Bonk *et al. Genet Sel Evol (2016) 48:36*
DOI 10.1186/s12711-016-0214-0

Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Mendelian sampling covariability of marker effects and genetic values

Sarah Bonk[1], Manuela Reichelt[1], Friedrich Teuscher[1], Dierck Segelke[2] and Norbert Reinsch[1*]

Abstract

Background: Measures of the expected genetic variability among full-sibs are of practical relevance, such as in the context of mating decisions. An important application field in animal and plant breeding is the selection and allocation of mates when large or small amounts of genetic variability among offspring are desired, depending on user-specific goals. Estimates of the Mendelian sampling variance can be obtained by simulating gametes from parents with known diplotypes. Knowledge of recombination rates and additive marker effects is also required. In this study, we aimed at developing an exact method that can account for both additive and dominance effects.

Results: We derived parent-specific covariance matrices that exactly quantify the within-family (co-)variability of additive and dominance marker effects. These matrices incorporate prior knowledge of the parental diplotypes and recombination rates. When combined with additive marker effects, they allow the exact derivation of the Mendelian sampling (co-)variances of (estimated) breeding values for several traits, as well for the aggregate genotype. A comparative analysis demonstrated good average agreement between the exact values and the simulation results for a practical dataset (74,353 German Holstein cattle).

Conclusions: The newly derived method is suitable for calculating the exact amount of intra-family variation of the estimated breeding values and genetic values (comprising additive and dominance effects).

Background

The degree of genetic variability among full-sibs is known as Mendelian sampling variance. This variability is due to the inheritance of random samples of alleles from both parents. For a quantitative trait, the amount of this variability depends on the parental degree of heterozygosity, $1 - F_{♂}$, where $F_{♂}$ ($F_{♀}$) is the inbreeding coefficient of an individual's sire (dam), which is derived from the pedigree. Under additivity and with unlinked loci, the Mendelian sampling variance is the sum of two parental contributions, $\frac{1}{4}\sigma_a^2(1 - F_{♂}) + \frac{1}{4}\sigma_a^2(1 - F_{♀})$, where σ_a^2 is the additive genetic variance [1]. The latter expression is of general importance in quantitative genetics, especially in the context of estimating genetic parameters and in genetic evaluations. In certain models (e.g. [2, 3]), it is used explicitly for the relative weighting of observations from progeny of inbred versus non-inbred parents. Moreover, the inverse Mendelian sampling variance plays a pivotal role in direct inversion of the numerator relationship matrix [4].

New methods to track Mendelian sampling variance are based on the availability of phased genotypes (diplotypes) at genetic markers across the genome as a byproduct of genomic selection (e.g. [5, 6]). Single nucleotide polymorphism (SNP) diplotypes of parents differ in terms of three features: the degree of heterozygosity, the genotypes at homozygous loci, and the linkage phase between loci. All of these features have consequences for the variability of gametes that are generated by a particular individual, and thus for the variance among the progeny in a family. A small within-family genetic variation contributes to phenotypic uniformity, which is desired e.g. for birth-weight of piglets (e.g. [7]), while a large Mendelian sampling variability may increase selection opportunity between sibs [6]. When phased genotypes are available, it is possible to simulate a large sample of the population of gametes of a selection candidate by

*Correspondence: reinsch@fbn-dummerstorf.de
[1] Institute of Genetics and Biometry, Leibniz-Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
Full list of author information is available at the end of the article

considering recombination within chromosomes, as was demonstrated in a study on 58,035 Holsteins [5]. However, to the best of our knowledge, exact formulae for calculating the Mendelian sampling variance from phased genotypes have not been reported previously.

In this study, we provide the requisite formulae for the exact calculation of within-family genetic variation. The within-family covariance matrix between the additive and dominance effects of all markers can be derived exactly from phased SNP-genotypes and the known genetic distances between markers. Conversion to the within-family variance of (estimated) additive and dominance values for a trait is then achieved via the (estimated) additive and dominance marker effects. We provide a comparison between the results obtained by simulations and the exact method, as well as a brief discussion of the application of this method to the allocation of mates.

Methods

In the following, a breeding population is assumed, where phased SNP-genotypes are available for all potential mating partners, as well as estimates of the SNP-effects for all traits. Furthermore, it is assumed that the genetic distances between all markers are known in terms of their recombination rates, which are summarized in a comprehensive genetic map for all SNPs.

