'\mathbf{Q}_\alpha^{-1} \tag{6}$$

$$\mathbf{D}_g = \mathbf{T}_\delta\mathbf{T}_\delta' = \mathbf{Q}_\delta^{-1}\mathbf{W}_\delta\mathbf{W}_\delta'\mathbf{Q}_\delta^{-1} \tag{7}$$

In Equations 6–7, subscript 'g' is used to distinguish $\mathbf{A}_g$ and $\mathbf{D}_g$ from the $\mathbf{A}$ and $\mathbf{D}$ matrices calculated from pedigree data [20]. In addition to representing a number of definitions of genomic relationships, $\mathbf{T}_\alpha$ and $\mathbf{T}_\delta$ are used to define equivalent models to achieve computing efficiency.

Two Equivalent Mixed Models, Two Sets of Complementary Formulations

With the $\mathbf{T}$ matrices of Equations 4–5, two equivalent mixed models with complementary computing advantages, Model 1 and Model 2, can be defined. Model 1 can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZT}_\alpha\alpha + \mathbf{ZT}_\delta\delta + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zd} + \mathbf{e} \tag{8}$$

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{ZA}_g\mathbf{Z}'\sigma_\alpha^2 + \mathbf{ZD}_g\mathbf{Z}'\sigma_\delta^2 + \mathbf{I}_N\sigma_e^2 \tag{9}$$

where $\mathbf{a} = \mathbf{T}_\alpha\alpha =$ genomic breeding values,$\mathbf{d} = \mathbf{T}_\delta\delta =$ genomic dominance deviations, $\mathrm{Var}(\mathbf{a}) = \mathbf{A}_g\sigma_\alpha^2$, and $\mathrm{Var}(\mathbf{d}) = \mathbf{D}_g\sigma_\delta^2$. Model 2 can be rewritten as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\alpha + \mathbf{Z}_2\delta + \mathbf{e} \tag{10}$$

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}_1\mathbf{Z}_1'\sigma_\alpha^2 + \mathbf{Z}_2\mathbf{Z}_2'\sigma_\delta^2 + \mathbf{I}_N\sigma_e^2 \tag{11}$$

(11)where $\mathbf{Z}_1 = \mathbf{ZT}_\alpha$ and $\mathbf{Z}_2 = \mathbf{ZT}_\delta$. Model 1 of Equation 8 and Model 2 of Equation 10 have the same mathematical expectation, i.e., $E(\mathbf{y}) = \mathbf{Xb}$. The two equivalent models of Equations 8–11 can generate four sets of formulations with identical results of GBLUP, reliability and GREML. Each model can use a conditional expectation (CE) or mixed model equations (MME) to calculate GBLUP. The CE set of Model 1 is the best for large number of markers ($m>q$) and the MME set of Model 2, to be referred as the QM set (QM meaning $q>m$), is the best for large number of individuals ($q>m$). The MME set of Model 1 (to be referred to as MQ, with MQ meaning $m>q$) has no computing advantage because the matrix size is twice as large as that of CE and requires the inverses of the relationship matrices. The CE set of Model 2 (CE2) also has no computational advantage because CE2 requires more memory than QM if $m>q$. These two sets (MQ and CE2) are not considered further. In the following, we focus on the CE and QM sets of solutions, where each set consists of GBLUP, reliability of GBLUP and GREML formulations. We first present these three types of formulations in each set, CE for $m>q$ or QM for $q>m$, and then summarize the main features of the CE and QM sets.

GBLUP-CE, Reliability and GREML-CE for $m>q$

The CE form of GBLUP from Model 1 can be calculated as:

$$\hat{\mathbf{a}} = \sigma_\alpha^2\mathbf{A}_g\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_\alpha^2\mathbf{A}_g\mathbf{Z}'\mathbf{Py} \tag{12}$$

$$\hat{\mathbf{d}} = \sigma_\delta^2\mathbf{D}_g\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_\delta^2\mathbf{D}_g\mathbf{Z}'\mathbf{Py} \tag{13}$$

where $\hat{\mathbf{a}} =$ GBLUP of breeding values, $\hat{\mathbf{d}} =$ GBLUP of dominance deviations, $\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is best linear unbiased estimator (BLUE) of $\mathbf{b}$, $\mathbf{V}$ is defined by Equation 9, and

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1} \tag{14}$$

We refer to the GBLUP of Equations 12–13 as GBLUP-CE. The GBLUP of genotypic values is calculated as $\hat{\mathbf{g}} = \hat{\mathbf{a}} + \hat{\mathbf{d}}$. The reliability measures of $\hat{\mathbf{a}}$, $\hat{\mathbf{d}}$ and $\hat{\mathbf{g}}$ for individuals with phenotypic observations (individuals in training data set) are:

$$R_{ai}^2 = \sigma_\alpha^2\left(\mathbf{A}_g\mathbf{Z}'\mathbf{PZA}_g\right)_{ii}/a_{ii}\ ,$$

$$R_{di}^2 = \sigma_\delta^2\left(\mathbf{D}_g\mathbf{Z}'\mathbf{PZD}_g\right)_{ii}/d_{ii}\ ,$$

$$R_{gi}^2 = (\mathbf{G}_\alpha\mathbf{Z}'\mathbf{PZG}_\alpha + \mathbf{G}_\alpha\mathbf{Z}'\mathbf{PZG}_\delta + \mathbf{G}_\delta\mathbf{Z}'\mathbf{PZG}_\alpha + \mathbf{G}_\delta\mathbf{Z}'\mathbf{PZG}_\delta)_{ii} \\ /\left(a_{ii}\sigma_\alpha^2 + d_{ii}\sigma_\delta^2\right)$$

where $R_{ai}^2 =$ the reliability of GBLUP of breeding values ($\hat{\mathbf{a}}$) for individual i, $R_{di}^2 =$ the reliability of GBLUP for dominance deviations ($\hat{\mathbf{d}}$) for individual i, $R_{gi}^2 =$ the reliability of GBLUP for genotypic values ($\hat{\mathbf{g}}$), $a_{ii} =$ diagonal element i of $\mathbf{A}_g$, $d_{ii} =$ diagonal element i of $\mathbf{D}_g$, $\mathbf{G}_\alpha = \sigma_\alpha^2\mathbf{A}_g$, $\mathbf{G}_\delta = \sigma_\delta^2\mathbf{D}_g$, and $\mathbf{P}$ is given by Equation 14. Note that $a_{ii} = d_{ii} = 1$ for Definition III but the a_{ii} and d_{ii} values generally are not '1′ for Definitions I and II. The average of a_{ii} values and the average of d_{ii}values are '1′ under Definitions II and III, and are expected to be '1′ under Definition I although the observed average a_{ii} and d_{ii}values under Definition I may deviate from '1′. For individuals without phenotypic observations (individuals in validation data set), formulations of GBLUP-CE and associated reliability measures are given in Text S1: Part B. GREML-CE via the EM type algorithm [20–22] are:

$$\sigma_\alpha^{2(i+1)} = \sigma_\alpha^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{ZA}_g\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}/tr(\mathbf{P}^{(i)}\mathbf{ZA}_g\mathbf{Z}') \tag{15}$$

$$\sigma_\delta^{2(i+1)} = \sigma_\delta^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{ZD}_g\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}/tr(\mathbf{P}^{(i)}\mathbf{ZD}_g\mathbf{Z}') \tag{16}$$

$$\sigma_e^{2(i+1)} = \sigma_e^{2(i)} \mathbf{y}\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y}/tr(\mathbf{P}^{(i)}) \tag{17}$$

