Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction

David Habier,*,†,1 Rohan L. Fernando,* and Dorian J. Garrick*
*Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, Iowa 50011, and †DuPont Pioneer, Johnston, Iowa 50131

ABSTRACT Genomic best linear unbiased prediction (BLUP) is a statistical method that uses relationships between individuals calculated from single-nucleotide polymorphisms (SNPs) to capture relationships at quantitative trait loci (QTL). We show that genomic BLUP exploits not only linkage disequilibrium (LD) and additive-genetic relationships, but also cosegregation to capture relationships at QTL. Simulations were used to study the contributions of those types of information to accuracy of genomic estimated breeding values (GEBVs), their persistence over generations without retraining, and their effect on the correlation of GEBVs within families. We show that accuracy of GEBVs based on additive-genetic relationships can decline with increasing training data size and speculate that modeling polygenic effects via pedigree relationships jointly with genomic breeding values using Bayesian methods may prevent that decline. Cosegregation information from half sibs contributes little to accuracy of GEBVs in current dairy cattle breeding schemes but from full sibs it contributes considerably to accuracy within family in corn breeding. Cosegregation information also declines with increasing training data size, and its persistence over generations is lower than that of LD, suggesting the need to model LD and cosegregation explicitly. The correlation between GEBVs within families depends largely on additive-genetic relationship information, which is determined by the effective number of SNPs and training data size. As genomic BLUP cannot capture short-range LD information well, we recommend Bayesian methods with *t*-distributed priors.

GENOMIC best linear unbiased prediction (BLUP) is a statistical method that has been used to predict height in humans (Yang *et al.* 2010) and breeding values for selection in animal and plant breeding (VanRaden 2008). It uses a so-called genomic relationship matrix that describes genetic relationships between individuals calculated from genotypes at single-nucleotide polymorphisms (SNPs). In genomic selection applications (Meuwissen *et al.* 2001), those individuals comprise both training individuals that are phenotyped for a quantitative trait and genotyped at SNPs and selection candidates that are genotyped only.

Genomic BLUP differs from the traditional pedigree BLUP (Henderson 1975) in the replacement of the pedigree relationship matrix with a genomic relationship matrix. Coefficients of the pedigree relationship matrix describe additive-genetic relationships (Malécot 1948) between individuals at quantitative trait loci (QTL) conditional on pedigree information, but it is not obvious to what extent the genomic relationship matrix explains genetic covariances between individuals at QTL. Despite this, several authors called the genomic relationship matrix the actual (Hill and Weir 2011) or realized relationship matrix (Goddard 2009; Hayes *et al.* 2009b; Lee *et al.* 2010) as it describes *identity-by-descent* at SNPs (Hayes *et al.* 2009b), assuming an ancient founder population. However, these terms are misleading because only genetic relationships at QTL matter in quantitative-genetic analyses.

To understand better how genomic relationships capture relationships at QTL, we propose to apply concepts of pedigree analyses that define founders in a recent past generation. Based on these concepts, we show that coefficients of the genomic relationship matrix do not explain genetic covariances between individuals at QTL unless either there is linkage disequilibrium (LD) between QTL and SNPs measured in founders or selection candidates are related by pedigree to the training individuals. The latter results in cosegregation of alleles at QTL and SNPs that are

doi: 10.1534/genetics.113.152207
Manuscript received December 13, 2012; accepted for publication April 21, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152207/-/DC1.
Additional data deposited in the Dryad Repository: http://dx.doi.org/10.5061/dryad.r7g12
[1]Corresponding author: DuPont Pioneer, 8305 NW 62nd Ave., PO Box 7060, Johnston, IA 50131-7060. E-mail: dhabier@gmail.com

linked, also known as *linkage* information, and in additive-genetic relationships at QTL captured by SNPs. These three types of information affect differently the persistence of accuracy of genomic estimated breeding values (GEBVs) from BLUP over generations (Habier *et al.* 2007), realized selection intensities, and inbreeding. The contributions of these parameters to accuracy of GEBVs depending on training data size, extent of LD, and mating design have not been demonstrated; a better understanding will allow us to optimize statistical models, training data, and selection strategies.

LD observed in the training data were first believed to be the only source of information until Habier *et al.* (2007) and Gianola *et al.* (2009) demonstrated that SNP genotypes also capture pedigree relationships. Habier *et al.* (2007) partitioned the observed accuracy of GEBVs into a part due to LD in the training data and a remainder due to pedigree relationships. Accuracy due to LD is the component of accuracy that persists over generations without retraining and provides the accuracy for individuals that are unrelated to the training individuals. Compared to Bayesian methods with *t*-distributed priors (Meuwissen *et al.* 2001), accuracy due to LD tends to be lower with genomic BLUP (Habier *et al.* 2007, 2010a, 2011).

Goddard (2009) presented formulas for calculating the accuracy due to LD, but derivations assume that the markers completely capture the variability at the QTL. Nevertheless, that accuracy was calculated as a function of the effective number of chromosomal segments, which was estimated only from effective population size and genome length. Real data analyses have shown that accuracy due to LD varies for quantitative traits with similar heritability (Habier *et al.* 2010a, 2011), and thus different genetic architectures cannot be described by those and similar formulas (Daetwyler *et al.* 2008, 2010). Also, modeling pedigree relationships between training individuals and selection candidates is not straightforward if only LD parameters are used to explain accuracy.

Cosegregation is traditionally exploited in linkage analyses. The advantage of cosegregation information is the ability to explain both rare allelic variants and structural variations if they segregate within families. Several authors (Goddard 2009; Hayes *et al.* 2009b; Habier *et al.* 2010a; Goddard *et al.* 2011) assumed it is utilized in genomic BLUP, but that has never been formally proven in the presence of LD and pedigree relationships or quantified. A statistical method that explicitly models both LD and cosegregation was proposed for genomic selection (Calus *et al.* 2008), but it did not outperform a Bayesian method similar to BayesA (Meuwissen *et al.* 2001). The question remains, how much cosegregation is captured implicitly by genomic BLUP compared to methods that model LD and cosegregation explicitly (*e.g.*, Meuwissen *et al.* 2002; Fernando 2003; Pérez-Enciso 2003; Legarra and Fernando 2009)?

This article has two objectives: (1) to present concepts that allow us to disentangle LD, cosegregation, and additive-genetic relationships and (2) to study the contributions of these parameters to accuracy of GEBVs depending on SNP density, training data size, and extent of LD. Dairy cattle and corn breeding scenarios were simulated to evaluate accuracy of GEBVs both within and across families obtained by different types of information, discrepancy between accuracy of GEBVs due to additive-genetic relationships and accuracy of traditional pedigree-based selection indexes, persistence of accuracy due to LD and due to cosegregation from one generation to the next without retraining, and the effect of each type of information on the correlation of GEBVs within families. Accuracies within families for the case of linkage equilibrium between QTL and SNPs were used to demonstrate unambiguously that genomic BLUP captures cosegregation, as there are no additive-genetic relationships within family. In addition, formulas for the covariance between true and estimated breeding values were derived for a simplified scenario to prove that all three sources of information are utilized by genomic BLUP.

Theory

Genetic model

Trait phenotypes of training individuals are simulated by the assumed true genetic model

$$\mathbf{y} = 1\mu + \mathbf{Wa} + \mathbf{e} \tag{1}$$

(Goddard 2009; Hayes *et al.* 2009b: Goddard *et al.* 2011), where $\mathbf{y}$, $\mathbf{a}$, and $\mathbf{e}$ are vectors containing trait phenotypes, additive QTL effects, and residual effects, respectively; μ is the overall mean; and $\mathbf{W}$ is a matrix of genotype scores at biallelic QTL. Each score is coded as the number of one of the two alleles at a locus adjusted by twice the frequency of the counted allele in founders. Both QTL and residual effects are treated as random with mean zero and with variance–covariance matrices $\mathbf{I}\sigma_a^2$ and $\mathbf{I}\sigma_e^2$, respectively. The aim of the following statistical analysis is to use $\mathbf{y}$ for estimating the true breeding value of an individual i given by $g_i = \mathbf{w}_i'\mathbf{a}$, where $\mathbf{w}_i'$ contains QTL genotype scores.

Statistical model

Phenotypes generated by the genetic model are used in

$$\mathbf{y} = 1\mu + \mathbf{g} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{g}$ and $\boldsymbol{\varepsilon}$ are vectors containing breeding values and residual effects, respectively. Breeding values in $\mathbf{g}$ are random with mean zero and variance–covariance matrix $\mathbf{G}\sigma_\beta^2$, where $\mathbf{G} = \mathbf{ZZ}'$, $\mathbf{Z}$ is a matrix of genotype scores at K SNPs (VanRaden 2008), $\sigma_\beta^2 = \sigma_A^2/2\sum_{k=1}^{K} p_k(1-p_k)$ (Habier *et al.* 2007), $\sigma_A^2 = \sigma_a^2 \sum_{q=1}^{N_{qtl}} 2p_q(1-p_q)$ is the additive-genetic variance (Gianola *et al.* 2009, Equation 18), σ_a^2 is the variance of additive QTL effects with mean zero, p_q is the allele frequency at QTL q in founders, and p_k is the allele frequency at SNP k in founders. The genotype score of a training

individual at SNP k is the number of one of its alleles adjusted by $2p_k$. Residual effects have mean zero and variance $\mathbf{I}\sigma^2_\epsilon$.

Statistical methods

Following Henderson (1973), the breeding value of individual i can be estimated by BLUP as

$$\hat{g}_i = \mathbf{G}_{i-}(\mathbf{G} + \mathbf{I}\lambda)^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}),$$

where $\mathbf{G}_{i-} = \mathbf{z}_i'\mathbf{Z}'$ is a vector of genomic relationships between individual i and the training individuals, z_i' is a vector of adjusted SNP genotype scores of individual i, and $\lambda = \sigma^2_\epsilon/\sigma^2_\beta$. The overall mean, μ, is estimated by generalized least squares.

The three types of quantitative-genetic information

In pedigree analyses that model LD and cosegregation explicitly (*e.g.*, Pérez-Enciso 2003), genotype scores are realized values of random processes that start with the sampling of founder alleles and continue with the transmission of those alleles from generation to generation down the pedigree. Founder alleles from different loci, but on the same gamete, are not sampled independently if loci are in LD; and nonfounder alleles from different loci, but on the same gamete, are not transmitted independently if loci are linked. We define the following:

Linkage disequilibrium: Statistical dependency between alleles at two or more loci on the same gamete. It is measured only in founders and therefore summarizes historic population events and describes genetic relationships between founders.

Cosegregation: Deviation from independent segregation of alleles on the same gamete if loci are linked. In other words, it describes the inheritance of alleles at linked loci. Thus, it is unnecessary to measure LD either in nonfounder generations or within families, because such LD is sufficiently explained by LD in founders and cosegregation.

Additive-genetic relationships: Statistical dependency between alleles from the same locus but from two different gametes. In genomic BLUP, any SNP can contribute additive-genetic relationship information between two individuals at QTL because, if there is a possibility that the SNP alleles on the two gametes can be traced back to a common founder allele, the same would be true at any QTL.

In pedigree analyses, these principles used to model dependence between allele states at different loci on a gamete are analogous to those used to model additive-genetic covariances between pedigree members, using a single additive-genetic variance defined in the founders together with the additive-genetic relationship matrix constructed for the pedigree. In many analyses, however, the pedigree is ignored and only a single value of LD is used to characterize the dependence between alleles at different loci. In modeling covariances, this is analogous to ignoring the pedigree and estimating a single additive-genetic variance for the entire pedigree, which is not done in practice. In other situations, LD is defined for each family. This is analogous to defining a family-specific additive-genetic variance, which also is not done.

Table 1 Simulated designs that differ in the quantitative-genetic information available for genomic prediction

	QTL and SNPs are			
Design	Linked	In LD[a]	SNPs are in LD[a]	Related[b]
LD only	Yes	Yes	Yes	No
RS only	No	No	Yes	Yes
RS + CS	Yes	No	Yes	Yes
RS + CS + LD	Yes	Yes	Yes	Yes

[a] LD measured in founders.
[b] Training and validation individuals are related.

Simulations

The aim was to study contributions of LD, cosegregation, and additive-genetic relationships to accuracy of GEBVs in dairy cattle and corn breeding scenarios. Factors analyzed were SNP density, training data size, extent of LD, and different relationships between training and validation individuals.

Designs for analyzing different types of information

Four designs were considered that differ in the types of genetic information utilized in genomic BLUP. As summarized in Table 1, these designs utilized (1) only founder LD (LD only), (2) only additive-genetic relationships (RS) (RS only), (3) additive-genetic relationships and cosegregation (CS) (RS + CS), and (4) all three sources of information (RS + CS + LD). In LD only, training and validation individuals were unrelated, whereas in all other cases each validation individual had the same number of relatives in training. In RS only, QTL were located on different chromosomes than SNPs to avoid linkage between these two types of loci, and chromosomes carrying the QTL were simulated independently from the chromosomes with SNPs to exclude LD between QTL and SNPs. In designs with cosegregation or LD, all loci were located on the same chromosomes to ensure linkage. In the RS + CS design, QTL and SNPs were in linkage equilibrium by resampling founder alleles at QTL, using founder allele frequencies. Importantly, SNPs were always in LD, because this has a large effect on capturing information from additive-genetic relationships and cosegregation. In RS + CS + LD, QTL and SNPs were in LD.

Pedigree structure

Two types of pedigrees were simulated as summarized in Table 2: one represents a cross-validation scenario from dairy cattle breeding, and the other one is a *top-cross* design

Table 2 Dairy cattle and corn breeding scenarios simulated in combination with short- and long-range LD, the four information designs, various SNP densities, and 20 QTL per chromosome

Pedigree type	No. chromosomes	h^2	Family type	Family size in training[a]	Scenario	No. families	Training size	No. replicates
Dairy cattle	29	0.5	Half sibs	7	1	14	98	750
					2	143	1001	300
					3	285	1995	75
Corn	10	0.33	Full sibs	30	4	15	450	1000
					5	60	1800	200

[a] Corresponds to the number of relatives of a validation individual in training.

