GSE Genetics Selection Evolution

RESEARCH Open Access

Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action

Jian Zeng[1], Ali Toosi[2], Rohan L Fernando[1*], Jack CM Dekkers[1] and Dorian J Garrick[1]

Abstract

Background: Genomic selection is an appealing method to select purebreds for crossbred performance. In the case of crossbred records, single nucleotide polymorphism (SNP) effects can be estimated using an additive model or a breed-specific allele model. In most studies, additive gene action is assumed. However, dominance is the likely genetic basis of heterosis. Advantages of incorporating dominance in genomic selection were investigated in a two-way crossbreeding program for a trait with different magnitudes of dominance. Training was carried out only once in the simulation.

Results: When the dominance variance and heterosis were large and overdominance was present, a dominance model including both additive and dominance SNP effects gave substantially greater cumulative response to selection than the additive model. Extra response was the result of an increase in heterosis but at a cost of reduced purebred performance. When the dominance variance and heterosis were realistic but with overdominance, the advantage of the dominance model decreased but was still significant. When overdominance was absent, the dominance model was slightly favored over the additive model, but the difference in response between the models increased as the number of quantitative trait loci increased. This reveals the importance of exploiting dominance even in the absence of overdominance. When there was no dominance, response to selection for the dominance model was as high as for the additive model, indicating robustness of the dominance model. The breed-specific allele model was inferior to the dominance model in all cases and to the additive model except when the dominance variance and heterosis were large and with overdominance. However, the advantage of the dominance model over the breed-specific allele model may decrease as differences in linkage disequilibrium between the breeds increase. Retraining is expected to reduce the advantage of the dominance model over the alternatives, because in general, the advantage becomes important only after five or six generations post-training.

Conclusion: Under dominance and without retraining, genomic selection based on the dominance model is superior to the additive model and the breed-specific allele model to maximize crossbred performance through purebred selection.

Background

Numerous studies have shown encouraging results of applying genomic selection (GS) in purebred populations [1-6]. However, except for dairy cattle, most animals used in livestock production systems are crossbreds, with advantages of heterosis and breed complementarity. For such systems, the breeding goal in purebreds should be to optimize the performance of crossbred descendents.

GS has advantages in selection for crossbred performance over conventional methods [7-9] as reported in [9]. In particular, training on crossbred data for GS accounts for genetic differences between purebred and crossbred animals and genotype by environment effects. Different GS models have been proposed and used to select purebreds for crossbred performance [8-11]. In most studies, additive gene action or perfect knowledge of gene substitution effects or both have been assumed. It has been

*Correspondence: rohan@iastate.edu
[1]Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA, USA
Full list of author information is available at the end of the article

argued that dominance is the likely genetic basis of heterosis [12,13], therefore explicitly including dominance in the GS model may be beneficial for selection of purebreds for crossbred performance.

Allele substitution effects are a function of additive and dominance effects, and of allele frequencies [12]. In the case of crossbred records, Dekkers [8] proposed a model that fits breed-specific allele substitution effects (BSAM), where the substitution effects at the QTL (quantitative trait loci) for paternal and maternal alleles would be different if the parental breeds differ in allele frequencies at the QTL. It has been shown that, when additive and dominance effects of the QTL are known without error, BSAM can give greater response to selection than the usual additive model that fits a common substitution effect for each QTL [10,11]. When the QTL genotype is not observed and marker genotypes are fitted in the model, linkage disequilibrium (LD) between SNPs and the QTL may not be consistent between the parental breeds. This difference in LD will also contribute to differences in allele substitution effects between breeds. When SNP effects must be estimated, the advantage of fitting BSAM over the additive GS model was not always observed under additive inheritance [9], which suggests that differences in LD between breeds may not be as important as the presence of dominant gene action in practice.

Dekkers [14] showed that, to maximize performance in the first generation of crossbreds, the allele substitution effect for an identified QTL must be derived based on allele frequencies in the selected mates. However, allele frequencies of selected mates cannot be observed prior to computation of the substitution effects that are needed for selection. Thus, Dekkers [14] proposed an iterative algorithm to compute substitution effects based on selected mates.

A model that explicitly includes dominance effects (the dominance model) provides estimates of both additive and dominance effects and therefore enables the computation of allele substitution effects using appropriate allele frequencies. Once estimates of SNP effects are obtained from training, they can be successively applied over generations with updated allele frequencies to develop prediction equations specific to that generation.

The BSAM model gives allele substitution effects for each parental breed that depend on allele frequencies among individuals from the other breed that were used to produce the training population. It is not straightforward to apply the iterative algorithm of Dekkers to BSAM. In addition, the substitution effects from BSAM cannot be used to recompute the appropriate substitution effects in the subsequent generations, as allele frequencies change due to selection. Furthermore, breed origin of SNP alleles must be known or inferred for the BSAM model [8,10], but such knowledge is not needed for the dominance model.

The primary objective of this study was to assess the performance of the dominance model in comparison with the additive model or BSAM for GS on purebreds for crossbred performance. Substitution effects from the dominance model were computed based on allele frequencies of unselected mates. Ibánẽz-Escriche et al. [9] compared BSAM to the additive model under additive gene action alone and Kinghorn et al. [10,11] made this comparison with dominant gene action when QTL effects were assumed known. Thus, a secondary objective of this study was to compare BSAM to the additive model under dominance when SNP effects must be estimated. Model performance was evaluated by computer simulation based on response to 20 generations of selection in a two-way crossbreeding program.

Methods

Simulations

Comparisons between the dominance, BSAM and additive models were made for four scenarios of gene action. To clearly detect an advantage of including dominance in the model, the dominance variance V_D and heterosis H were chosen to be large in scenario 1, allowing for overdominance. In scenarios 2 and 3, V_D and H were restricted to more realistic values with (scenario 2) or without (scenario 3) overdominance. In scenario 4, V_D was reduced to zero to examine any disadvantage of using the dominance model when gene action is purely additive. Changes to V_D and H were achieved by changing the size and proportion of beneficial dominance effects. Other parameters, including locus positions and LD between loci and allele frequencies were held constant between the four scenarios. A total of 16 random simulations were carried out for each scenario.

Genome and trait phenotypes

A genome was simulated with either one or ten chromosomes. Each chromosome was one Morgan and consisted of 100 randomly distributed QTL and 1000 SNP markers that were almost evenly spaced. All loci were biallelic with starting allele frequencies of 0.5 and a reversible mutation rate of 2.5×10^{-5}. A binomial map function was used to model recombination with interference on a chromosome [15]. A base population of 500 unrelated individuals was randomly mated for 1000 discrete generations to create LD between loci in the founders of different breeds.

The additive effect a of a QTL is defined as half the difference in genotypic value between alternate homozygotes and the dominance effect d as the deviation of the value of the heterozygote from the mean of the two homozygotes [12]. Bennewitz and Meuwissen [16] evaluated QTL mapping results from many studies in pigs for meat

quality and carcass traits and concluded that an exponential distribution with rate parameter 5.81 is an adequate "generating mechanism" for the absolute values of the additive effects of QTL. That distribution was used here to generate the unsigned value of the additive effect for each QTL, and a positive or negative sign was assigned to each additive effect with equal probability. Although the relationship between additive and dominance effects of QTL has been studied [16-18], a consistent relationship has not been observed. Thus, in scenarios 1 and 2 with overdominance, we assumed, for simplicity, that the dominance effects were independent of additive effects. The absolute values of dominance effects were independently sampled from the same exponential distribution as was used for the additive effects. This was considered reasonable because the exponential distribution has a high probability for the occurrence of small effects, which is also plausible for dominance effects. In scenario 3 with no overdominance, the dominance effect of a QTL was sampled from a uniform distribution that ranged from zero to the absolute value of the QTL's additive effect. In order for the trait to manifest positive heterosis, only 30% of the sampled dominance effects were negative in scenarios 1 and 2, and this fraction was further reduced to 20% in scenario 3 for the sake of no overdominance. The resulting distribution of dominance coefficients, defined as the ratio of the dominance effect over the absolute value of additive effect, was similar to what has been observed for real data [16].

The QTL effects were scaled (as described in Appendix A) such that the relative contributions of the additive and dominance effects to the genetic variability of the trait were 2:1, 4:1 and 1:0 in the scenarios where V_D was set to be large, realistic or null, respectively. After scaling, 40-45% of QTL showed partial dominance and 30-35% overdominance when overdominance was present. Trait phenotypes were simulated by adding a standard normal residual effect to the genotypic value of each animal. The variance of the residual effects was chosen such that broad sense heritability h^2_{bs} of the trait was 0.5 in the founders. As a result, narrow sense heritability h^2_{ns} was 0.33 for the scenario with large V_D, 0.4 with realistic V_D, and 0.5 with $V_D = 0$.

Breed formation

Breeds A and B were simulated by randomly sampling 100 animals from the founders in generation -55 and random mating for 54 additional generations to mimic recent breed formation (Figure 1). In the founder generation, 100 QTL and 1000 SNPs were randomly chosen from those with a minor allele frequency greater than 0.1. To guarantee that the number of such loci to choose from would be sufficient, five times as many loci were simulated in the base population. This procedure ensured that most QTL chosen to define the trait and SNPs chosen for inclusion in the analysis segregated in both breeds.

In generation -1, the genetic disparity between breeds was due to differences in LD and in allele frequencies. Averaged over simulations, the heterozygosity of each breed was about 0.3, and the mean difference in allele frequencies between breeds was about 0.3. In each simulation, while the same set of QTL characterized the trait, the contribution of QTL effects to the phenotypic variability differed between breeds due to disparities in allele frequencies. Between simulations, the observed values of variance components in a given breed varied due to genetic drift during the 54 generations of random mating after breed separation.

