Da *BMC Genetics* (2015) 16:144
DOI 10.1186/s12863-015-0301-1

BMC Genetics

METHODOLOGY ARTICLE Open Access

Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers

Yang Da

Abstract

Background: The amount of functional genomic information has been growing rapidly but remains largely unused in genomic selection. Genomic prediction and estimation using haplotypes in genome regions with functional elements such as all genes of the genome can be an approach to integrate functional and structural genomic information for genomic selection. Towards this goal, this article develops a new haplotype approach for genomic prediction and estimation.

Results: A multi-allelic haplotype model treating each haplotype as an 'allele' was developed for genomic prediction and estimation based on the partition of a multi-allelic genotypic value into additive and dominance values. Each additive value is expressed as a function of h − 1 additive effects, where h = number of alleles or haplotypes, and each dominance value is expressed as a function of h(h − 1)/2 dominance effects. For a sample of q individuals, the limit number of effects is 2q − 1 for additive effects and is the number of heterozygous genotypes for dominance effects. Additive values are factorized as a product between the additive model matrix and the h − 1 additive effects, and dominance values are factorized as a product between the dominance model matrix and the h(h − 1)/2 dominance effects. Genomic additive relationship matrix is defined as a function of the haplotype model matrix for additive effects, and genomic dominance relationship matrix is defined as a function of the haplotype model matrix for dominance effects. Based on these results, a mixed model implementation for genomic prediction and variance component estimation that jointly use haplotypes and single markers is established, including two computing strategies for genomic prediction and variance component estimation with identical results.

Conclusion: The multi-allelic genetic partition fills a theoretical gap in genetic partition by providing general formulations for partitioning multi-allelic genotypic values and provides a haplotype method based on the quantitative genetics model towards the utilization of functional and structural genomic information for genomic prediction and estimation.

Keywords: Haplotype, Genomic selection, Variance component, Heritability, BLUP, REML

Correspondence: yda@umn.edu
Department of Animal Science, University of Minnesota, Saint Paul, MN, USA

Background

Genomic best linear unbiased prediction (GBLUP) using genome-wide single nucleotide polymorphism (SNP) markers can utilize a wealth of theoretical results and computational strategies of best linear unbiased prediction (BLUP) [1] that has become a standard approach for genetic evaluation, with dairy cattle having the most widespread use of BLUP worldwide [2–5]. The implementation of GBLUP within the BLUP framework is made possible by a genomic relationship matrix that replaces the pedigree relationship matrix in BLUP [6]. With genomic relationship matrix established, genomic estimation of variance components can also readily use the method of restricted maximum likelihood estimation (REML) [7], to be referred to as GREML (genomic REML). Using a quantitative genetics model as the unifying model, genomic relationship matrix is formulated by equaling the covariance of genomic values between two individuals to the corresponding pedigree covariance [8, 9]. Previously defined genomic relationships based on standardization of SNP coding [6, 8, 10, 11] can be considered as special cases of this unifying approach [9]. The quantitative genetics model partitions a genotypic value as the summation of a common mean, breeding value and dominance deviation [12–18]. Using matrix notations, this partition can be expressed as: $\mathbf{g} = \mathbf{1}\mu + \mathbf{a} + \mathbf{d} = \mathbf{1}\mu + \mathbf{W}_\alpha\boldsymbol{\alpha} + \mathbf{W}_\delta\boldsymbol{\delta}$, where μ = common mean, $\mathbf{1}$ = column vector of 1's, $\mathbf{a}$ = breeding values (additive values), $\mathbf{d}$ = dominance deviations (dominance values), $\boldsymbol{\alpha}$ = SNP additive effects, $\boldsymbol{\delta}$ = SNP dominance effects, $\mathbf{W}_\alpha$ = model matrix of $\boldsymbol{\alpha}$ as a function of SNP allele frequencies, and $\mathbf{W}_\delta$ = model matrix of $\boldsymbol{\delta}$ as a function of SNP allele frequencies. With the factorization of $\mathbf{a} = \mathbf{W}_\alpha\boldsymbol{\alpha}$ and $\mathbf{d} = \mathbf{W}_\delta\boldsymbol{\delta}$, genomic additive relationship is a function of $\mathbf{W}_\alpha\mathbf{W}_\alpha'$ and genomic dominance relationship is a function of $\mathbf{W}_\delta\mathbf{W}_\delta'$ [9]. This approach for defining genomic relationships was only available for bi-allelic loci. Although SNPs are bi-allelic loci, the issue of multi-allelic loci for genomic prediction and estimation arises if each haplotype is treated as an 'allele' and the haplotype block containing the haplotypes is treated as a 'locus'. For a multi-allelic locus, the partition of a genotypic value into additive and dominance values ($\mathbf{g} = \mathbf{1}\mu + \mathbf{a} + \mathbf{d}$) was available [17] and the multi-allelic factorization of $\mathbf{a} = \mathbf{W}_\alpha\boldsymbol{\alpha}$ and $\mathbf{d} = \mathbf{W}_\delta\boldsymbol{\delta}$ was available for three alleles [19]. However, general factorization formulations for an arbitrary number of alleles were unavailable, and a method using such multi-allelic haplotype model for genomic prediction and estimation was unavailable.

Haplotype analysis is advantageous over single-locus analysis for several reasons: a haplotype is a functional unit [20], a haplotype contains combined effects of tightly linked cis-acting causal variants [21, 22], a phenotype is affected by multiple causal loci with weak LD (LD = linkage disequilibrium) [23], or a genomic region is subjected to selection with stronger LD than genome regions unaffected by selection [24, 25]. Haplotype analysis has been widely used in genetic and genomic studies [22, 26–28]. Relatively limited studies were available on using haplotypes compared to the literature on using single SNPs for genomic prediction. Methods to define haplotype blocks for genomic prediction included a constant number of SNPs per SNP block [29, 30], fixed block length [31], or LD blocks [32]. Haplotype coding methods for genomic prediction and estimation included 2-1-0 copies of a haplotype in the two-haplotype genotype [30, 33], or maternal or paternal haplotype [29]. Haplotype mixed model methods based on the quantitative genetics model with multi-allelic factorization of additive and dominance values were unavailable for genomic prediction and estimation. Functional genomic information has been growing rapidly but remains largely unused in genomic selection. Simulation study showed that genomic prediction using causal mutations could substantially improve prediction accuracy [34], and using SNPs in transcriptional regions [35] or location specific priors based on QTL mapping results [36] improved prediction accuracy. Haplotype analysis can be a useful tool to account for joint allelic effects unaccounted for by single-SNP analysis and we have obtained encouraging preliminary results of using haplotype analysis of functional genomic information [37, 38].

The purpose of this article is to develop a quantitative genetics based multi-allelic haplotype model as an alternative method to single-SNP analysis towards the integration of functional and structural genomic information for genomic selection. This development includes deriving general multi-allelic partition of genotypic values with factorization for defining genomic relationships using haplotypes, and deriving mixed model formulations for genomic prediction and estimation that can use haplotypes separately or jointly with single SNPs.

Methods

Allelic mean and population mean of multi-allelic genotypic values

A set of m SNP markers are assumed available, and r haplotype blocks are defined from some of the m SNPs across the genome. Each haplotype block is treated as a 'locus' and each haplotype within the haplotype block is treated as an 'allele'. Each locus (haplotype block) is assumed to have h alleles (haplotypes) denoted by $A_i, \ldots, A_h$, with allele frequency of p_i for A_i, $i = 1, \ldots, h$, and $\sum_{i=1}^{h} p_i = 1$. The allelic array in the population is $\sum_{i=1}^{h} p_i A_i$. Let P_{ij} = frequency of

A_iA_j genotype, $\sum_{i=1}^{h}\sum_{j=1}^{h}P_{ij}A_iA_j$ = the genotypic array of the population, and g_{ij} = genotypic value of A_iA_j genotype, i,j = 1,…,h. Hardy-Weinberg equilibrium (HWE) is assumed so that the genotypic array of the population is the squared allelic array, i.e., $\sum_{i=1}^{h}\sum_{j=1}^{h}P_{ij}A_iA_j = (\sum_{i=1}^{h}p_iA_i)^2$. Allele frequency of A_i is calculated as:

$$p_i = P_{ii} + \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{h} P_{ij} \tag{1}$$

The allelic mean of A_i allele is the weighted mean of all genotypic values with the A_i allele, with each genotypic value weighted by the number of copies of the A_i allele the genotype carries. The general expression of the allelic mean without requiring HWE is a conditional mean [13] and simplifies to a weighted average of genotypic values with allele frequencies as the weights under the HWE assumption [13, 17], i.e.,

$$\begin{aligned}\mu_i &= \left[2P_{ii}g_{ii} + \sum_{j\neq i}^{h}P_{ij}g_{ij}\right] / \left[2P_{ii} + \sum_{j\neq i}^{h}P_{ij}\right]\\ &= \sum_{j=1}^{h}p_jg_{ij}\end{aligned} \tag{2}$$

The population mean is the mean of all genotypic values in the population. The general formula without requiring HWE and its expression as a weighted average of allelic means with allele frequencies as the weights requiring HWE are:

$$\begin{aligned}\mu &= \sum_{i=1}^{h}\sum_{j=1}^{h}P_{ij}g_{ij}\\ &= \sum_{i=1}^{h}p_i^2g_{ii} + 2\sum_{i=1}^{h-1}\sum_{j=i+1}^{h}p_ip_jg_{ij}\\ &= \sum_{k=1}^{h}p_k\mu_k\end{aligned} \tag{3}$$

The expressions of $\mu_i = \sum_{j=1}^{h}p_jg_{ij}$ and $\mu = \sum_{k=1}^{h}p_k\mu_k$ play an important role in the derivations to factorize additive and dominance values and in defining fundamental genetic parameters of quantitative traits.