In the first part of this section, we only consider additive marker effects that contribute to the genomic breeding value of an individual:

$$g = \mathbf{c}'\mathbf{m}, \tag{1}$$

where $\mathbf{c}'$ is a row vector of genotype indicators (see Eq. 3) and $\mathbf{m}$ is a vector of marker effects. We show that the within-family covariance matrix $\mathbf{\Omega}$ of $\mathbf{c}$, with dimensions equal to the number of marker effects, can be expressed as the sum of two parent-specific covariance matrices $\mathbf{\Omega}^{♂}$ and $\mathbf{\Omega}^{♀}$. They define independent parental contributions to the Mendelian sampling variance of genomic breeding values of a trait:

$$\sigma_g^2 = \mathbf{m}'\mathbf{\Omega}\mathbf{m}. \tag{2}$$

Subsequently we demonstrate that this additive property of the covariance matrix vanishes when dominance marker effects are included.

Definitions and the case of pure additivity

Throughout this study, we exploit the fact that correlations between marker effects are equivalent to correlations between genotype indicators. For the additive effect a (half the average phenotypic difference between homozygotes) at a two-allelic (A, B) locus i, the genotype indicator $c_{a,i}$ is given by:

$$c_{a,i} = \begin{cases} 1, & \text{for genotype } AA \\ 0, & \text{for genotypes } AB \text{ or } BA \\ -1, & \text{for genotype } BB \end{cases}, \tag{3}$$

and for dominance effects by:

$$c_{d,i} = \begin{cases} -1, & \text{for genotypes } AA \text{ or } BB \\ 1, & \text{for genotypes } AB \text{ or } BA \end{cases}, \tag{4}$$

for $1 \le i \le n$, where n is the number of markers. See the discussion for a treatment of the coding of additive and dominance effects. Derivation of the covariance is based on the linkage disequilibrium (LD) among all the gametes produced by a parent with a particular diplotype.

A parent has one of the 16 possible diplotypes when two bi-allelic markers are involved, which are represented in Table 1. Pairs of additive and dominance genotype indicators are given in columns 2 to 5. The gametes generated by the parent (comprising one allele at the first locus and one allele at the second locus) follow a probability distribution that depends on the diplotype of the parent. Columns 6 to 9 shows the probabilities of gametes, to which we apply the concept of LD. For each parent, LD can be determined as:

$$D_{i,j} = p_{A-A}p_{B-B} - p_{A-B}p_{B-A}, \tag{5}$$

where the lower index indicates gametes (see Table 1, column 10).

In Table 1 the probability of the appearance of allele A at the ith locus is denoted by p_i. Note that this definition applies to the diplotypes as well as to the gametes of the parent. The values for p_i and p_j are given in columns 11 and 12 of Table 1. All entries in Table 1 apply to both the sire (upper index ♂) and the dam (upper index ♀).

The joint distribution of genotypes at two bi-allelic marker loci among offspring is outlined in Table 2, using the frequencies of the parental gametes from Table 1. All nine two-locus genotypes are enumerated, together with their underlying diplotypes (the upper haplotype is paternal; columns 1–3). The probability of each ordered diplotype is the product of the two gametic probabilities, and the probability of a two-locus genotype is the sum of the probabilities for all of its possible underlying diplotypes (Table 2, last column). In the next step, the sex-specific probabilities (indexed as ♂ or ♀) of the parental gametes are expressed as functions of the parent-specific LD-parameters and allele frequencies:

$$p_{A-A}^{♂} = D_{i,j}^{♂} + p_i^{♂}p_j^{♂}, \tag{6}$$

$$p_{A-B}^{♂} = -D_{i,j}^{♂} + p_i^{♂}\left(1 - p_j^{♂}\right), \tag{7}$$

$$p_{B-A}^{♂} = -D_{i,j}^{♂} + \left(1 - p_i^{♂}\right)p_j^{♂}, \tag{8}$$

Table 1 Ten classes of parental diplotypes with different two-locus genotypes and distributions of produced gametes