GBLUP-QM, Reliability and GREML-QM for *q*>*m*

The mixed model equations for predicting SNP additive effects ($\boldsymbol{\alpha}$) and dominance effects ($\boldsymbol{\delta}$) based on Model 2 are:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{I}_m\lambda\alpha & \mathbf{Z}_1'\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{X} & \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 + \mathbf{I}_m\lambda\delta \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\alpha} \\ \hat{\delta} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \end{pmatrix} \tag{18}$$

where $\mathbf{I}_m = m \times m$ identity matrix, $\lambda_\alpha = \sigma_e^2/\sigma_\alpha^2$ and $\lambda_\delta = \sigma_e^2/\sigma_\delta^2$. To reduce the size of Equation 18, equations for $\hat{\mathbf{b}}$ can be absorbed, and Equation 18 after the absorption reduces to:

$$\begin{pmatrix} \mathbf{Z}_1{}'\mathbf{M}\mathbf{Z}_1 + \mathbf{I}_\mathrm{m}\lambda_\alpha & \mathbf{Z}_1{}'\mathbf{M}\mathbf{Z}_2 \\ \mathbf{Z}_2{}'\mathbf{M}\mathbf{Z}_1 & \mathbf{Z}_2{}'\mathbf{M}\mathbf{Z}_2 + \mathbf{I}_\mathrm{m}\lambda_\delta \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\delta} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1'\mathbf{M}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{M}\mathbf{y} \end{pmatrix} \tag{19}$$

where $\mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$. The GBLUP of breeding values and dominance deviations for all individuals with phenotypic observations can be calculated as:

$$\hat{\mathbf{a}} = \mathbf{T}_\alpha\hat{\alpha} = \sigma_\alpha^2\mathbf{T}_\alpha(\mathbf{Z}\mathbf{T}_\alpha)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_\alpha^2\mathbf{A}_g\mathbf{Z}'\mathbf{P}\mathbf{y} \tag{20}$$

$$\hat{\mathbf{d}} = \mathbf{T}_\delta\hat{\delta} = \sigma_\delta^2\mathbf{T}_\delta(\mathbf{Z}\mathbf{T}_\delta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_\delta^2\mathbf{D}_g\mathbf{Z}'\mathbf{P}\mathbf{y} \tag{21}$$

where $\mathbf{T}_\alpha$ is defined by Equation 4, $\mathbf{T}_\delta$ by Equation 5, and $\hat{\alpha}$ and $\hat{\delta}$ are solutions to Equation 16. We refer to the approach of Equations 19–21 as GBLUP-QM. The comparison between Equations 20–21 and Equations 12–13 shows that GBLUP-CE and GBLUP-QM are mathematically identical. Reliabilities of GBLUP-QM from Equations 19–21 are:

$$R_{ai}^2 = 1 - \lambda_\alpha\left(\mathbf{T}_\alpha\mathbf{C}^{\alpha\alpha}\mathbf{T}_\alpha'\right)_{ii}/a_{ii}$$

$$R_{di}^2 = 1 - \lambda_\delta\left(\mathbf{T}_\delta\mathbf{C}^{\delta\delta}\mathbf{T}_\delta'\right)_{ii}/d_{ii}$$

$$R_{gi}^2 = 1 - \sigma_e^2\left(\mathbf{T}_\alpha\mathbf{C}^{\alpha\alpha}\mathbf{T}_\alpha' + \mathbf{T}_\alpha\mathbf{C}^{\alpha\delta}\mathbf{T}_\delta' + \mathbf{T}_\delta\mathbf{C}^{\delta\alpha}\mathbf{T}_\alpha' + \mathbf{T}_\delta\mathbf{C}^{\delta\delta}\mathbf{T}_\delta'\right)_{ii} \\ /\left(a_{ii}\sigma_\alpha^2 + d_{ii}\sigma_\delta^2\right)$$

where $\mathbf{T}_\alpha$ and $\mathbf{T}_\delta$ are defined by Equations 4–5, and $\mathbf{C}^{\alpha\alpha}$, $\mathbf{C}^{\alpha\delta}$, $\mathbf{C}^{\delta\alpha}$and $\mathbf{C}^{\delta\delta}$ are submatrices that satisfy:

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{Z}_1{}'\mathbf{M}\mathbf{Z}_1 + \mathbf{I}_m\lambda_\alpha & \mathbf{Z}_1{}'\mathbf{M}\mathbf{Z}_2 \\ \mathbf{Z}_2{}'\mathbf{M}\mathbf{Z}_1 & \mathbf{Z}_2{}'\mathbf{M}\mathbf{Z}_2 + \mathbf{I}_m\lambda_\delta \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}^{\alpha\alpha} & \mathbf{C}^{\alpha\delta} \\ \mathbf{C}^{\delta\alpha} & \mathbf{C}^{\delta\delta} \end{pmatrix} \tag{22}$$

For individuals without phenotypic observations (individuals in validation data set), formulations of GBLUP-QM and associated reliability measures are given in Text S1: Part B. GREML-QM formulations via EM type algorithm are:

$$\sigma_\alpha^{2(i+1)} = \hat{\alpha}^{(i)}\hat{\alpha}^{(i)}/[m - tr(\mathbf{C}^{\alpha\alpha(i)})\lambda_\alpha^{(i)}] \tag{23}$$

$$\sigma_\delta^{2(i+1)} = \hat{\delta}^{(i)}\hat{\delta}^{(i)}/[m - tr(\mathbf{C}^{\delta\delta(i)})\lambda_\delta^{(i)}] \tag{24}$$

$$\sigma_e^2(i+1) = \hat{\mathbf{e}}^{(i)}\hat{\mathbf{e}}^{(i)}/\{N - [r - tr(\mathbf{C}^{\alpha\alpha(i)})\lambda_\alpha^{(i)} - tr(\mathbf{C}^{\delta\delta(i)})\lambda_\delta^{(i)}]\} \tag{25}$$

where r is the rank of the coefficient matrix of Equation 18, $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}_1\hat{\alpha} - \mathbf{Z}_2\hat{\delta}$ and $\mathbf{C}^{\alpha\alpha}$ and $\mathbf{C}^{\delta\delta}$ are defined by Equation 22.