Multi-allelic effect, additive effect, additive value

The allelic effect (average effect) of allele A_i (i = 1,…h) is the deviation of the allelic mean from the population mean. From Eqs. 2 and 3, the allelic effect of A_i is:

$$a_i = \mu_i - \mu = \sum_{j\neq i}^{h}p_j\left(\mu_i - \mu_j\right) = \sum_{j\neq i}^{h}p_j\alpha_{ij} \tag{4}$$

where α_{ij} is the additive effect or the average effect of gene substitution that is the difference between the allelic effects of the two alleles defined by Eq. 4, i.e.,

$$\alpha_{ij} = a_i - a_j = \mu_i - \mu_j = \sum_{k=1}^{h}p_k\left(g_{ik} - g_{jk}\right) = -\alpha_{ji} \tag{5}$$

For h alleles, h(h – 1)/2 α_{ij} parameters of Eq. 5 are possible but these parameters are not independent for all ij values. An example of this dependency is:

$$\alpha_{ij} = \alpha_{1j} - \alpha_{1i} \tag{6}$$

Based on Eq. 6, h-1 independent additive effects can be defined:

$$\alpha_{1k} = a_1 - a_k = \mu_1 - \mu_k, k = 2, \ldots h \tag{7}$$

where μ_1 = allelic mean of allele 1 that is used as the reference allele (e.g., defining the most frequent allele as 'allele 1'). It is readily seen that $\alpha_{ii} = 0$. The derivation process will allow the presence of α_{ii} but the final results will be based on the h−1 independent additive effects of α_{lk} defined by Eq. 7. All the h(h – 1)/2 possible α_{ij} parameters can be expressed in terms of the h−1 independent α_{lk} parameters through Eq. 6. The additive value (breeding value) of genotype A_iA_j is the summation of the two allelic effects of the genotype, i.e.,

$$a_{ij} = a_i + a_j \tag{8}$$

Each additive value defined by Eq. 8 will be shown to be a function of all h−1 additive effects defined by Eq. 7.

Dominance effect and dominance value

Dominance effect of A_iA_j genotype (δ_{ij}) is the deviation of the heterozygous genotypic value from the average of the two homozygous genotypic values, i.e.,

$$\delta_{ij} = g_{ij} - \frac{1}{2}\left(g_{ii} + g_{jj}\right) \tag{9}$$

With the above definition, dominance effect is the unique effect of a heterozygous genotype. Therefore, the number of dominance effects is the same as number of heterozygous genotypes, and the maximum number of dominance effects is h(h – 1)/2. It is readily seen from Eq. 9 that $\delta_{ii} = 0$. The derivation process will allow the presence of δ_{ii} but the final results will not have δ_{ii}. Dominance value or dominance deviation is the deviation of the genotypic value from the common mean and additive value, i.e.,

$$d_{ij} = g_{ij} - \mu - a_{ij} \tag{10}$$

An important difference between 'dominance value' and 'dominance effect' is that a homozygous genotype may have non-zero dominance value but always has zero dominance effect. Each dominance value defined by Eq. 10 will be shown to be a function of all h(h – 1)/2 dominance effects defined by Eq. 9.

Multi-allelic partition of genotypic value and variance

The genotypic value of a multi-allelic genotype has the same partition as for a bi-allelic locus [17], i.e.,

$$g_{ij} = \mu + a_{ij} + d_{ij} \tag{11}$$

with $E(a_{ij}) = 0$ and $E(d_{ij}) = 0$. The multi-allelic genotypic variance (σ_g^2) also has the same partition as for a bi-allelic locus [17], i.e., $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$, where σ_a^2 = additive variance, and σ_d^2 = dominance variance. The multi-allelic haplotype model to be developed starts with the factorization of the additive and dominance values in Eq. 11.

Results and discussion

Factorization of additive and dominance values

From Eqs. 4–7, an allelic effect can be expressed as:

$$\begin{aligned} a_i &= \mu_i - \mu = \sum\nolimits_{k\neq i}^{h} p_k \alpha_{ik} = \sum\nolimits_{k\neq i}^{h} p_k(\alpha_{1k} - \alpha_{1i}) \\ &= -(1-p_i)\alpha_{1i} + \sum\nolimits_{k\neq i}^{h} p_k \alpha_{1k} \end{aligned} \tag{12}$$

where α_{lk} is defined by Eq. 7. Equation 12 shows that an allelic effect is a function of all h-1 parameters of additive effects denoted by α_{lk}. The additive values (breeding values) of A_iA_j and A_iA_i genotypes can be expressed as:

$$\begin{aligned} a_{ij} = a_i + a_j &= \left[-(1-p_i)\alpha_{1i} + \sum\nolimits_{k\neq i}^{h} p_k \alpha_{1k}\right] \\ &\quad + \left[-\left(1-p_j\right)\alpha_{1j} + \sum\nolimits_{k\neq i}^{h} p_k \alpha_{1k}\right] \\ &= -(1-2p_i)\alpha_{1i} - \left(1-2p_j\right)\alpha_{1j} + 2\sum\nolimits_{k\neq ij}^{h} p_k \alpha_{1k} \end{aligned} \tag{13}$$

$$a_{ii} = 2a_i = -2(1-2p_i)\alpha_{1i} + 2\sum\nolimits_{k\neq ij}^{h} p_k \alpha_{1k} \tag{14}$$

In Eqs. 13 and 14, $\alpha_{li} = 0$ if $i = 1$ and $\alpha_{1j} = 0$ if $j = 1$. From Eqs. 1–3 and 9–10, the dominance value of the A_iA_j genotype can be expressed as

$$\begin{aligned} d_{ij} &= g_{ij} - \mu - a_i - a_j = g_{ij} - \mu_i - \mu_j + \mu = \left(g_{ij} - \mu_i\right) - \left(\mu_j - \mu\right) \\ &= \sum\nolimits_{k\neq j}^{h} p_k \left(g_{ij} - g_{ik}\right) - \sum\nolimits_{k\neq j}^{h} p_k \left(\mu_j - \mu_k\right) \\ &= \sum\nolimits_{k\neq j}^{h} p_k \left[\left(g_{ij} - \mu_j\right) - (g_{ik} - \mu_k)\right] \\ &= \sum\nolimits_{k\neq j}^{h} p_k \left[\sum\nolimits_{f\neq i}^{h} p_f \left(g_{ij} - g_{jf}\right) - \sum\nolimits_{f\neq i}^{h} p_f (g_{ik} - g_{kf})\right] \\ &= \sum\nolimits_{k\neq j}^{h} p_k \sum\nolimits_{f\neq i}^{h} p_f \left(g_{ij} - g_{ik} - g_{jf} + g_{kf}\right) \end{aligned} \tag{15}$$

In Eq. 15, the quantity $g_{ij} - g_{ik} - g_{jf} + g_{kf}$ has two positive terms and two negative terms, and each subscript is associated with a positive term and a negative term. Using this fact and the definition of dominance effect (δ_{ij}) of Eq. 9 with $\delta_{ii} = 0$, $g_{ij} - g_{ik} - g_{jf} + g_{kf}$ can be expressed as:

$$g_{ij} - g_{ik} - g_{jf} + g_{kf} = \delta_{ij} - \delta_{ik} - \delta_{jf} + \delta_{kf} \tag{16}$$