In livestock, dominance can explain up to about 10% of phenotypic variation [19] and heterosis from a breed cross can be up to 10% [20]. In scenario 1, where V_D was large, V_D was on average 16.7% in the pure breeds and H was on average 31%. Under more realistic settings, $\overline{V}_D = 10\%$ and $\overline{H} = 9.1\%$ in scenario 2, where overdominance was allowed, and $\overline{V}_D = 4.5\%$ and $\overline{H} = 8.4\%$ in scenario 3, where overdominance was not allowed. When the genome was extended from one to ten chromosomes, these values were kept about the same by reducing the proportion of beneficial dominance effects.

Crossbreeding program

A two-way crossbreeding program with 20 generations of selection was simulated, as illustrated in Figure 1. The goal was to improve crossbred performance through selection in both parental breeds, starting from 1000 animals of sire breed (A) and 1000 animals of dam breed (B) in generation 0. The selection criteria was the rank of the individual's genomic estimated breeding values (GEBV). The SNP effects for the prediction of GEBV were estimated only once, using 1000 crossbred AB_0 animals in generation 0. These estimates of SNP effects were then repeatedly applied to predict GEBV in the following 20 generations of selection. In generation 1 through 20, 600 animals were selected from 1000 candidates in each parental breed based on their GEBV, of which the top 100 were used as males and the other 500 as females. As described in detail later, with the dominance model, GEBV were based on the breed-specific allele substitution effects that were recomputed for each generation, using allele frequencies in the opposite breed in that generation. The selected animals were randomly mated within each breed to produce 1000 purebred replacement animals for the next generation. Meanwhile, the 100 selected males of breed A were also randomly mated to the 500 selected females of breed B to produce 1000 crossbred progeny. The phenotypic mean of crossbreds was computed in each generation of selection (AB_1 – AB_{20}) to evaluate the cumulative response to selection.

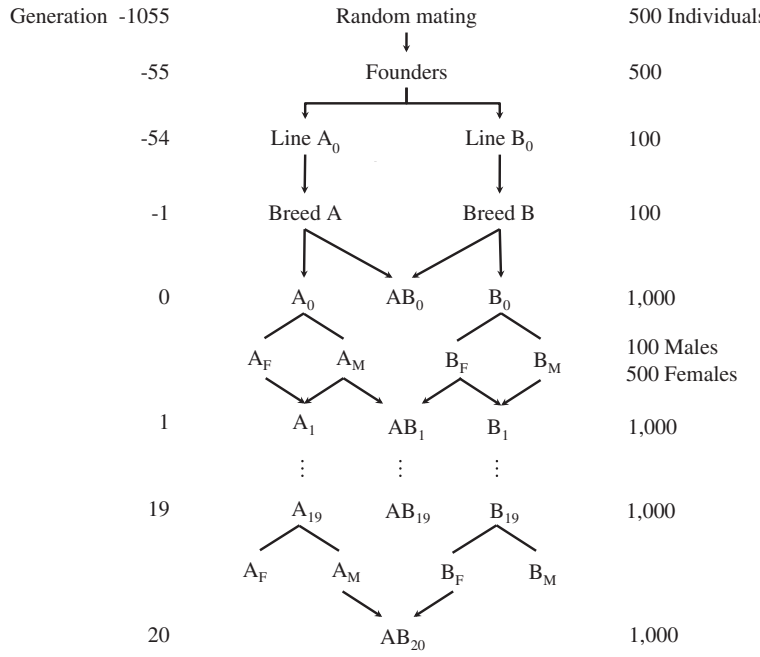

Figure 1 Schematic representation of the simulated population history and the two-way crossbreeding program. The crossbreeding program consisted of 20 generations of purebred selection for crossbred performance; crossbred AB_0 is the training population; A_M and B_M represent the selected breed A and B males, A_F and B_F the selected breed A and B females; lines with arrows denote reproduction, while lines without arrows denote selection.

Starting from the same set of purebred selection candidates in generation 0, the subsequent 20 generations of selection were repeated 100 times to increase power to detect differences between the GS models in cumulative response. A mixed linear model (see Appendix B) was used to test for differences between GS models by generation 20. Note that the 16 random simulations each with 100 repetitions account for differences in purebreds due to genetic drift. In sum, the collected data consisted of 1600 observations in each generation of selection.

The cumulative response in the i^{th} generation of selection, where $i = 1, ..., 20$, was computed as

$$R_i = \frac{\mu_i - \mu_0}{\sigma_0},$$

where μ_i is the phenotypic mean of crossbreds in generation i and σ_0 is the phenotypic standard deviation of crossbreds in generation 0.

Statistical models

Additive model

The following mixed linear model was used to estimate SNP effects assuming additive gene action:

$$y_i = \mu + \sum_{j=1}^{k} X_{ij}\alpha_j + e_i,$$

where y_i is the phenotype of animal i, μ is the overall mean, X_{ij} is the copy number of a given allele of SNP j centered by the mean, α_j is the allele substitution effect for SNP j, and e_i is the residual effect for animal i. The prior specification for model parameters and the sampling strategy followed the BayesCπ method proposed by [6]. In order to concentrate the signal and reduce noise, only a proportion of SNPs was assumed to have a non-null effect. Conditional on σ_α^2, the variance of random substitution effects for all SNPs, α_j had a mixture prior of a normal distribution and a point mass at zero:

$$\alpha_j|\sigma_\alpha^2 = \begin{cases} 0 & \text{with probability } \pi \\ \sim N(0, \sigma_\alpha^2) & \text{with probability } 1-\pi. \end{cases} \quad (1)$$

The proportion π of SNPs that have null effects on the trait was considered unknown, with a uniform prior between 0 and 1. A scaled inverse Chi-square distribution with degrees of freedom $\nu_\alpha = 4$ and scale parameter S_α^2 was specified as a prior for $\sigma_\alpha^2 \sim \nu_\alpha S_\alpha^2 \chi_{\nu_\alpha}^{-2}$. The value of S_α^2 was chosen based on the following relationship between the expectation of a scaled inverse Chi-square variable and its scale parameter:

$$\mathrm{E}(\sigma^2) = \frac{S^2\nu}{\nu - 2}, \quad (2)$$

and $\mathrm{E}(\sigma_\alpha^2)$ was obtained following [21] such that $\mathrm{E}(\sigma_\alpha^2) = \frac{V_A}{k(1-\pi_0)\mathrm{E}(2pq)}$, where k is the total number of SNPs, π_0 is

the chosen probability that a SNP has no effect prior to the analysis, $p = 1 - q$ is the allele frequency, and V_A is the additive genetic variance for the trait that is explained by all SNPs. The residual e_i had a normal prior with variance also following a scaled inverse Chi-square distribution: $e_i|\sigma_e^2 \sim N(0, \sigma_e^2)$ and $\sigma_e^2 \sim \nu_e S_e^2 \chi_{\nu_e}^{-2}$, where $\nu_e = 4$. The value of S_e^2 was obtained from (2), with $\mathrm{E}(\sigma_e^2) = V_E$, where V_E is the residual variance that cannot be explained by the SNPs. True values were given to V_A and V_E in this study.

Dominance model

The dominance model, as shown below, simultaneously fits additive and dominance effects of SNPs:

$$y_i = \mu + \sum_{j=1}^{k}(X_{ij}a_j + W_{ij}d_j) + e_i, \tag{3}$$

where y_i, μ, X_{ij} are as defined in the additive model, W_{ij} is the indicator variable for the heterozygous genotype of SNP j that is centered by its mean, a_j is the additive effect, and d_j the dominance effect for SNP j, and e_j is the residual. Given the assumption that epistasis is absent, the residual term in the dominance model only contains non-genetic effects, while that of the additive model also includes dominance deviations. The model specification for the dominance model is similar to that of the additive model. Conditional on π_a (the probability that a_j is zero) and σ_a^2 (the variance of a_j when it is nonzero), the prior for a_j is a mixture of normals, as given in the additive model (1). Similarly, the prior for d_j is also a mixture of normals, given π_d and σ_d^2, with corresponding definitions. However, in order to account for the directionality of dominance, the normal component of the prior for d_j has an unknown nonzero mean μ_d:

$$d_j|\mu_d, \sigma_d^2 = \begin{cases} 0 & \text{with probability } \pi_d \\ \sim N(\mu_d, \sigma_d^2) & \text{with probability } 1 - \pi_d. \end{cases} \tag{4}$$

For convenience, the prior of μ_d was assumed to depend on the variance:

$$\mu_d|\sigma_d^2 \sim N(\eta, \sigma_d^2/\phi),$$

where η is our prior belief about μ_d and ϕ is the "prior sample size", which expresses the strength of the prior belief in terms of σ_d^2. Here, ϕ was set to 10, a small number relative to 100 QTL, to allow the data to "dominate" the posterior distribution of μ_d. The value of η was chosen as described below.

Let the allele frequency at SNP j be p_j^S in sires and p_j^D in dams in generation -1. Assuming Hardy-Weinberg equilibrium in the parental populations, heterosis (H) in the training crossbred AB_0 is a function of the dominance effects and the difference in allele frequencies in the parental populations ($\Delta = p_j^S - p_j^D$) [12]:

$$H = \sum\nolimits_j d_j \Delta^2.$$

Assuming independence between dominance effects and allele frequencies and ignoring selection, this can be written as

$$H = k(1 - \pi_{d,0})\mathrm{E}(d)\mathrm{E}(\Delta^2),$$

where $\pi_{d,0}$ is a chosen value for the proportion of SNPs that have nonzero dominance effects. Assuming that each QTL is associated with at least one SNP, $\pi_{d,0}$ should be at most 0.9, as 100 QTL and 1000 SNPs were simulated. Rearranging gives

$$\eta = \mathrm{E}(d) = \frac{H}{k(1 - \pi_{d,0})\mathrm{E}(\Delta^2)}.$$

The variance components σ_a^2 and σ_d^2 were assumed to have independent scaled inverse Chi-square distributions. As shown in (2), specification of the hyper parameters S_a^2 and S_d^2 requires knowing $\mathrm{E}(\sigma_a^2)$ and $\mathrm{E}(\sigma_d^2)$. The following describes how $\mathrm{E}(\sigma_a^2)$ or $\mathrm{E}(\sigma_d^2)$ were computed based on the known quantities V_A and V_D.