Parental diplotype	Genotype indicators				Probabilities of gametes				Characterizing parameters		
	$c_{a,i}$	$c_{a,j}$	$c_{d,i}$	$c_{d,j}$	p_{A-A}	p_{A-B}	p_{B-A}	p_{B-B}	$D_{i,j}$	p_i	p_j
$\frac{A-A}{A-A}$	1	1	−1	−1	1	0	0	0	0	1	1
$\frac{A-A}{A-B}/\frac{A-B}{A-A}$	1	0	−1	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	1	$\frac{1}{2}$
$\frac{A-A}{B-A}/\frac{B-A}{A-A}$	0	1	1	−1	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	1
$\frac{A-B}{A-B}$	1	−1	−1	−1	0	1	0	0	0	1	0
$\frac{B-A}{B-A}$	−1	1	−1	−1	0	0	1	0	0	0	1
$\frac{A-A}{B-B}/\frac{B-B}{A-A}$	0	0	1	1	$\frac{1-\theta_{i,j}}{2}$	$\frac{\theta_{i,j}}{2}$	$\frac{\theta_{i,j}}{2}$	$\frac{1-\theta_{i,j}}{2}$	$\frac{1-2\theta_{i,j}}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
$\frac{A-B}{B-A}/\frac{B-A}{A-B}$	0	0	1	1	$\frac{\theta_{i,j}}{2}$	$\frac{1-\theta_{i,j}}{2}$	$\frac{1-\theta_{i,j}}{2}$	$\frac{\theta_{i,j}}{2}$	$-\frac{1-2\theta_{i,j}}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
$\frac{A-B}{B-B}/\frac{B-B}{A-B}$	0	−1	1	−1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0
$\frac{B-A}{B-B}/\frac{B-B}{B-A}$	−1	0	−1	1	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$
$\frac{B-B}{B-B}$	−1	−1	−1	−1	0	0	0	1	0	0	0

Different diplotypes with identical genotypes at markers *i* and *j* are separated by a slash. Genotype indicators are given for additive ($c_{a,i}$, $c_{a,j}$) and dominance ($c_{d,i}$, $c_{d,j}$) effects at both parental marker genotypes. The probabilities of gametes are specific for each diplotype class and they can be summarized by three characteristic parameters: LD $D_{i,j}$ and the probabilities p_i, p_j for an *A*-allele at locus *i* or *j*. $\theta_{i,j}$ is the recombination fraction

Table 2 Two-locus genotype probabilities in a full-sib family

L1	L2	Diplotypes	Probabilities
BB	*BB*	$\frac{B-B}{B-B}$	$p^{♂}_{B-B}p^{♀}_{B-B}$
	AB/BA	$\frac{B-A}{B-B}/\frac{B-B}{B-A}$	$p^{♂}_{B-A}p^{♀}_{B-B}+p^{♂}_{B-B}p^{♀}_{B-A}$
	AA	$\frac{B-A}{B-A}$	$p^{♂}_{B-A}p^{♀}_{B-A}$
AB/BA	*BB*	$\frac{A-B}{B-B}/\frac{B-B}{A-B}$	$p^{♂}_{A-B}p^{♀}_{B-B}+p^{♂}_{B-B}p^{♀}_{A-B}$
	AB/BA	$\frac{A-A}{B-B}/\frac{B-B}{A-A}/\frac{A-B}{B-A}/\frac{B-A}{A-B}$	$p^{♂}_{A-A}p^{♀}_{B-B}+p^{♂}_{B-B}p^{♀}_{A-A}+p^{♂}_{A-B}p^{♀}_{B-A}+p^{♂}_{B-A}p^{♀}_{A-B}$
	AA	$\frac{A-A}{B-A}/\frac{B-A}{A-A}$	$p^{♂}_{A-A}p^{♀}_{B-A}+p^{♂}_{B-A}p^{♀}_{A-A}$
AA	*BB*	$\frac{A-B}{A-B}$	$p^{♂}_{A-B}p^{♀}_{A-B}$
	AB/BA	$\frac{A-A}{A-B}/\frac{A-B}{A-A}$	$p^{♂}_{A-A}p^{♀}_{A-B}+p^{♂}_{A-B}p^{♀}_{A-A}$
	AA	$\frac{A-A}{A-A}$	$p^{♂}_{A-A}p^{♀}_{A-A}$

Nine classes of two-locus genotypes (L1, L2) in the offspring, which all correspond to ordered diplotypes (separated by a slash, where the upper haplotype is paternal) and the probability of each class as a function of the frequencies of parental gametes (superscripts indicate the sex of the parent and subscripts indicate the haplotypes of gametes)

and

$$p^{♂}_{B-B} = D^{♂}_{i,j} + \left(1-p^{♂}_i\right)(1-p^{♂}_j\right), \tag{9}$$

for a sire. Expressions for a dam are obtained in the same manner by replacing ♂ with ♀.