Combining Eqs. 15 and 16 with Eq. 10 and using $p_j = 1 - \Sigma_{k\neq j}^{h} p_k$ (Eq. 1) yields:

$$\begin{aligned} d_{ij} &= \sum\nolimits_{k\neq j}^{h} p_k \sum\nolimits_{f\neq i}^{h} p_f \left(\delta_{ij} - \delta_{ik} - \delta_{jf} + \delta_{kf}\right) \\ &= \sum\nolimits_{k\neq j}^{h} p_k \left[\sum\nolimits_{f\neq i}^{h} p_f \left(\delta_{ij} - \delta_{ik}\right) - \sum\nolimits_{f\neq i}^{h} p_f \left(\delta_{jf} - \delta_{kf}\right)\right] \\ &= \sum\nolimits_{k\neq j}^{h} p_k \left[(1-p_i)\left(\delta_{ij} - \delta_{ik}\right) - \sum\nolimits_{f\neq i}^{h} p_f \delta_{jf} + \sum\nolimits_{f\neq i}^{h} p_f \delta_{kf}\right] \\ &= (1-p_i)\left(1-p_j\right)\delta_{ij} - (1-p_i)\sum\nolimits_{k\neq j}^{h} p_k \delta_{ik} \\ &\qquad - \sum\nolimits_{k\neq j}^{h} p_k \left(\sum\nolimits_{f\neq i}^{h} p_f \delta_{jf} - \sum\nolimits_{f\neq i}^{h} p_f \delta_{kf}\right) \\ &= (1-p_i)\left(1-p_j\right)\delta_{ij} - (1-p_i)\sum\nolimits_{k\neq j}^{h} p_k \delta_{ik} - \left(1-p_j\right)\sum\nolimits_{f\neq i}^{h} p_f \delta_{jf} \\ &\qquad + \sum\nolimits_{k\neq j}^{h} p_k \sum\nolimits_{f\neq i}^{h} p_f \delta_{kf} \end{aligned} \tag{17}$$

In Eq. 17,

$$\begin{aligned} \sum\nolimits_{k\neq j}^{h} p_k \sum\nolimits_{f\neq i}^{h} p_f \delta_{kf} &= p_i p_j \delta_{ij} + p_i \sum\nolimits_{f\neq i,k}^{h} p_f \delta_{jf} + p_j \sum\nolimits_{k\neq j,f}^{h} p_k \delta_{jk} \\ &\qquad + \sum\nolimits_{k\neq i,j}^{h} p_k \sum\nolimits_{f\neq k}^{h} p_f \delta_{kf} \\ &= p_i p_j \delta_{ij} + p_i \sum\nolimits_{k\neq i,j}^{h} p_k \delta_{ik} + p_j \sum\nolimits_{f\neq i,j}^{h} p_f \delta_{jf} \\ &\qquad + 2\sum\nolimits_{k\neq i,j}^{h-1} p_k \sum\nolimits_{f=k+1}^{h} p_f \delta_{kf} \end{aligned} \tag{18}$$

Combining Eqs. 17 and 18 yields:

$$\begin{aligned} d_{ij} = g_{ij} - \mu - a_i - a_j &= \left[1 - p_i\left(1-p_j\right) - p_j(1-p_i)\right]\delta_{ij} \\ &\quad -(1-2p_i)\sum\nolimits_{k\neq i,j}^{h} p_k \delta_{ik} - \left(1-2p_j\right)\sum\nolimits_{f\neq i,j}^{h} p_f \delta_{jf} \\ &\quad + 2\sum\nolimits_{k\neq i,j}^{h-1} p_k \sum\nolimits_{f=k+1}^{h} p_f \delta_{kf} \end{aligned} \tag{19}$$

$$\begin{aligned} d_{ii} &= g_{ii} - \mu - 2a_i \\ &= -2(1-p_i)\sum\nolimits_{k\neq i}^{h} p_k \delta_{ik} \\ &\quad + 2\sum\nolimits_{k\neq i}^{h-1} p_k \sum\nolimits_{f=k+1}^{h} p_f \delta_{kf} \end{aligned} \tag{20}$$

Equations 13 and 14 show that each additive value is a function of all h – 1 additive effects defined by Eq. 7, and Eqs. 19–20 show that each dominance value is a function of all h(h – 1)/2 dominance effects defined by Eq. 9. Equations 13 and 14 provide the additive coding and Eqs. 19 and 20 provide the dominance coding of each multi-allelic genotype for the mixed model implementation.

Multi-allelic haplotype model based on multi-allelic genetic partition

Using the results of factorization of additive and dominance values given by Eqs. 13–14 and 19–20, the multi-allelic haplotype model treating each haplotype as an 'allele' by Eq. 11 can be expressed as:

$$g_{ij} = \mu + a_{ij} + d_{ij} = \mu + \sum\nolimits_{k=2}^{h} w_{\alpha}^{ij,k} \alpha_{1k} + \sum\nolimits_{k=1}^{h-1} \sum\nolimits_{f=k+1}^{h} w_{\delta}^{ij,kf} \delta_{kf} \tag{21}$$

In $w_{\alpha}^{ij,k}$, superscripts ij are for the genotype of A_iA_j and superscript k is for α_{1k}. In $w_{\delta}^{ij,kf}$, superscripts ij are for d_{ij} and superscripts kf are for δ_{kf}. From Eqs. 13 and 14, the additive coding ($w_{\alpha}^{ij,k}$) of a multi-allelic genotype is:

$$w_{\alpha}^{ij,k} = 2p_k \text{ for } i, j \neq k \left(a_{ij} \text{ and } \alpha_{1k} \text{ do not share allele k}\right) \tag{22}$$

$$w_{\alpha}^{ij,k} = -(1-2p_k) \text{ for } i \neq j \text{ but } i = k \text{ or } j = k \quad \left(a_{ij} \text{ and } \alpha_{1k} \text{ share allele } k, i \neq k\right) \tag{23}$$

$$w_{\alpha}^{ij,k} = -2(1-p_k) \text{ for } i = j = k \quad \left(a_{ij} \text{ and } \alpha_{1k} \text{ share allele k}, \ i = j\right) \tag{24}$$

From Eqs. 19 and 20, the dominance coding ($w_{\delta}^{ij,kf}$) of a multi-allelic genotype is:

$$w_{\delta}^{ij,kf} = 1-p_i\left(1-p_j\right) - p_j(1-p_i) \text{ for } ij = kf \quad \left(d_{ij} \text{ and } \delta_{kf} \text{ share 2 alleles}\right) \tag{25}$$

$$w_{\delta}^{ij,kf} = -p_k(1-2p_i) \text{ for } i \neq j \text{ and } i = f \quad \left(d_{ij} \text{ and } \delta_{kf} \text{ share allele f}, \ i \neq j\right) \tag{26}$$

$$w_{\delta}^{ij,kf} = -p_f\left(1-2p_j\right) \text{ for } i \neq j \text{ and } j = k \quad \left(d_{ij} \text{ and } \delta_{kf} \text{ share allele k}, \ i \neq j\right) \tag{27}$$

$$w_{\delta}^{ij,kf} = -2p_k(1-p_i) \text{ for } i = j \text{ and } i = f \quad \left(d_{ij} \text{ and } \delta_{kf} \text{ share allele f}, \ i = j\right) \tag{28}$$

$$w_{\delta}^{ij,kf} = 2p_k p_f \text{ for } i, j \neq k, f \quad (d_{ij} \text{ and } \delta_{kf} \text{ share no allele}, \ i = j \text{ or } i \neq j) \tag{29}$$

For convenience of computer programming, Eqs. 22–24 can be characterized by whether a_{ij} and α_{lk} share no common allele (Eq. 22), or 1 common allele when $i \neq j$ (Eq. 23) or 1 common allele when $i = j$ (Eq. 24). Similarly, between d_{ij} and δ_{kf}, Eq. 25 shares two common alleles, Eqs. 26 and 27 share 1 common allele with $i \neq j$, Eq. 28 shares one common allele with $i = j$, and Eq. 29 share no common allele. In Eqs. 25–29, p_i or p_j is the allele frequency of the shared allele between d_{ij} and δ_{kf} and p_k or p_f is the allele frequency of the non-shared allele between d_{ij} and δ_{kf}. From Eqs. 21–29, the multi-allelic haplotype model for $h(h+1)/2$ possible genotypic values (**g**) of a given haplotype block with h haplotypes can be expressed as:

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{a}_h + \mathbf{d}_h = \mathbf{1}\mu + \mathbf{W}_{\alpha h}\alpha_h + \mathbf{W}_{\delta h}\delta_h \tag{30}$$

where μ = common mean, $\mathbf{1} = [h(h+1)/2] \times 1$ column vector of 1's, $\mathbf{a}_h = \mathbf{W}_{\alpha h}\boldsymbol{\alpha}_h = [h(h+1)/2] \times 1$ column vector of additive values (breeding values), $\mathbf{d}_h = \mathbf{W}_{\delta h}\boldsymbol{\delta}_h = [h(h+1)/2] \times 1$ column vector of dominance values (dominance deviations), $\mathbf{W}_{\alpha h} = [h(h+1)/2] \times (h-1)$ model matrix of $\boldsymbol{\alpha}_h$with $w_{\alpha}^{ij,k}$ defined by Eqs. 22–24, $d_h = [h(h+1)/2] \times 1$ column vector of dominance values (dominance deviations), $\mathbf{W}_{\delta h} = [h(h+1)/2] \times [h(h-1)/2]$ matrix of $\boldsymbol{\delta}_h$ with $w_{\delta}^{ij,kf}$ defined by Eqs. 25–29, and $\boldsymbol{\alpha}_h = (h-1) \times 1$ column vector with α_{lk} defined by Eq. 7, and $\boldsymbol{\delta}_h = [h(h-1)/2] \times 1$ column vector with δ_{kf} defined by Eq. 9.