(Falconer and Mackay 1996, p. 276) from corn breeding similar to those in Albrecht *et al.* (2011). The dairy cattle pedigree consisted of 14, 143, or 285 families, each having 7 half sibs in training and 10 in validation. Hence, each validation individual had 7 half sibs in training. In the LD-only design, none of the validation individuals was related to the training individuals, while the half-sib structure in training was retained to capture the same LD information as in the other information designs. The total numbers of training individuals were 98, 1001, and 1995, according to the number of half-sib families in the pedigree.

The corn breeding pedigree consisted of either 15 or 60 families, each having 60 doubled haploids (Bernardo 2010) that descended from two inbred parents. Each doubled haploid was crossed to a single inbred called tester (Bernardo 2010) that is used across all families to generate hybrids. Half of the hybrids were used for training and the other half for validation, so that each validation hybrid had 30 closely related hybrids in training. In the LD-only design, training and validation hybrids were unrelated. In total, the training set consisted of either 450 or 1800 hybrids depending on the number of families simulated (Table 2). Persistence of LD and cosegregation information without retraining was evaluated by simulating validation hybrids that were derived from the next generation of doubled-haploid families (Figure 1). These doubled haploids were the grand-progeny of the original inbred parents, where one parent was a founder, while the other parent was a full sib of those doubled haploids that had training hybrids. A family of the next generation had 30 doubled haploids, each having one validation hybrid, where the tester was the same in both generations.

Genome structure

The number of chromosomes, their length, and the number of SNPs per chromosome differed for the two types of pedigrees. These data were provided by DuPont Pioneer for maize and by the US Department of Agriculture (USDA) for dairy cattle (G. Wiggans, personal communication) as presented in Supporting Information, File S2 and File S3, respectively. The ten *Zea mays* chromosomes with 55,843 SNPs were used in the corn breeding scenario, and the 29 bovine autosomes with 47,833 SNPs were used in the dairy cattle scenario. SNPs were evenly spaced and 200 QTL were randomly positioned on each chromosome in addition to the SNPs.

LD was simulated by starting with a base population of 1500 individuals in linkage and Hardy–Weinberg equilibria and allele frequencies of 0.5. As outlined in Table 3, this population was randomly mated, excluding selfing, for 1000 discrete generations to generate short-range LD due to genetic drift between biallelic loci. Afterward, the population was reduced to a size of 100 individuals and randomly mated for another 15 discrete generations to extend the range of LD. The same simulation scheme was used by Habier *et al.* (2010b), showing good agreement between simulated LD and LD observed in real dairy cattle populations (De Roos *et al.* 2008). Simulations using short-range LD were generated by omitting the last 15 generations with 100 individuals. For comparisons of accuracies from the pedigree-based selection index with accuracies of GEBVs due to additive-genetic relationships, founders of simulated pedigrees were not allowed to be closely related. Therefore, in the scenario with long-range LD (short-range LD) the 100 (1500) individuals from generation 1015 (1000) were randomly mated to create 10,000 offspring, which were then randomly mated for another 2 discrete generations, while retaining a constant population size (Table 3). Founders of pedigrees used to simulate training and validation individuals were drawn without replacement from the last generation of 10,000 individuals. The number of crossovers in meiosis was simulated by a binomial mapping function with mean of 1 crossover per morgan (Karlin 1984), crossover positions were uniform, and the mutation rate was 2.5×10^{-5} as in other simulations (Habier *et al.* 2007,

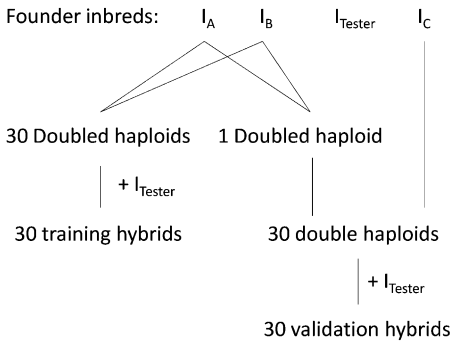


Figure 1 Top-cross design showing one of the families used in cross-validations with training hybrids descending from the first generation of doubled haploids and validation hybrids descending from the next generation of doubled haploids. Both training and validation hybrids come from the same inbred tester, I_{Tester}.

Table 3 Number of generations of random mating and population size used to generate short-range and long-range LD

Short-range LD		Long-range LD	
Generation	Size	Generation	Size
1	1,500	1	1,500
1,000	1,500	1,000	1,500
1,001	10,000	1,001	100
1,003 (founders)	10,000	1,015	100
		1,016	10,000
		1,018 (founders)	10,000

2009; Daetwyler *et al.* 2010; Calus and Veerkamp 2011; Bastiaansen *et al.* 2012). The average LD between adjacent SNPs in the scenarios long-range LD and short-range LD was 0.21 and 0.15, respectively, and LD between QTL and SNPs is depicted in Figure 2.

SNP density was varied by sampling a subset of SNPs for each analysis from the total number of available SNPs, where the number of SNPs per chromosome was proportional to chromosome length as shown in File S2 and File S3. In the corn breeding designs, 195, 996, 4995, 9995, 19,995, and 39,996 SNPs were used in the statistical analysis, while in the dairy cattle designs 564, 2885, 14,483, 28,984, 43,483, and 47,831 SNPs were used. From the 200 QTL that were initially positioned on each chromosome, only 20 were randomly selected in each replicate and effects were sampled from a standard normal distribution. These QTL effects were standardized such that the additive-genetic variance was 1 in founders of the dairy cattle pedigrees and 2 in inbred founders of the corn breeding pedigrees.

Phenotypes

Phenotypes of training individuals followed the genetic model of Equation 1, where the residual effects were sampled from a standard normal distribution. Consequently, heritability was 0.5 in cattle scenarios, whereas only one-third of the variance of hybrid phenotypes was due to additive-genetic effects as a result of the single tester.

Evaluation criteria: Accuracy of GEBVs was defined as correlation between GEBVs and true breeding values of validation individuals and was estimated both within and across families. Across families means that one correlation was calculated using validation individuals from all families, whereas within families means that a correlation was calculated for each family. These accuracies were calculated for each replicate of the simulation and averaged across replicates. Accuracy of the pedigree-based selection index (pedigree index) was calculated for the dairy cattle scenario by

$$\rho_{g_i,\hat{I}_i} = 0.5\sqrt{\frac{n}{n+(4-h^2)/h^2}}$$

(Mrode 2005, p. 9), where n is the number of half sibs of validation individual i in training, and h^2 is the heritability. Another criterion of interest regarding the avoidance of inbreeding and improving selection intensities in breeding schemes is the correlation between GEBVs of selection candidates from the same family, $\rho_{\hat{g}_i\hat{g}'_i}$ (Hill 1976). It can be estimated by the intraclass correlation (Snedecor and Cochran 1967) with

$$\hat{\rho}_{\hat{g}_i\hat{g}_{i'}} = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_w},$$

where σ^2_b and σ^2_w are variances between and within families, respectively, estimated by a one-way ANOVA (Snedecor and Cochran 1967). This intraclass correlation is used here to determine the role of additive-genetic relationships in the presence of LD and cosegregation as detailed in *Discussion*.

Results

Dairy cattle pedigree

Accuracies of GEBVs across families obtained with long-range LD and 1001 training individuals are shown in Figure 3. Accuracies increased with SNP density and reached plateaus in all four information designs, where those of RS only and RS + CS plateaued at a lower SNP density than those of LD only and RS + CS + LD. Even at the highest SNP density, accuracy for RS only was 0.14 (±0.002), lower than that for the pedigree index. Cosegregation information contributed only little to accuracy, both across (RS + CS) and within half-sib families (results not shown). With LD, genomic BLUP outperformed pedigree index with an accuracy of 0.56 (±0.002) when each validation individual had seven half sibs in training (RS + CS + LD), while the accuracy for an unrelated validation individual was 0.09 (±0.002) lower (LD only). Figure 4A depicts accuracies of GEBVs with increasing training set size obtained with long-range LD and 47,831 SNPs. As expected, accuracy due to LD increased with training set size (LD only), but accuracy due to additive-genetic relationships declined (RS only), thereby increasing the discrepancy to accuracy of pedigree index. Similarly, cosegregation contributes less with increasing training set size, both across (Figure 4A, RS + CS) and within families (results not shown). As a result, the difference in accuracy for validation individuals that were either related or unrelated to the training data decreased from 0.21 (±0.005) to 0.05 (±0.003) (LD only *vs.* RS + CS + LD). Accuracies from LD only were lower for short-range LD than those for long-range LD, while those from RS only and RS + CS were higher (Figure 4B *vs.* 4A).

Intraclass correlations were 0.25 (±0.002) at low SNP density, increased with increasing SNP density (results not shown), and quickly plateaued at values shown in Figure 5 for different training set sizes and information designs using long-range LD. In general, a low intraclass correlation is favorable in mass selection for low inbreeding and high realized selection intensities. For LD only, intraclass correlations were 0.29 (±0.002) across all training set sizes, but

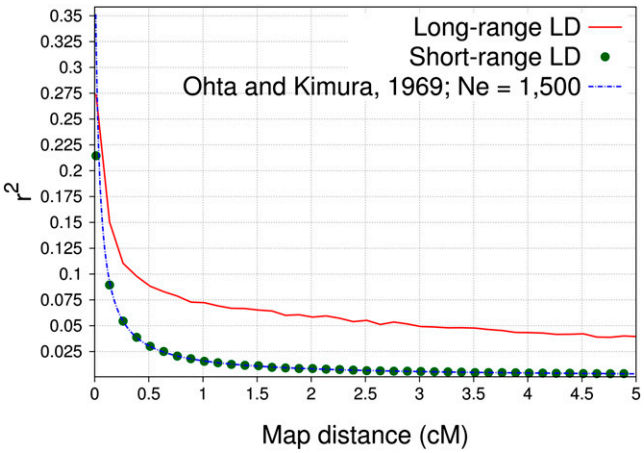


Figure 2 Linkage disequilibrium between QTL and SNPs measured as r^2 against map distance in centimorgans for the two scenarios long-range and short-range LD and using formulas by Ohta and Kimura (1969) with an effective population size of 1500.

when training and validation individuals were related, they decreased with increasing training set size irrespective of LD and cosegregation from 0.81 (±0.002) to 0.37 (±0.002). Thus, intraclass correlations were very similar at a given training set size for the information designs RS only, RS + CS, and RS + CS + LD. The explanation is that the variance of GEBVs both within and between families increased with LD and cosegregation information.

Corn breeding pedigree

Accuracies of GEBVs within and across families increased with SNP density and plateaued at different levels, depending on information utilized. In designs with LD, plateaus were reached at higher SNP densities than in RS only and RS + CS, especially as training set size increased (results not shown). Figure 6A depicts accuracies of GEBVs across families obtained with long-range LD and 39,996 SNPs for 15 and 60 doubled-haploid families used in cross-validations. Accuracies were much higher for validation individuals that were related to the training data (LD only *vs.* other designs), because 30 related training hybrids per validation individual provided extensive additive-genetic relationships, resulting in an accuracy of 0.56 (±0.003) for RS only. Cosegregation information increased that accuracy considerably by 0.09 (±0.004), while LD added only little more (RS only *vs.* RS + CS and RS + CS *vs.* RS + CS + LD). As training set size increased from 450 to 1800 hybrids, accuracies of LD only and RS + CS + LD increased by 0.2 (±0.003) and 0.05 (±0.003), respectively, whereas those of RS only and RS + CS remained constant (Figure 6A). Accuracies within families are depicted in Figure 6B for all information designs but RS only, because additive-genetic relationships explain no variation of GEBVs within families. For 15 families, cosegregation provided more information than LD (RS + CS *vs.* LD only). The difference between the designs LD only and RS + CS + LD shows the contribution of cosegregation in the presence of LD, which was 0.15 (±0.003). As training set size increased, more LD information and less cosegregation information were exploited, so that accuracies of LD only and RS + CS + LD increased by 0.19 (±0.003) and 0.11 (±0.002), respectively, while the accuracy of RS + CS decreased by 0.07 (±0.002). Consequently, the contribution of cosegregation in the presence of LD decreased by 0.09 (±0.002) to 0.06 (±0.002) (LD only *vs.* RS + CS + LD).

Figure 7 depicts accuracies within families for validation hybrids of both the same generation as the 450 training hybrids and the next generation. Accuracy of LD only remained constant from one generation to the next, whereas the accuracies of RS + CS and RS + CS + LD dropped by 0.17 (±0.007) and 0.11 (±0.006), respectively, due to the decline of cosegregation information.

Discussion

The objectives of this article were (1) to present concepts that allow us to disentangle LD in founders, cosegregation, and additive-genetic relationships and (2) to study their contributions to accuracy of GEBVs by simulation of four designs that differ in the types of information available (Table 1). In addition, formulas were derived in File S1 and File S4 for a simplified scenario proving that the three types of information contribute to accuracy of GEBVs. In the following, mechanisms that lead to the results are elaborated and then consequences for practical application of genomic BLUP are discussed.

Concepts and simulation designs

The concepts presented here can be applied to any statistical method. As for all pedigree analyses, contributions attributed to the three types of information depend on pedigree depth. Generally, cosegregation is expected to become more important as pedigree depth increases, but this requires further investigations. Here, the pedigree consisted of only one nonfounder generation in training, which allowed us to evaluate cosegregation information from half- and full-sib families. If the pedigree had more nonfounder generations in training, cosegregation would also comprise information from more distant relatives. Thus, a better understanding can be gained by varying pedigree depth.

The LD-only design is a realistic scenario, whereas RS only and RS + CS seem contrived. However, RS only always occurs in reality when a SNP on one chromosome explains variation at a QTL on another chromosome that is not explained by LD and cosegregation. Also, intraclass correlations have shown that the findings from RS only are relevant for realistic scenarios, which are detailed later. As for RS + CS, LD patterns vary across the genome; hence there may be QTL that are in low LD with SNPs.