These terms also allow us to rewrite the genotype probabilities of Table 2 as functions of the LD-parameters and allele frequencies (see Table 3). We consider that they are arranged in a three-by-three matrix $\mathbf{Z} = (z_{s,q})$, the rows and columns of which pertain to the genotypes at the first locus and the genotypes at the second locus, respectively (the order of genotypes is *BB*, *AB* / *BA*, and *AA* for both loci); for example, the probability of *BB*, *BB* is:

$$\begin{aligned} z_{1,1} &= p^{♂}_{B-B}p^{♀}_{B-B} \\ &= \left[D^{♂}_{i,j} + \left(1-p^{♂}_i\right)\left(1-p^{♂}_j\right)\right] \\ &\quad \cdot \left[D^{♀}_{i,j} + \left(1-p^{♀}_i\right)\left(1-p^{♀}_j\right)\right]. \end{aligned} \tag{10}$$

Table 3 Two-locus genotype probabilities as functions of characteristic parameters

L1	L2	Probabilities
BB	*BB*	$[D_{i,j}^{♂}+(1-p_i^{♂})(1-p_j^{♂})][D_{i,j}^{♀}+(1-p_i^{♀})(1-p_j^{♀})]$
	AB/BA	$[-D_{i,j}^{♂}+(1-p_i^{♂})p_j^{♂}][D_{i,j}^{♀}+(1-p_i^{♀})(1-p_j^{♀})]+[D_{i,j}^{♂}+(1-p_i^{♂})(1-p_j^{♂})][-D_{i,j}^{♀}+(1-p_i^{♀})p_j^{♀}]$
	AA	$[-D_{i,j}^{♂}+(1-p_i^{♂})p_j^{♂}][-D_{i,j}^{♀}+(1-p_i^{♀})p_j^{♀}]$
AB/BA	*BB*	$[-D_{i,j}^{♂}+p_i^{♂}(1-p_j^{♂})][D_{i,j}^{♀}+(1-p_i^{♀})(1-p_j^{♀})]+[D_{i,j}^{♂}+(1-p_i^{♂})(1-p_j^{♂})][-D_{i,j}^{♀}+p_i^{♀}(1-p_j^{♀})]$
	AB/BA	$[D_{i,j}^{♂}+p_i^{♂}p_j^{♂}][D_{i,j}^{♀}+(1-p_i^{♀})(1-p_j^{♀})]+[D_{i,j}^{♂}+(1-p_i^{♂})(1-p_j^{♂})][D_{i,j}^{♀}+p_i^{♀}p_j^{♀}]$ $+[-D_{i,j}^{♂}+p_i^{♂}(1-p_j^{♂})][-D_{i,j}^{♀}+(1-p_i^{♀})p_j^{♀}]+[-D_{i,j}^{♂}+(1-p_i^{♂})p_j^{♂}][-D_{i,j}^{♀}+p_i^{♀}(1-p_j^{♀})]$
	AA	$[D_{i,j}^{♂}+p_i^{♂}p_j^{♂}][-D_{i,j}^{♀}+(1-p_i^{♀})p_j^{♀}]+[-D_{i,j}^{♂}+(1-p_i^{♂})p_j^{♂}][D_{i,j}^{♀}+p_i^{♀}p_j^{♀}]$
AA	*BB*	$[-D_{i,j}^{♂}+p_i^{♂}(1-p_j^{♂})][-D_{i,j}^{♀}+p_i^{♀}(1-p_j^{♀})]$
	AB/BA	$[D_{i,j}^{♂}+p_i^{♂}p_j^{♂}][-D_{i,j}^{♀}+p_i^{♀}(1-p_j^{♀})]+[-D_{i,j}^{♂}+p_i^{♂}(1-p_j^{♂})][D_{i,j}^{♀}+p_i^{♀}p_j^{♀}]$
	AA	$[D_{i,j}^{♂}+p_i^{♂}p_j^{♂}][D_{i,j}^{♀}+p_i^{♀}p_j^{♀}]$

Nine classes of two-locus genotypes (L1, L2) for offspring in a full-sib family and the probability of each class as a function of the parental LD *D* and the *A*-allele frequency p_i at locus *i* (the superscripts indicate the sex of the parent and the subscripts indicate the locus)

The covariance between the additive genotype indicators in Eq. 3 is then determined by:

$$\operatorname{cov}\left(c_{a,i},c_{a,j}\right)=\sum_{s=1}^{3}\sum_{q=1}^{3}c_{a,s}c_{a,q}z_{s,q}-\sum_{s=1}^{3}c_{a,s}z_{s,\bullet}\sum_{q=1}^{3}c_{a,q}z_{\bullet,q}, \tag{11}$$

where the dot indicates summation over all assigned components, and $z_{s,\bullet}$ and $z_{\bullet,q}$ are the marginal genotype probabilities for the first and second locus, respectively. Furthermore, $c_{a,1}=-1$ since the first row and the first column contain genotype *BB*; $c_{a,2}=0$ since the second row and the second column contain heterozygous genotypes; and by analogy, $c_{a,3}=1$. After simplification (using Mathematica [8]), the result obtained for the off-diagonal elements of the covariance matrix $\mathbf{\Omega}$ is:

$$\operatorname{cov}\left(c_{a,i},c_{a,j}\right)=D_{i,j}^{♂}+D_{i,j}^{♀}. \tag{12}$$