Numerical example of multi-allelic genetic partition

A hypothetical numerical example is used to illustrate the genetic partition of multi-allelic genotypic values described by Eqs. 21–30. Four haplotypes as 'alleles' are assumed with frequencies in Table 1 and genotypic values in Table 2. The common mean of

Table 1 Four hypothetical haplotypes and their frequencies (h = 4)

Haplotype	1	2	3	4
Frequency	0.4	0.3	0.2	0.1

Table 2 Genotypic values of haplotype genotypes ($g_{ij} = g_{ji}$)

Haplotype	1	2	3	4
1	$g_{11} = 25$	$g_{12} = 18$	$g_{13} = 15$	$g_{14} = 10$
2		$g_{22} = 30$	$g_{23} = 33$	$g_{24} = 40$
3			$g_{33} = 17$	$g_{34} = 12$
4				$g_{44} = 35$

the genotypic values using Eq. 3 is: $\mu = 22.09$. The additive effects of the four haplotypes defined by Eqs. 5–7, are:

$$\boldsymbol{\alpha}_h' = [-7.4 \quad -1.1 \quad -2.5]',$$

and the dominance effects defined by Eq. 9 are:

$$\boldsymbol{\delta}_h' = [-9.5 \quad -6 \quad -20 \quad 9.5 \quad 7.5 \quad -14]'.$$

Using Eqs. 13–14 and 22–24, the additive values (breeding values) are:

$$\mathbf{a}_h = \begin{bmatrix} a_{11} \\ a_{22} \\ a_{33} \\ a_{44} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{23} \\ a_{24} \\ a_{34} \end{bmatrix} = \begin{bmatrix} 2p_2 & 2p_3 & 2p_4 \\ -2(1-p_2) & 2p_3 & 2p_4 \\ 2p_2 & -2(1-p_3) & 2p_4 \\ 2p_2 & 2p_3 & -2(1-p_4) \\ -(1-2p_2) & 2p_3 & 2p_4 \\ 2p_2 & -(1-2p_3) & 2p_4 \\ 2p_2 & 2p_3 & -(1-2p_4) \\ -(1-2p_2) & -(1-2p_3) & 2p_4 \\ -(1-2p_2) & 2p_3 & -(1-2p_4) \\ 2p_2 & -(1-2p_3) & -(1-2p_4) \end{bmatrix} \begin{bmatrix} \alpha_{12} \\ \alpha_{13} \\ \alpha_{14} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 & 0.2 \\ -1.4 & 0.4 & 0.2 \\ 0.6 & -1.6 & 0.2 \\ 0.6 & 0.4 & -1.8 \\ -0.4 & 0.4 & 0.2 \\ 0.6 & -0.6 & 0.2 \\ 0.6 & 0.4 & -0.8 \\ -0.4 & -0.6 & 0.2 \\ -0.4 & 0.4 & -0.8 \\ 0.6 & -0.6 & -0.8 \end{bmatrix} \begin{bmatrix} -7.4 \\ -1.1 \\ -2.5 \end{bmatrix} = \begin{bmatrix} -5.38 \\ 9.42 \\ -3.18 \\ -0.38 \\ 2.02 \\ -4.28 \\ -2.88 \\ 3.12 \\ 4.52 \\ -1.78 \end{bmatrix}.$$

Using Eqs. 19–20 and 25–29, the dominance values (dominance deviations) are:

$$\mathbf{d}_h = \begin{bmatrix} d_{11} \\ d_{22} \\ d_{33} \\ d_{44} \\ d_{12} \\ d_{13} \\ d_{14} \\ d_{23} \\ d_{24} \\ d_{34} \end{bmatrix} = \begin{bmatrix} -2p_2(1-p_1) & -2p_3(1-p_1) & -2p_4(1-p_1) & 2p_2p_3 & 2p_2p_4 & 2p_3p_4 \\ -2p_1(1-p_2) & 2p_1p_3 & 2p_1p_4 & -2p_3(1-p_2) & -2p_4(1-p_2) & 2p_3p_4 \\ 2p_1p_2 & -2p_1(1-p_3) & 2p_1p_4 & -2p_2(1-p_3) & 2p_2p_4 & -2p_4(1-p_3) \\ 2p_1p_2 & 2p_1p_3 & -2p_1(1-p_4) & 2p_2p_3 & -2p_2(1-p_4) & -2p_3(1-p_4) \\ w_\delta^{12,12} & -p_3(1-2p_1) & -p_4(1-2p_1) & -p_3(1-2p_2) & -p_4(1-2p_2) & 2p_3p_4 \\ -p_2(1-2p_1) & w_\delta^{13,13} & -p_4(1-2p_1) & -p_2(1-2p_3) & 2p_2p_4 & -p_4(1-2p_3) \\ -p_2(1-2p_1) & -p_3(1-2p_1) & w_\delta^{14,14} & 2p_2p_3 & -p_2(1-2p_4) & -p_3(1-2p_4) \\ -p_1(1-2p_2) & -p_1(1-2p_3) & 2p_1p_4 & w_\delta^{23,23} & -p_4(1-2p_2) & -p_4(1-2p_3) \\ -p_1(1-2p_2) & 2p_1p_3 & -p_1(1-2p_4) & -p_3(1-2p_2) & w_\delta^{24,24} & -p_3(1-2p_4) \\ 2p_1p_2 & -p_1(1-2p_3) & -p_1(1-2p_4) & -p_2(1-2p_3) & -p_2(1-2p_4) & w_\delta^{34,34} \end{bmatrix} \begin{bmatrix} \delta_{12} \\ \delta_{13} \\ \delta_{14} \\ \delta_{23} \\ \delta_{24} \\ \delta_{34} \end{bmatrix}$$

$$= \begin{bmatrix} -0.36 & -0.24 & -0.12 & 0.12 & 0.06 & 0.04 \\ -0.56 & 0.16 & 0.08 & -0.28 & -0.14 & 0.04 \\ 0.24 & -0.64 & 0.08 & -0.48 & 0.06 & -0.16 \\ 0.24 & 0.16 & -0.72 & 0.12 & -0.54 & -0.36 \\ 0.54 & -0.04 & -0.02 & -0.08 & -0.04 & 0.04 \\ -0.06 & 0.56 & -0.02 & -0.18 & 0.06 & -0.06 \\ -0.06 & -0.04 & 0.58 & 0.12 & -0.24 & -0.16 \\ -0.16 & -0.24 & 0.08 & 0.62 & -0.04 & -0.06 \\ -0.16 & 0.16 & -0.32 & -0.08 & 0.66 & -0.16 \\ 0.24 & -0.24 & -0.32 & -0.18 & -0.24 & 0.74 \end{bmatrix} \begin{bmatrix} -9.5 \\ -6 \\ -20 \\ 9.5 \\ 7.5 \\ -14 \end{bmatrix} = \begin{bmatrix} 8.29 \\ -1.51 \\ -1.91 \\ 13.29 \\ -6.11 \\ -2.81 \\ -9.21 \\ 7.79 \\ 13.39 \\ -8.31 \end{bmatrix}$$

The genotypic values calculated as the summation of the additive and dominance values are:

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{a}_h + \mathbf{d}_h = \begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix}(22.09) + \begin{bmatrix}-5.38\\9.42\\-3.18\\-0.38\\2.02\\-4.28\\-2.88\\3.12\\4.52\\-1.78\end{bmatrix} + \begin{bmatrix}8.29\\-1.51\\-1.91\\13.29\\-6.11\\-2.81\\-9.21\\7.79\\13.39\\-8.31\end{bmatrix} = \begin{bmatrix}25\\30\\17\\35\\18\\15\\10\\33\\40\\12\end{bmatrix} = \begin{bmatrix}g_{11}\\g_{22}\\g_{33}\\g_{44}\\g_{12}\\g_{13}\\g_{14}\\g_{23}\\g_{24}\\g_{34}\end{bmatrix}.$$

By comparing with the genotypic values in Table 2, the above result verifies that the multi-allelic partition of $\mathbf{g} = \mathbf{1}\mu + \mathbf{a}_h + \mathbf{d}_h = \mathbf{1}\mu + \mathbf{W}_{\alpha h}\boldsymbol{\alpha}_h + \mathbf{W}_{\delta h}\boldsymbol{\delta}_h$ described by Eqs. 21–30 is correct. With the note that $g_{ij} = g_{ji}$, $a_{ij} = a_{ji}$ and $d_{ij} = d_{ji}$, the genotypic variance (σ_g^2), additive variance (σ_a^2) and dominance variance (σ_d^2) are:

$$\begin{aligned}\sigma_g^2 &= \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j g_{ij}^2 - \mu^2 = 71.0419\\ \sigma_a^2 &= \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j a_{ij}^2 = 20.1178\\ \sigma_d^2 &= \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j d_{ij}^2 = 50.9241\end{aligned}$$

It is readily seen that $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$.