Heritability Estimates

Three heritability estimates can be obtained from estimates of variance components: additive heritability or heritability in the narrow sense (h_α^2), dominance heritability (h_δ^2), and the total heritability or heritability in the broad sense (H^2). Let $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\delta^2 + \sigma_e^2 =$ phenotypic variance. Then, $h_\alpha^2 = \sigma_\alpha^2/\sigma_y^2$, $h_\delta^2 = \sigma_\delta^2/\sigma_y^2$, and $H^2 = h_\alpha^2 + h_\delta^2$. Note that the variances of additive and dominance effects (σ_α^2 and σ_δ^2) could be converted into the variances of breeding values and dominance deviations (σ_a^2 and σ_d^2) by $\sigma_a^2 = \sigma_\alpha^2(\sum_{i=1}^q a_{ii})/q$ and $\sigma_d^2 = \sigma_\delta^2(\sum_{i=1}^q d_{ii})/q$ based on $Var(\mathbf{a}) = \mathbf{A}_g\sigma_\alpha^2$ and $Var(\mathbf{d}) = \mathbf{D}_g\sigma_\delta^2$ defined in Equation 9. However, this type of conversion practically is unnecessary because the average a_{ii} and d_{ii}values are '1′ under Definitions II and III and are expected to be '1′ under Definition I of genomic additive and dominance relationships.

Main Features of the CE and QM Formulations

The CE and QM sets of formulations for GBLUP, reliability and GREML are mathematically identical, offer identical results, and offer complimentary computing efficiency. The CE set is designed for *m*>*q* and is the best approach for using a large number of markers for GBLUP and GREML, while GBLUP-QM is designed for *q*>*m* and is the best approach for using a large number of individuals in GBLUP and GREML. A simple rule for choosing between CE and QM is: use CE if *q*<*2m* or vice versa. This is because the size of the **V** matrix to be inverted is *q* for CE (assuming one observation per individual) and the size of the MME coefficient matrix of Equation 19 is *2m* for QM so that **V** become easier to invert than the MME coefficient matrix of Equation 19 for *q*<*2m*. Both sets do not require the inversions of the additive and dominance relationship matrices. The CE set uses relationship matrices explicitly whereas the QM set does so implicitly. Both sets are invariant to the invertibility of $\mathbf{A}_g$ and $\mathbf{D}_g$, i.e., both sets are applicable to singular $\mathbf{A}_g$ and $\mathbf{D}_g$, applicable to *m*>*q* where $\mathbf{A}_g$ and $\mathbf{D}_g$ are generally invertible, and applicable to *q*>*m* where $\mathbf{A}_g$ and $\mathbf{D}_g$ are non-invertible. The property of invariance to the invertibility of additive and dominance relationship matrices is a significant convenience because researchers do not have to require *m*>*q* and do not need to assess invertibility that is not guaranteed by *m*>*q*, e.g., the existence of identical twins results in non-invertible $\mathbf{A}_g$ and $\mathbf{D}_g$.

GBLUP for Validation Data Set, AI-REML, Computer Implementation

Formulations for GBLUP-CE and GBLUP-QM for individuals without phenotypic observations (individuals in validation data set) and reliability measures are given in Text S1: Part B. The EM type algorithm of Equations of 15–17 and 23–25 is known to be reliable but slow. The AI-REML algorithm [23–25] is fast but is not as reliable as EM type. The implementation of AI-REML for estimating additive, dominance and residual variance components is described in Text S1: Part C. All formulations for GBLUP, reliability, genomic relationships and GREML including AI-REML are implemented by the GVCBLUP package [26], which is freely available at http://animalgene.umn.edu.

Accuracy of GREML and GBLUP for Additive and Dominance Heritabilities

Simulation study with known true values of genetic effects and parameters is an effective approach to evaluate the accuracy of a new methodology because the observed GBLUP and GREML estimates can be compared with the true values. We generated a large number of simulated data sets based on a true dairy cattle SNP structure of 1654 Holstein cows assuming true additive and dominance heritability levels of 0, 0.05, 0.15 and 0.30, and we applied seven SNP sets to the simulated data, 1K causal variants, 1K SNP, 2K SNP and causal variants, 3K, 7K 40K SNP markers, and 41K SNP markers and causal variants. Detailed information about these marker sets and the procedure to generate the simulation data are described in Text S1: Part D. For the sample size of 1654 individuals in the simulation study with seven causal and SNP marker sets, GREML were able to capture small effects that each accounted for only 0.00005–0.0003 of the phenotypic variance with high accuracy and were able to distinguish between high and low heritability levels. However, dominance GREML was less accurate and required higher density of SNP markers than additive GREML (Table S1). These results were encouraging given the rapid data growth in genomic selection [27–29] that could substantially increase the GREML accuracy for both additive and dominance effects over the accuracies observed with our sample size.

GREML accuracy of causal variants. Causal SNP markers (1K_QTL, Table S1) had the best accuracy in almost all cases and had similar accuracies for both additive and dominance heritabilities except the case with $h_\alpha^2 = 0.05$ and $h_\delta^2 = 0.05$, where the estimate of dominance heritability was $\hat{h}_\delta^2 = 0.03 \pm 0.02$ and the estimate of additive heritability was $\hat{h}_\alpha^2 = 0.06 \pm 0.01$. Adding linked SNP makers to the causal SNP (2K and 41K in Table S1) decreased GREML accuracy in most cases. Causal SNP markers had nearly unbiased estimates of heritabilities (Figure 1) and had the smallest MSE of heritability estimates (Figure 2). The bias and MSE of variance components had similar patterns as those for heritabilities (data not shown).

GREML accuracy of linked SNP markers. Linked SNP markers were less accurate than causal SNP markers in nearly all cases but were still highly accurate for estimating additive variance. For additive effects, GREML using the 40K and 41K SNP sets had a tendency of slightly overestimating additive heritabilities and variance components. For dominance effects, the marker densities in this simulation study, 1K_SNP, 3K, 7K and 40K, were all insufficient to achieve accurate estimates of dominance heritabilities and variance components, although the 40K set was able to distinguish between high and low dominance heritabilites. Accuracy of dominance GREML increased as the density of linked SNP marker increased from 1K_SNP to 40K, indicating that further increase in marker density over 40K could improve the accuracy of dominance GREML (Table S1).