Note that the LD depends on the recombination fraction θ, which can be converted into a genetic distance x (in Morgan) using Haldane's mapping function [9]

$$x=-0.5\ln(1-2\theta). \tag{13}$$

To complete the covariance matrix $\mathbf{\Omega}=\operatorname{cov}(c_i,c_j)_{i,j}$, the variance in the genotype indicator (as defined by Eq. 3) at each locus i is expressed as a function of A-allele frequency p_i. The genotype frequencies at locus i are $(1-p_i^{♂})(1-p_i^{♀})$ for genotype BB, $p_i^{♂}(1-p_i^{♂})+p_i^{♀}(1-p_i^{♀})$ for AB / BA, and $p_i^{♂}p_i^{♀}$ for AA. Then,

$$\mathrm{E}\left(c_{a,i}\right)=-1+p_i^{♂}+p_i^{♀} \tag{14}$$

and

$$\mathrm{E}\left(c_{a,i}^2\right)=1-p_i^{♂}-p_i^{♀}+2p_i^{♂}p_i^{♀}, \tag{15}$$

and thus

$$\operatorname{var}\left(c_{a,i}\right)=\mathrm{E}\left(c_{a,i}^2\right)-\mathrm{E}^2\left(c_{a,i}\right)=\pi_i^{♂}+\pi_i^{♀} \tag{16}$$

with $\pi_i^{♂}=p_i^{♂}(1-p_i^{♂})$ and $\pi_i^{♀}=p_i^{♀}(1-p_i^{♀})$. Note that the only possible values for $\operatorname{var}(c_{a,i})$ are 0, $\frac{1}{4}$, or $\frac{1}{2}$, since $\pi_i^{♂}$ and $\pi_i^{♀}$ can only have values of 0 or $\frac{1}{4}$.

Now we have derived all of the elements of the covariance matrix $\mathbf{\Omega}$, so we can express the Mendelian variance for a particular trait as (Eq. 2):

$$\sigma_g^2=\mathbf{m}'\mathbf{\Omega}\mathbf{m}.$$

As already mentioned, this Mendelian sampling variance can be split into the sum of two independent parental contributions:

$$\sigma_g^2=\mathbf{m}'\mathbf{\Omega}\mathbf{m}=\mathbf{m}'\mathbf{\Omega}^{♂}\mathbf{m}+\mathbf{m}'\mathbf{\Omega}^{♀}\mathbf{m}, \tag{17}$$

where $\mathbf{\Omega}^{♂}$ and $\mathbf{\Omega}^{♀}$ represent parent-specific covariance matrices for the additive effects of single alleles based on the paternal and maternal gametes. The paternal (maternal) covariance matrix $\mathbf{\Omega}^{♂}$ ($\mathbf{\Omega}^{♀}$) contains off-diagonal elements that are equal to $D_{i,j}^{♂}$ ($D_{i,j}^{♀}$) and $\pi_i^{♂}$ ($\pi_i^{♀}$) on the diagonals, according to Eqs. 12 and 16.

The variance $\operatorname{var}(c_{a,i})$ becomes zero at loci for which both parents are homozygous. The corresponding rows and columns of the covariance matrix only contain zeroes, which causes a rank deficiency. The corresponding diagonal and off-diagonal elements in the correlation matrix $\mathbf{R}$ are defined as zero (although they are not defined in a strictly mathematical sense) for our purposes in order to maintain rank equality between **Ω** and **R**.

Joint additive and dominance genetic effects

The covariance between two dominance effects can be derived in a similar manner as that between additive effects. Again, we consider the genotype probabilities of Table 2 arranged in a three-by-three matrix $\mathbf{Z} = (z_{s,q})$ and we define: $c_{d,1} = -1$ since the first row and the first column contain genotype *BB*; $c_{d,2} = 1$ since the second row and the second column contain heterozygous genotypes; and by analogy, $c_{d,3} = -1$. Inserting these dominance codes into Eq. 11 and expanding yield:

$$\begin{aligned}\operatorname{cov}\left(c_{d,i}, c_{d,j}\right) &= 16D_{i,j}^{♂}D_{i,j}^{♀}\\ &\quad + 4D_{i,j}^{♂}\left(1-2p_i^{♀}\right)\left(1-2p_j^{♀}\right)\\ &\quad + 4D_{i,j}^{♀}\left(1-2p_i^{♂}\right)\left(1-2p_j^{♂}\right). \quad (18)\end{aligned}$$