Mixed model and multi-allelic genomic relationship matrices

A mixed model to implement the multi-allelic haplotype model of Eq. 30 can be established with appropriate changes of matrix dimensions for $\mathbf{W}_{\alpha h}$, $\mathbf{W}_{\delta h}$, $\mathbf{a}_h$, $\mathbf{d}_h$, $\boldsymbol{\alpha}_h$ and $\boldsymbol{\delta}_h$ in Eq. 30. A set of m SNP markers are assumed available, and r haplotype blocks of the m SNPs are defined across the genome. Haplotypes of all individuals are assumed known (e.g., constructed using a phasing or imputing software). Each haplotype block is treated as a 'locus' and each haplotype within a haplotype block is treated as an 'allele'. The i_{th} haplotype block has h_i haplotypes, h_i-1 additive effects, and $n_{\delta i}$ dominance effects or heterozygous genotypes. Let n_α = total number of additive effects of all r haplotype blocks, n_δ = total number of dominance effects (or heterozygous genotypes) of all r haplotype blocks. Then, $n_\alpha = \Sigma_{i=1}^{r} h_i - r$, and $n_\delta = \Sigma_{i=1}^{r} n_{\delta i}$. For a given sample of q individuals, the limit number of effects is 2q-1 for additive effects and is the number of heterozygous genotypes for dominance effects. For a sample with N observations on q individuals, the mixed model to implement the multi-allelic haplotype model of Eq. 30 can be expressed as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}(\mathbf{W}_{\alpha h}\boldsymbol{\alpha}_h + \mathbf{W}_{\delta h}\boldsymbol{\delta}_h) + \mathbf{e} \qquad (31)$$

where $\mathbf{Z} = N \times q$ incidence matrix allocating phenotypic observations to each individual = identity matrix for one observation per individual ($N = q$), $\boldsymbol{\alpha}_h = n_\alpha \times 1$ column vector of haplotype additive effects, $\mathbf{W}_{\alpha h} = q \times n_\alpha$ model matrix of $\boldsymbol{\alpha}_h$, $\boldsymbol{\delta}_h = n_\delta \times 1$ column vector for dominance effects of haplotype genotypes, $\mathbf{W}_{\delta h} = q \times n_\delta$ model matrix of $\boldsymbol{\delta}_h$, $\boldsymbol{\alpha}_s = m \times 1$ column vector of single-SNP additive effects, $b = c \times 1$ column vector of fixed effects such as heard-year-season in dairy cattle (c = number of fixed effects), and $X = N \times c$ model matrix of $\mathbf{b}$. To define two equivalent models with complementary computing advantages and identical GBLUP and GREML results, the mixed model of Eq. 31 needs to be expressed as [8]:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{Z}(\mathbf{T}_{\alpha h}\boldsymbol{\alpha}_h + \mathbf{T}_{\delta h}\boldsymbol{\delta}_h) + \mathbf{e}\\ &= \mathbf{X}\mathbf{b} + \mathbf{Z}(\mathbf{a}_h + \mathbf{d}_h) + \mathbf{e}\end{aligned} \qquad (32)$$

where $\mathbf{a}_h = \mathbf{T}_{\alpha h}\boldsymbol{\alpha}_h$ = multi-allelic genomic breeding values, $\mathbf{d}_h = \mathbf{T}_{\delta h}\boldsymbol{\delta}_h$ = multi-allelic genomic dominance values, and each $\mathbf{T}$ matrix can be defined by any of the six definitions of genomic relationships we previously discussed and implemented [9]. For simplicity of notations, the $\mathbf{T}$ matrices are defined as: $\mathbf{T}_{\alpha h} = \mathbf{W}_{\alpha h}/k_{\alpha h}^{1/2}$, $\mathbf{T}_{\delta h} = \mathbf{W}_{\delta h}/k_{\delta h}^{1/2}$, where $k_{\alpha h}$ = the average of diagonal elements of $\mathbf{W}_{\alpha h}\mathbf{W}_{\alpha h}'$, and $k_{\delta h}$ = the average of diagonal elements of $\mathbf{W}_{\delta h}\mathbf{W}_{\delta h}'$. The genomic relationship matrices of Eq. 31 can thus be defined as:

$$\begin{aligned}\mathbf{A}_{\mathrm{h}} &= \mathbf{T}_{\alpha\mathrm{h}}\mathbf{T}_{\alpha\mathrm{h}}{}' \\ &= \text{ multi-allelic genomic additive relationship matrix}\end{aligned} \tag{33}$$

$$\begin{aligned}\mathbf{D}_{\mathrm{h}} &= \mathbf{T}_{\delta\mathrm{h}}\mathbf{T}_{\delta\mathrm{h}}{}' \\ &= \text{ multi-allelic genomic dominance relationship matrix}\end{aligned} \tag{34}$$

Interpretation of multi-allelic and haplotype genomic relationship matrices

The multi-allelic genomic relationships of Eqs. 33 and 34 using multi-allelic markers such as microsatellite markers have the same interpretation and theoretical expectation as using SNP markers that are bi-allelic, e.g., a genomic additive relationship is expected to be twice the coancestry coefficient [8, 9]. Using either multi-allelic or bi-allelic markers under the assumption of no inbreeding, the theoretical expectation of genomic additive relationships is 0.5, 0.5, 0.25 and 0 for parent-offspring, full-sibs, half-sibs and unrelated individuals respectively, and the corresponding theoretical expectation of genomic dominance relationships is 0, 0.25, 0 and 0.

It is important to distinguish between single-locus multi-allelic markers such as microsatellite markers from haplotypes where each haplotype is treated as an 'allele' and each haplotype block is treated as a 'locus', because recombination between loci within a haplotype block generally exists, leading to lowered haplotype similarity than single-locus similarity among relatives. As the number of loci increases in each haplotype block, genomic relationships using haplotypes are expected to decrease from those using single-locus markers. Therefore, the utility of haplotype genomic relationships using Eqs. 33 and 34 is for genomic prediction using haplotypes, not for measuring relationships among individuals. The optimal block size and hence the number of haplotypes per block is an important issue for genomic prediction and could be determined by validation studies, as to be further discussed towards the end of this article.

Two equivalent mixed models with complementary computing strategies

To establish mixed models using multi-allelic markers or haplotypes, assumptions for the first and second moments of the mixed model of Eq. 32 are: $E(\mathbf{y}) = \mathbf{Xb}$, $E(\boldsymbol{\alpha}_{\mathrm{h}}) = E(\boldsymbol{\delta}_{\mathrm{h}}) = E(\boldsymbol{\alpha}_{\mathrm{s}}) = E(\boldsymbol{\delta}_{\mathrm{s}}) = 0$, $\mathrm{Var}(\boldsymbol{\alpha}_{\mathrm{h}}) = \sigma^2_{\alpha\mathrm{h}}\mathbf{I}_{\mathrm{n}\alpha}$, $\mathrm{Var}(\mathbf{a}_{\mathrm{h}}) = \mathbf{G}_{\alpha\mathrm{h}} = \sigma^2_{\alpha\mathrm{h}}\mathbf{A}_{\mathrm{h}}$, $\mathrm{Var}(\boldsymbol{\delta}_{\mathrm{h}}) = \sigma^2_{\delta\mathrm{h}}\mathbf{I}_{\mathrm{n}\delta}$, $\mathrm{Var}(\mathbf{d}_{\mathrm{h}}) = \mathbf{G}_{\delta\mathrm{h}} = \sigma^2_{\delta\mathrm{h}}\mathbf{D}_{\mathrm{h}}$, and $\mathrm{Var}(\mathbf{e}) = \mathbf{R} = \sigma^2_{\mathrm{e}}\mathbf{I}_{\mathrm{N}}$, where $\sigma^2_{\alpha\mathrm{h}}$ = variance of multi-allelic additive effects, $\sigma^2_{\delta\mathrm{h}}$ = variance of multi-allelic dominance effects, σ^2_{e} = residual variance, and $\mathbf{I}_{\mathrm{n}\alpha}$, $\mathbf{I}_{\mathrm{n}\delta}$, $\mathbf{I}_{\mathrm{m}}$ and $\mathbf{I}_{\mathrm{N}}$ are identity matrices of orders n_α, n_δ, m and N, respectively. All random effects are assumed to be uncorrelated so that the phenotypic variance-covariance matrix is:

$$\begin{aligned}\mathbf{V} = \mathrm{Var}(\mathbf{y}) &= \mathbf{Z}(\mathbf{G}_{\alpha\mathrm{h}} + \mathbf{G}_{\delta\mathrm{h}})\mathbf{Z}' + \sigma^2_{\mathrm{e}}\mathbf{I}_{\mathrm{N}} \\ &= \mathbf{Z}\left(\sigma^2_{\alpha\mathrm{h}}\mathbf{A}_{\mathrm{h}} + \sigma^2_{\delta\mathrm{h}}\mathbf{D}_{\mathrm{h}}\right)\mathbf{Z}' + \sigma^2_{\mathrm{e}}\mathbf{I}_{\mathrm{N}}\end{aligned} \tag{35}$$