GREML estimates for '0′ heritability. Estimating '0′ heritability generally is considerably more difficulty than estimating non-null heritability. Therefore, the accuracy in estimating '0′ heritability is a strong test for the accuracy of the GREML formulations. From the same simulation data set we generated above, we generated another set of simulation data requiring additive or dominance effects to be the only genetic effects such that $h_\alpha^2 = 0.00$ and/or $h_\delta^2 = 0.00$ to test the performance of GREML when the true heritability and variance component for one or both effects were null. The causal variants (503_A and 503_D) again had the highest accuracy in estimating '0′ heritabilities and variance components, with average heritability estimates in the range 0–0.01 for additive heritability and 0–0.02 for dominance heritability (Table S2). The 1K SNP set with half causal variants and half inter-QTL SNP (503_A +503_D) was virtually as accurate as the causal variants of 503_A or 503_D. The 41K set also included the causal variants but were not as accurate as the 1K set and overestimated dominance heritability by 0.05 when the true dominance heritability was '0′. The 40K inter-QTL SNP markers had the same overestimates as the 41K. The results of the 1K, 40K and 41K SNP sets showed that a large number of linked SNP markers decreased the GREML accuracy when the true dominance heritability was null. Overall, the GREML formulations were surprisingly accurate in estimating null additive and dominance heritabilities except the 40K and 41K marker sets for null dominance heritability.

Accuracy of GBLUP for breeding values, dominance deviations and genetic values. GBLUP of genotypic values ($\hat{\mathbf{g}}$) and GBLUP of breeding values ($\hat{\mathbf{a}}$) were less sensitive to marker density than GREML. GBLUP of dominance deviations ($\hat{\mathbf{d}}$) was sensitive to marker density as was dominance GREML. Observed and expected accuracies all increased as heritability levels increased. The benefit of using $\hat{\mathbf{g}}$ over $\hat{\mathbf{a}}$ or $\hat{\mathbf{d}}$ for predicting the total genotypic values increased as dominance heritability increased for a given additive heritability except the case $h_\delta^2 = 0.05$ (Figure 3).

GBLUP accuracy of causal variants and linked SNP markers. Causal variants had the best GBLUP accuracy for $\hat{\mathbf{a}}$, $\hat{\mathbf{d}}$ and $\hat{\mathbf{g}}$, but the accuracy for $\hat{\mathbf{d}}$ was lower than that for $\hat{\mathbf{a}}$ and $\hat{\mathbf{g}}$, unlike additive and dominance GREML that had similar accuracies using causal variants. The difference in observed GBLUP accuracy between $\hat{\mathbf{a}}$ and $\hat{\mathbf{d}}$ was the largest for low additive and dominance heritabilities at $h_\alpha^2 = h_\delta^2 = 0.05$ with $\hat{R}_a = 0.50$ and $\hat{R}_d = 0.36$, and was the smallest for high heritabilities at $h_\alpha^2 = 0.30$ and $h_\delta^2 = 0.30$ with $\hat{R}_a = 0.80$ and $\hat{R}_d = 0.76$. These results indicated that dominance GBLUP could be considerably more difficult than additive GBLUP for low dominance heritabilities. Observed accuracy of $\hat{\mathbf{g}}$ ($\hat{R}_g$) was higher than that of for $\hat{\mathbf{a}}$ ($\hat{R}_a$) i.e., $\hat{R}_g > \hat{R}_\alpha$ for all heritability levels in this simulation study except the case $h_\alpha^2 = h_\delta^2 = 0.05$ (Table S3).

For various densities of inter-QTL SNP markers ranging from 3K, 7K to 40K, $\hat{R}_g$ and $\hat{R}_a$ were relatively unchanged within each combination of additive and dominance heritability levels, indicating that increasing SNP density over 3K would achieve little improvement in $\hat{R}_g$and $\hat{R}_a$. For the 1K_SNP, $\hat{R}_g$and $\hat{R}_a$ were lower than the 3K, 7K and 40K by about 0.05. In contrast, $\hat{R}_d$ was similar for the 7K and 40K, had substantial decrease for the 1K_SNP and 3K, and was considerably lower than $\hat{R}_a$ across all heritability combinations. These results indicated that dominance GBLUP required higher density of SNP markers and was more difficult than additive GBLUP (Table S3).

Adding linked SNP makers to causal variants (1K_QTL + 1K_SNP, 1K_QTL +40K) had lower observed accuracies than causal variants alone. The decrease in $\hat{R}_a$ was 0.03 for adding the