Analogous to the additive effects, for the variance of the dominance effects we obtain:

$$\operatorname{var}\left(c_{d,i}\right) = 4\left(\pi_i^{♂} + \pi_i^{♀}\right) - 16\pi_i^{♂}\pi_i^{♀}, \quad (19)$$

which can take values 1 or 0. In particular, var($c_{d,i}$) is equal to 1 if at least one parent is heterozygous, and 0 if both parents are homozygous. Thus, we can conclude that if one marker is homozygous in both parents, the covariance between the dominance effects is equal to 0, which is analogous to previous considerations of the rank deficiency for additive effects. Furthermore, Eqs. 18 and 19 contain the products of the characteristic parameters $(D_{i,j}, p_i, p_j, \pi_i)$ for both the sire and the dam, which is why **Ω** and the Mendelian variance can no longer be split into a sum of two separate parental parts when dominance effects are included.

In order to determine the covariance cov($c_{a,i}, c_{d,j}$) between the additive and dominance indicators, we assign the additive indicators by Eq. 3 to the first (*i*th) locus and the dominance indicators by Eq. 4 to the second (*j*th) locus, i.e.,

$$\begin{aligned}\operatorname{cov}\left(c_{a,i}, c_{d,j}\right) &= \sum_{s=1}^{3}\sum_{q=1}^{3} c_{a,s}c_{d,q}z_{s,q}\\ &\quad - \sum_{s=1}^{3} c_{a,s}z_{s,\bullet}\sum_{q=1}^{3} c_{d,q}z_{\bullet,q}, \quad (20)\end{aligned}$$

which has to be expanded, resulting in:

$$\operatorname{cov}\left(c_{a,i}, c_{d,j}\right) = 2D_{i,j}^{♂}\left(1-2p_j^{♀}\right) + 2D_{i,j}^{♀}\left(1-2p_j^{♂}\right). \quad (21)$$

Exchanging the loci yields:

$$\begin{aligned}\operatorname{cov}\left(c_{d,i}, c_{a,j}\right) &= \operatorname{cov}\left(c_{a,j}, c_{d,i}\right)\\ &= 2D_{i,j}^{♂}\left(1-2p_i^{♀}\right)\\ &\quad + 2D_{i,j}^{♀}\left(1-2p_i^{♂}\right). \quad (22)\end{aligned}$$

The rank deficiencies in **Ω** arise from variances equal to 0 as well as from perfect correlations, which is demonstrated by the two examples of joint correlation matrices for additive and dominance effects shown in the upper part of Fig. 1, as well as their corresponding parental diplotypes. The number of markers is 16 in both examples, which yield correlation matrices with dimensions 32 × 32. Some diagonal elements of the off-diagonal blocks, which contain correlations between additive and dominance effects, indicate a perfect dependency between the additive and dominance effects at the same locus with correlations of either 1 or −1. After the redundant rows and columns have been deleted from the dominance part of the matrix, the correlation matrices obtained (Fig. 1, bottom) exhibit a block diagonal structure. The remaining covariance matrix for dominance effects corresponds to loci for which both parents are heterozygous (five and seven for the example in Fig. 1). The first block remains unchanged numerically, but it has a different interpretation because it now represents the correlation matrix for a vector $\mathbf{m}_{a^*}$, which is defined as:

$$\mathbf{m}_{a^*} = \begin{bmatrix}\mathbf{I}_n & \mathbf{H}_n\end{bmatrix}\begin{bmatrix}\mathbf{m}_a\\ \mathbf{m}_d\end{bmatrix}, \quad (23)$$

where $\mathbf{m}_a$ and $\mathbf{m}_d$ are vectors of the additive and dominance effects for all markers in order of their map position and $\mathbf{I}_n$ is an identity matrix of order *n*. **H** is a diagonal matrix with elements:

$$h_{i,i} = \begin{cases} 0, & \text{if both parents are heterozygous}\\ & \text{or homozygous}\\ 1, & \text{if one parent is heterozygous}\\ & \text{and the other is } BB\\ -1, & \text{if one parent is heterozygous}\\ & \text{and the other is } AA \end{cases}. \quad (24)$$

Now, $m_{a^*,i}$ has the form:

$$m_{a^*,i} = \begin{cases} m_{a,i}, & \text{if both parents are}\\ & \text{heterozygous}\\ & \text{or homozygous}\\ m_{a,i} + m_{d,i}, & \text{if one parent is heterozygous}\\ & \text{and the other is } BB\\ m_{a,i} - m_{d,i}, & \text{if one parent is heterozygous}\\ & \text{and the other is } AA \end{cases}. \quad (25)$$