Information from LD

In the LD-only design, extent and amount of LD determine both SNP density at which the plateau is reached (Figure 3) and accuracy of GEBVs at the plateau (Figure 4A *vs.* 4B). For

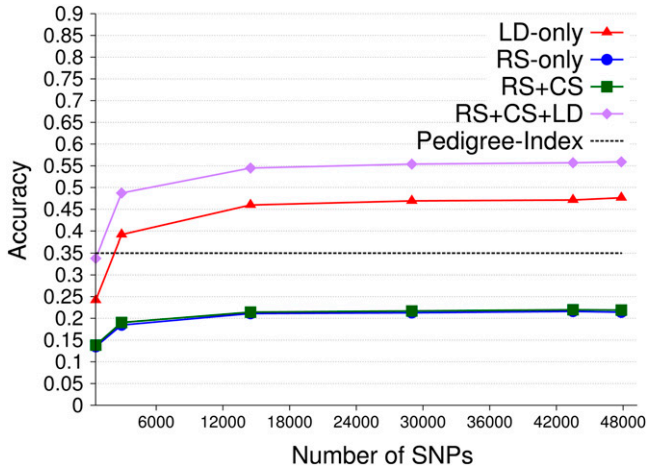


Figure 3 Accuracy of GEBVs obtained by genomic BLUP, long-range LD, and 47,831 SNPs for the four information designs and accuracy of pedigree index using 1001 training individuals structured into 143 half-sib families and a heritability of 0.5. Each validation individual had 7 half sibs in training in all designs but LD only. The number of replicates was 300.

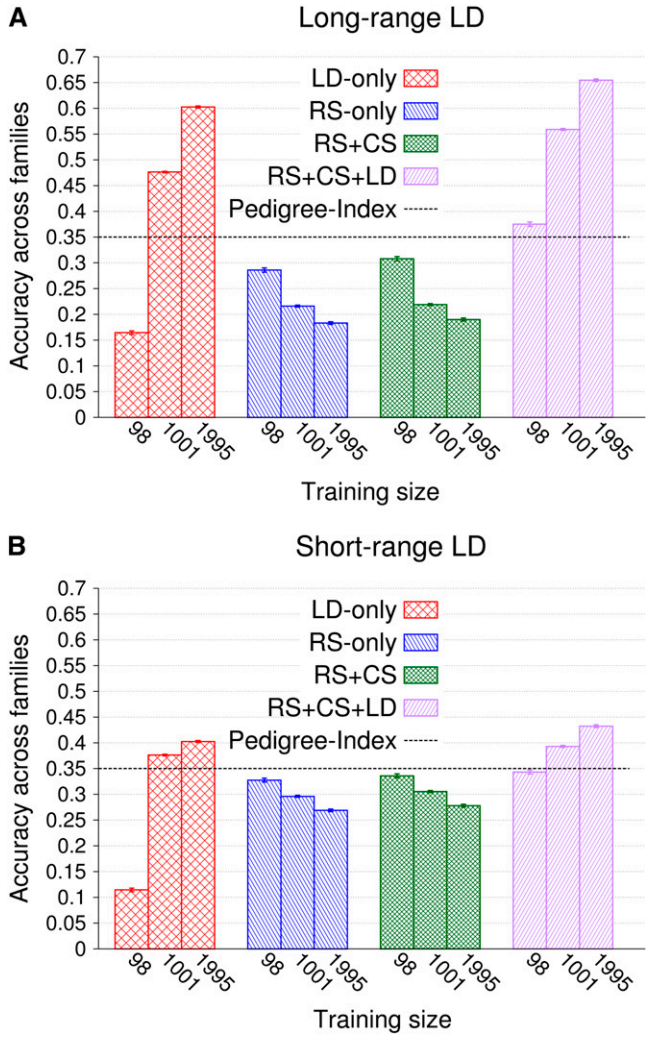


Figure 4 (A and B) Accuracy of GEBVs and standard errors obtained by genomic BLUP and 47,831 SNPs for the four information designs according to training data size and extent of LD and accuracy of pedigree index using a heritability of 0.5. Training data were structured into half-sib families of size seven, and each validation individual had seven half sibs in training in all designs but LD only. The numbers of replicates for training data sizes 98, 1001, and 1995 were 750, 300, and 75, respectively.

short-range LD as in humans (Reich *et al.* 2001) that SNP density is expected to be higher and accuracy to be lower than for long-range LD found in animals and plants under selection (Andreescu *et al.* 2007). The reason is that with increasing SNP density the shrinkage of SNP effects in genomic BLUP becomes stronger. An extreme case was described by Fernando *et al.* (2007) in which loci were in linkage equilibrium, but QTL were included together with SNPs in genomic BLUP and BayesB (Meuwissen *et al.* 2001). While BayesB found the QTL and provided high accuracies, the accuracies of genomic BLUP decreased with increasing SNP density until approaching the accuracy of pedigree BLUP. This does not happen under realistic conditions, because with increasing SNP density more SNPs support the same QTL and compensate shrinkage, which results in a balance expressed by the accuracy at the plateau. Under long-range LD, QTL effects are captured by more SNPs than with short-range LD, so that shrinkage has a smaller impact. Even if QTL are included in the statistical model for genomic BLUP, accuracy of GEBVs obtained under short-range LD did not change for training data sizes used in this study. In humans, training data sizes are much larger, but also millions of SNPs that affect shrinkage are used. Therefore, as the amount of LD information is responsible for *missing heritability* in humans (Yang *et al.* 2010), we argue that Bayesian methods with *t*-distributed priors, which are expected to exploit LD better than genomic BLUP (Fernando *et al.* 2007; Habier *et al.* 2007, 2010a, 2011), are more suitable than genomic BLUP. This disagrees with results from Ober *et al.* (2012), but that study used only 124 training individuals.

Information from additive-genetic relationships and cosegregation

It can be shown that additive-genetic relationships are captured best when SNPs are in linkage equilibrium, segregate independently, and have a minor allele frequency of 0.5. These SNPs are referred here to as *ideal* SNPs. In Habier *et al.* (2007, 2010a), the accuracy of GEBVs due to additive-genetic relationships obtained by genomic BLUP approached the accuracy of pedigree BLUP when the number of simulated ideal SNPs exceeded the number of training individuals. In reality, however, SNPs are in LD, linked on a limited number of chromosomes, and the average minor allele frequency is <0.5. Thus, the actual number of SNPs in the model is not informative about the ability to explain additive-genetic relationships. Therefore, we define the effective number of SNPs (M_{SNP}) as the number of ideal SNPs that gives the same accuracy due to additive-genetic relationships as the actual number of SNPs in the model for a given cross-validation scenario. The effective number of

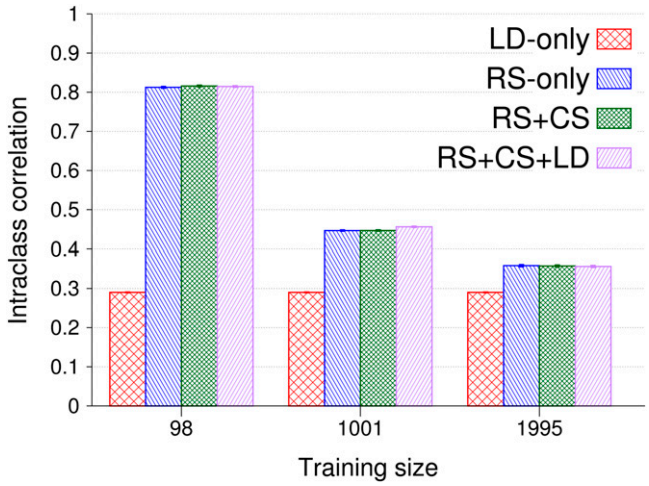


Figure 5 Intraclass correlations and standard errors for GEBVs within half-sib families obtained by genomic BLUP, long-range LD, and 47,831 SNPs according to training data size and information design. Training data were structured into half-sib families of size seven, and each validation individual had seven half sibs in training in all designs but LD only. The numbers of replicates for training data sizes 98, 1001, and 1995 were 750, 300, and 75, respectively.

SNPs was estimated here by additional simulations of the RS-only design, using only ideal SNPs. For accuracies of 0.22 and 0.3 obtained with long-range and short-range LD (Figure 4, A and B, RS only), respectively, using 1001 training individuals and 47,831 actual SNPs, M_{SNP} was ~2000 for long-range LD and 6000 for short-range LD. Thus, M_{SNP} decreases with increasing range of LD and therefore depends on effective population size. Also, M_{SNP} is smaller than the number of SNPs on high-density SNP chips, which explains why the accuracy due to additive-genetic relationships did not improve beyond a certain SNP density (Figure 3). Cosegregation is also captured best when SNPs are in linkage equilibrium and have high minor allele frequencies, because each family can have most distinctive SNP haplotypes around QTL. With increasing range of LD, however, SNP haplotypes become similar across families, so that the ability to capture cosegregation decreases.

Information with increasing training data size

The increase in accuracy due to LD with training data size is well known, whereas the decrease in accuracy of GEBVs due to additive-genetic relationships in the dairy cattle design is new. This decrease is related to the effective number of SNPs and the level of additive-genetic relationships between individuals. In RS only, consider the genomic relationship matrix as an estimate of the pedigree relationship matrix (Habier *et al.* 2007; Gianola *et al.* 2009). Deviations between the coefficients of these two matrices cause the linear combinations of phenotypes for estimating GEBVs to become inaccurate. For example, phenotypes of training individuals that are unrelated to the validation individual contribute to the GEBV of a validation individual in genomic BLUP. As training data size increases, more erroneous contributions occur. The decay of accuracy with increasing training data size was lower with short-range LD than with long-range LD, because the effective number of SNPs was larger under short-range LD, resulting in smaller deviations between those two matrices. In the corn breeding scenarios, accuracies due to additive-genetic relationships did not decrease with increasing training data size (Figure 6) because the level of pedigree relationships was higher than in dairy cattle designs; with a higher level, the importance of inaccurately weighted phenotypes relative to the phenotypes of related training individuals becomes smaller.

In practice, decreasing information from additive-genetic relationships may result in a smaller increase of accuracy observed with real data or even a decay if LD information cannot compensate for that loss. This can be suspected from results of Habier *et al.* (2011), because the increase in accuracy due to LD from 4000–6500 training bulls was small. Also, combining training data from different breeds (Hayes *et al.* 2009a) may risk a reduction in accuracy due to additive-genetic relationships. In contrast, if training data sets from the same breed but different breeding regions are combined, accuracy due to additive-genetic relationships can increase if individuals are closely related as for example in dairy cattle (Lund *et al.* 2011). The decline in accuracy due to additive-genetic relationships with increasing training data size may be avoided by simultaneously fitting polygenic effects via traditional pedigree relationships together with genomic breeding values using Bayesian methods (Calus and Veerkamp 2007). This simultaneous inference of effects is necessary because the partitioning of the genetic variance into polygenic and genomic components depends on training data size.

Cosegregation information decreased with increasing training data size (Figure 6) for similar reasons to those described for additive-genetic relationships: genomic relationships estimate covariances between individuals at QTL, where deviations from true covariances result in prediction errors of GEBVs. Modeling LD and cosegregation explicitly may avoid this decline in accuracy due to cosegregation. Also, multiallelic markers such as copy number variants, which are available from sequence data, may enhance utilization of additive-genetic relationships and cosegregation.

Contributions to accuracy of GEBVs

The four information designs allowed us to evaluate the maximal contribution of LD, cosegregation, and additive-genetic relationships. The level of accuracy due to LD depends on genome structure (Figure 4) and hardly on the number of QTL (Daetwyler *et al.* 2010). However, this may not be misinterpreted such that genetic architecture does not affect the accuracy due to LD. Results of Habier *et al.* (2010a, 2011) showed that even traits with similar heritability, such as fat and protein yield, can have very different accuracies due to LD. One explanation is that QTL of different traits are in different LD with SNPs, because the genome consists of long- and short-range LD patterns

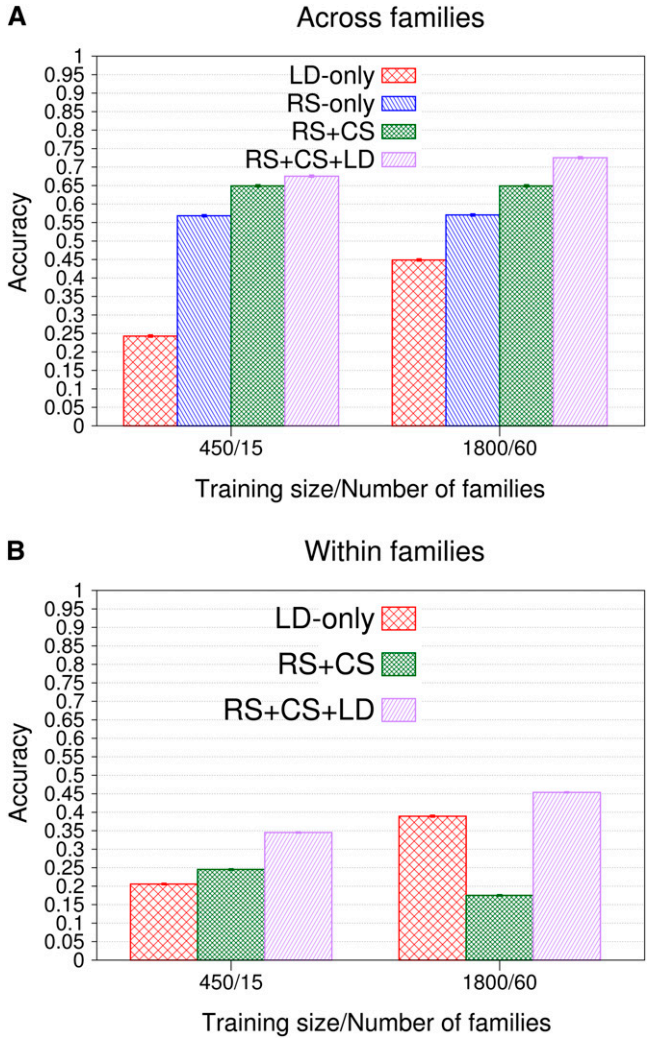


Figure 6 (A and B) Accuracy of GEBVs and standard errors within and across families obtained by genomic BLUP, long-range LD, and 39,991 SNPs for the four information designs using 450 and 1800 training hybrids structured into families of size 30. Each validation hybrid had 30 related hybrids in training in all designs but LD only. The numbers of replicates for training data sizes 450 and 1800 were 1000 and 200, respectively.