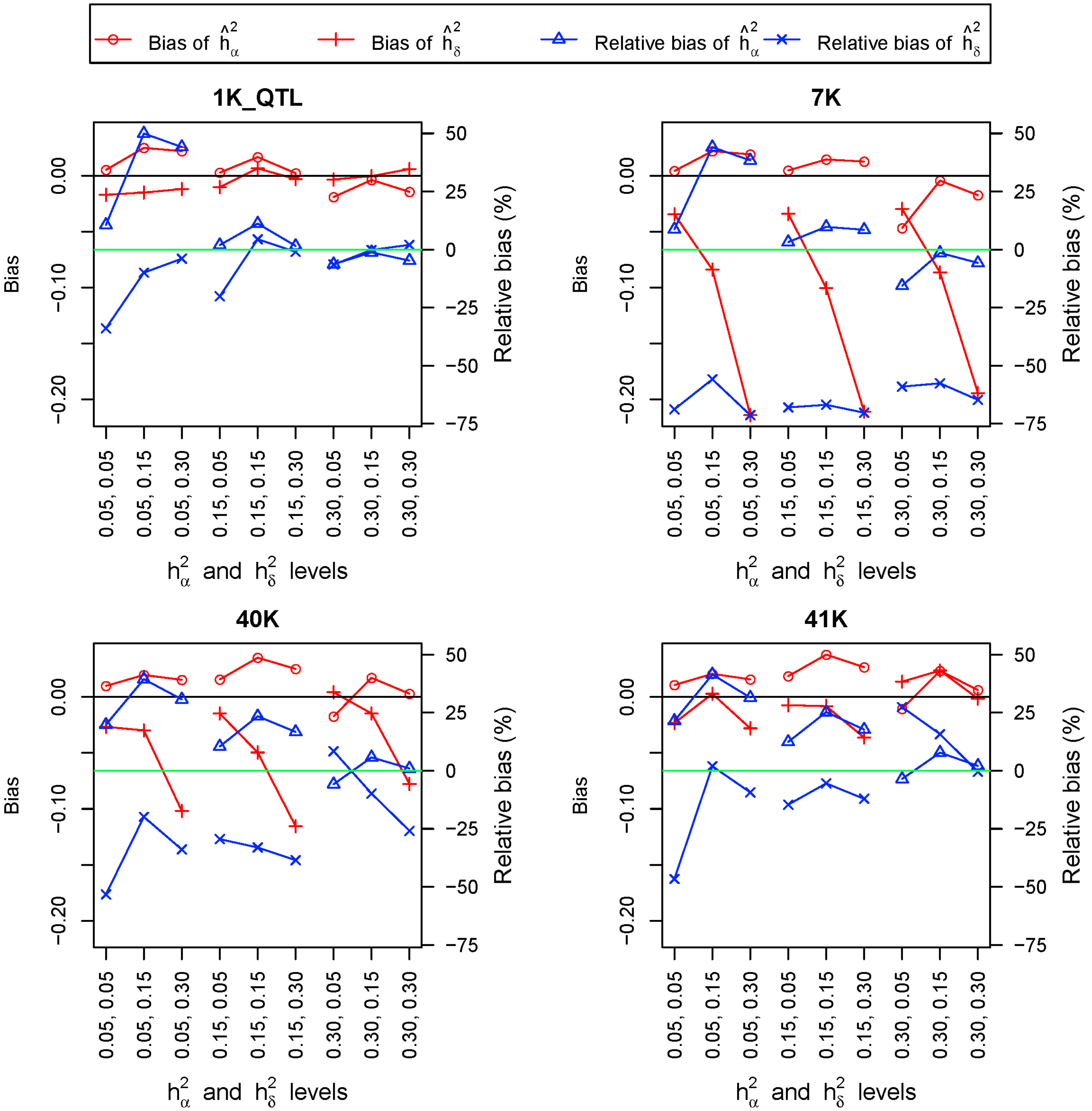

Figure 1. Bias and relative bias of GREML estimates of additive and dominance heritabilities. On the X-axis, heritiabilities of the top row are dominance heritabilities and those of the bottom row are additive heritabilities. (n = 10 repeats).
doi:10.1371/journal.pone.0087666.g001

1K_SNP to the 1K_QTL and was 0.06 for adding the 40K to the 1K_QTL. The decreases in $\hat{R}_d$were even larger, 0.06 and 0.13, respectively. These decreases were relatively constant across heritability levels (Table S3). However, any marker set with causal variants, the 2K or 41K, was more accurate than linked SNP only, the 1K_SNP, 3K, 7K or 40K.

Predicted and observed GBLUP accuracies. Predicted accuracy for breeding values (R_a) and for genotypic values (R_g) agreed well with the observed accuracies ($\hat{R}_a$and $\hat{R}_g$) across all heritability levels used in this study. For dominance deviations, predicted accuracy (R_d) and observed accuracy ($\hat{R}_d$) agreed well except $h^2_\alpha = h^2_\delta = 0.05$, where R_d was substantially lower than observed accuracies. In real data sets, observed accuracies measured by $\hat{R}_a$, $\hat{R}_d$and $\hat{R}_g$ are unavailable. The good agreements between predicted and observed accuracies indicated that predicted accuracy could reliably represent the observed accuracy in real data.

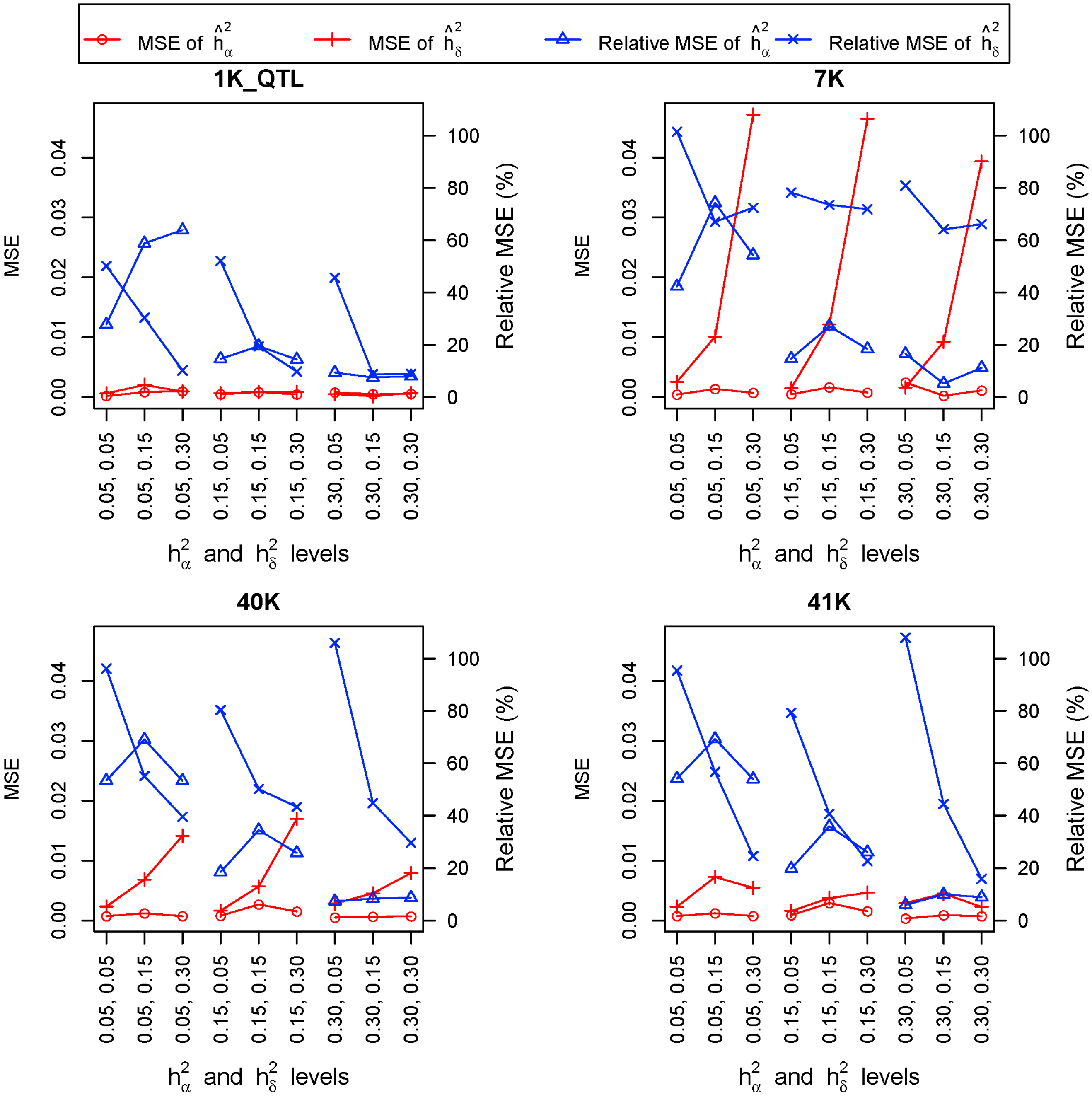

Figure 2. Mean square error (MSE) and relative MSE of GREML estimates of additive and dominance heritabilities. On the X-axis, heritiabilities of the top row are dominance heritabilities and those of the bottom row are additive heritabilities. (n = 10 repeats).
doi:10.1371/journal.pone.0087666.g002

Comparison of Genomic Additive and Dominance Relationships with Expected Relationships