Given independence between SNPs holds and in the absence of selection [12]:

$$V_D = \sum\nolimits_j (2p_j q_j d_j)^2. \tag{5}$$

Assuming independence between effects and allele frequencies, this can be written as

$$\begin{aligned} V_D &= k(1 - \pi_{d,0})\mathrm{E}[\,(2pq)^2]\,\mathrm{E}(d^2) \\ &= k(1 - \pi_{d,0})\mathrm{E}[\,(2pq)^2]\,\{(1 + 1/\phi)\sigma_d^2 + [\,\mathrm{E}(d)]^2\,\}. \end{aligned}$$

Rearranging and replacing $\mathrm{E}(d)$ by η gives

$$\sigma_d^2 = \left(\frac{V_D}{k(1 - \pi_{d,0})\mathrm{E}[\,(2pq)^2]} - \eta^2\right)/(1 + 1/\phi). \tag{6}$$

Also [12]:

$$V_A = \sum\nolimits_j (2p_j q_j \alpha_j^2). \tag{7}$$

Under the same assumptions as made in (5), in addition to additive effects having mean zero and being independent of dominance effects, this becomes

$$V_A = k(1 - \pi_{a,0})\mathrm{E}(2pq)\mathrm{E}(\alpha^2), \tag{8}$$

where

$$\begin{aligned} \mathrm{E}(\alpha^2) &= \mathrm{E}\{[\,a + (1 - 2p)d]^2\,\} \\ &= \sigma_a^2 + \mathrm{E}[\,(q - p)^2]\,(\sigma_d^2 + \eta^2). \end{aligned}$$

Substituting this in (8) and rearranging gives

$$\sigma_a^2 = \frac{V_A}{k(1-\pi_{a,0})\mathrm{E}(2pq)} - \mathrm{E}[\,(1-2p)^2\,]\,(\sigma_d^2+\eta^2). \quad (9)$$

Values for S_a^2 and S_d^2 can now be calculated by substituting (9) and (6) in (2), respectively.

Breed-specific SNP allele model

As shown in [9] (with a slightly different notation), BSAM fits SNP allele states in the following model:

$$y_i = \mu + \sum_{j=1}^{k}(X_{ij}^A\alpha_j^A + X_{ij}^B\alpha_j^B) + e_i, \quad (10)$$

where X_{ij}^A and X_{ij}^B, with value (0, 1), are the breed-specific copy numbers of a given allele at SNP j of breed origin A or B that animal i received from its sire or dam, and α_j^A and α_j^B are the breed-specific substitution effects for the alleles of breed origin A and B. The other parameters are defined as in the additive and dominance models. In BSAM, the SNP allele effects have breed-specific variances $\sigma^2_{\alpha^A}$ and $\sigma^2_{\alpha^B}$, and breed-specific parameters π_{α^A} and π_{α^B}. The same prior as used in the additive model is used for $\sigma^2_{\alpha^A}$ and $\sigma^2_{\alpha^B}$.

Inference for model parameters

Markov chain Monte Carlo (MCMC) sampling was used to draw inferences from the posterior distributions of parameters. Gibbs sampling was used to sample parameters from their full conditional distributions, which are derived for some key model parameters in Appendix C. Since the implementation of a Gibbs sampler in the additive model has been well described by [6], here we focus on the algorithm for the dominance model. The decision to include a SNP in the model was separately sampled for the additive and dominance effects in the dominance model. Similarly, in BSAM, the decision to include a SNP in the model was separately sampled for the sire and dam breed-specific allele substitution effects.

The analyses were implemented by modifying *GenSel* [22] to allow dominance and allele specific effects. The Markov chain used for inference consisted of 11 000 samples, with the first 1000 discarded as a burn-in. Longer chains did not improve prediction accuracy. Parameters were estimated from the mean of the resulting 10 000 posterior samples.

True and genomic estimated breeding values

For animal i from breed r, the true breeding value is given by

$$TBV_i^r = \sum_{t=1}^{m} T_{it}\alpha_t^r, \quad (11)$$

where T_{it} is the QTL genotype, coded as 0, 1, or 2, and α_t^r is the true allele substitution effect for QTL t, and the GEBV is given by

$$GEBV_i^r = \sum_{j=1}^{k} Z_{ij}\hat{\alpha}_j^r, \quad (12)$$

where Z_{ij} is the marker genotype and $\hat{\alpha}_j^r$ is the estimated allele substitution effect for SNP j. The definition of α_t^r for a purebred animal with a breeding goal of maximizing the performance of the crossbred descendents is described next.

Suppose Q_1 and Q_2 are two alleles at a QTL. Let p^S denote the frequencies of Q_2 in the sire breed and p^D denote the frequencies of Q_2 in the dam breed. The genotypic values (G) of genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2 are $0, a+d$ and $2a$, respectively. The average effect of a Q_1 allele from the sire is defined as the expected genotypic value of a crossbred offspring that received Q_1 from the sire minus the crossbred population mean. Let S denote the allele that animal i inherited from its sire. Based on the above definition,

$$\begin{aligned}\alpha_1^S &= \mathrm{E}(G|S=Q_1) - \mu \\ &= (a+d)p^D - \mu.\end{aligned}$$

Similarly, the average effect of a Q_2 allele from the sire is $\alpha_2^S = (a+d)(1-p^D) + 2ap^D - \mu$. The difference between the two average effects gives the substitution effect for the sire:

$$\begin{aligned}\alpha^S &= \alpha_2^S - \alpha_1^S \\ &= a + (1-2p^D)d.\end{aligned}$$

Similarly, the substitution effect for the dam is $\alpha^D = a + (1-2p^S)d$. As a result, the allele substitution effects for a purebred parent used for crossbreeding are breed-specific and defined in terms of the allele frequencies in the breed of the other parent.

In summary, for a purebred r, α_t^r in (11) is defined as

$$\alpha_t^r = a_t + (1-2p_t^{r'})d_t, \quad (13)$$

where r' is the breed of the other parent of the crossbreds. In BSAM, α_j^r is directly estimated for prediction of $GEBV_i^r$ in (12), while it is indirectly estimated from the dominance model by combining the estimates of a_j and d_j with the current value of $p_j^{r'}$ from breed r' in (13). The additive model does not estimate breed-specific substitution effects. Instead, it estimates a common α_j for SNP j, which uses the allele frequency in the crossbreds used for training.

Results

Cumulative response to selection

Figure 2 depicts cumulative responses to GS for the additive, BSAM and dominance models under the four scenarios, when the genome consisted of one chromosome, 100 QTL and 1000 SNPs. The cumulative response to selection at generation 20 by the BSAM and the dominance model compared to the additive model is also given in Table 1. In scenario 1, where the dominance model was most favored, the dominance model had a substantially greater cumulative response than BSAM and the additive model. By generation 20, the dominance model had additional responses of 14.9% over BSAM and of 22.4% over the additive model ($P < 10^{-16}$). In scenario 2, however, this advantage was reduced, since the proportion of dominance variance and heterosis decreased from 16.7% and 31% to about 10% for both. As a result, in generation 20, the advantage of the dominance model was reduced from