(Qanbari *et al.* 2010). However, the difference in accuracy between those two traits was even larger with BayesA and BayesB (Habier *et al.* 2011), which have different shrinkage mechanisms than genomic BLUP. Thus, an explanation may be epistasis.

Accuracy due to additive-genetic relationships and cosegregation can be regarded as lower bounds if accuracy due to LD is small as for somatic cell score in dairy cattle (Habier *et al.* 2011). However, that accuracy depends largely on the number of close relatives in training such as parents and siblings; for more distant relatives, the effective number of SNPs may not be sufficient. This may be an emerging problem in dairy cattle, because information from parents and siblings will not be available anymore due to accelerated selection cycles by genomic selection. In dairy cattle, cosegregation from half sibs plays a minor role, because selection candidates have a limited number of half sibs in training (Habier *et al.* 2010a), a few selection candidates have only one or two full sibs, and parents do not provide cosegregation information. Additionally, the number of half-sib families is large, which is unfavorable for capturing cosegregation (Figure 4). In plants, a top-cross design provides extensive additive-genetic relationship information to the accuracy across families due to many training hybrids that are related to a validation hybrid. For the same reasons, cosegregation contributes notably to the accuracy within family. Therefore, if LD information does not increase further with training data size, large full- or half-sib families can be generated for training to exploit cosegregation information.

Correlation between GEBVs within family

Intraclass correlations were evaluated to estimate correlations between GEBVs within half-sib families, which give additional insight into the use of additive-genetic relationships in the presence of LD and cosegregation. There are at least two notions: the accuracy of GEBVs is either (1) due to LD plus a remainder due to additive-genetic relationships and cosegregation (LD only *vs.* RS + CS + LD) or (2) due to additive-genetic relationships plus a remainder due to LD and cosegregation (RS only *vs.* RS + CS + LD). If accuracy due to LD is high, the first notion, which was suggested by Habier *et al.* (2007), can underestimate the importance of additive-genetic relationships captured by SNPs for genomic prediction as demonstrated by the intraclass correlations. These were similar for all information designs in which validation individuals were related to the training data despite LD and cosegregation information and decreased with increasing training data size (Figure 5). Thus, the correlation of GEBVs within families depends mostly on additive-genetic relationships captured by SNPs, and therefore they are similarly important for realized selection intensities and expected inbreeding in genomic selection with and without cosegregation and LD information.

Persistence of LD and cosegregation information

Persistence of accuracy across generations without retraining is an important criterion in genomic selection. Results of the corn breeding scenarios showed that LD is more persistent than cosegregation from a single full-sib family (Figure 7). Thus, the decay of accuracy within families for individuals from the first few generations after training may be due to the decay of cosegregation information caused by recombinations of haplotypes surrounding QTL (Figure 7, RS + CS + LD). The surprisingly large decay of cosegregation information may indicate that cosegregation was explained by rather large chromosome segments. However, if the training data contain multigeneration families, persistence of cosegregation information might be different. As capturing additive-genetic relationships becomes more difficult with decreasing genetic relatedness

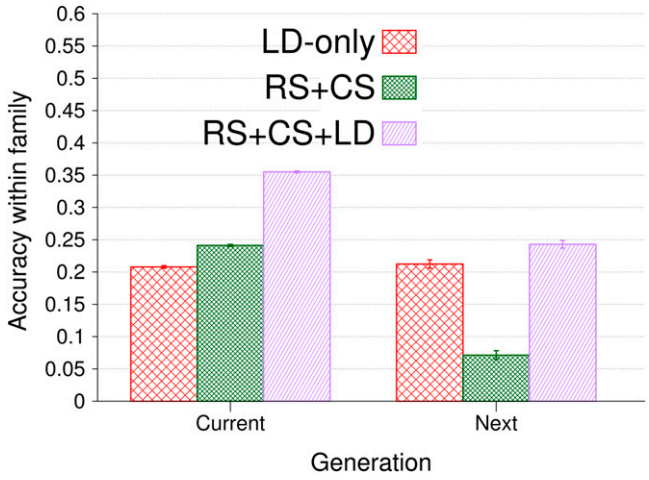


Figure 7 Accuracy of GEBVs and standard errors within families obtained by genomic BLUP, long-range LD, and 39,991 SNPs for validation hybrids of the same (current) and following (next) generation and for the designs LD only, RS + CS, and RS + CS + LD, using 450 training hybrids. Each validation hybrid had 30 related hybrids in training in all designs but LD only. The number of replicates was 1000.

between training and validation individuals when training data size is large, capturing cosegregation may be even more difficult. This also suggests modeling LD and cosegregation explicitly (*e.g.*, Calus *et al.* 2008) if LD information is captured at least as well as with BayesA and BayesB.

Conclusions

We showed that genomic BLUP exploits LD, cosegregation, and additive-genetic relationships captured by SNPs in analyses that explicitly define LD and cosegregation information. We demonstrated that additive-genetic relationship information can decline with increasing training data size, depending on extent of LD and level of additive-genetic relationships. This suggests that polygenic effects should be modeled jointly with either SNP effects or genomic breeding values by a pedigree relationship matrix using Bayesian methods. The correlation of genomic estimated breeding values within families—an important parameter in breeding schemes—depends largely on additive-genetic relationship information, which is determined by the effective number of SNPs and training data size. Little cosegregation information comes from half sibs in current dairy cattle breeding designs, but cosegregation information from full sibs can contribute considerably to accuracy within families in corn breeding. However, its persistence is lower than that of LD information because cosegregation information declines quickly with increasing training data size and over generations without retraining. Thus, LD and cosegregation should be modeled explicitly. As genomic BLUP is not suitable to capture LD information in genome regions in which LD decays rapidly with map distance, we recommend Bayesian methods with *t*-distributed priors.

Acknowledgments

This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2012-67015-19420 and by National Institutes of Health grant R01GM099992.

Literature Cited

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont *et al.*, 2007 Linkage disequilibrium in related breeding lines of chickens. Genetics 177: 2161–2169.

Bastiaansen, J., A. Coster, M. Calus, J. van Arendonk, and H. Bovenhuis, 2012 Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44: 3.

Bernardo, R., 2010 *Breeding for Quantitative Traits in Plants*, Ed. 2. Stemma Press, Woodbury, Minnesota.

Calus, M., and R. Veerkamp, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. 124: 362–368.

Calus, M., and R. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genet. Sel. Evol. 43: 26.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553–561.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: e3395.

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

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503–1512.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Prentice-Hall, Englewood Cliffs, NJ.

Fernando, R. L., 2003 Statistical issues in marker assisted selection, pp. 101–108 in *8th Genetic Prediction Workshop of The Beef Improvement Federation*. Edited by L. V. Cundiff.

Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers, and L. R. Totir, 2007 Genomic selection. Acta Agric. Scand. Anim. Sci. 57: 192–195.

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

Goddard, M. E., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245–257.

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

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2009 Genomic selection using low-density marker panels. Genetics 182: 343–353.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010a The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Habier, D., L. R. Totir, and R. L. Fernando, 2010b A two-stage approximation for analysis of mixture genetic models in large pedigrees. Genetics 185: 655–670.

Habier, D., R. Fernando, K. Kizilkaya, and D. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hayes, B., P. Bowman, A. Chamberlain, K. Verbyla, and M. Goddard, 2009a Accuracy of genomic breeding values in multibreed dairy cattle populations. Genet. Sel. Evol. 41: 51.

Hayes, B., M. Goddard, and P. Vissher, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91: 47–60.

Henderson, C. R., 1973 *Sire evaluation and genetic trends*, pp. 10–41 in *Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*. American Society of Animal Science and American Dairy Science Association, Champaign, IL.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.

Hill, W. G., 1976 Order statistics of correlated variables and implications in genetic selection programs. Biometrics 32: 889–902.

Hill, W., and B. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93: 47–64.

Karlin, S., 1984, Theoretical aspects of genetic map functions in recombination processes, pp. 209–228 in *Human Population Genetics: The Pittsburgh Symposium*, edited by A. Chakravarti. Van Nostrand Reinhold, New York.

Lee, S., M. Goddard, P. Visscher, and J. van der Werf, 2010 Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genet. Sel. Evol. 42: 22.

Legarra, A., and R. Fernando, 2009 Linear models for joint association and linkage QTL mapping. Genet. Sel. Evol. 41: 43.

Lund, M., A. de Roos, A. de Vries, T. Druet, V. Ducrocq *et al.*, 2011 A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43: 43.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson et Cie., Paris.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161: 373–379.

Mrode, R. A., 2005 *Linear Models for the Prediction of Animal Breeding Values*, Ed. 2. CABI Publishing, Wallingford, Oxfordshire, UK.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 8: e1002685.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63: 229–238.

Pérez-Enciso, M., 2003 Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163: 1497–1510.

Qanbari, S., E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner *et al.*, 2010 The pattern of linkage disequilibrium in German Holstein cattle. Anim. Genet. 41: 346–356.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature 411: 199–204.

Snedecor, G., and W. Cochran, 1967 *Statistical Methods*, Ed. 6. Iowa State University Press, Ames, IA.

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

Yang, J., B. Benyamin, B. Mc, H. A. Evoy, S. Gordon *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152207/-/DC1

Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction

David Habier, Rohan L. Fernando, and Dorian J. Garrick

DOI: 10.1534/genetics.113.152207

File S1

Covariance between true and estimated breeding values for simplified scenarios

This supplement is divided into two parts. In the first part, formulas for the covariance between true and estimated breeding values of a validation individual that were derived for simplified scenarios are summarized and explained for readers that want to avoid detailed derivations. These formulas are illustrated by examples, verified by simulations, and briefly discussed. In the second part of this supplement, detailed derivations are given starting with a crude formula for the covariance between true and estimated breeding values when the training data set contains only one individual. It reveals a connection between the QTL alleles in the phenotype of the training individual and the QTL alleles in the true breeding value of the validation individual via their SNP alleles. This formula was further evaluated for simplified scenarios, such that the covariance becomes a function of allele frequencies, linkage disequilibrium, recombination frequency, shrinkage of SNP effects in the statistical method, and the relatedness of training and validation individuals. The genetic and statistical models underlying the derivations were described in the THEORY section of the main manuscript.

Part 1

Summary of deterministic formulas

The aim was to demonstrate that Genomic-BLUP, a method that does not explicitly condition on SNP genotypes and pedigree information as for example the co-segregation approach of FERNANDO and GROSSMAN (1989), exploits information from linkage disequilibrium (LD) between QTL and SNP alleles of founders, co-segregation (CS) of QTL and SNP alleles at linked loci, and additive-genetic relationships at QTLs (RS) captured by SNPs. For illustration purposes, the genetic model had only 1 QTL, the statistical model had only 1 SNP, and the training data had only one individual. To accommodate only one training individual, selection index methodology was used instead of BLUP, which assumes that the overall mean is known. This sets the focus on the quantitative-genetic information captured by the statistical method for prediction of $\hat{g}_i$ rather than for the estimation of μ, while affecting the accuracy of $\hat{g}$ only marginally. This allows us to evaluate the maximal information content of a single phenotype for genomic prediction.

For simplicity, but without loss of generality for this scenario, the QTL effect was assumed to be fixed. Three scenarios were considered: 1) training and validation individuals are unrelated, 2)

training and validation individuals are half sibs, and QTL and SNP are in linkage equilibrium, and 3) training and validation individuals are half sibs, and loci are in LD. When training and validation individuals are unrelated, and both are assumed to be founders, the covariance between true and estimated breeding values of the validation individual is

$$Cov(g_i, \hat{g}_i) = 4a^2 r^2 Var(w_f) Var(z_f) f_3(p, \lambda)$$

where a is the QTL effect, r^2 denotes LD expressed as the squared correlation between QTL and SNP alleles of founders, $Var(w_f)$ and $Var(z_f)$ are variances of founder allele states at the QTL and SNP, respectively, $f_3(p, \lambda)$ is a function of allele frequency, p, at the SNP, and shrinkage parameter, λ, as defined in the statistical methods in the main manuscript. This function was derived in part 2 of this supplement and can be written as

$$f_3(p, \lambda) = \frac{2p(1-p)}{4p^2 + \lambda} + \frac{1 - 4p(1-p)}{(1-2p)^2 + \lambda} + \frac{2p(1-p)}{4(1-p)^2 + \lambda}.$$

When training and validation individuals are half sibs through a common sire, and the two loci are in linkage equilibrium, the covariance becomes

$$Cov(g_i, \hat{g}_i) = a^2 0.25[2(1-c)^2 + 2c^2] Var(w_f) Var(z_f) f_3(p, \lambda)$$

where 0.25 results from the fact that only the paternal gametes of the training and validation individuals contribute to the covariance (each gamete is drawn with probability of 0.5), and $c \in [0, 0.5]$ is the recombination frequency between QTL and SNP. In $[2(1-c)^2 + 2c^2]$, the term $(1-c)^2$ represents the case where both individuals receive the same non-recombinant gamete from their sire (Figure 1), whereas c^2 means that both individuals received the same recombinant gamete (Figure 2).

Figure 1: Training and validation individuals received the same non-recombinant gamete.