For genomic additive and dominance relationships, Definitions I-III had nearly identical results. The 1K, 3K, 7K and 41K marker sets had similar results of relationships (data not shown). For the 41K results with the removal of three full-sib outliers and nine half-sib outliers, additive and dominance relationships agreed well with theoretical expectations (Figure 4). For full-sibs, genomic additive and dominance relationships were nearly identical to theoretical expectations. Average genomic additive relationships was 0.471 for Definition I, 0.478 for Definition II, and 0.488 for Definition III, while the mean value of pedigree coancestry coefficients for full-sibs was 0.262, i.e., genomic additive relationships were about twice as large as pedigree coancestry coefficients. The mean dominance correlation for full-sibs was 0.245 for Definition I, 0.248 for Definition II and 0.254 for Definition III, compared to the expected full-sib dominance correlation of 0.25 assuming no inbreeding. The 1654 cows used in this comparison of genomic and pedigree relationships in fact were all related [30]. Therefore, the true full-sib dominance relationships should have been above 0.25. For half-sibs, Definitions I-III had mean additive relationship of 0.213–0.221 and the average of '2×(pedigree coancestry coefficient)' was 0.282. Genomic dominance relation-

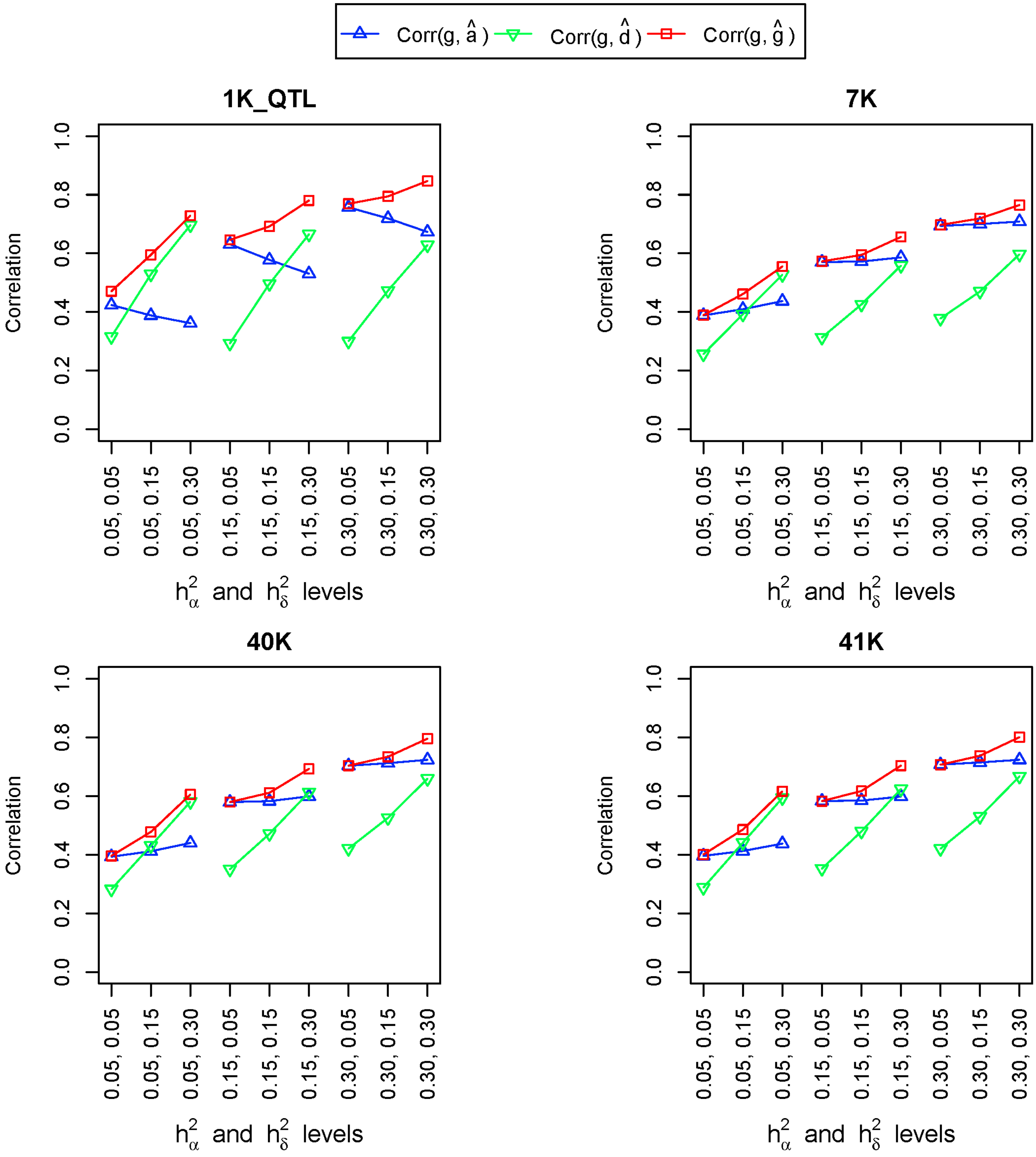

Figure 3. Correlation between the true genotypic values and GBLUP of breeding values, dominance deviations and genetic values. $Corr(g,\hat{a})$ is the correlation between true genotypic values and GBLUP of breeding values, $Corr(g,\hat{d})$ is the correlation between true genotypic values and GBLUP of dominance deviations, and $Corr(g,\hat{g})$ is the correlation between true genotypic values and GBLUP of genotypic values. On the X-axis, heritabilities of the top row are dominance heritabilities and those of the bottom row are additive heritabilities. (n = 10 repeats).
doi:10.1371/journal.pone.0087666.g003

ships were null for half-sibs and unrelated individuals, and genomic additive relationships for unrelated individuals were also null, as expected (Figure 4). The observed similarity between Definitions I-III likely was due to the underlying similarity of the three definitions: Definition I and II are the same if the expected and observed SNP variances are the same, and Definitions II and III are the same if all diagonal elements are the same. Definitions I and II make slightly less modification to the original mixed model