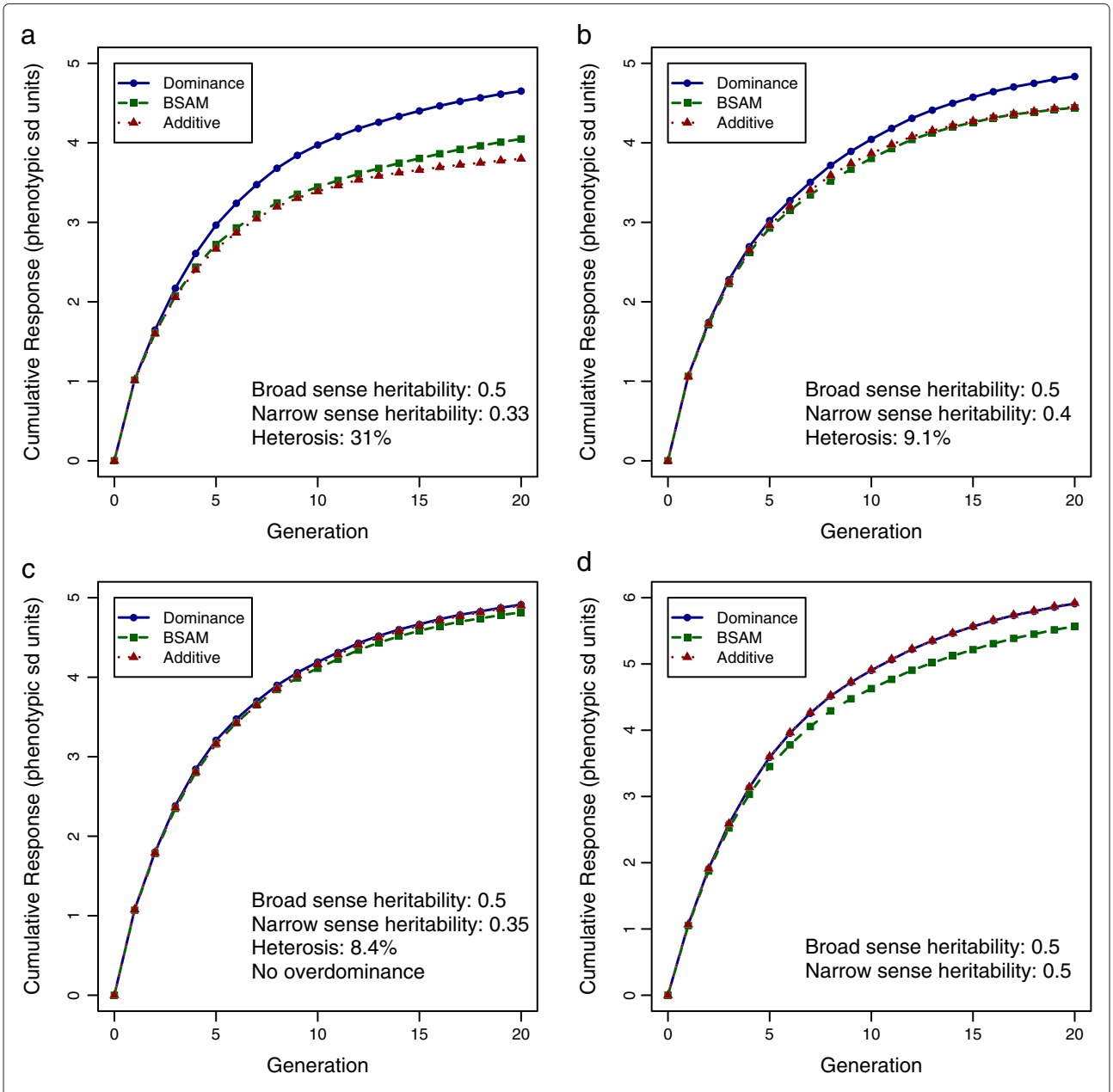

Figure 2 Cumulative response to genomic selection with one chromosome. Cumulative response to GS was computed using the dominance model, BSAM and the additive model in the four scenarios, when there was one chromosome, 100 QTL and 1000 SNPs; the plotted cumulative responses are means from 1600 replicates, standardized by the phenotypic standard deviation of crossbreds in generation 0: (**a**) results from scenario 1, where V_D was large with overdominance, (**b**) results from scenario 2, where V_D was realistic with overdominance, (**c**) results from scenario 3, where V_D was realistic without overdominance and (**d**) results from scenario 4, where dominance was absent.

Table 1 Cumulative response to genomic selection at generation 20 by the BSAM and dominance models compared to the additive model

	One chromosome		Ten chromosomes	
Scenario	Dominance model	BSAM	Dominance model	BSAM
1	22.4%*	6.5%*	10.1%*	1.7%
2	8.6%*	0.3%	2.1%	-3.1%
3	0.2%	-1.7%	1.0%	-4.4%
4	-0.1%	-5.9%*	-0.7%	-6.6%*

*Significant difference at the 0.01 level.

Cumulative responses were measured as the mean advantage of the additive model based on 1600 replicates of the simulations; the four scenarios are: (1) large overdominance, (2) realistic overdominance, (3) realistic incomplete dominance and (4) additive; in all scenarios, there was either one chromosome, 100 QTL and 1000 SNPs or ten chromosomes, 1000 QTL and 10 000 SNPs.

14.9% to 8.9% for BSAM and from 22.4% to 8.6% for the additive model. However, the differences in the cumulative response at generation 20 between the dominance model and the BSAM and additive model were still significant ($P < 10^{-9}$). The advantage of BSAM over the additive model was 6.5% ($P = 9 \times 10^{-4}$) in scenario 1 but this advantage was not significant ($P = 0.84$) in scenario 2. In scenario 3, where overdominance was absent, the proportion of V_D was only 5% and the realized heterosis was 8.4%. In this situation, responses to all three GS models were not significantly different ($P > 0.37$). In scenario 4, where there was no dominance, the dominance model still had a response as high as the additive model, which was 6.2% greater than the response for BSAM in generation 20 ($P = 4 \times 10^{-8}$).

Figure 3 shows the results with ten chromosomes, 1000 QTL and 10 000 SNPs in the genome. In scenario 1, where the dominance variance and the heterosis were set to be large, the dominance model had a clear advantage over BSAM and the additive model ($P < 10^{-10}$). In the other scenarios, the dominance model had either the highest response or a response equal to that of the additive model, whereas BSAM was inferior to the additive model in most situations. Even in scenario 1, with large dominance variance, BSAM had merely a small non-significant advantage ($P = 0.16$) over the additive model. It can be seen from Table 1 that with more chromosomes and loci, the advantage of the dominance model over the additive model by generation 20 decreased, except in scenario 3, which had only incomplete dominance. The performance of BSAM was negatively affected by the increase in the number of loci in all scenarios. Although the differences between models became less significant in the simulations with ten chromosomes, the ranking of the models was consistent with those from the simulations with one chromosome.

Changes in breed averages and heterosis in response to selection

From the definition of heterosis, cumulative response to selection in crossbred performance (CR) can be written as

$$CR = BA + H,$$

where BA denotes the breed average and H the heterosis manifested in the crossbreds. Thus, the observed advantage of the dominance model in some scenarios may be due to greater response in BA or in H, or in both. The relative contributions of BA and H to CR were investigated by partitioning CR into the response in BA and H for each of the 20 generations, and the differences between models were shown by plotting the results of selection on GEBV from one model against those of another model (Figure 4). Only results from scenario 1 with a single chromosome were used for illustration. In Figure 4, the advantage of a model in heterosis was always accompanied by some cost to purebred improvement. The dominance model had a lower response in BA, especially compared to the additive model. However, this lower response was more than compensated by increased heterosis, which resulted in a greater overall CR. By generation 20, the dominance model had a BA that was 0.35 phenotypic standard deviation (sd) lower than that of the additive model. This loss, however, was made up by an advantage of 1.2 phenotypic sd in heterosis, summing up to a total benefit of 0.95 phenotypic sd in CR for the dominance over the additive model (Figure 2a). In contrast, the advantage of BSAM over the additive model in heterosis was almost cancelled out by a comparable loss in response in BA (Figure 4c). This explains why BSAM had a limited advantage in CR over the additive model.

Response to selection in heterozygosity

When overdominance is present, crossbred performance is maximized when alternate alleles are fixed in the two pure breeds. Then, all crossbreds will be heterozygous for the over-dominant QTL. Genomic selection had a dramatic effect on heterozygosity of over-dominant QTL in crossbreds (Figure 5). In scenario 1, the dominance and BSAM models steadily increased heterozygosity over the 20 generations. However, the rate of increase was smaller for BSAM, for which heterozygosity stabilized to a lower level than for the dominance model after about 12 generations of selection because there was no retraining of the

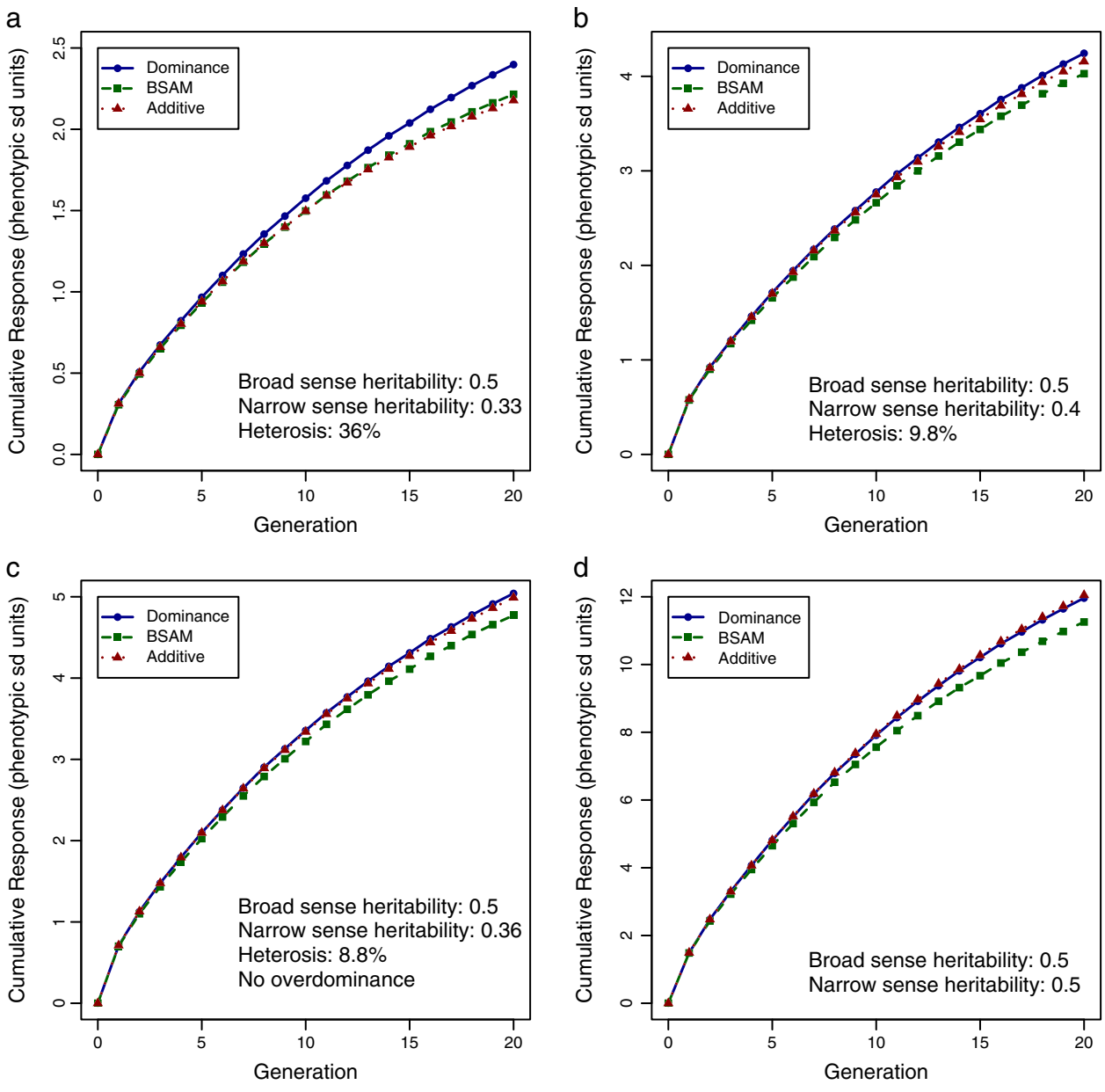

Figure 3 Cumulative response to genomic selection with ten chromosomes. Cumulative response to GS was computed using the dominance model, BSAM and the additive model in the four scenarios, when there were ten chromosomes, 1000 QTL and 10 000 SNPs; the plotted cumulative responses are means from 1600 replicates, standardized by the phenotypic standard deviation of crossbreds in generation 0: (**a**) results from scenario 1, where V_D was large with overdominance, (**b**) results from scenario 2, where V_D was realistic with overdominance, (**c**) results from scenario 3, where V_D was realistic without overdominance and (**d**) results from scenario 4, where dominance was absent.

prediction model. With the additive model, heterozygosity increased up to generation 6 and then dropped in subsequent generations.