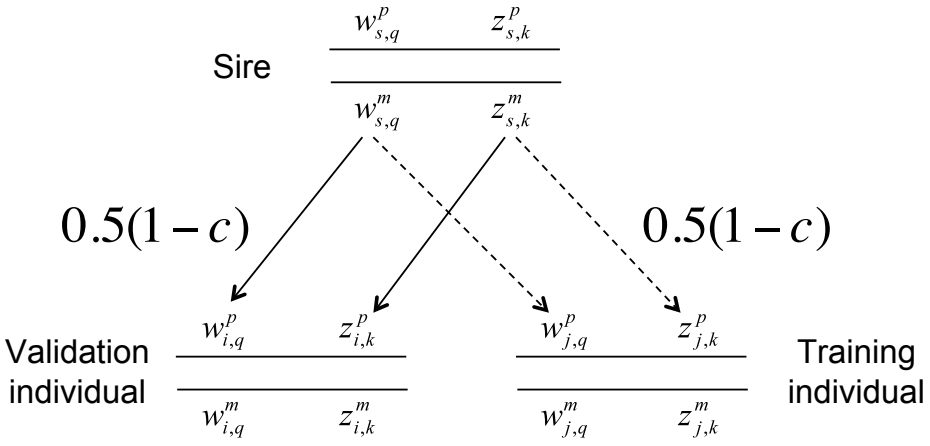
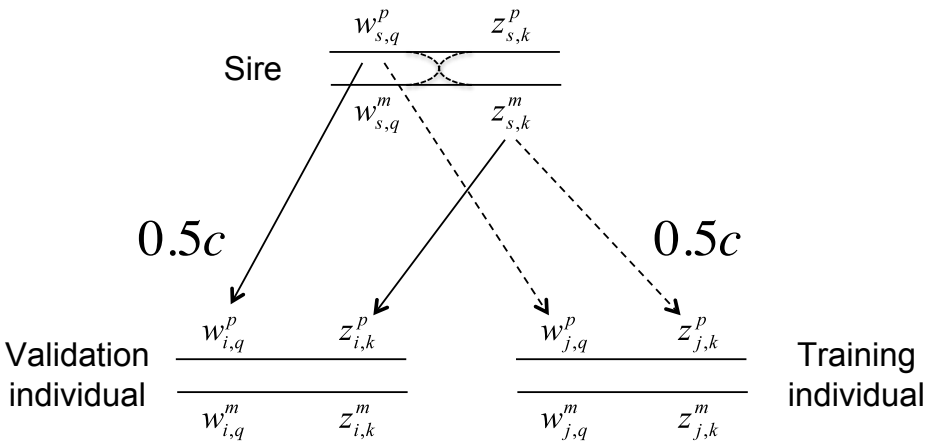


Figure 2: Training and validation individuals received the same recombinant gamete.



The term $[2(1-c)^2 + 2c^2]$ increases with decreasing c from 1 to 2, where 1 is obtained when the two loci are unlinked, which can be interpreted as the scenario where only RS information contributes to the covariance. Thus, the last equation can be modified to disentangle information from RS and CS by writing

$$Cov(g_i, \hat{g}_i) = a^2 0.25\Big[1 + [2(1-c)^2 + 2c^2 - 1]\Big] Var(w_f) Var(z_f) f_3(p, \lambda) \tag{1}$$

where the first 1 within the bracket defines the part due to RS, and $[2(1-c)^2 + 2c^2 - 1]$, which takes values between 0 and 1, defines the part due to CS. Hence, consistent with the definition of RS in the main manuscript, RS always contribute to the covariance whether or not the two loci are linked, whereas CS is considered as additional information when QTL and SNP are linked. Consequently, in this specific scenario, CS contributes at most as much information as RS when $c = 0$. When LD between QTL and SNP exists, the covariance is

$$\begin{aligned} Cov(g_i, \hat{g}_i) = a^2 0.25\Big[&[2(1-c)^2 + 2c^2][Var(w_f)Var(z_f) + D^2] f_3(p, \lambda) \\ &2(1-c)^2 D(1-2p_q)[f_8(p,\lambda) + f_9(p,\lambda)] + \\ &4(1-c)cD(1-2p_q) f_{10}(p,\lambda)\Big] + \\ &a^2 D^2 (2-c) f_3(p,\lambda) + a^2 0.5(1-c) D^2 [f_3(p,\lambda) + f_6(p,\lambda) + f_7(p,\lambda)], \end{aligned} \tag{2}$$

where D denotes LD expressed as the covariance between founder allele states at QTL and SNP. The functions $f_6(\cdot)$ and $f_7(\cdot)$ can be found on page 16 and 17, respectively, while $f_8(\cdot)$, $f_9(\cdot)$, and $f_{10}(\cdot)$ are presented on page 19. Close inspection reveals that equation (2) contains the covariance for the case of linkage equilibrium in equation (1). Thus, all three types of quantitative-genetic information contribute additively to the covariance between true and estimated breeding values.

The contribution of LD to the covariance between true and estimated breeding values can be illustrated by the notion that the genomic relationship between training and validation individuals becomes more informative for the genetic relationship between both individuals at the QTL. LD contributes through both the maternal (founder) gametes of the two individuals and the sire gametes, but in this illustration we focused only on the sire gametes. Let's denote the two allele states at the SNP M and m, and those at the QTL Q and q. Assume a recombination frequency of zero and a founder allele frequency at the SNP of 0.5. Under complete LD, there are only two founder haplotypes, $M-Q$ and $m-q$. If the sire is heterozygous, the training and validation individuals receive either identical or different gametes from the sire. This means that the contribution of the paternal gamete to the genomic relationship between the two individuals either increases or decreases the resemblance at both the SNP and the QTL. Therefore, the genomic relationship becomes informative for the genetic relationship at the QTL. Under linkage equilibrium, in contrast, there are four founder haplotypes each with a frequency of 0.25. If the two sire gametes carry the same SNP allele but different QTL alleles, and training and validation individuals receive different gametes from the sire, the contribution of the paternal gametes to the genomic relationship does not contribute to a resemblance at the QTL. Thus, the linkage equilibrium case is less informative.

Simulations

Simulations were conducted to verify formulas derived for the covariance between true and estimated breeding values of a validation individual, and to show that this covariance is proportional to the accuracy of GEBVs. The simulated scenarios were the same as in the deterministic derivations above when training and validation individuals were half sibs. Founder gametes were drawn from a joint distribution of allele states at one QTL and one SNP by first sampling the SNP allele from a Bernoulli distribution with probability equal to the founder allele frequency at the SNP. The QTL allele was then sampled by the conditional allele frequency at the QTL given the sampled SNP allele. Similarly, the paternal gametes of the two half sibs were drawn by first sampling the maternal or paternal SNP allele of the sire with probability of 0.5. The allele state for the QTL is then derived from either the maternal or paternal gamete of the sire by sampling the origin of the QTL allele from a conditional probability given the origin at the SNP and the recombination frequency between SNP and QTL. The R-code used for this simulation is presented in supplemental file 4.

Results

The covariance between true and estimated breeding values of a validation individual having one half sib in training was almost identical in simulations and deterministic calculations, across all r^2 values and recombination frequencies (Table 1). The covariance increased with increasing LD and decreasing recombination frequency, and was proportional to the accuracy of GEBVs. The contribution of RS to this covariance decreased with decreasing recombination frequency, while that of CS increased. In linkage equilibrium, RS explained 100% of the covariance when loci were unlinked ($c = 0.5$), whereas RS and CS contributed equally to the covariance under complete

linkage ($c = 0$). For moderate to high LD between QTL and SNP, the information from LD explained most of the covariance; slightly more when both loci were unlinked ($r^2 > 0$, $c = 0.5$ vs. $c = 0$). The last scenario in Table 1, where $c = 0$ and $r^2 = 1$, should not be misinterpreted as the case in which SNP and QTL are identical, because neither LD nor CS are defined as there is only one locus.

Table 1: Covariance between true and estimated breeding values of a validation individual obtained by deterministic formulas ($Cov(g_i, \hat{g}_i)$) and simulations ($\bar{x}$) based on 1,000,000 replicates, contributions to this covariance by LD, co-segregation (CS), and additive-genetic relationship (RS), both absolute and in percent (%), and accuracy of GEBVs ($\hat{\rho}_{g_i\hat{g}_i}$) for a training data set containing a half sib of the validation individual according to recombination frequency, c, and LD measured as r^2 at a SNP allele frequency of 0.25, resulting in QTL allele frequency p_{QTL}.

r^2	p_{QTL}	c	$Cov(g_i, \hat{g}_i)$	$\bar{x}$ $(s.e.)$	RS (%)	CS (%)	LD (%)	$\hat{\rho}_{g_i\hat{g}_i}$
0	0.5	0.5	0.0268	0.0268 ($7{\cdot}10^{-4}$)	0.0268 (100)	0 (0)	0 (0)	0.04
0	0.5	0.1	0.0439	0.0445 ($7{\cdot}10^{-4}$)	0.0268 (61)	0.0171 (39)	0 (0)	0.07
0	0.5	0.0	0.0536	0.0546 ($7{\cdot}10^{-4}$)	0.0268 (50)	0.0268 (50)	0 (0)	0.08
0.3	0.35	0.5	0.0953	0.0947 ($8{\cdot}10^{-4}$)	0.0268 (28)	0 (0)	0.0685 (72)	0.14
0.3	0.35	0.1	0.1479	0.1476 ($8{\cdot}10^{-4}$)	0.0268 (18)	0.0171 (12)	0.1039 (70)	0.22
0.3	0.35	0.0	0.1691	0.1693 ($8{\cdot}10^{-4}$)	0.0268 (16)	0.0268 (16)	0.1155 (68)	0.25
1	0.25	0.5	0.2571	0.2583 ($9{\cdot}10^{-4}$)	0.0268 (10)	0 (0)	0.2304 (90)	0.38
1	0.25	0.1	0.3943	0.3936 ($1{\cdot}10^{-3}$)	0.0268 (7)	0.0171 (4)	0.3504 (89)	0.59
1	0.25	0	0.4429	0.4449 ($1{\cdot}10^{-3}$)	0.0268 (6)	0.0268 (6)	0.3893 (88)	0.67

Discussions

Deterministic formulas proved that LD, CS, and RS are utilized by Genomic-BLUP. The simple scenario even showed that they contribute additively to the covariance between true and estimated breeding values, but further analyses are required to answer whether this holds under realistic scenarios with many SNPs and training individuals. Applying different recombination frequencies and LD parameters to these formulas revealed that each SNP may have a different pattern of how each type of information contributes to the covariance, depending on map distance to the QTL and extent of LD. For example, in linkage equilibrium and if SNP and QTL are unlinked, RS contributes 100%, whereas if LD is high and loci are linked, LD contributes more than 65% (Table 1). A decreasing recombination frequency increases both CS and LD information, because training

and validation individuals receive the same non-recombinant haplotype with higher probability (Equation 2), which increases not only genetic similarity at the SNP, but also at the QTL.

Extending these formulas to more loci is straightforward, but requires assumptions about the decay of LD, and extending them to more than two training individuals poses difficulties connected to the inverse within BLUP equations. Although a training data set of only one individual seems unrealistic, and is only feasible by replacing BLUP with selection index methodology, these formulas emphasize the notion that genetic covariances originate from LD, CS, and RS. In addition, they set the focus on the informativeness of a single observation used for genomic prediction. The phenotype of a clone of a validation individual, such as an identical twin, inbred, or hybrid, sets the upper bound for the accuracy of GEBVs from a single record; this accuracy equals to the square root of heritability as the phenotype is identical to an own performance of the validation individual.

Definitions for LD, CS, and RS are identical to methods that model them explicitly conditional on pedigree information, map positions, and SNP genotypes (e.g., HABIER *et al.*, 2010). In contrast, the deterministic formulas of this study were derived by averaging over possible SNP genotypes. A comprehensive comparison of Genomic-BLUP with those methods may further improve our understanding of information utilized in genomic prediction.

References

FERNANDO, R. L., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21**: 467–477.

HABIER, D., L. R. TOTIR, and R. L. FERNANDO, 2010 A two-stage approximation for analysis of mixture genetic models in large pedigrees. Genetics **185**: 655–670.

Part 2

Detailed derivations

Genotype scores as random variables

In agreement with definitions of the genetic and statistical models in the main manuscript, the adjusted genotype score of individual i at locus k is

$$G_{i,k} = (S^m_{i,k} - p_k) + (S^p_{i,k} - p_k), \tag{3}$$

where $S^m_{i,k}$ and $S^p_{i,k}$ denote maternal (m) and paternal (p) allele states, respectively, which are treated as Bernoulli random variables. For a clear distinction of adjusted genotype scores and adjusted allele states for QTLs and SNPs, those for QTLs are denoted by

$$W_{i,q} = w^m_{i,q} + w^p_{i,q},$$

whereas those for SNPs are denoted by

$$Z_{i,k} = z^m_{i,k} + z^p_{i,k}.$$

As described in the main manuscript, the genotype scores of QTLs and SNPs from the genetic and statistical model, respectively, are realized values of random processes that start with sampling of founder alleles, and continue with transmitting those alleles from generation to generation down the pedigree. To describe these random processes deterministically, both training and validation individuals are assumed to be randomly drawn conditional on a pedigree. Consequently, genotype scores in $\mathbf{W}$, $\mathbf{Z}$, $\boldsymbol{w}'_i$, and $\boldsymbol{z}'_i$, as described the main manuscript, are random variables, whose joint distribution allows inferring the accuracy of $\hat{g}_i$. This accuracy is defined here as the correlation between true and estimated breeding values of validation individual i as

$$\rho_{g_i\hat{g}_i} = \frac{Cov(g_i, \hat{g}_i)}{\sqrt{Var(g_i)Var(\hat{g}_i)}}. \tag{4}$$

Inference about $\rho_{g_i\hat{g}_i}$ is conditional on genomic positions of loci, founder allele frequencies, LD in founders, and pedigree information. Assumptions about the QTL effects in vector $\boldsymbol{a}$ with mean $\boldsymbol{\mu}_a$ and variance-covariance matrix $\mathbf{V}_a$ are more general in the following derivations than in previous equations. Furthermore, Hardy-Weinberg equilibrium is assumed in these derivations.