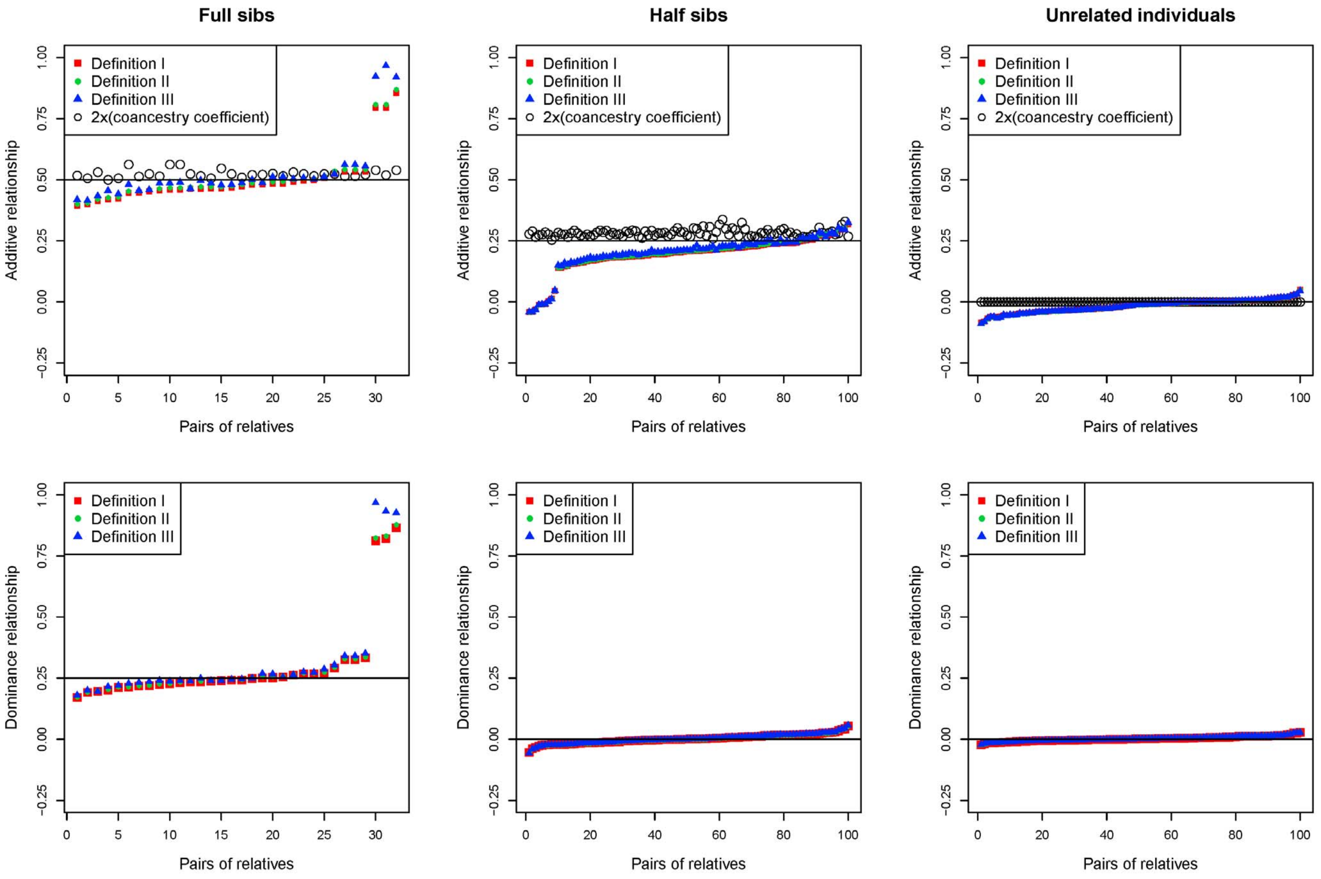

Figure 4. Genomic additive and dominance relationships among full-sibs, half-sibs and unrelated individuals.
doi:10.1371/journal.pone.0087666.g004

of Equation 3, whereas off-diagonal elements of Definition III as measures of genomic relatedness among individuals are mathematically comparable.

Genomic relationships have a distinct advantage over pedigree relationships: the calculation of genomic relationships does not need to know the pedigree. This advantage is important for assessing relatedness among individuals in species where pedigree information is unavailable or difficult to collect such as in wildlife species. Two important differences exist between relationships based on markers and relationships based on pedigree information. The first difference is that marker density affects the invertibility of genomic relationship matrices, which are non-invertible when $q>m$. In contrast, pedigree relationship matrices are positive definite in the absence of identical twins. Our cattle data showed that the invertibility of a genomic relationship matrix should not affect the use of genomic relationships as measures of genomic relatedness among individuals, because the genomic relationships calculated from genomic relationship matrices that were invertible or non-invertible had nearly identical values that were consistent with theoretical expectations (data not shown). The additive and dominance relationship matrices were non-invertible for the 1K SNP set, and were invertible for the 3K, 7K and 41K sets after removing a potentially identical twin or duplicated individual. The second difference is the range of relationship values. Genomic relationships by Definitions I-III could take negative values whereas pedigree relationships are non-negative. However, no negative values were observed for full-sib genomic relationships. Negative genomic additive relationships with small absolute values near ‘0′ were observed for unrelated individuals and some half-sibs, and negative dominance relationships with small absolute values near ‘0′ were observed among half-sib (Figure 4). In all those situations, the expected relationships were ‘0′. Therefore, negative genomic relationships close to ‘0′ could be interpreted as no correlation. It remains to be seen whether genomic relationship measures could detect true ‘negative genomic correlations’ (if such correlations exist) that are impossible to detect using pedigree information.

Table 1. Estimated genomic additive and dominance heritabilities from a swine nucleus line.

	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5
h^2_α	0.03 (0.07*)	0.27 (0.16)	0.22 (0.38)	0.35 (0.58)	0.38 (0.62)
h^2_δ	7.22×10^{-7}	0.02	0.07	0.01	0.05
H^2	0.03	0.29	0.29	0.36	0.44

*Value in each () is the pedigree-based heritability estimate [31].
doi:10.1371/journal.pone.0087666.t001

Effect of Genomic Relationship Definitions on GBLUP and GREML Accuracies

Simulation results showed that the methods to normalize $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ (Definitions I and II of genomic relationships) had no effect on GBLUP accuracy, i.e., the original mixed model was just as accurate, as shown by the $\hat{R}_a$ and $\hat{R}_d$values (Table S4).

Definition III of genomic relationships had the same accuracy as Definitions I-II for breeding values and had slightly lower accuracy for dominance deviations for one case only at $h_\alpha^2 = h_\delta^2 = 0.30$, with $\hat{R}_d = 0.75$ for Definition III and $\hat{R}_d = 0.76$ for Definitions I and II (Table S4). For GREML, normalization or transformation of the $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$ matrices was necessary. Without such normalization or transformation, diagonal values in $\mathbf{W}_\alpha\mathbf{W}_\alpha^{'}$ and $\mathbf{W}_\delta\mathbf{W}_\delta^{'}$increased and estimates of variance components decreased as the number of SNP markers increased regardless of the true heritability level, so that hertitability estimates based on such variance component estimates became meaningless, as shown by the comparison of GREML estimates and the corresponding heritability estimates in Table S4. For GREML estimation of additive and dominance variances, Definitions I-III had similar estimates that were consistent with the true values.