Figure 6 shows the response to selection in allele frequencies of two over-dominant QTL in the two parental breeds, where the plotted values are the means of 100 replicates of the selection process in a given random simulation. The advantage of the dominance model over the additive model had two components. First, the rate of fixation of alternate alleles was faster with the dominance model. Second, the same allele was more often fixed in both parental breeds with the additive model, which is undesirable for over-dominant QTL. The greater efficiency of the dominance model in fixing alternate alleles in the two breeds at over-dominant QTL explains the greater heterosis in Figure 4a.

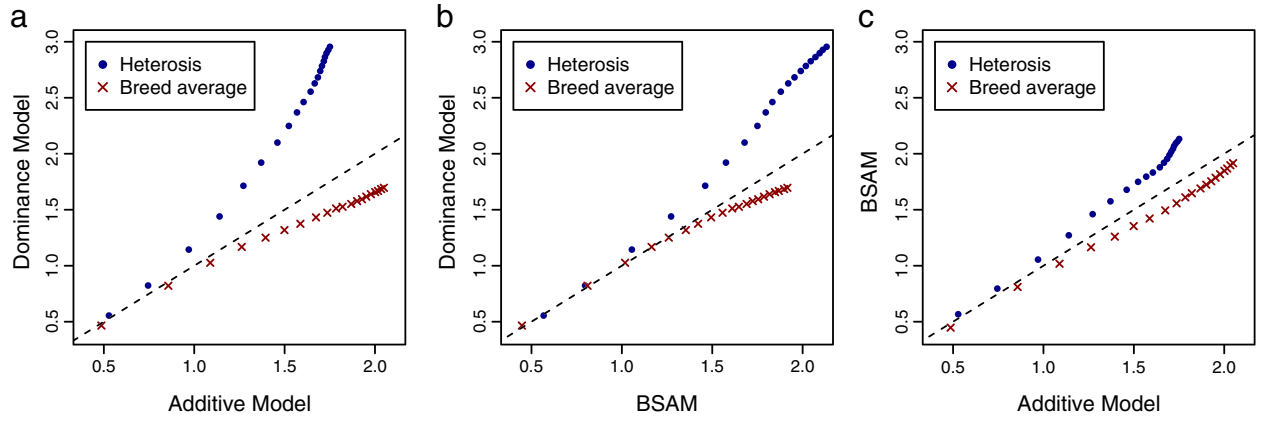

Figure 4 Cumulative response for breed average and heterosis in crossbreds. Cumulative response for breed average (red crosses) and heterosis in crossbreds (blue dots) was computed from scenario 1, with large V_D and allowing overdominance, when there was one chromosome, 100 QTL and 1000 SNPs; the plotted cumulative responses are means from 1600 replicates, standardized by the phenotypic standard deviation of crossbreds in generation 0: (**a**) cumulative response using the dominance model (y-axis) plotted against response using the additive model (x-axis), (**b**) cumulative response using the dominance model against response using BSAM and (**c**) cumulative response using BSAM against the additive model; the broken line is y=x.

Discussion

Including dominance in addition to additive effects in the model was advantageous for response to selection when dominant gene action was present. Even when all gene action was purely additive, using the dominance model did not show a negative effect. However, an advantage of BSAM was observed only when dominance variance and heterosis were large, which confirms our hypothesis that the superiority of BSAM over the additive model may be primarily due to dominance effects, rather than differences in LD between the parental breeds.

Comparison between the dominance model and the additive model

It has been shown [14] that for a two-way cross and ignoring selection, the allele substitution effects for QTL or markers in one parental breed depend on the allele frequencies in the other parental breed. Thus, in the

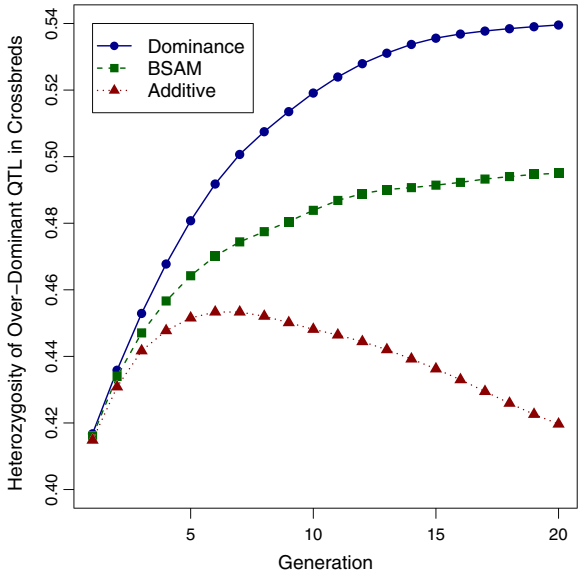

Figure 5 Changes in heterozygosity for over-dominant QTL in crossbreds over generations. The plotted values are changes in heterozygosity for over-dominant QTL in crossbreds over generations of selection, under the dominance model, BSAM and the additive model, in scenario 1, with one chromosome and large dominance, averaged over 1600 replicates.

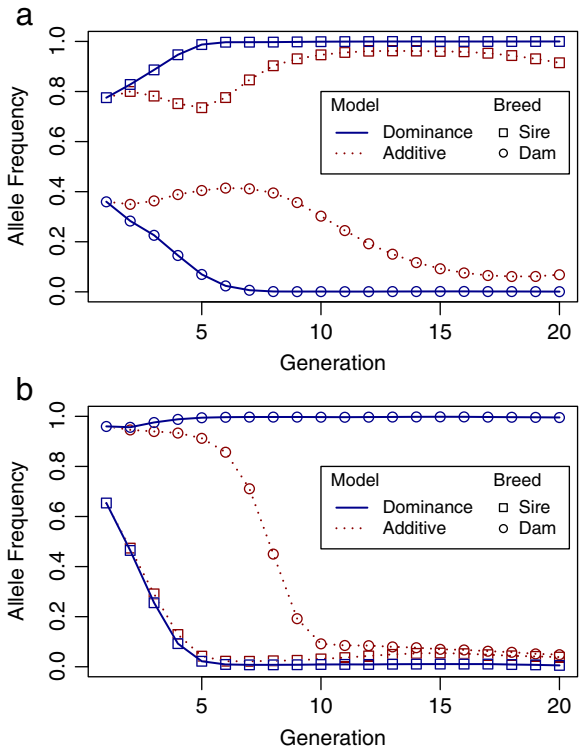

Figure 6 Changes in allele frequencies of two over-dominant QTL in the parental breeds over generations. The plotted values are changes in allele frequencies of two over-dominant QTL with major dominance effects in the sire and dam breeds over generations of selection, under the additive model and the dominance model, in scenario 1, with one chromosome and large dominance; results are means from 100 replicates in a given random simulation: (**a**) shows alternate alleles approaching fixation in the sire and dam breeds more rapidly with the dominance than the additive model and (**b**) shows the same allele approaching fixation in both parental breeds with the additive model, in contrast to the dominance model.

computation of substitution effects, failure to use the appropriate allele frequencies may result in a loss of response to selection. Dekkers [23] showed that, with the availability of only purebred data, selection within a breed using QTL allele substitution effects that are based on the allele frequencies from the opposite breed gave substantially greater response to selection than using allele frequencies from the same breed.

In the additive model, a single substitution effect was estimated for each SNP, assuming it is the same for both parental breeds. Selection on GEBV derived using such allele substitution effects is expected to fix the favorable allele in both breeds (Figure 6a). Exceptions to this could be genetic drift or the marker and QTL being in LD with opposite phases in the two parental breeds (Figure 6b). When the two breeds have opposite LD phases, and a common nonzero substitution effect is estimated for a SNP in the additive model, the allele frequencies of associated QTL will move in opposite directions in the two breeds.

In the dominance model, breed-specific allele substitution effects were computed using estimated dominance and additive effects, together with the appropriate allele frequencies from the other parental breed. When overdominance is present, the allele substitution effect $\alpha = a + (1 - 2p)d$ may have opposite signs in the parental breeds, depending on allele frequencies p in the two breeds [12]. In this case, the two parental breeds are expected to be fixed for alternate alleles of over-dominant QTL, which increases the frequency of favorable heterozygotes in crossbred progeny. Note that under the additive model, fixation of the favorable allele in both breeds would result in lower heterozygosity in the crossbreds. This explains why the dominance model resulted in substantially greater heterosis than the additive model (Figure 4a). Recall that the purebred gain was lower with the dominance model than with the additive model because the unfavorable allele was moved towards fixation in one parental breed at some loci. However, the crossbred gain was more than compensated by the larger amount of heterosis in the crossbreds. The dominance model was mostly favored in crossbred gain in scenario 1, where the dominance variance was large (Figure 2a), because in this scenario the difference in allele substitution effects between breeds was large.

When dominance is incomplete, the breed-specific substitution effects for a locus will have the same sign in both breeds, as can been seen from (13). In the long term, and ignoring drift, the favorable allele will be fixed in both breeds with any GS model. Thus, differences in response to selection in scenario 3 were not significant between GS models (Figure 2c). However, Kinghorn et al. [11] observed that with many QTL, even incomplete dominance had a significant influence on the divergence of allele frequencies in the parental breeds because the effect of drift is not negligible with a large number of loci. This phenomenon was also confirmed in our study, in which the advantage of the dominance model over the additive model in cumulative response to selection increased from 0.2% to 1.0% in the presence of only incomplete dominance (Table 1, scenario 3), when the number of loci increased ten-fold. Although the signs of the substitution effects are the same in the parental breeds with incomplete dominance, their magnitude can be very different depending on the allele frequencies. In the additive model, where a common substitution effect is used, loci with negligible substitution effects will be under higher selection pressure than in the dominance model, where breed-specific substitution effects are used. When the number of loci is large, relative to the additive model, the dominance model will put little or no selection pressure on loci with small or negligible substitution effects, resulting in greater genetic progress. Thus, the ability to exploit dominance for loci without overdominance may still be very important.