To simply notations for the two equivalent mixed models, terms in Eqs. 32–35 are re-written as $\boldsymbol{\alpha}_{\mathrm{h}} = \boldsymbol{\tau}_1$, $\boldsymbol{\delta}_{\mathrm{h}} = \boldsymbol{\tau}_2$; $\mathbf{T}_{\alpha\mathrm{h}} = \mathbf{T}_1$, $\mathbf{T}_{\delta\mathrm{h}} = \mathbf{T}_2$; $\mathbf{u}_{\mathrm{i}} = \mathbf{T}_{\mathrm{i}}\boldsymbol{\tau}_{\mathrm{i}}$, i = 1,2; $\mathbf{A}_{\mathrm{h}} = \mathbf{S}_1$, $\mathbf{D}_{\mathrm{h}} = \mathbf{S}_2$; and $\sigma^2_{\alpha\mathrm{h}} = \sigma^2_1$, $\sigma^2_{\delta\mathrm{h}} = \sigma^2_2$. Then, Eqs. 32 and 35 can be expressed as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}\sum\nolimits_{i=1}^{2}\mathbf{T}_{\mathrm{i}}\boldsymbol{\tau}_{\mathrm{i}} + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}\sum\nolimits_{\mathrm{i}=1}^{2}\mathbf{u}_{\mathrm{i}} \tag{36}$$

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathbf{Z}\left(\sum\nolimits_{\mathrm{i}=1}^{2}\sigma^2_{\mathrm{i}}\mathbf{S}_{\mathrm{i}}\right)\mathbf{Z}' + \sigma^2_{\mathrm{e}}\mathbf{I}_{\mathrm{N}}. \tag{37}$$

By defining $\mathbf{Z}_{\mathrm{i}} = \mathbf{Z}\mathbf{T}_{\mathrm{i}}$, an equivalent model of Eqs. 36 and 37 can be re-written as:

$$\mathbf{y} = \mathbf{Xb} + \sum\nolimits_{\mathrm{i}=1}^{2}\mathbf{Z}_{\mathrm{i}}\boldsymbol{\tau}_{\mathrm{i}} + \mathbf{e} \tag{38}$$

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \sum\nolimits_{\mathrm{i}=1}^{2}\sigma^2_{\mathrm{i}}\mathbf{Z}_{\mathrm{i}}\mathbf{Z}_{\mathrm{i}}{}' + \sigma^2_{\mathrm{e}}\mathbf{I}_{\mathrm{N}}. \tag{39}$$

Equations 36 and 37 will be referred to as Model-I, and Eqs. 38 and 39 as Model-II. Model-I and Model-II are equivalent models because both models have identical $E(\mathbf{y})$ and $\mathbf{V}$, but these two models have different computational advantages that can be complementary to each other. For each model, two methods can be established for genomic prediction and estimation: the method of conditional expectation (CE) and the method of mixed model equations (MME), yielding a total of four methods for the two equivalent models. Model-I using CE is the best method for large numbers of SNP markers and multiple genetic factors, Model-II using MME is the best method for large numbers of individuals, and Model-I using MME and Model-II using CE have no computing advantage. Therefore, Model-I using CE and Model-II using MME will be used for genomic prediction and estimation. Using our previous naming of these two methods, GBLUP and GREML of Model-I using CE will be referred to as the CE set of formulations, and GBLUP and GREML of Model-II using MME as the QM set of formulation, where QM means '$q > m$'. These two methods yield identical results of prediction and estimation and are applicable to singular genomic relationship matrices. Assuming one observation per individual, CE based on Eqs. 36 and 37 is approximately easier to compute than QM based on Eqs. 38 and 39 if $q < c + n_\alpha + n_\delta$ according to the size of the largest matrix to invert for each method (Table 3). Model-I using MME has no computing advantage over Model-I using CE due to the large coefficient matrix of MME and the requirement for full-rank relationship matrices; and Model-II using CE has no computing advantage over Model-I using CE due to the large $\mathbf{T}$ matrices to store in memory.

Table 3 Comparison of computational feasibility of four methods from the two equivalent models with haplotypes and SNPs for GBLUP and GREML

		Method of for calculating GBLUP	
		Conditional expectation (CE)	Mixed model equations (MME)
Model I, Eqs. 36 and 37	Largest matrix to invert	**V**, phenotypic variance-covariance matrix	**C**, coefficient matrix of MME
	Size of largest matrix to invert	q × q, assuming one observation per individual	c + 2q for **C**
	Largest matrix to store in memory	q × q **P** matrix	c + 2q for **C**
	Applicable to singular genomic relationship matrices	Yes, inverse relationship matrices avoided	No, inverse relationship matrices required
Model II, Eqs. 38 and 39	Largest matrix to invert	**V**, phenotypic variance-covariance matrix	**C**, coefficient matrix of MME
	Size of largest matrix to invert	q × q, assuming one observation per individual	c + n_α + n_δ for **C**
	Largest matrix to store in memory	q × n_α and q × n_δ **T** matrices, q × q **P** matrix	c + n_α + n_δ for **C**
	Applicable to singular genomic relationship matrices	Yes, inverse relationship matrices avoided	Yes, inverse relationship matrices avoided

Genomic best linear unbiased prediction of genetic values (GBLUP)

Using the CE method of Model-I (Eqs. 36 and 37), GBLUP of the $\mathrm{i_{th}}$ type of genetic values for individuals in the training population is obtained as:

$$\hat{\mathbf{u}}_\mathrm{i} = \sigma_\mathrm{i}^2\mathbf{S}_\mathrm{i}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\mathbf{b}}\right) = \sigma_\mathrm{i}^2\mathbf{S}_\mathrm{i}\mathbf{Z}'\mathbf{P}\mathbf{y} = \mathbf{S}_\mathrm{i}\boldsymbol{\varepsilon}_\mathrm{i}, \mathrm{i} = 1,2 \tag{40}$$

where $\widehat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ = best linear unbiased estimator (BLUE) of fixed non-genetic effects, $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$, and $\boldsymbol{\varepsilon}_\mathrm{i} = \sigma_\mathrm{i}^2\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\mathbf{b}}\right) = \mathbf{Z}'\mathbf{P}\mathbf{y} = \mathrm{q} \times 1$ column vector of regressed phenotypic values of the training population as a regression of the $\mathrm{i_{th}}$ type of genetic values on the phenotypic values in the training population. Two equivalent methods with identical results can be used to predict genetic values of individuals without phenotypic observations (validation population): placing all individuals with or without records in the same mixed model by setting to zero the **Z** matrix for the validation population, or calculate predictions separately based on the regressed phenotypic values of the training population [8, 39]. Using this second method, GBLUP of the $\mathrm{i_{th}}$ type of genetic values for individuals in the validation population is calculated as:

$$\hat{\mathbf{u}}_\mathrm{i0} = \sigma_\mathrm{i}^2\mathbf{S}_\mathrm{i01}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\mathbf{b}}\right) = \sigma_\mathrm{i}^2\mathbf{S}_\mathrm{i01}\mathbf{Z}'\mathbf{P}\mathbf{y} = \mathbf{S}_\mathrm{i01}\boldsymbol{\varepsilon}_\mathrm{i} \tag{41}$$

where $\mathbf{S}_\mathrm{i01} = \mathrm{q}_0 \times \mathrm{q}$ genomic relationship matrix between the training and validation populations for the $\mathrm{i_{th}}$ type of genetic values (q_0 = number of individuals in the validation population).

