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.

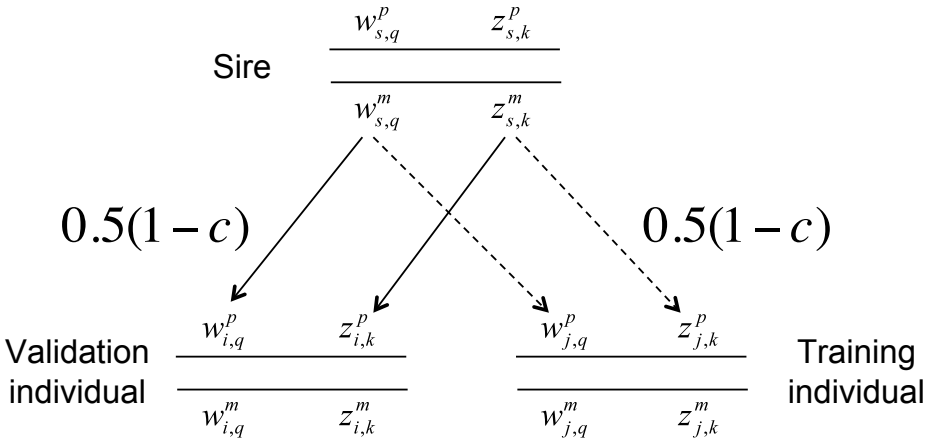
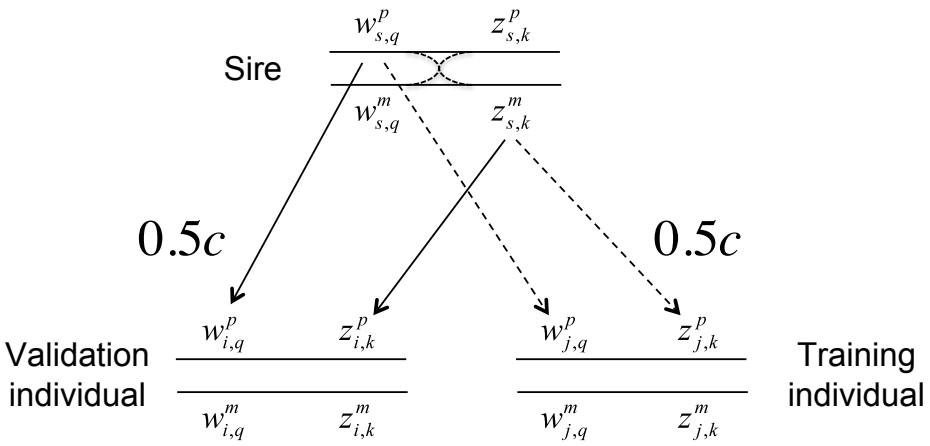


Figure 2: Training and validation individuals received the same recombinant gamete.



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

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