The advantage of the dominance model over the additive model is attributable to the use of breed-specific allele substitution effects to calculate GEBV of purebreds for crossbred performance with the dominance model. Kinghorn et al. [10,11] observed an advantage of using breed-specific allele substitution effects for crossbreeding (reciprocal recurrent genomic selection, RRGS) in an ideal situation, where additive and dominance effects at the QTL were known without error. In the presence of overdominance, RRGS had greater cumulative response to selection than the additive model, regardless of whether recalculation of substitution effects of QTL alleles was performed only in the first generation, every five generations or every generation [10]. In the absence of overdominance, the advantage of RRGS over the additive model was still observed with retraining each generation, especially for many QTL [11].

Comparison between BSAM and the additive model

In BSAM, for the first generation after training, the estimated breed-specific allele substitution effects of SNPs account for differences in LD and allele frequencies between breeds. However, with repeated selection over generations, the breed-specific allele substitution effects must be re-estimated in BSAM to accommodate changes in allele frequency and LD. Ibánẽz-Escriche et al. [9] showed that, when using crossbred data to predict GEBV of purebred descendants, the additive model had a greater accuracy of prediction than BSAM if the breeds were related and the training population size was small relative to the number of markers. However, pure additive inheritance was assumed in their study. Under dominant gene action, we expect BSAM to have an additional advantage over the additive model because, as can be seen from (13), apart from different LD, breed-specific allele substitution effects will be different when allele frequencies differ between breeds. However, in this study, BSAM had a lower response to selection than the additive model, except when the dominance effects were large and the number of SNPs was not greater than the number of observations. When the number of SNPs was ten times greater than the number of observations, the advantage of BSAM over the additive model was not significant.

One reason for the inferior performance of BSAM in most scenarios is model complexity. If a SNP is segregating in both parental breeds, then the SNP will have a nonzero substitution effect in the crossbreds for both breeds. Thus, when most SNPs are segregating in both breeds, we expect BSAM to have about twice the number of nonzero SNP effects in the model as the additive model. As shown in Table 2, the posterior mean of the number of effects in the additive model was proportional to the magnitude of the dominance variance relative to the additive variance, but this relationship was not observed in BSAM, for which the number of nonzero effects hardly changed between scenarios. In scenario 1, where the dominance variance was about half of the additive variance, many markers were needed in the additive model to jointly pick up the QTL effects. In this case, BSAM was superior to the additive model since only a few additional effects were fitted in the model. In scenario 4, where the size of additive variance was maximum, the number of effects in the additive model was reduced to about the same as the number of QTL, while the number of effects in BSAM was still higher than twice the number of QTL because the allele substitution effects were expected to be nonzero for both breeds. When the number of chromosomes and loci increased ten-fold, the performance of BSAM compared with the additive model became even worse due to the increased model complexity (Table 1).

Another possible reason to explain the lower response of BSAM relative to the additive model was given by [9]. Consider a locus that is segregating in breed A but fixed in breed B. Because the additive model regresses phenotypes only on segregating alleles, the common substitution effect for this locus is actually the effect specific to breed A, just as in BSAM. However, BSAM includes an additional substitution effect for breed B, for which the allele is fixed. The substitution effect estimated from BSAM for

Table 2 Posterior means of the number of nonzero SNP effects fitted in the model

		BSAM			Dominance model		
Scenario	Additive model	α^S	α^D	Total	a	d	Total
1	183.1	73.2	131.0	204.2	284.3	58.4	342.7
2	166.3	88.2	151.3	239.5	181.8	58.1	239.9
3	130.9	68.7	114.4	183.1	149.4	34.8	184.2
4	105.4	90.9	144.8	235.7	116.3	4.1	120.4

Numbers in the table are means from 16 random simulations, where there was one chromosome, 100 QTL and 1000 SNPs; α^S and α^D are the breed-specific allele substitution effects for the sire and dam breeds in BSAM; a and d are the additive and dominance SNP effects in the dominance model; numbers in column "Total" in BSAM and the dominance model are the sum of the number of nonzero SNP effects of different types in the corresponding model.

the breed B allele will only add noise to the prediction of GEBV. Therefore, when several loci are nearly fixed in one of the parental breeds but segregating in the other, the additive model is expected to show an advantage over BSAM.

Comparison between BSAM and the dominance model

The number of effects in BSAM equals the number of breeds times the number of SNPs. When only two breeds are involved in the cross, the dominance model is expected to have the same number of parameters as BSAM if the number of SNPs included in the model is fixed. However, with separate probability of inclusion parameters π_a and π_d for additive and dominance effects, the number of dominance effects that are fitted in the model only depends on the actual number of dominance effects for the trait. As dominance variance was lowered from large to null, the posterior mean of the number of nonzero dominance effects in the model decreased from 58.4 to 4.1 (Table 2). Thus, in contrast to BSAM, model complexity does not seem to be an issue for the dominance model. Even when dominance was absent, the response in the dominance model could not be distinguished from that of the additive model (Figure 2d) because only a few nonzero dominance effects were fitted in the model. In this regard, the dominance model is more robust than BSAM.

Even when additive and dominance effects are consistent between breeds, allele substitution effects will be breed-specific if allele frequencies differ between breeds. In such a case, the estimates of breed-specific allele substitution effects in the dominance model are expected to be more accurate than those in BSAM for the following four reasons:

First, the estimates of additive and dominance effects from the dominance model are combined with the observed allele frequencies in the opposite parental breed to calculate the breed-specific allele substitution effects. In BSAM, however, breed-specific allele substitution effects are estimated directly. Thus, the allele frequencies used in BSAM are based on the frequencies in the training population of the alleles inherited from the opposite parental breed. Note that the alleles inherited by the training population are a random sample of those from the parental population, and therefore their frequencies will deviate from those of the parental population. Thus, the use of observed allele frequencies from the parental population to compute breed-specific allele substitution effects favors the dominance model over BSAM.

Second, in the dominance model, as selection progresses and allele frequencies change due to selection, the observed allele frequencies in each generation are combined with the estimates of additive and dominance effects obtained in training to compute the current values of the breed-specific allele substitution effects. However, with the additive model and with BSAM, the allele substitution effects estimated in training are repeatedly used to compute GEBV of selection candidates, ignoring changes in allele frequencies. As a result, drops in accuracy of GEBV in generations of selection were greater for the additive model and BSAM than for the dominance model (results not shown). Thus, use of the dominance model is expected to require less frequent retraining than use of BSAM or the additive model. This is appealing for traits that are difficult or expensive to measure.

Third, under the assumption that additive and dominance effects of QTL are consistent across breeds, as explained below, the goodness of fit to training data is expected to be better for the dominance model than for BSAM. The expected phenotypic value for a given genotype ij at a QTL is,

$$E(y|ij) = G_{ij} = \begin{cases} G_{00}, & ij = 00 \\ G_{00} + a + d, & ij = 01 \text{ or } 10 \\ G_{00} + 2a, & ij = 11. \end{cases} \quad (14)$$

In the dominance model given by (3), these expected values are modeled exactly in terms of a and d as $\mu + Xa + Wd$, where $X = W = 0$ for $ij = 00$, $X = W = 1$ for $ij = (01$

or 10), or $X = 2$ and $W = 0$ for $ij = 11$. In BSAM, these conditional expectations are modeled as

$$G_{00} = \mu + \alpha_1^A + \alpha_1^B + \epsilon_{00}, \quad (15)$$

$$G_{01} = \mu + \alpha_1^A + \alpha_2^B + \epsilon_{01}, \quad (16)$$

$$G_{10} = \mu + \alpha_2^A + \alpha_1^B + \epsilon_{10},$$

$$G_{11} = \mu + \alpha_2^A + \alpha_2^B + \epsilon_{11}.$$

Subtracting (16) from (15) gives

$$G_{01} - G_{00} = \alpha_2^B - \alpha_1^B + \epsilon_{01} - \epsilon_{00}.$$

As $\alpha^B = \alpha_2^B - \alpha_1^B$, based on (13),

$$\begin{aligned} G_{01} - G_{00} &= \alpha_2^B - \alpha_1^B + \epsilon_{01} - \epsilon_{00} \\ &= a + d - 2p^A d + \epsilon_{01} - \epsilon_{00}. \end{aligned}$$

Comparing this to (14), where $G_{01} - G_{00} = a + d$, implies

$$\epsilon_{01} - \epsilon_{00} = 2p^A d.$$

Thus, deviates ϵ_{00} and ϵ_{01} cannot both be zero. Similarly, it can be shown that $\epsilon_{10} - \epsilon_{00} = 2p^B d$ and $\epsilon_{11} - \epsilon_{00} = 2(p^A + p^B - 1)d$. The nonzero differences between the deviates imply that at least some deviates are not equal to zero. More simply, $\mu + \alpha_1^A + \alpha_2^B$ in the BSAM for G_{01} may not be equal to $\mu + \alpha_2^A + \alpha_1^B$ for G_{10}, although G_{01} is equal to G_{10}. In other words, the genotypic value is not exactly modeled in BSAM.

Forth, implementation of BSAM requires knowing the breed origins of SNP alleles but this is not required for the dominance model. In this study, breed origins of SNP alleles were assumed to be known without error but this may not hold for real data, which would introduce errors to the estimation of breed-specific SNP effects in BSAM.

However, the advantages of the dominance model may not hold if additive and dominance effects of SNPs are not consistent between breeds. More precisely, although a and d at the QTL may be consistent between breeds, the SNP effects may differ between breeds due to differences in LD. When these differences of SNP effects between breeds are large, the dominance model may become inferior to BSAM, for which breed-specific allele substitution effects account for differences in LD in addition to differences in allele frequencies. The magnitude and phase of LD, especially long-range LD, has been found to be inconsistent between breeds in livestock [24-26]. However, whether those differences in LD are large enough to enable BSAM to outperform the dominance model needs further investigation with real data.