As a practical consequence, we can use $\mathbf{m}_d{}^*$ (i.e., $\mathbf{m}_d$ with all elements $m_{d,i}$ eliminated, where one parent is

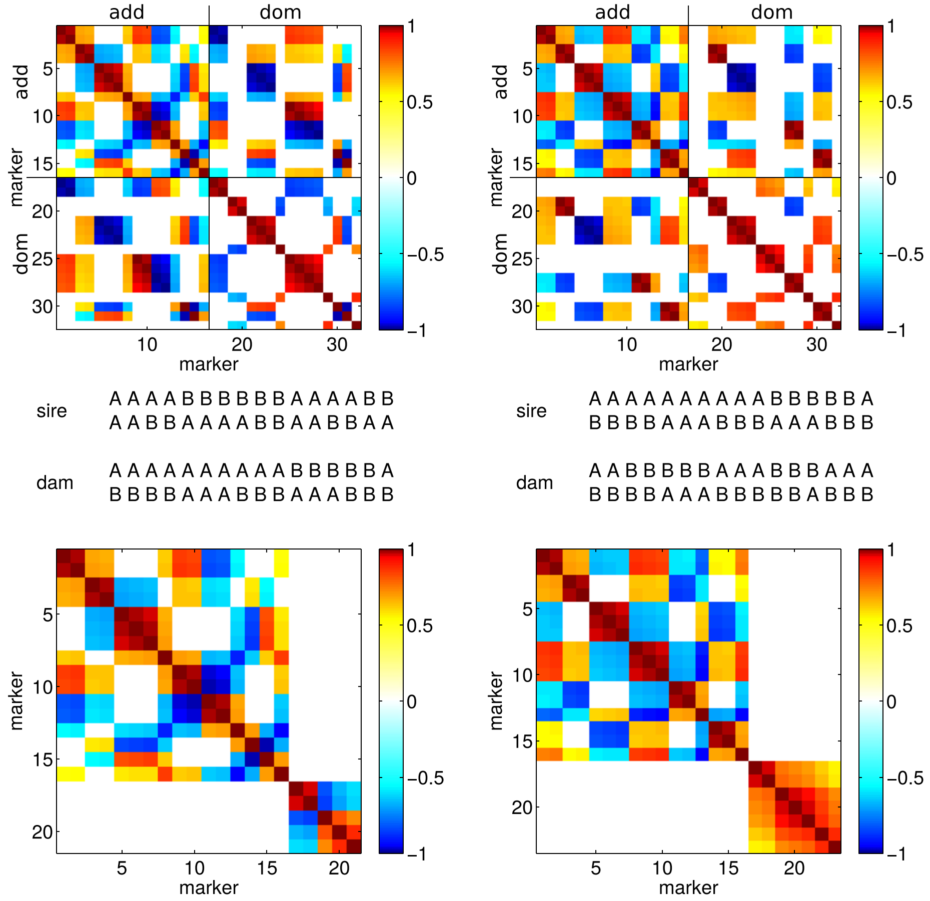

Fig. 1 Correlation matrices for two matings (*top*) with their respective phased genotypes (*middle*). The marker distance is 1 cM. The dimension of both matrices is 32 due to the additive and dominance effects at the 16 markers. Block diagonal matrices (*bottom*) remain after the *rows* and *columns* of dominance effects are deleted when there is a perfect correlation between a dominance and additive effect at the same locus

homozygous at locus i) and $\mathbf{m}_a{}^*$. Then, σ_g^2 can be computed as:

$$\sigma_g^2 = \left[\,\mathbf{m}'_{a^*}\;\; \mathbf{m}'_{d^*}\,\right] \boldsymbol{\Omega}^* \begin{bmatrix} \mathbf{m}_{a^*} \\ \mathbf{m}_{d^*} \end{bmatrix}, \tag{26}$$

where $\boldsymbol{\Omega}^*$ has a block-diagonal structure with unequally sized blocks, as in the example shown in Fig. 1.