Allele origin variables

Co-segregation can be described as statistical dependency between allele origin states at two or more loci on the same gamete of a non-founder. Each origin state for an allele of individual i at locus k, $O_{i,k} \in m, p$, describes from which gamete of the parent it was received, i.e., either from the maternal (m), or from the paternal (p) gamete. Let $Pr(\boldsymbol{O}^x_i) = Pr(O^x_{i,k} = x_k, 1, \ldots, K)$ be the joint probability of origin states at K linked loci on a haplotype of individual i that was received from either the mother ($x = m$) or the father ($x = p$). If those K loci are ordered by their chromosomal positions, this probability can be written as

$$Pr(\boldsymbol{O}^x_i) = Pr(O^x_{i,1} = x_1) \prod_{k=2}^{K} Pr(O^x_{i,k} = x_k | O^x_{i,k-1} = x_{k-1}).$$

The probability of the allele origin for the first locus, $Pr(O^x_{i,1} = x_1)$, is 0.5, expressing equal chance of coming from either the maternal or paternal gamete of the parent. The conditional probability $Pr(O^x_{i,k} = x_k | O^x_{i,k-1} = x_{k-1})$ equals $c_{k,k-1}$ if $x_k \neq x_{k-1}$ and it equals $1 - c_{k,k-1}$ if $x_k = x_{k-1}$, where $c \in [0, 0.5)$ is the recombination frequency between loci k and $k-1$. Note that co-segregation only occurs if $c < 0.5$, which distinguishes it from additive-genetic relationships. The assumptions underlying the last equation are identical to those of Haldane's mapping function.

Covariance between true and estimated breeding values

Genotype scores and allele states are identified in the equation of the statistical method as follows. According to selection index methodology, $\hat{g}_i$ can be calculated by

$$\hat{g}_i = \mathbf{G}_{i-}(\mathbf{G} + \mathbf{I}\lambda)^{-1}(\boldsymbol{y} - 1\boldsymbol{\mu}).$$

As the training data set contains only one individual, this equation becomes

$$\begin{aligned}\hat{g}_i &= \boldsymbol{z}_i'\boldsymbol{z}_j(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}(y - \mu) \\ &= \boldsymbol{z}_i'\boldsymbol{z}_j(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}(\boldsymbol{w}_j'\boldsymbol{a} + e_j),\end{aligned}$$

where $\boldsymbol{z}_i$ is the vector of SNP genotypes for validation individual i, and $\boldsymbol{z}_j$, $\boldsymbol{w}_j$, and e_j denote the vector of SNP genotypes, the vector of QTL genotypes, and the residual effect, respectively, all for training individual j. The covariance between true and estimated breeding values of the validation individual can be written as

$$Cov(g_i, \hat{g}_i) = E(g_i\hat{g}_i) - E(g_i)E(\hat{g}_i).$$

The expected value of the true breeding value, $E(g_i)$, is zero. This is true irrespective of whether the validation individual is founder or non-founder, because every non-founder allele can be traced back to a founder allele, which has expected value of zero after adjusting its allele state by the founder allele frequency. Therefore, the covariance can be calculated by

$$Cov(g_i, \hat{g}_i) = E(g_i\hat{g}_i).$$

Assuming that residual effects are uncorrelated to genotype scores and QTL effects, and replacing g_i and $\hat{g}_i$ by their causal components as defined in the genetic and statistical models, respectively, the expected value of the cross-product can be evaluated as

$$Cov(g_i, \hat{g}_i) = E(\boldsymbol{w}_i'\boldsymbol{a}\boldsymbol{a}'\boldsymbol{w}_j(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}\boldsymbol{z}_j'\boldsymbol{z}_i).$$

QTL effects and genotype scores are assumed independent, hence they can be evaluated separately, which is achieved by introducing the trace function to rotate vector $\boldsymbol{w}_i'$ at the end of the product:

$$\begin{aligned}Cov(g_i, \hat{g}_i) &= E(\text{tr}\{\boldsymbol{a}\boldsymbol{a}'\boldsymbol{w}_j(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}\boldsymbol{z}_j'\boldsymbol{z}_i\boldsymbol{w}_i'\}) \\ &= \text{tr}\{E(\boldsymbol{a}\boldsymbol{a}')E(\boldsymbol{w}_j(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}\boldsymbol{z}_j'\boldsymbol{z}_i\boldsymbol{w}_i')\},\end{aligned}$$

where $E(\boldsymbol{a}\boldsymbol{a}') = \boldsymbol{\mu}_a\boldsymbol{\mu}_a' + \mathbf{V}_a$. Note that $(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}$ and the dot product $\boldsymbol{z}_j'\boldsymbol{z}_j$ are scalars, allowing us to rearrange these terms in the expected value function so that

$$\begin{aligned}Cov(g_i, \hat{g}_i) &= \text{tr}\{(\boldsymbol{\mu}_a\boldsymbol{\mu}_a' + \mathbf{V}_a)E(\boldsymbol{w}_j\boldsymbol{w}_i'(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}\boldsymbol{z}_j'\boldsymbol{z}_i)\} \\ &= \text{tr}\{(\boldsymbol{\mu}_a\boldsymbol{\mu}_a' + \mathbf{V}_a)E(\boldsymbol{w}_j\boldsymbol{w}_i'(\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda)^{-1}\sum_{k=1}^{N_{snp}} Z_{j,k}Z_{i,k})\}.\end{aligned}$$

The vector product $\boldsymbol{w}_j\boldsymbol{w}_i'$ is an $N_{qtl} \times N_{qtl}$ matrix of scalar products between QTL genotypes of training individual j and validation individual i, which gives

$$Cov(g_i, \hat{g}_i) = \text{tr}\Big\{(\boldsymbol{\mu}_a\boldsymbol{\mu}_a' + \mathbf{V}_a)\Big[\sum_{k=1}^{N_{snp}} E\Big(\frac{W_{j,q}Z_{j,k}Z_{i,k}W_{i,r}}{\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda}\Big)\Big]_{qr}\Big\},$$

where q and r denote row and column indices, respectively, of a matrix. Observe the difference between treating QTL effects as fixed or random: if they are random with mean zero and variance-covariance matrix $\mathbf{V}_a = \mathbf{I}\sigma_a^2$, the trace function evaluates only the diagonal elements of the $q \times r$-matrix, i.e., where $q = r$. Genotypes are finally replaced by maternal and paternal allele states adjusted by founder allele frequencies, resulting in

$$Cov(g_i, \hat{g}_i) = \text{tr}\Big\{(\boldsymbol{\mu}_a\boldsymbol{\mu}_a' + \mathbf{V}_a)\Big[\sum_{k=1}^{N_{snp}}\sum_{x_1=1}^{2}\sum_{x_2=1}^{2}\sum_{x_3=1}^{2}\sum_{x_4=1}^{2} E\Big(\frac{w_{j,q}^{x_1}z_{j,k}^{x_2}z_{i,k}^{x_3}w_{i,r}^{x_4}}{\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda}\Big)\Big]_{qr}\Big\}, \tag{5}$$

where x_1, x_2, x_3, and x_4 denote parental origins of alleles; maternal origin is coded either with index 1 or m, and paternal origin either with index 2 or p. The four allele states in the expected value function reveal a connection between a QTL allele in the phenotype of the training individual and a QTL allele in the true breeding value of the validation individual through the SNP alleles of both individuals.

The expected value in the last equation is resolved by tracing non-founder alleles back to founder alleles using allele origin variables, and then averaging over all possible founder allele states:

$$E\Big(\frac{w_{1,q}^{x1}z_{1,j}^{x_2}z_{i,j}^{x_3}w_{i,r}^{x_4}}{\boldsymbol{z}_j'\boldsymbol{z}_j + \lambda}\Big) = \sum_{O} Pr(\boldsymbol{O})\Big[\sum_{\boldsymbol{z}_f}\Big(\frac{E(w_{f_1,q}^{\kappa_1}z_{f_2,j}^{\kappa_2}z_{f_4,j}^{\kappa_3}w_{f_3,r}^{\kappa_4}|\boldsymbol{z}_f)}{\boldsymbol{z}_f'\boldsymbol{z}_f + \lambda}\Big)Pr(\boldsymbol{z}_f|\boldsymbol{O})\Big],$$

where $\boldsymbol{O}$ is a vector of allele origins, $\boldsymbol{z}_f$ is a vector of SNP genotypes of founders, and κ_1, κ_2, κ_3, and κ_4 denote gametes of the founders f_1, f_2, f_3, and f_4, respectively.

Further evaluations for simplified scenarios

A scenario was employed in which the genetic model has only one QTL, the statistical model has only one SNP, and the training data set contains only one individual that can be differently related to the validation individual. Also, without loss of generality for this simple scenario, we assume a fixed QTL effect. Strategies applied to resolve the expected value function of equation (5) are as follows:

1. Identify cases in which at least one of the four alleles in the numerator is independent of all other alleles in both numerator and denominator; these cases can be ignored.

2. Trace back non-founder alleles to founder alleles, considering all possible paths, and calculate the joint probability of all allele origin states involved in each path. Use this probability to weigh each path.

3. Repeat step 1 for founder alleles.

4. Express the product of adjusted QTL and SNP alleles coming from the same gamete as $E(w^x_{i,q}z^x_{i,k}) = D_{q,k}$, if these two alleles are independent of the remaining alleles.

5. Express the product of two adjusted QTL alleles from the same gamete as $E([w^x_{i,k}]^2) = Var(w^x_{i,q})$, if they are independent of the remaining alleles.

6. Use the conditional expectation of an adjusted QTL allele given a SNP allele on the same gamete, $E(w^x_{i,q}|z^x_{i,k})$, which is a function of $D_{q,k}$, if a QTL allele is not independent of alleles in the denominator. Further, if two identical QTL alleles are not independent of alleles in the denominator, evaluate them by $E([w^x_{i,q}]^2|z^x_{i,k})$. These conditional expectations were derived below.

7. Evaluate the remaining SNP alleles of founders by averaging over possible allele states.

Training and validation individuals are founders

Equation (5), which can be written here as

$$Cov(g_i, \hat{g}_i) = a^2 \sum_{x_1=1}^{2}\sum_{x_2=1}^{2}\sum_{x_3=1}^{2}\sum_{x_4=1}^{2} E\Big(\frac{w^{x_1}_{i,q}z^{x_2}_{i,k}w^{x_3}_{j,q}z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big),$$

can be simplified by recognizing that the four founder gametes of the two individuals are independent, resulting in

$$Cov(g_i, \hat{g}_i) = a^2 \sum_{x_1=1}^{2}\sum_{x_2=1}^{2}\sum_{x_3=1}^{2}\sum_{x_4=1}^{2} E(w^{x_1}_{i,q}z^{x_2}_{i,k})E\Big(\frac{w^{x_3}_{j,q}z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big).$$

Further, if QTL and SNP alleles of the validation individual come from different gametes, i.e., $x_1 \neq x_2$, these two alleles are independent, therefore $E(w^{x_1}_{i,q}z^{x_2}_{i,k}) = 0$ and

$$Cov(g_i, \hat{g}_i) = a^2 \sum_{x_1=1}^{2} E(w^{x_1}_{i,q}z^{x_1}_{i,k}) \sum_{x_3=1}^{2}\sum_{x_4=1}^{2} E\Big(\frac{w^{x_3}_{j,q}z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big).$$

As allele states were adjusted by founder allele frequencies, $E(w^{x_1}_{i,q}z^{x_1}_{i,k}) = D_{q,k}$, so that

$$Cov(g_i, \hat{g}_i) = a^2 2D_{q,k} \sum_{x_3=1}^{2}\sum_{x_4=1}^{2} E\Big(\frac{w^{x_3}_{j,q}z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big).$$

Next, the QTL allele of the training individual, $w^{x_3}_{j,q}$, is evaluated separately from the ratio of adjusted SNP allele states such that the LD parameter $D_{q,k}$ is identified; we write

$$Cov(g_i, \hat{g}_i) = a^2 2D_{q,k} \sum_{x_3=1}^{2}\sum_{x_4=1}^{2}\sum_{Z_{j,k}} E(w^{x_3}_{j,q}|z^{x_3}_{j,k}) \frac{z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda} Pr(z^p_{j,k})Pr(z^m_{j,k}),$$

where $z_{j,k}^{x_3}$ in $E(w_{j,q}^{x_3}|z_{j,k}^{x_3})$ is either $z_{j,k}^{m}$ or $z_{j,k}^{p}$, which are both given in the denominator of the ratio. Also, we sum over all possible SNP genotypes of the training individual, $Z_{j,k}$, which contains the adjusted SNP alleles $z_{j,k}^{m}$ and $z_{j,k}^{p}$. The conditional expectation term can be evaluated as

$$\begin{aligned} E(w_{j,q}^{x_3}|z_{j,k}^{x_3}) &= E(S_{j,q}^{x_3} - p_q|S_{j,k}^{x_3} - p_k) \\ &= E(S_{j,q}^{x_3}|S_{j,k}^{x_3}) - p_q \\ &= Pr(S_{j,q}^{x_3}|S_{j,k}^{x_3}) - p_q. \end{aligned}$$

The conditional probability is derived from the definition of LD given by

$$D_{q,k} = Pr(S_{j,q}^{x_3}, S_{j,k}^{x_3}) - Pr(S_{j,q}^{x_3})Pr(S_{j,k}^{x_3}).$$

Dividing both sides of this equation by $Pr(S_{j,k}^{x_3})$ and rearranging gives

$$Pr(S_{j,q}^{x_3}|S_{j,k}^{x_3}) = \frac{D_{q,k}}{Pr(S_{j,k}^{x_3})} + Pr(S_{j,q}^{x_3}).$$

One must now recognize that the absolute value of $D_{q,k}$ is identical irrespective of whether $S_{j,k}^{x_3}$ equals to 0 or 1, but its sign is different for these two SNP allele states. In our derivations the sign is positive for $S_{j,k}^{x_3} = 1$ and negative for $S_{j,k}^{x_3} = 0$. The sign of $D_{q,k}$ must be considered in further derivations, so we define

$$1^{S_{j,k}^{x_3}} = \begin{cases} -1 & S_{j,k}^{x_3} = 0 \\ 1 & S_{j,k}^{x_3} = 1. \end{cases}$$

The conditional expectation can now be written as,

$$E(w_{j,q}^{x_3}|z_{j,k}^{x_3}) = D_{q,k}\frac{1^{S_{j,k}^{x_3}}}{Pr(S_{j,k}^{x_3})},$$

which allows us to separate $D_{q,k}$ from the sum over $Z_{j,k}$, while leaving $\frac{1^{S_{j,k}^{x_3}}}{Pr(S_{j,k}^{x_3})}$ within this sum. The covariance becomes

$$Cov(g_i, \hat{g}_i) = a^2 2D_{q,k}^2 \sum_{x_3=1}^{2}\sum_{x_4=1}^{2}\sum_{Z_{j,k}} \frac{1^{(S_{j,k}^{x_3})}}{Pr(S_{j,k}^{x_3})}\frac{z_{j,k}^{x_4}}{(z_{j,k}^{m} + z_{j,k}^{p})^2 + \lambda} Pr(z_{j,k}^{p})Pr(z_{j,k}^{m}),$$

which is free of QTL alleles, and leaves us with averaging over SNP allele states of the training individual for different origin states of the variables x_3 and x_4. If x_3 = x_4, i.e., the QTL allele, $w_{j,q}^{x_3}$, and the SNP allele on the numerator, $z_{j,k}^{x_4}$, are from the same gamete of the training individual, the last summation term is calculated as shown in Table 1.