The benefit of using one model over another depended on the number and size of QTL and the density of SNPs relative to the size of the training population. When the genome was extended from one to ten chromosomes, there was a ten-fold increase in the number of SNPs but the SNP density remained the same. In addition, with the values of variance components held constant, each QTL explained a smaller proportion of the total genetic variance for the larger genome, which made it more difficult to capture the QTL effects through SNPs. This explains the observed reduction of differences in response between models. This suggests that the preference of the dominance model is expected to hold for a larger genome but the amount of benefit will decrease when the genetic architecture is more polygenic. Furthermore, a larger training population will be needed.

In this study, SNP effects were estimated only once and applied successively over 20 generations of selection. This is rarely done in practice and retraining is usually carried out after each generation of selection. With retraining, the advantage of the dominance model is expected to be much smaller or may even be ignored because in that case, the additive and BSAM models are also expected to give estimated allele substitution effects using allele frequencies from the current or recent generations. However, if the training population consists of individuals from multiple generations, the estimates of substitution effects in the additive model or BSAM will depend on allele frequencies across multiple generations, which may not be appropriate to predict crossbred performance for the purebred candidates.

Although the dominance model was studied here when purebreds were selected for crossbred performance, the advantages of this model over the additive model also apply to selection for purebred performance. Furthermore, the estimates of a and d from the dominance model can be used to predict GEBV in any population with SNP genotypes, provided the LD in the training and candidate populations are about the same. It has been found that the accuracy of GEBV in Holstein-Friesian cattle with training in Jerseys and vice versa was as low as -0.1 to 0.3 over traits [27], and the correlations of SNP allele frequencies between Holsteins, Jerseys, and Brown Swiss cattle were as low as 0.65 to 0.67 [28]. Thus, differences in allele frequencies between breeds, along with dominant gene action, can be attributed to the low accuracy of prediction across breeds in dairy cattle. In this situation, the dominance model is expected to give better results, especially with a high marker density, since then LD would be more consistent between breeds.

Besides dominance, other types of non-additive effects may also contribute to heterosis, such as epistasis, imprinting, etc. The dominance model can be further extended to account for imprinting if the heterozygotes are phased. Epistatic interactions between loci are partially included in BSAM as a component of allele substitution effects but will be misspecified in the dominance model, which will impair accuracy of selection. Thus, BSAM may be more robust in the presence of epistasis.

Conclusions

When dominance, particularly overdominance, is the key driver of heterosis, using a dominance model for GS is expected to result in greater cumulative response to selection of purebred animals for crossbred performance than either BSAM or the additive model. The extent of this additional response to selection depends on the size of dominance effects at the QTL and the power of detecting the dominance effects through SNP genotypes. Also, when there are many loci, it is important to exploit dominance, even in the absence of overdominance. When BayesCπ is used, the dominance model is robust because, even in the absence of dominance, it does not give a lower response than the additive model. BSAM is favored over the additive model only when dominance effects are large enough to overwhelm its shortcomings. Furthermore, implementation of BSAM requires that the breed origins of SNP alleles are known. Our results suggest that in the presence of dominant gene action, relative to BSAM and the additive model, GS with the dominance model is superior to maximize crossbred performance through purebred selection, especially when no retraining is carried out at each generation.

Appendix A

Scaling procedure for QTL additive and dominance effects

Let V_A and V_D denote the observed additive and dominance genetic variance of the trait. Assuming no genotype by genotype interactions among QTL that define the trait, the genetic variance components can be written as the sum of the variance explained by each QTL [12]:

$$V_A = \sum_j 2p_jq_j\alpha_j^2, \tag{17}$$

$$V_D = \sum_j (2p_jq_jd_j)^2, \tag{18}$$

where $p_j = 1-q_j$ is the observed allele frequency for QTL j, d_j is the QTL dominance effect, and α_j is the QTL allele substitution effect defined as

$$\alpha_j = a_j + (q_j - p_j)d_j,$$

where a_j is the QTL additive effect.

For scenarios allowing overdominance

Let V_A^* and V_D^* denote the corresponding desired genetic variance components and a_j^*, d_j^*, α_j^* the corresponding scaled QTL effects. Let

$$s = \frac{V_D^*}{V_D}. \tag{19}$$

From (18) and (19), we have

$$\sum_j (2p_jq_jd_j^*)^2 = s\sum_j (2p_jq_jd_j)^2.$$

Thus,

$$d_j^* = \sqrt{s}d_j.$$

Similar to $\sqrt{s}$ the scalar for dominance effects, we have a scalar t for additive effects such that

$$a_j^* = ta_j, \tag{20}$$

and

$$V_A^* = \sum_j 2p_jq_j(\alpha_j^*)^2 = \sum_j 2p_jq_j(a_j^* + (q_j - p_j)d_j^*)^2. \tag{21}$$

Substituting (20) in (21) and rearranging it, we have

$$t^2\sum_j 2p_jq_ja_j^2 + t\sum_j (q_j - p_j)a_jd_j + \sum (q_j - p_j)^2d^{*2} - V_A^* = 0.$$

This can be seen as a quadratic equation with variable t unknown. Thus, the scalar t for the additive effects can be obtained by solving this equation.

For scenarios without overdominance

Since d is already smaller than a for each QTL, to obtain the desirable additive genetic variance V_A^*, we have a scalar c where

$$c = \frac{V_A^*}{V_A}.$$

Based on (17), the new substitution effect α_j^* for QTL j is simply

$$\alpha_j^* = \sqrt{c}\,\alpha_j.$$

Thus, the scaled additive and dominance effects are respectively, $a_j^* = \sqrt{c}a_j$ and $d_j^* = \sqrt{c}d_j$. With this approach, the additive genetic variance is the one desired but the dominance genetic variance is not under control. However, as desired, there would be no overdominance affecting the trait.

Appendix B

Hypothesis test for the GS model effect

The objective was to test if any difference in cumulative response to GS observed at generation 20 of selection among the additive model, BSAM, and the dominance model is statistically significant. As described, data were collected from 16 simulations each with 100 replicates resulting in a total of 1600 observations. The following mixed linear model was used to fit the data:

$$y_{ij} = \mu + m_i + b_j + e_{ij},$$

where y_{ij} is the response from GS model i in simulation j, m_i is the fixed effect for GS model $i = \{1, 2, 3\}$, $b_j \sim N(0, \sigma_b^2)$ is the random blocking effect for simulation set

$j = \{1, ..., 16\}$, and $e_{ij} \sim N(0, \sigma_e^2)$ is the residual. A set of t-tests was used for testing the null hypothesis that 1) $m_1 = m_2$, 2) $m_1 = m_3$, or 3) $m_2 = m_3$.

Appendix C

Full conditionals for some key model parameters

Let β_j denote either a_j or d_j in the dominance model. The mixture prior for β_j allows it to be included or not in the model. Let $\delta_{j,\beta}$ denote a model inclusion indicator variable defined as

$$\delta_{j,\beta} = \begin{cases} 1 \text{ then } \beta_j \sim \text{Normal} \\ 0 \text{ then } \beta_j = 0 \end{cases}$$

with a prior probability $1 - \pi_\beta$ that $\delta_{j,\beta} = 1$. It should be clear that the model allows $\delta_{j,a} \neq \delta_{j,d}$. The indicator $\delta_{j,\beta}$ and the effect β_j can be sampled jointly by first sampling $\delta_{j,\beta}$ from its marginal distribution and then sampling β_j conditional on $\delta_{j,\beta}$. The posterior "marginal" probability that $\delta_{j,\beta} = 1$ respecting to β_j is calculated by

$$Pr(\delta_{j,\beta} = 1|\boldsymbol{y}, \boldsymbol{\theta}_{else}) = \frac{f(\boldsymbol{y}|\delta_{j,\beta} = 1, \boldsymbol{\theta}_{else}) Pr(\delta_{j,\beta} = 1)}{\sum_{\delta_{j,\beta}} f(\boldsymbol{y}|\delta_{j,\beta}, \boldsymbol{\theta}_{else}) Pr(\delta_{j,\beta})}, \quad (22)$$

where $\boldsymbol{\theta}_{else}$ denotes other parameters besides $\delta_{j,\beta}$ and β_j. The "likelihood" in (22) can be obtained by integrating β_j out from the sampling distribution as described below. Let

$$\boldsymbol{v} = \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{a} - \sum_{j' \neq j} \boldsymbol{W}_{j'} d_{j'} = \boldsymbol{W}_j d_j + \boldsymbol{e}.$$

Then,

$$\begin{aligned} f(\boldsymbol{y}|\delta_{j,d} = 1, \boldsymbol{\theta}_{else}) &= \int f(\boldsymbol{y}|\delta_{j,d} = 1, d_j, \boldsymbol{\theta}_{else}) f(d_j)\, \mathrm{d}d_j \\ &= (2\pi)^{-\frac{n}{2}} (\sigma_e^2)^{-\frac{n}{2}} (\sigma_d^2)^{-\frac{1}{2}} (\frac{C_{j,d}}{\sigma_e^2})^{-\frac{1}{2}} \quad (23) \\ &\quad \exp\{-\frac{\boldsymbol{v}'\boldsymbol{v} + \lambda\mu_d^2 - C_{j,d}(\hat{d}_j + \lambda C_{j,d}^{-1}\mu_d)^2}{2\sigma_e^2}\}, \quad (24) \end{aligned}$$

where $C_{j,d} = \boldsymbol{W}_j'\boldsymbol{W}_j + \frac{\sigma_e^2}{\sigma_d^2}$ and $\hat{d}_j = \frac{\boldsymbol{W}_j\boldsymbol{v}}{C_{j,d}}$ are, respectively, the coefficient of the mixed-model equation and the BLUP estimate for d_j. Similarly, let