Using the QM method (MME method of Model-II of Eqs. 38 and 39), genomic prediction first calculates the GBLUP of haplotype effects and then calculates GBLUP of genetic values. GBLUP of haplotype effects is obtained from solving the following MME:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\mathrm{g} \\ \mathbf{Z}_\mathrm{g}{}'\mathbf{X} & \mathbf{Z}_\mathrm{g}{}'\mathbf{Z}_\mathrm{g} + \bigoplus_{\mathrm{i}=1}^{2}(\lambda_\mathrm{i}\mathbf{I}_\mathrm{ti}) \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \hat{\boldsymbol{\tau}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_\mathrm{g}{}'\mathbf{y} \end{pmatrix} \tag{42}$$

where $\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{\tau}}_1, , \hat{\boldsymbol{\tau}}_2)$, $\mathbf{Z}_\mathrm{g} = (\mathbf{Z}_1, \mathbf{Z}_2)$, $\lambda_\mathrm{i} = \sigma_\mathrm{e}^2/\sigma_\mathrm{i}^2$, t = n_α, n_δ, m and N for i = 1,2, respectively, and ⊕ denotes direct sum that defines a block diagonal matrix. With haplotype and SNP effects from Eq. 42, GBLUP of the $\mathrm{i_{th}}$ type of genetic values for individuals in the training and validation populations are obtained as:

$$\hat{\mathbf{u}}_\mathrm{i} = \mathbf{T}_\mathrm{i}\hat{\boldsymbol{\tau}}_\mathrm{i} \tag{43}$$

$$\hat{\mathbf{u}}_\mathrm{i0} = \mathbf{T}_\mathrm{i0}\hat{\boldsymbol{\tau}}_\mathrm{i} \tag{44}$$

where $\mathbf{T}_\mathrm{i0}$ = the $\mathbf{T}_\mathrm{i}$ matrix calculated using SNPs of the validation population. Equations 43 and 44 yield identical results as those of Eqs. 40 and 41. The prediction of total genotypic values in either training or validation population can be obtained from Eqs. 40 and 41 or 43 and 44 as: $\hat{\mathbf{g}} = \sum_{\mathrm{i}=1}^{2}\hat{\mathbf{u}}_\mathrm{i}$ = predicted genotypic values of all individuals, and $\hat{\mathbf{g}}_0 = \sum_{\mathrm{i}=1}^{2}\hat{\mathbf{u}}_\mathrm{i0}$ = predicted genotypic values of the validation population. Prediction reliabilities of additive, dominance and genotypic predictions as the squared correlations between the genomic and true values has the same formulations as the R_ai^2, R_di^2 and R_gi^2 formulae in [8], and prediction accuracy is obtained as the square root of the reliability estimate.

Genomic restricted maximum likelihood estimation (GREML) of variance components

Using the CE method of Model-I (Eqs. 36 and 37), the EM type GREML estimates of variance components are:

$$\sigma_i^{2(k+1)} = \sigma_i^{2(k)} \mathbf{y}\mathbf{P}^{(k)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\mathbf{P}^{(k)}\mathbf{y}/\mathrm{tr}\left(\mathbf{P}^{(k)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\right),\quad i = 1,2 \tag{45}$$

$$\sigma_e^{2(k+1)} = \sigma_e^{2(k)} \mathbf{y}\mathbf{P}^{(k)}\mathbf{P}^{(k)}\mathbf{y}/\mathrm{tr}\left(\mathbf{P}^{(k)}\right) \tag{46}$$

where k = iteration number. Using the QM method (Eqs. 38 and 39), the EM type GREML estimates of variance components are

$$\sigma_i^{2(k+1)} = \hat{\boldsymbol{\tau}}_i^{(k)}\hat{\boldsymbol{\tau}}_i^{(k)}/\left[m - \mathrm{tr}\left(\mathbf{C}^{ii(k)}\right)\lambda_i^{(k)}\right] \tag{47}$$

$$\sigma_e^{2(k+1)} = \hat{\mathbf{e}}^{(k)\prime}\hat{\mathbf{e}}^{(k)}/\left\{N - \left[r - \sum_{i=1}^{4}\mathrm{tr}\left(\mathbf{C}^{ii(k)}\lambda_i^{(k)}\right)\right]\right\} \tag{48}$$

where r is the rank of the coefficient matrix of Eq. 42, $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \sum_{i=1}^{2}\mathbf{Z}_i\hat{\boldsymbol{\tau}}_i$, and $\mathbf{C}^{ii}$ is defined by:

$$\mathbf{H}^{-1} = \left(\mathbf{Z}_g{}'\mathbf{M}\mathbf{Z}_g + \bigoplus_{i=1}^{2}\lambda_i\mathbf{I}_{ti}\right)^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}$$

where $\mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, and $ti = n_\alpha$ for $i = 1$ and $ti = n_\delta$ for $i = 2$.

The EM-REML of Eqs. 45–48 are known to be slow but reliable to yield non-negative estimates of variance components. The AI-REML algorithm is fast but may be sensitive to starting values of variance components and may fail for extreme heritability levels. Formulations of AI-REML for the multi-allelic haplotype model in this article are straightforward extensions of the formulations we implemented for GVCBLUP [40].

Integration of haplotype and single SNP effects in genomic prediction and estimation

Haplotype analysis and single SNP analysis can be analyzed jointly for genomic prediction in the same mixed model by adding single SNP effects from our previous work [8] to the mixed model of Eq. 31, i.e.,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}(\mathbf{T}_{\alpha h}\boldsymbol{\alpha}_h + \mathbf{T}_{\delta h}\boldsymbol{\delta}_h + \mathbf{T}_{\alpha s}\boldsymbol{\alpha}_s + \mathbf{T}_{\delta s}\boldsymbol{\delta}_s) + \mathbf{e} \tag{49}$$

$$\begin{aligned}\mathbf{V} &= \mathrm{Var}(\mathbf{y}) \\ &= \mathbf{Z}\left(\sigma_{\alpha h}^2\mathbf{A}_h + \sigma_{\delta h}^2\mathbf{D}_h + \sigma_{\alpha s}^2\mathbf{A}_s + \sigma_{\delta s}^2\mathbf{D}_s\right)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N\end{aligned} \tag{50}$$

where $\boldsymbol{\alpha}_s = m \times 1$ column vector of SNP additive effects, $\mathbf{T}_{\alpha s} = q \times m$ model matrix of $\boldsymbol{\alpha}_s$, $\boldsymbol{\delta}_s = m \times 1$ column vector of SNP dominance effects, $\mathbf{T}_{\delta s} = q \times m$ model matrix of $\boldsymbol{\delta}_s$, $\mathrm{Var}(\boldsymbol{\alpha}_s) = \sigma_{\alpha s}^2\mathbf{I}_m$, $\mathrm{Var}(\mathbf{a}_s) = \mathbf{G}_{\alpha s} = \sigma_{\alpha s}^2\mathbf{A}_s$, $\mathrm{Var}(\boldsymbol{\delta}_s) = \sigma_{\delta s}^2\mathbf{I}_m$, $\mathrm{Var}(\mathbf{d}_s) = \mathbf{G}_{\delta s} = \sigma_{\delta s}^2\mathbf{D}_s$, $\mathbf{A}_s$ = genomic additive relationship matrix, and $\mathbf{D}_s$ = SNP genomic dominance relationship matrix, and where $\mathbf{A}_s = \mathbf{T}_{\alpha s}\mathbf{T}_{\alpha s}{}'$ and $\mathbf{D}_s = \mathbf{T}_{\delta s}\mathbf{T}_{\delta s}{}'$. Let $\boldsymbol{\alpha}_s = \boldsymbol{\tau}_3$, $\boldsymbol{\delta} = \boldsymbol{\tau}_4$; $\mathbf{u}_i = \mathbf{T}_i\boldsymbol{\tau}_i$, $i = 1,\ldots,4$; $\mathbf{A}_s = \mathbf{S}_3$, $\mathbf{D}_h = \mathbf{S}_4$; and $\sigma_{\alpha s}^2 = \sigma_3^2$, $\sigma_{\delta s}^2 = \sigma_4^2$. The GBLUP and GREML formulations to jointly include haplotype and single SNP additive and dominance effects essentially entails to extending the range of the subscript i from 2 to 4 for Eqs. 38–50.

GREML estimation using the joint mixed model with haplotype and SNP effects offer flexibility to estimate the heritability for various types of functional genomic information in any given autosome regions based on

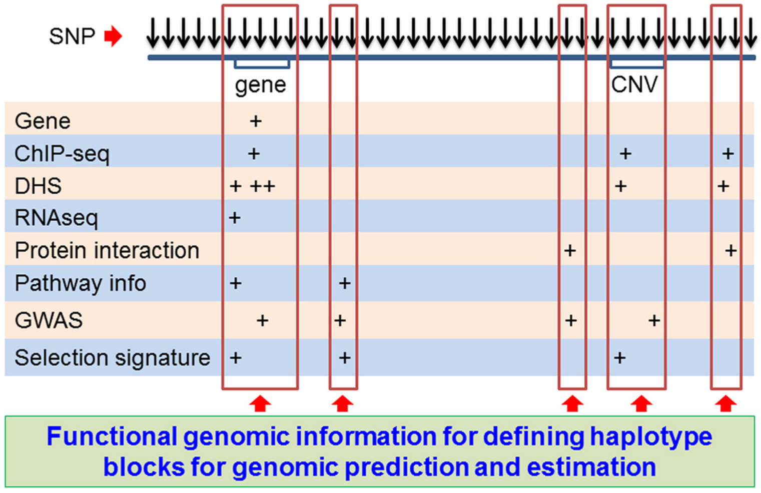
SNP
gene
CNV
Gene
ChIP-seq
DHS
RNAseq
Protein interaction
Pathway info
GWAS
Selection signature
+
+ ++
Functional genomic information for defining haplotype blocks for genomic prediction and estimation