Table 2: Averaging over SNP allele states of the training individual when $x_3 = x_4 = m$.

$S^m_{j,k}$	$S^p_{j,k}$	$z^m_{j,k}$	$z^p_{j,k}$	$Pr(z^m_{j,k})$	$Pr(z^p_{j,k})$	$1^{S^m_{j,k}}$	$Pr(S^m_{j,k})$	Terms of $\sum_{Z_{j,k}}$
0	0	$-p_k$	$-p_k$	$1-p_k$	$1-p_k$	-1	$1-p_k$	$\frac{p_k(1-p_k)}{4p_k^2+\lambda}$
0	1	$-p_k$	$1-p_k$	$1-p_k$	p_k	-1	$1-p_k$	$\frac{p_k^2}{(1-2p_k)^2+\lambda}$
1	0	$1-p_k$	$-p_k$	p_k	$1-p_k$	1	p_k	$\frac{(1-p_k)^2}{(1-2p_k)^2+\lambda}$
1	1	$1-p_k$	$1-p_k$	p_k	p_k	1	p_k	$\frac{p_k(1-p_k)}{4(1-p_k)^2+\lambda}$

Terms in Table 1 are a function of the allele frequency at SNP k, p_k, and the shrinkage parameter λ; thus we define the sum of all terms as

$$\begin{aligned} f_1(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^{x_4}_{j,k})}}{Pr(S^{x_4}_{j,k})} \frac{z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda} Pr(z^p_{j,k}) Pr(z^m_{j,k}) \\ &= \frac{p_k(1-p_k)}{4p_k^2+\lambda} + \frac{1-2p_k(1-p_k)}{(1-2p_k)^2+\lambda} + \frac{p_k(1-p_k)}{4(1-p_k)^2+\lambda}, \end{aligned}$$

If $x_3 \neq x_4$, i.e., the QTL allele and the SNP allele on the numerator come from different gametes of the training individual,

$$\begin{aligned} f_2(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^{x_3}_{j,k})}}{Pr(S^{x_3}_{j,k})} \frac{z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda} Pr(z^p_{j,k}) Pr(z^m_{j,k}) \\ &= \frac{p_k(1-p_k)}{4p_k^2+\lambda} - \frac{2p_k(1-p_k)}{(1-2p_k)^2+\lambda} + \frac{p_k(1-p_k)}{4(1-p_k)^2+\lambda}. \end{aligned}$$

As a result we obtain

$$\begin{aligned} Cov(g_i,\hat{g}_i) &= 4a^2 D^2_{q,k}[f_1(p_k,\lambda)+f_2(p_k,\lambda)] \\ &= 4a^2 r^2_{q,k} Var(w^x_{i,q}) Var(z^x_{i,k}) f_3(p_k,\lambda) \end{aligned}$$

where $r^2_{q,k}$ is the usual r^2 measure for LD between QTL q and SNP k, $Var(w^x_{i,q}) = p_q(1-p_q)$ and $Var(z^x_{i,k}) = p_k(1-p_k)$ are variances of founder allele states at the QTL and SNP, respectively, x denotes either maternal or paternal origin, and

$$\begin{aligned} f_3(p_k,\lambda) &= f_1(p_k,\lambda) + f_2(p_k,\lambda) \\ &= \frac{2p_k(1-p_k)}{4p_k^2+\lambda} + \frac{1-4p_k(1-p_k)}{(1-2p_k)^2+\lambda} + \frac{2p_k(1-p_k)}{4(1-p_k)^2+\lambda}. \end{aligned}$$

Training and validation individuals are half sibs

Consider two half sibs descending from a common sire and unknown dams, where the sire is assumed to be a founder. One half sib is used in training, while the other one is in validation. The two maternal gametes of these half sibs are considered independent founder gametes; thus, they only contribute to the covariance between true and estimated breeding values of the validation individual if there is LD between QTL and SNP.

a) Linkage equilibrium

Knowing that the covariance can only come through the paternal alleles, both QTL and SNP alleles of the validation individual and the QTL allele of the training individual must have paternal origin; this is $x_1 = x_2 = x_3 = p$ in equation (5). The two QTL alleles must be paternal so that they trace back onto the same gamete of the sire, and the maternal SNP allele of the validation individual is independent to all other alleles in that equation. The SNP allele of the training individual, however, is contained in both numerator and denominator of that equation, having either maternal or paternal origin, and therefore must be evaluated by averaging over maternal SNP allele states of the training individual and SNP allele states of the sire. Consequently, equation (5) becomes,

$$Cov(g_i, \hat{g}_i) = a^2 \sum_{x_4=1}^{2} E\Big(\frac{w_{i,q}^{p} w_{j,q}^{p} z_{i,k}^{p} z_{j,k}^{x_4}}{(z_{j,k}^{m} + z_{j,k}^{p})^2 + \lambda}\Big).$$

This equation is further evaluated by tracing all four paternal alleles of the two half sibs back to the common sire. Among the 16 possible combinations, only four contribute to the covariance between true and estimated breeding values, which are those where both the QTL alleles and the SNP alleles are *ibd*. This occurs either if the same gamete is transmitted without recombination from the sire to both half sibs (Figure 1), or if both half sibs receive the same recombinant gamete (Figure 2). The first case has probability $0.25(1-c)^2$, and the second case has probability $0.25c^2$, where c is the recombination frequency. In the following equation, the paternal alleles of training and validation individuals are replaced by the maternal and paternal alleles of sire, s. In terms 1 and 3 of the following equation, training and validation individuals received the same non-recombinant maternal and paternal gamete of the sire, respectively, and in terms 2 and 4 both

individuals received the same recombinant gamete.

$$
\begin{aligned}
Cov(g_i, \hat{g}_i) = a^2 \Big[& 0.25(1-c)^2 E([w^m_{s,q}]^2) \sum_{Z_{j,k}} \frac{z^m_{s,k} z^m_j}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda} Pr(z^m_{s,k}) Pr(z^m_{j,k}) \\
& + 0.25c^2 E([w^m_{s,q}]^2) \sum_{Z_{j,k}} \frac{z^p_{s,k} z^m_j}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda} Pr(z^p_{s,k}) Pr(z^m_{j,k}) \\
& + 0.25(1-c)^2 E([w^p_{s,q}]^2) \sum_{Z_{j,k}} \frac{[z^p_{s,k}]^2}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda} Pr(z^p_{s,k}) Pr(z^m_{j,k}) \\
& + 0.25c^2 E([w^p_{s,q}]^2) \sum_{Z_{j,k}} \frac{[z^m_{s,k}]^2}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda} Pr(z^m_{s,k}) Pr(z^m_{j,k}) \Big].
\end{aligned}
$$

The two QTL alleles of the sire were separated from the ratio because there is no LD, meaning that QTL and SNP alleles from the same founder gamete are independent. Adding up terms and replacing $E([w^x_{s,q}]^2)$ with $Var(w^x_{s,q})$, where x denotes one of the two sire gametes,

$$
\begin{aligned}
Cov(g_i, \hat{g}_i) &= a^2 0.25[2(1-c)^2 + 2c^2] Var(w^x_{s,k}) \\
&\Big[\sum_{Z_{j,k}} \frac{[z^x_{s,k}]^2}{(z^m_{j,k} + z^x_{s,k})^2 + \lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\
&+ \sum_{Z_{j,k}} \frac{z^x_{s,k} z^m_j}{(z^m_{j,k} + z^x_{s,k})^2 + \lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \Big] \\
&= a^2 0.25[2(1-c)^2 + 2c^2] Var(w^x_{s,q}) [f_4(p_k, \lambda) + f_5(p_k, \lambda)],
\end{aligned}
$$

where

$$
\begin{aligned}
f_4(p_k, \lambda) &= \sum_{Z_{j,k}} \frac{[z^x_{s,k}]^2}{(z^m_{j,k} + z^x_{s,k})^2 + \lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\
&= Var(z^x_{s,k}) \Big[\frac{p_k(1-p_k)}{4p_k^2 + \lambda} + \frac{1 - 2p_k(1-p_k)}{(1-2p_k)^2 + \lambda} + \frac{p_k(1-p_k)}{4(1-p_k) + \lambda} \Big] \\
&= Var(z^x_{s,k}) f_1(p_k, \lambda),
\end{aligned}
$$

with $Var(z^x_{s,k}) = p_k(1-p_k)$ and

$$
\begin{aligned}
f_5(p_k, \lambda) &= \sum_{Z_{j,k}} \frac{z^x_{s,k} z^m_j}{(z^m_{j,k} + z^x_{s,k})^2 + \lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \Big] \\
&= Var(z^x_{s,k}) \Big[\frac{p_k(1-p_k)}{4p_k^2 + \lambda} - \frac{2p_k(1-p_k)}{(1-2p_k)^2 + \lambda} + \frac{p_k(1-p_k)}{4(1-p_k)^2 + \lambda} \Big] \\
&= Var(z^x_{s,k}) f_2(p_k, \lambda).
\end{aligned}
$$

In summary, using $f_3(p_k, \lambda) = f_1(p_k, \lambda) + f_2(p_k, \lambda)$, we obtain

$$Cov(g_i, \hat{g}_i) = a^2 0.25[2(1-c)^2 + 2c^2] Var(w^x_{s,q}) Var(z^x_{s,k}) f_3(p_k, \lambda).$$

Note the resemblance to the equation when training and validation individuals are founders.

b) Linkage disequilibrium

Equation (5) can be simplified by recognizing that QTL and SNP alleles of the validation individual must come from the same parental gamete, i.e., $x_1 = x_2$, because any combination where one allele is maternal and the other one is paternal leaves at least one allele of the validation individual independent of all other alleles in the covariance function. Hence, equation (5) becomes

$$Cov(g_i, \hat{g}_i) = a^2 \sum_{x_1=1}^{2} \sum_{x_3=1}^{2} \sum_{x_4=1}^{2} E\Big(\frac{w^{x_1}_{i,q} z^{x_1}_{i,k} w^{x_3}_{j,q} z^{x_4}_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big).$$

Three components of that covariance are derived separately, where the first two involve neither RS nor CS, because not all alleles have paternal origin. In the first component, QTL and SNP alleles of the validation individual come from the maternal (founder) gamete, which is independent of the alleles from the training individual; we obtain

$$\begin{aligned} c_1 &= a^2 E(w^m_{i,q} z^m_{i,k}) \sum_{x_3=1}^{2} \sum_{x_4=1}^{2} E\Big(\frac{w^{x_3}_{j,q} z^{x_4}_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big) \\ &= a^2 D_{q,k} \Big[E\Big(\frac{w^m_{j,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big) + E\Big(\frac{w^m_{j,q} z^p_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big) \\ &+ E\Big(\frac{w^p_{j,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big) + E\Big(\frac{w^p_{j,q} z^p_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda} \Big) \Big]. \end{aligned}$$

The paternal allele of training individual j, $z^p_{j,k}$, in the first two terms within the bracket of the last equation is replaced by either the maternal or paternal allele of sire, s. Each of these two cases has probability 0.5. In the last two terms of the last equation, the training individual must have received one of the two non-recombinant gametes from the sire with probability $1 - c$, because otherwise the QTL allele is independent of all other alleles. After replacing all paternal alleles of the training individuals by sire alleles, the first two terms can be evaluated as shown above for the scenario where training and validation individuals were founders. Consequently, these four terms can be evaluated to

$$\begin{aligned} c_1 &= a^2 D_{q,k} \Big[D_{q,k} f_1(p_k, \lambda) + D_{q,k} f_2(p_k, \lambda) + D_{q,k}(1-c) f_2(p_k, \lambda) + D_{q,k}(1-c) f_1(p_k, \lambda) \Big] \\ &= a^2 D^2_{q,k} (2-c) f_3(p_k, \lambda). \end{aligned}$$

To derive the second component, we consider cases in equation (5) that are restricted to the paternal alleles of validation individual i and the maternal QTL allele of training individual j:

$$\begin{aligned} c_2 &= a^2 \sum_{x_4=1}^{2} E\Big(\frac{w^p_{i,q} z^p_{i,k} w^m_{j,q} z^{x_4}_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big) \\ &= a^2\Big[E\Big(\frac{w^p_{i,q} z^p_{i,k} w^m_{j,q} z^m_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big) + E\Big(\frac{w^p_{i,q} z^p_{i,k} w^m_{j,q} z^p_{j,k}}{(z^m_{j,k}+z^p_{j,k})^2+\lambda}\Big)\Big]. \end{aligned}$$