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{W}\boldsymbol{d} - \sum_{j' \neq j} \boldsymbol{X}_{j'} a_{j'} \\ &= \boldsymbol{X}_j a_j + \boldsymbol{e}. \end{aligned}$$

We have

$$f(\boldsymbol{y}|\delta_{j,a} = 1, \boldsymbol{\theta}_{else}) = (2\pi)^{-\frac{n}{2}} (\sigma_e^2)^{-\frac{n}{2}} (\sigma_a^2)^{-\frac{1}{2}} (\frac{C_{j,a}}{\sigma_e^2})^{-\frac{1}{2}} \exp\{-\frac{\boldsymbol{u}'\boldsymbol{u} - C_{j,a}(\hat{a}_j)^2}{2\sigma_e^2}\}, \quad (25)$$

where $C_{j,a} = \boldsymbol{X}_j'\boldsymbol{X}_j + \frac{\sigma_e^2}{\sigma_a^2}$ and $\hat{a}_j = \frac{\boldsymbol{X}_j\boldsymbol{u}}{C_{j,a}}$. For $\delta_{j,\beta} = 0$, the "likelihood" is simply

$$f(\boldsymbol{y}|\delta_{j,\beta} = 0, \boldsymbol{\theta}_{else}) = (2\pi)^{-\frac{n}{2}} (\sigma_e^2)^{-\frac{n}{2}} \exp\{-\frac{\boldsymbol{w}'\boldsymbol{w}}{2\sigma_e^2}\}, \quad (26)$$

where $\boldsymbol{w}$ is either $\boldsymbol{v}$ or $\boldsymbol{u}$ corresponding to $\delta_{j,d}$ or $\delta_{j,a}$. Substituting (24-26) in (22) gives the marginal distribution of $\delta_{j,\beta}$ respecting to β_j.

Given that $\delta_{j,d} = 1$, d_j has a normal prior centered at μ_d, otherwise it is zero. Let $\boldsymbol{d}_{-j}$ denote the other additive effects besides that for SNP j, then the full conditional for d_j is

$$\begin{aligned} f(d_j|\boldsymbol{y}, \mu, \boldsymbol{d}_{-j}, \boldsymbol{a}, \mu_d, \sigma_d^2, \sigma_e^2) &\propto f(\boldsymbol{y}|\mu, \boldsymbol{d}_{-j}, \boldsymbol{a}, \sigma_e^2) f(d_j|\mu_d, \sigma_d^2) \\ &\propto \exp\left\{-\frac{C_{j,d}\left\{d_j - \left(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d\right)\right\}^2}{2\sigma_e^2}\right\}, \end{aligned}$$

which is a normal $N(\hat{d}_j + \frac{\lambda}{C_{j,d}}\mu_d, \frac{\sigma_e^2}{C_{j,d}})$ where $\lambda = \frac{\sigma_e^2}{\sigma_d^2}$.

The variance variable for the dominance effect depends on the data only through the dominance effects:

$$f(\sigma_d^2|\boldsymbol{y}, \mu, \boldsymbol{a}, \boldsymbol{d}, \sigma_e^2, S_d) \propto f(\boldsymbol{d}|\sigma_d^2) f(\sigma_d^2|S_d).$$

Thus, due to the conjugacy, the full conditional for σ_d^2 is also a scaled inverse Chi-square $\sim \tilde{S}_d \chi_{\tilde{\nu}_d}^{-2}$ where $\tilde{\nu}_d = k + \nu_d$ and $\tilde{S}_d = \frac{(\boldsymbol{d} - \mathbf{1}\mu_d)'(\boldsymbol{d} - \mathbf{1}\mu_d) + \nu_d S_d}{\tilde{\nu}_d}$ given the sampled value of μ_d.

As shown, the prior of μ_d depends on the value of σ_d^2, therefore the full conditional for μ_d is

$$\begin{aligned} f(\mu_d|\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{a}, \boldsymbol{d}, \sigma_d^2, \sigma_e^2, \eta, \phi) &\propto f(\boldsymbol{d}|\mu_d, \sigma_d^2) f(\mu_d|\eta, \phi, \sigma_d^2) \\ &\propto \exp\{-\frac{(\mu_d - \frac{\mathbf{1}'\boldsymbol{d} + \phi\eta}{k + \phi})^2}{2\sigma_d^2/(k + \phi)}\}, \end{aligned}$$

which is a normal $\sim N\left(\frac{\mathbf{1}'\boldsymbol{d} + \phi\eta}{k + \phi}, \frac{\sigma_d^2}{k + \phi}\right)$.

The full conditionals for the other parameters including additive effect for SNP j and its variance variable are as illustrated in [6].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JZ and RLF designed the study and developed the algorithm for the dominance model. JZ and AT wrote the simulation programs under RLF's guidance. JZ carried out the simulation studies, performed the statistical analysis, and drafted the manuscript. JCMD and DJG advised on the design of the study, the development of the algorithm, and the analysis of the simulations. RLF, JCMD and DJG revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The research was funded by USDA-NRI Grant 2007-35205-17862.

Author details
[1]Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA, USA. [2]Monsanto Company, 3302 SE Convenience Blvd., Ankeny, IA, USA.

Received: 8 November 2012 Accepted: 15 March 2013
Published: 26 April 2013

References

1. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157:**1819–1829.
2. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: reliability of genomic predictions for North American Holstein bulls.** *J Dairy Sci* 2009, **92:**16–24.
3. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: progress and challenges.** *J Dairy Sci* 2009, **92:**433–443.
4. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177:**2389–2397.
5. Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G: **The impact of genetic relationship information on genomic breeding values in German Holstein cattle.** *Genet Sel Evol* 2010, **42:**5.
6. Habier D, Fernando RL, Kizilkaya K, Garrick DJ: **Extension of the Bayesian alphabet for genomic selection.** *BMC Bioinformatics* 2011, **12:**186.
7. Daetwyler HD, Villanueva B, Bijma P, Woolliams JA: **Inbreeding in genome-wide selection.** *J Anim Breed Genet* 2007, **124:**369–376.
8. Dekkers JCM: **Marker-assisted selection for commercial crossbred performance.** *J Anim Sci* 2007, **85:**2104–2114.
9. Ibanez-Escriche N, Fernando RL, Toosi A, Dekkers JCM: **Genomic selection of purebreds for crossbred performance.** *Genet Sel Evol* 2009, **41:**12.
10. Kinghorn BP, Hickey JM, van der Werf JHJ: **Reciprocal recurrent genomic selection (RRGS) for total genetic merit in crossbred individuals.** In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig*. Paper 36; 2010. http://www.kongressband.de/wcgalp2010/assets/html/0036.htm.
11. Kinghorn BP, Hickey JM, van der Werf JHJ: **Long-range phasing and use of crossbred data in genomic selection.** In *Proceedings of the 7th European Symposium on Poultry Genetics: 5-7 October 2011; Peebles Hydro, near Edinburgh*. Paper 1; 2011:7–9. http://www.roslin.ed.ac.uk/7espg/assets/7espg-edited-proceedings.pdf.
12. Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*. 4th edition. Harlow: Longmans Green; 1996.
13. Charlesworth D, Willis JH: **The genetics of inbreeding depression.** *Nat Rev Genet* 2009, **10:**783–796.
14. Dekkers JCM: **Breeding values for identified quantitative trait loci under selection.** *Genet Sel Evol* 1999, **31:**421–436.
15. Karlin S: **Theoretical aspects of genetic map functions in recombination processes.** In *Human Population Genetics: The Pittsburgh Symposium*. Edited by Chakravarti A. New York: Van Nostrand Reinhold; 1984:209–228.
16. Bennewitz J, Meuwissen THE: **The distribution of QTL additive and dominance effects in porcine F2 crosses.** *J Anim Breed Genet* 2010, **127:**171–179.
17. Kacser H, Burns JA: **The molecular basis of dominance.** *Genetics* 1981, **97:**639–666.
18. Caballero A, Keightley PD: **A pleiotropic nonadditive model of variation in quantitative traits.** *Genetics* 1994, **138:**883–900.
19. Misztal I, Lawlor TJ, Gengler N: **Relationships among estimates of inbreeding depression, dominance and additive variance for linear traits in Holsteins.** *Genet Sel Evol* 1997, **29:**319–326.
20. Cundiff LV, Gregory KE: **What is systematic crossbreeding?** In *Proceedings of the National Cattlemen's Beef Association: 11 February 1999*. Charlotte; 1999. [http://beefmagazine.com/mag/beef_systematic_crossbreeding]
21. Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR: **Genomic selection.** *Acta Agric Scand A Anim Sci* 2008, **57:**182–195.
22. Fernando RL, Garrick DJ: **GenSel - User manual for a portfolio of genomic selection related analyses.** [http://bigs.ansci.iastate.edu/bigsgui]
23. Dekkers JCM, Chakraborty R: **Optimizing purebred selection for crossbred performance using QTL with different degrees of dominance.** *Genet Sel Evol* 2004, **36:**297–324.
24. de Roos APW, Hayes BJ, Goddard ME: **Reliability of genomic predictions across multiple populations.** *Genetics* 2009, **183:**1545–1553.
25. de Roos APW, Hayes BJ, Spelman RJ, Goddard ME: **Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle.** *Genetics* 2008, **179:**1503–1512.
26. Meadows JRS, Chan EKF, Kijas JW: **Linkage disequilibrium compared between five populations of domestic sheep.** *BMC Genet* 2008, **9:**61.
27. Harris B, Johnson D, Spelman R: **Genomic selection in New Zealand and the implications for national genetic evaluation.** In *Proceedings of the Interbull Meeting: 16-19 June 2008; Niagara Falls*. Edited by Sattler JD: Department of Animal Breeding and Genetics; 2008:325–330.
28. VanRaden P, Olson K, Wiggans G, Cole J, Tooker M: **Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss.** *J Dairy Sci* 2011, **94:**5673–5682.

doi:10.1186/1297-9686-45-11
Cite this article as: Zeng *et al.*: Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genetics Selection Evolution* 2013 **45**:11.