Random and Directional Dominance Effects

Random additive and dominance effects with zero means were assumed in the simulation study reported in the section of Results. Under these assumptions, dominance effects were more difficult to predict and estimate in two aspects: the current densities of inter-QTL SNP markers up to 40K were insufficient to achieve accuracies comparable to those for additive effects, and causal variants had lower accuracy of dominance GBLUP than the accuracy of additive GBLUP, although causal variants had similar accuracy for estimating additive and dominance variance components. The simulation results indicated that the number of SNP markers needed in the absence of causal variants would be considerably greater than 40K to achieve accuracies of dominance GBLUP and GREML comparable to the accuracies of additive GBLUP and GREML. High density of SNP markers could also compensate the lower accuracies of causal variants, whether or not causal variants were among the SNP markers. The simulation data set assuming positive dominance deviation for each heterozygous genotype (Text S1: Part D) showed that dominance GBLUP had similar accuracies to additive GBLUP (Table S5).

Taken all evidence together, genomic prediction and variance component estimation of dominance effects was more difficult than those of additive effects in populations where additive and dominance effects had similar distributions and heritabilities but could achieve similar accuracies as those for additive effects if heterosis exists.

An Application to Estimate Genomic Additive and Dominance Heritabilities in a Swine Population

We applied our methodology to a publically available swine genomics data set with anonymous genome-wide SNP markers and phenotypes with the SNP locations and true trait names masked [31] to compare genomic additive heritability with the reported heritability estimated using pedigree information and to explore whether the swine phenotypes had dominance effects. The data set included 3534 animals from a single PIC nucleus pig line with genotypes from the Illumina PorcineSNP60 chip [32]. Genotyped animals had phenotypes for five purebred traits (phenotypes in a single nucleus line), with additive heritability estimated from pedigree data ranging from 0.07 to 0.62 (Table 1). Genotypes were filtered by requiring minor allele frequency (MAF) >0.001 and proportion of missing SNP genotypes <0.100. Markers on the X or Y chromosome were excluded. The total number of available autosome markers used in our analysis was 52,842, with missing genotypes imputed using software AlphaImpute [33]. The results showed that estimates of genomic additive heritability of 0.22–0.38 were substantially lower than the pedigree estimates of 0.38–0.62 for traits 3–4, the genomic additive heritability (0.27) was higher than the pedigree estimate (0.16) for trait 2, and was in agreement with the pedigree estimate for trait 1, 0.03 versus 0.07. Only traits 3 and 5 had small dominance heritabiities, 0.07 for trait 3 and 0.05 for trait 5. The genomic estimates reported here provide useful information to breeders about the underlying true genetic factors and about the potential true heritability levels of the five traits.

Conclusions

The genomic model based on the partition of a genotypic value into breeding value and dominance deviation with additive and dominance relationship matrices calculated using SNP markers parallels the traditional quantitative genetics model that calculates additive and dominance relationships using pedigree information. The GREML and GBLUP methods based on equivalent models with complementary computing advantages and identical mathematical results provide an efficient approach for the genomic estimation of variance components and heritabilities and for the genomic prediction of additive and dominance effects using SNP markers. These methods were able to capture small additive and dominance effects and were able to differentiate different levels of additive and dominance heritabilities. GBLUP of total genetic value that includes additive and dominance effects can be an effective tool to predict an individual's total genetic potential for a phenotype.

Supporting Information

Table S1 GREML estimates of variance components and heritabilities of additive and dominance effects (mean ± standard deviation, n = 10 repeats).
(PDF)

Table S2 GREML estimates and GBLUP accuracy for simulation data with additive or dominance effects only (mean ± standard deviation, n = 10 repeats).
(PDF)

Table S3 GBLUP Accuracies for breeding values, dominance deviations and genotypic values (mean ± standard deviation, n = 10 repeats).
(PDF)

Table S4 GREML estimates of variance components and GBLUP accuracies with and without genomic relationships for phenotypes with additive and dominance effects of 1006 QTL (mean ± standard deviation, n = 10 repeats).
(PDF)

Table S5 GBLUP accuracies in simulation data assuming random additive effects and directional dominance effects values (mean ± standard deviation, n = 10 repeats).
(PDF)

Text S1 Proofs, formulations and simulation study. Part A: Derivations for the traditional quantitative genetics model of SNP markers with unequal and equal allele frequencies. Part B: Genomic prediction of additive and dominance effects for individuals without phenotypic observations. Part C: AI-REML implementation. Part D: Simulation study to evaluate GREML and GBLUP accuracies.
(PDF)

Author Contributions

Conceived and designed the experiments: YD. Performed the experiments: YD CW SW GH. Analyzed the data: CW SW GH. Contributed reagents/materials/analysis tools: CW SW. Wrote the paper: YD CW GH.

References

1. Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
2. VanRaden P (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414–4423.
3. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, et al. (2009) Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92: 16–24.
4. Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53: 876–883.
5. Goddard M, Hayes B (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128: 409–421.
6. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42: 565–569.
7. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics (4th edition). Harlow, Essex, UK: Longmans Green.
8. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11: 446–450.
9. Vineis P, Pearce N (2010) Missing heritability in genome-wide association study research. Nature Reviews Genetics 11: 589–589.
10. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences 109: 1193–1198.
11. Misztal I, Lawlor T, Fernando R (1997) Dominance models with method R for stature of Holsteins. Journal of dairy science 80: 975–978.
12. Sun C, Vanraden PM, O'Connell JR, Weigel KA, Gianola D (2013). Mating programs including genomic relationships and dominance effects. Journal of Dairy Science 96: 1–10.
13. Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol 42: 33.
14. Ober U, Erbe M, Long N, Porcu E, Schlather M, et al. (2011) Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188: 695–708.
15. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS one 7: e45293.
16. Zeng J, Pszczola M, Wolc A, Strabel T, Fernando RL, et al. Genomic breeding value prediction and QTL mapping of QTLMAS2011 data using Bayesian and GBLUP methods; 2012. BioMed Central Ltd. S7.
17. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.
18. Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.
19. Wright S (1922) Coefficients of inbreeding and relationship. The American Naturalist 56: 330–338.
20. Henderson C (1985) Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of animal science 60: 111–117.
21. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological): 1–38.
22. Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72: 320–338.
23. Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics: 1440–1450.
24. Johnson D, Thompson R (1995) Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of dairy science 78: 449–456.
25. Lee SH, van der Werf JH (2006) 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 38: 1–19.
26. Wang C, Prakapenka D, Wang S, Runesha HB, Da Y (2013) GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. Version 3.7. Department of Animal Science, University of Minnesota. [http://animalgene.umn.edu]
27. Hayes B, Bowman P, Chamberlain A, Goddard M (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92: 433–443.
28. Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166–177.
29. Wiggans G, VanRaden P, Cooper T (2011) The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science 94: 3202.
30. Ma L, Wiggans GR, Wang S, Sonstegard TS, Yang J, et al. (2012) Effect of sample stratification on dairy GWAS results. BMC Genomics 13: 536.
31. Cleveland MA, Hickey JM, Forni S (2012) A common dataset for genomic analysis of livestock populations. *G3:* Genes| Genomes| Genetics 2: 429–435.
32. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, et al. (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PloS one 4: e6524.
33. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA (2012): A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol, 44: 9.