The paternal alleles are replaced by alleles of the sire, where each of the two terms in the equation above has four cases, each with probability $0.25(1-c)$. It results from sampling a non-recombinant paternal gamete for the validation individual with probability $0.5(1-c)$ and a paternal allele for the training individual from either the maternal or paternal QTL allele of the sire with probability 0.5. Again, the validation individual must have received a non-recombinant gamete to exploit LD information, or otherwise its alleles are independent from other alleles. Thus,

$$\begin{aligned} c_2 = a^2 0.25(1-c)\Big[&E\Big(\frac{w^m_{s,q} z^m_{s,k} w^m_{j,q} z^m_{j,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big) + E\Big(\frac{w^p_{s,q} z^p_{s,k} w^m_{j,q} z^m_{j,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big) + \\ &E(w^m_{s,q} z^m_{s,k}) E\Big(\frac{w^m_{j,q} z^m_{j,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big) + E(w^p_{s,q} z^p_{s,k}) E\Big(\frac{w^m_{j,q} z^m_{j,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big) + \\ &E\Big(\frac{w^m_{s,q} z^m_{s,k} w^m_{j,q} z^m_{s,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big) + E(w^p_{s,q} z^p_{s,k}) E\Big(\frac{w^m_{j,q} z^m_{s,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big) + \\ &E(w^m_{s,q} z^m_{s,k}) E\Big(\frac{w^m_{j,q} z^p_{s,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big) + E\Big(\frac{w^p_{s,q} z^p_{s,k} w^m_{j,q} z^p_{s,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big)\Big] \\ = a^2 0.25(1-c) D^2_{q,k}[&f_6(p_k,\lambda) + f_6(p_k,\lambda) + f_1(p_k,\lambda) + f_1(p_k,\lambda) \\ &+ f_7(p_k,\lambda) + f_2(p_k,\lambda) + f_2(p_k,\lambda) + f_7(p_k,\lambda)] \\ = a^2 0.5(1-r) D^2_{q,k}[&f_3(p_k,\lambda) + f_6(p_k,\lambda) + f_7(p_k,\lambda)], \end{aligned}$$

where

$$\begin{aligned} f_6(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^x_{s,k})}}{Pr(S^x_{s,k})} \frac{1^{(S^m_{j,k})}}{Pr(S^m_{j,k})} \frac{z^x_{s,k} z^m_{j,k}}{(z^m_{j,k}+z^x_{s,k})^2+\lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\ &= \frac{p_k^2}{4p_k^2+\lambda} + \frac{2p_k(1-p_k)}{(1-2p_k)^2+\lambda} + \frac{(1-p_k)^2}{4(1-p_k)^2+\lambda} \end{aligned}$$

and

$$\begin{aligned} f_7(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^x_{s,k})}}{Pr(S^x_{s,k})} \frac{1^{(S^m_{j,k})}}{Pr(S^m_{j,k})} \frac{[z^x_{s,k}]^2}{(z^m_{j,k}+z^x_{s,k})^2+\lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\ &= \frac{p_k^2}{4p_k^2+\lambda} - \frac{p_k^2+(1-p_k)^2}{(1-2p_k)^2+\lambda} + \frac{(1-p_k)^2}{4(1-p_k)^2+\lambda}. \end{aligned}$$

For the third component, we derive the formula for the case when QTL alleles of both individuals and the SNP allele of the validation individual have paternal origin:

$$
\begin{aligned}
c_3 &= a^2 \sum_{x_4=1}^{2} E\Big(\frac{w^p_{i,q} z^p_{i,k} w^p_{j,q} z^{x_4}_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda}\Big) \\
&= a^2 \Big[E\Big(\frac{w^p_{i,q} z^p_{i,k} w^p_{j,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda}\Big) + E\Big(\frac{w^p_{i,q} z^p_{i,k} w^p_{j,q} z^p_{j,k}}{(z^m_{j,k} + z^p_{j,k})^2 + \lambda}\Big)\Big].
\end{aligned}
$$

The paternal alleles are replaced by alleles of the sire multiplied by their corresponding origin probabilities, which depend on whether the recombinant or non-recombinant gametes were transmitted to the two half sibs. There are ten cases: all four possible cases in which both individuals received non-recombinant gametes with probability of $0.25(1-c)^2$; all four possible case in which both individuals received recombinant gametes with probability of $0.25c^2$, and two out of four possible cases in which the validation individual received the non-recombinant gamete and the training individual the recombinant gamete with probability of $0.25(1-c)c$. The following equation contains these ten cases, where each of them has two terms, because $z^{x_4}_{j,k}$ can either be $z^m_{j,k}$ or $z^p_{j,k}$. The numerator of the first term reveals the type of gametes received from the sire for both individuals as the first two alleles belong to the validation individual and the last two to the training individual. It follows

$$
\begin{aligned}
c_3 = a^2 0.25 \Big[&(1-c)^2 [E\Big(\frac{w^m_{s,q} z^m_{s,k} w^m_{s,q} z^m_{s,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^m_{s,q} z^m_{s,k} w^m_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big)] + \\
&(1-c)^2 [E\Big(\frac{w^p_{s,q} z^p_{s,k} w^p_{s,q} z^p_{s,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^p_{s,q} z^p_{s,k} w^p_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big)] + \\
&(1-c)^2 [E\Big(\frac{w^m_{s,q} z^m_{s,k} w^p_{s,q} z^p_{s,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^m_{s,q} z^m_{s,k} w^p_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big)] + \\
&(1-c)^2 [E\Big(\frac{w^p_{s,q} z^p_{s,k} w^m_{s,q} z^m_{s,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^p_{s,q} z^p_{s,k} w^m_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big)] + \\
&c^2 [E\Big(\frac{w^m_{s,q} z^p_{s,k} w^m_{s,q} z^p_{s,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^m_{s,q} z^p_{s,k} w^m_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big)] + \\
&c^2 [E\Big(\frac{w^p_{s,q} z^m_{s,k} w^p_{s,q} z^m_{s,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^p_{s,q} z^m_{s,k} w^p_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big)] + \\
&c^2 [E\Big(\frac{w^m_{s,q} z^p_{s,k} w^p_{s,q} z^m_{s,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^m_{s,q} z^p_{s,k} w^p_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^m_{s,k})^2 + \lambda}\Big)] + \\
&c^2 [E\Big(\frac{w^p_{s,q} z^m_{s,k} w^m_{s,q} z^p_{s,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big) + E\Big(\frac{w^p_{s,q} z^m_{s,k} w^m_{s,q} z^m_{j,k}}{(z^m_{j,k} + z^p_{s,k})^2 + \lambda}\Big)] +
\end{aligned}
\tag{6}
$$

$$(1-c)c[E\Big(\frac{w^m_{s,q}z^m_{s,k}w^m_{s,q}z^p_{s,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big)+E\Big(\frac{w^m_{s,q}z^m_{s,k}w^m_{s,q}z^m_{j,k}}{(z^m_{j,k}+z^p_{s,k})^2+\lambda}\Big)]+$$
$$(1-c)c[E\Big(\frac{w^p_{s,q}z^p_{s,k}w^p_{s,q}z^m_{s,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big)+E\Big(\frac{w^p_{s,q}z^p_{s,k}w^p_{s,q}z^m_{j,k}}{(z^m_{j,k}+z^m_{s,k})^2+\lambda}\Big)].$$

The following result is used to resolve the expected value functions in terms 1 to 4 of the equation above:

$$\begin{aligned}
E([w^{x_1}_{i,q}]^2|z^{x_1}_{i,k}) &= E([S^{x_1}_{i,q}-p_q]^2|S^{x_1}_{i,k}-p_k)\\
&= E([S^{x_1}_{i,q}]^2|S^{x_1}_{i,k}) - 2E(S^{x_1}_{i,q}|S^{x_1}_{i,k})p_q + p_q^2\\
&= E(S^{x_1}_{i,q}|S^{x_1}_{i,k}) - 2E(S^{x_1}_{i,q}|S^{x_1}_{i,k})p_q + p_q^2\\
&= E(S^{x_1}_{i,q}|S^{x_1}_{i,k})(1-2p_q) + p_q^2\\
&= Pr(S^{x_1}_{i,q}|S^{x_1}_{i,k})(1-2p_q) + p_q^2\\
&= (\frac{D_{q,k}\cdot 1^{S^{x_1}_{i,k}}}{Pr(S^{x_1}_{i,k})} + p_q)(1-2p_q) + p_q^2\\
&= \frac{D_{q,k}\cdot 1^{S^{x_1}_{i,k}}}{Pr(S^{x_1}_{i,k})}(1-2p_q) + Var(w^{x_1}_{s,q}).
\end{aligned}$$

For terms 19 to 22 we use

$$\begin{aligned}
E([w^{x_1}_{i,q}]^2 z^{x_1}_{i,k}) &= E([S^{x_1}_{i,q}-p_q]^2[S^{x_1}_{i,k}-p_k])\\
&= D_{q,k}(1-2p_w).
\end{aligned}$$

Equation (6) evaluates to

$$\begin{aligned}
c_3 = a^2 0.25\Big[\;&\\
&(1-c)^2[D_{q,k}(1-2p_q)[f_8(p_k,\lambda)+f_9(p_k,\lambda)] + Var(w^x_{s,q})[f_4(p_k,\lambda)+f_5(p_k,\lambda)]+\\
&(1-c)^2[D_{q,k}(1-2p_q)[f_8(p_k,\lambda)+f_9(p_k,\lambda)] + Var(w^x_{s,q})[f_4(p_k,\lambda)+f_5(p_k,\lambda)]+\\
&(1-c)^2 D^2_{q,k}[f_1(p_k,\lambda)+f_2(p_k,\lambda)]+\\
&(1-c)^2 D^2_{q,k}[f_1(p_k,\lambda)+f_2(p_k,\lambda)]+\\
&c^2 Var(w^x_{s,q})[f_4(p_k,\lambda)+f_5(p_k,\lambda)]+\\
&c^2 Var(w^x_{s,q})[f_4(p_k,\lambda)+f_5(p_k,\lambda)]+\\
&c^2 D^2_{q,k}[f_1(p_k,\lambda)+f_2(p_k,\lambda)]+\\
&c^2 D^2_{q,k}[f_1(p_k,\lambda)+f_2(p_k,\lambda)]+\\
&2(1-c)cD_{q,k}(1-2p_q)f_{10}(p_k,\lambda)+\\
&2(1-c)cD_{q,k}(1-2p_q)f_{10}(p_k,\lambda)\\
&\Big]
\end{aligned}$$

where

$$
\begin{aligned}
f_8(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^x_{s,k})}}{Pr(S^x_{s,k})} \frac{[z^x_{s,k}]^2}{(z^m_{j,k}+z^x_{s,k})^2+\lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\
&= \frac{-p_k^2(1-p_k)}{4p_k^2+\lambda} + \frac{(1-p_k)^3-p_k^3}{(1-2p_k)^2+\lambda} + \frac{p_k(1-p_k)^2}{4(1-p_k)^2+\lambda}, \\
f_9(p_k,\lambda) &= \sum_{Z_{j,k}} \frac{1^{(S^x_{s,k})}}{Pr(S^x_{s,k})} \frac{z^x_{s,k} z^m_{j,k}}{(z^m_{j,k}+z^x_{s,k})^2+\lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\
&= p_k(1-p_k)\Big[\frac{-p_k}{4p_k^2+\lambda} + \frac{2p_k-1}{(1-2p_k)^2+\lambda} + \frac{1-p_k}{4(1-p_k)^2+\lambda}\Big], \\
f_{10}(p,\lambda) &= \sum_{Z_{j,k}} \frac{z^x_{s,k}}{(z^m_{j,k}+z^x_{s,k})^2+\lambda} Pr(z^x_{s,k}) Pr(z^m_{j,k}) \\
&= p_k(1-p_k)\Big[\frac{p_k-1}{4p_k^2+\lambda} + \frac{1-2p_k}{(1-2p_k)^2+\lambda} + \frac{p_k}{4(1-p_k)^2+\lambda}\Big].
\end{aligned}
$$

Now we can set $f_1(p_k,\lambda)+f_2(p_k,\lambda) = f_3(p_k,\lambda)$ and $f_4(p_k,\lambda)+f_5(p_k,\lambda) = Var(z_f) f_3(p_k,\lambda)$, and add up terms, which gives

$$
\begin{aligned}
c_3 = a^2 0.25\Big[&[2(1-c)^2+2c^2][Var(w^x_{s,q})Var(z_f)+D^2_{q,k}] f_3(p_k,\lambda) \\
&2(1-c)^2 D_{q,k}(1-2p_q)[f_8(p_k,\lambda)+f_9(p_k,\lambda)]+ \\
&4(1-c)cD_{q,k}(1-2p_q) f_{10}(p_k,\lambda)\Big].
\end{aligned}
$$

In summary, i.e., $c_1+c_2+c_3$, we obtain

$$
\begin{aligned}
Cov(g_i,\hat{g}_i) = a^2 0.25\Big[&[2(1-c)^2+2c^2][Var(w^x_{s,q})Var(z^x_{s,k})+D^2_{q,k}] f_3(p_k,\lambda) \\
&2(1-c)^2 D_{q,k}(1-2p_q)[f_8(p_k,\lambda)+f_9(p_k,\lambda)]+ \\
&4(1-c)cD_{q,k}(1-2p_q) f_{10}(p_k,\lambda)\Big]+ \\
&a^2 D^2_{q,k}(2-r) f_3(p_k,\lambda) + a^2 0.5(1-r) D^2_{q,k}[f_3(p_k,\lambda)+f_6(p_k,\lambda)+f_7(p_k,\lambda)].
\end{aligned}
$$

File S2
Maize chromosome data provided by DuPont Pioneer

File S3
Dairy Cattle chromosome data provided by G. Wiggans

File S4
r file

Files S2-S4 are available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152207/-/DC1.