Fig. 1 Integration of functional and structural genomic information for genomic selection. Haplotype blocks are defined using functional genomic information and are analyzed using the multi-allelic haplotype model in this article for genomic prediction and estimation. Single SNPs as structural genomic information can be used jointly with the haplotype analysis. (DHS = DNase I hypersensitive site)

formulations we implemented in GVCBLUP [40], e.g., the additive and dominance heritabilities of haplotype blocks of all genes, all LD blocks, or all single SNPs. The heritability estimate for each type of genetic effects is: $h_i^2 = \sigma_i^2/\sigma_y^2$, where $\sigma_y^2 = \sum_{i=1}^{4}\sigma_i^2 + \sigma_e^2$ = phenotypic variance. The total heritability of all types of genetic effects is the summation of all effect heritabilities, i.e., $H^2 = \sum_{i=1}^{4} h_i^2$. Genomic heritability estimation has flexibility unavailable from heritability estimation using pedigree relationships: the heritability estimation for a single SNP, a chromosome region, or a set of selected SNPs. Using the GREML formulae of Eqs. 35 and 36, the heritability for haplotype block j or SNP set j can be estimated as: $h_{ij}^2 = \left(\hat{\boldsymbol{\tau}}_{ij}'\hat{\boldsymbol{\tau}}_{ij}/\hat{\boldsymbol{\tau}}_i'\hat{\boldsymbol{\tau}}_i\right)h_i^2$, where $\hat{\boldsymbol{\tau}}_{ij}$ = subset j of $\hat{\boldsymbol{\tau}}_i$, i = 1,…,4. Given sufficient computing power and sample sizes for extensive validation studies, these heritability estimates could help identify genomic regions and genes relevant to phenotypes within the framework of genomic prediction.

Defining haplotype blocks using functional genomic information

The multi-allelic haplotype model can be used for the integration of functional genomic information with genomic prediction and estimation. This integration defines haplotype blocks using functional genomic information under the hypothesis that a chromosome region with functional information required more than a single point to affect a phenotype, followed by genomic prediction and estimation using a haplotype analysis such as the methods developed in this article. Each gene could be a 'natural haplotype block' and the use of gene blocks improved the prediction accuracy for some human phenotypes in our preliminary results [37]. Other types of functional information can also be used to define haplotype blocks, including ChIP-seq sites, DNA methylation sites, CNV, protein interaction, pathway information, GWAS results and selection signatures (Fig. 1). Other than 'natural haplotype blocks', the optimal block sizes for functional information with best prediction accuracy could be determined by extensive validation studies.

Rare haplotypes, missing genotypic values

The mixed model approach outlined above allows rare haplotypes. In the extreme case of rare haplotypes with one observation per haplotype or haplotype frequency of 1/h when h is large, the multi-allelic model with the mixed model implementation still is applicable for additive effects and values. Missing genotypic values is a problem for dominance effects and values. The dominance effect defined by Eq. 9 requires the availability of all three genotypic values of a haplotype pair. Consequently, dominance effect is undefined with any missing genotypic value. We currently recommend ignoring any haplotype pair with missing genotypic value or values for defining dominance effects. For large haplotype blocks, nearly all individuals could be heterozygous so that such large blocks may not contribute to genomic prediction and estimation of dominance effects and values. This loss of dominance information should be a factor to consider in defining the block size.

Conclusions

A multi-allelic haplotype model for genomic prediction and estimation is established using the quantitative genetics model that partitions a multi-allelic genotypic value into additive and dominance values, factorizes each additive value into a product between a function of allele frequencies and additive effect, and factorizes each dominance value into a product between a function of allele frequencies and dominance effect. Haplotype genomic additive and dominance relationship matrices and formulations are then derived for GBLUP and GREML utilizing haplotypes in haplotype blocks. These results fill a gap in the theory of quantitative genetics for multi-allelic genetic partition and provide a haplotype approach within the theory of quantitative genetics towards the integration of functional and structural genomic information for genomic selection.

Availability of supporting data

The only data set used in this article is shown in Tables 1–2.

Abbreviations

SNP: single nucleotide polymorphism; BLUP: best unbiased linear prediction; GBLUP: genomic BLUP; REML: restricted maximum likelihood estimation; GREML: genomic REML; EM: expectation-maximization; AI-REML: average information REML; CE: conditional expectation; MME: mixed model equations.

Competing interests

The author declares to have no competing interests.

Acknowledgements

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. Dzianis Prakapenka and Chunkao Wang implemented the methodology in this article by the GVCHAP computer program. Cheng Tan and Dzianis Prakapenka evaluated the methodology. John R. Garbe provided summary and discussion of human functional genomic information. Li Ma processed a dataset for methodology evaluation.

Received: 6 May 2015 Accepted: 27 November 2015
Published online: 18 December 2015

References

1. Henderson C. Applications of Linear Models in Animal Breeding. Guelph: University of Guelph; 1984.
2. Fikse W, Philipsson J. Development of international genetic evaluations of dairy cattle for sustainable breeding programs. Anim Genet Resour Inf. 2007; 41:29–43.
3. Powell R, VanRaden P. International dairy bull evaluations expressed on national, subglobal, and global scales. J Dairy Sci. 2002;85(7):1863–8.

4. VanRaden P. Invited Review: Selection on Net Merit to Improve Lifetime Profit. J Dairy Sci. 2004;87(10):3125–31.
5. Wiggans G, Misztal I, Van Vleck L. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J Dairy Sci. 1988;71:54–69.
6. VanRaden P. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
7. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. Biometrika. 1971;58(3):545–54.
8. 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.
9. Wang C, Da Y. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. PLoS ONE. 2014;9:e114484.
10. Hayes B, Goddard M. Genome-wide association and genomic selection in animal breeding. Genome. 2010;53(11):876–83.
11. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
12. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans Roy Soc Edinb. 1918;52(02):399–433.
13. Fisher RA. Average excess and average effect of a gene substitution. Ann Eugen. 1941;11(1):53–63.
14. 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.
15. Kempthorne O. The correlation between relatives in a random mating population. Proc R Soc Lond B Biol Sci. 1954;143(910):103–13.
16. Lynch M, Walsh B. Genetics and analysis of quantitative traits, Sinauer Sunderland, Massachusetts; 1998.
17. Kempthorne O. An introduction to genetic statistics. New York: Wiley; 1957.
18. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th ed. Harlow, Essex: Longmans Green; 1996.
19. Álvarez-Castro JM, Yang R-C. Multiallelic models of genetic effects and variance decomposition in non-equilibrium populations. Genetica. 2011;139(9):1119–34.
20. Vormfelde SV, Brockmöller J: On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and-genomics. Nature Reviews Genetics 2007, 8(12), doi:10.1038/nrg1916-c1.
21. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781–91.
22. Garnier S, Truong V, Brocheton J, Zeller T, Rovital M, Wild PS, et al. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. PLoS Genet. 2013;9(1):e1003240.
23. Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol. 2002;23(3):221–33.
24. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–5.
25. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–8.
26. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210–23.
27. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.
28. Von Holdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature. 2010;464(7290):898–902.
29. Calus M, De Roos A, Veerkamp R. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 2008;178(1):553–61.
30. Villumsen T, Janss L, Lund M. The importance of haplotype length and heritability using genomic selection in dairy cattle. J Anim Breed Genet. 2009;126(1):3–13.
31. Sun X, L. FR, Garrick DJ, Dekkers JCM: Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes. Proceedings, 10th World Congress of Genetics Applied to Livestock Production Vancouver, BC, Canada https://asas.org/docs/default-source/wcgalp-proceedings-oral/209_paper_9178_manuscript_1682_0.pdf?sfvrsn=2 [Last accessed December 8 2015].
32. Cuyabano BC, Su G, Lund MS. Selection of haplotype variables from a high-density marker map for genomic prediction. Genet Sel Evol. 2015;47(1):1–11.
33. Mulder HA, Calus MP, Veerkamp RF. Prediction of haplotypes for ungenotyped animals and its effect on marker-assisted breeding value estimation. Genet Sel Evol. 2010;42:10.
34. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics. 2010;185(2):623–31.
35. Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114–29.
36. Brøndum RF, Su G, Lund MS, Bowman PJ, Goddard ME, Hayes BJ. Genome position specific priors for genomic prediction. BMC Genomics. 2012;13(1):543.
37. Da Y, Wang C, Tan C, Prakapenka D, Shigematsu M, Garbe J, Ma L: Multi-allelic haplotype model for genomic prediction and estimation. Abstract P1176. Plant and Animal Genome XXIII, January 10–14, 2015. San Diego. https://pag.confex.com/pag/xxiii/webprogram/Paper14435.html [Last accessed December 8 2015].
38. Tan C, Prakapenka D, Wang C, Ma L, Garbe JR, Hu X, Da Y: Integration of haplotype analysis of functional genomic information with single SNP analysis improved accuracy of genomic prediction. ADSA/ASAS 2015, Orlando, July 12–16 2015. Abstract M84. http://m.jtmtg.org/abs/t/65063. [Last accessed December 8 2015].
39. Henderson C. Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci. 1977;60(5):783–7.
40. 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.