Mendelian covariance with multiple traits

The Mendelian covariance between traits (t_k, t_l) is also of interest, especially when we aim at determining the Mendelian variance of the aggregate genotype (multiple trait breeding goal). If $\mathbf{m}_k$ and $\mathbf{m}_l$ are the vectors of marker effects for traits k and l, then the Mendelian covariance $\sigma_g(t_k, t_l)$ between these two traits is

$$\sigma_g(t_k, t_l) = \mathbf{m}'_k \boldsymbol{\Omega} \mathbf{m}_l. \tag{27}$$

The Mendelian variances and covariances for several traits can be collected in the Mendelian covariance matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma_g^2(t_1) & \sigma_g(t_1,t_2) & \cdots & \sigma_g(t_1,t_N) \\ \sigma_g(t_1,t_2) & \sigma_g^2(t_2) & \cdots & \sigma_g(t_2,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_g(t_1,t_N) & \sigma_g(t_2,t_N) & \cdots & \sigma_g^2(t_N) \end{bmatrix}, \tag{28}$$

which is given by:

$$\mathbf{V} = \mathbf{M}'\boldsymbol{\Omega}\mathbf{M}, \tag{29}$$

where each column $\mathbf{m}_k$ of $\mathbf{M}$ is a vector of marker effects for trait k. The Mendelian sampling variance for the aggregate genotype can then be obtained from $\mathbf{V}$ and the vector $\mathbf{f}$ of the economic weights for all traits:

$$\sigma_{gT}^2 = \mathbf{f}'\mathbf{V}\mathbf{f}. \tag{30}$$

This quantity has a pivotal role in mating decisions because the total breeding value (defined as the linear combination of single breeding values $\mathbf{f}'\mathbf{t}$,

$\mathbf{t} = (t_1, \ldots, t_N)$) is the most important criterion for selection.

Practical application

We compared the exact method with a recently published simulation approach [5]. The Mendelian sampling variance of gametes was calculated with both methods for each animal from a dataset that included the diplotypes of 74,353 male and female German Holstein cattle. Identical sets of recombination rates and estimates of additive marker effects were used for both methods. These parameters were derived from routine genomic evaluation data. This comparison was done for four traits: fat yield (FKG), protein yield (PKG), somatic cell score (SCS), and the direct genetic effect on stillbirth (SBd).

The exact gametic Mendelian variances $\sigma^2_{\hat{g}}$ of the estimated gametic values were calculated for each trait and each animal as:

$$\sigma^2_{\hat{g}} = \hat{\mathbf{m}}'\boldsymbol{\Omega}^{\male}\hat{\mathbf{m}} \quad \text{or} \quad \sigma^2_{\hat{g}} = \hat{\mathbf{m}}'\boldsymbol{\Omega}^{\female}\hat{\mathbf{m}} \tag{31}$$

according to Eq. 2, where $\hat{\mathbf{m}}$ is the vector of the estimated additive marker effects. The simulation method used 100,000 randomly generated gametes for each animal. Thus, the estimate of the Mendelian variability for the estimated gametic values was:

$$\widehat{\sigma^2}_{\hat{g}} = \sum_{o=1}^{K} \frac{\left(\mathbf{w}_o'\hat{\mathbf{m}}\right)^2}{K} - \left(\sum_{o=1}^{K} \frac{\mathbf{w}_o'\hat{\mathbf{m}}}{K}\right)^2, \tag{32}$$

where $K = 100{,}000$ is the number of simulated gametes per individual and $\mathbf{w}_o$ is the vector of the genotype indicators for the oth simulated gamete in the individual under consideration.

The Mendelian covariances between traits were also obtained using the exact method by applying Eq. 29. Furthermore, the four traits were combined with weights of 0.1, 0.4, 0.375, and 0.175 for the traits FKG, PKG, SCS, and SBd, respectively, and the (gametic) Mendelian variances were computed for this aggregate genotype. The covariances and aggregate genotypes were not implemented in the simulation method, so no comparisons could be made with the simulation method for these quantities.

Results

Scatter plots of the estimated $\widehat{\sigma^2}_{\hat{g}}$-values against their exact $\sigma^2_{\hat{g}}$ counterparts are in Fig. 2. The slopes of the linear regression lines were close to unity for all of the traits and the intercepts were small, which indicated a good average agreement between both methods. The variation around the regression line was due to the Monte Carlo error of the simulation.

To facilitate a better comparison between the trait combinations, the Mendelian covariances were transformed into correlations and their distributions are in Fig. 3. Interestingly, the sign and magnitude of these correlations exhibited a very high amount of variation between individuals. The Mendelian correlation between FKG and PKG was an exception because most of the values were positive and the distribution was bimodal. This bimodality is a consequence of the *DGAT1* gene, for which heterozygous animals led to the smaller peak at correlations below 0.5 and homozygotes were responsible for the larger peak with correlations above 0.5.

The coefficient of variation of the Mendelian variances of the aggregate genotype was 18.7 %, which was similar to the coefficient of variation of the Mendelian variances of PKG (19.0 %), SCS (20.2 %), and SBd (21.5 %), but somewhat smaller than that of FKG (30.3 %).

Discussion

The derived covariance matrices depend on prior information on the order and genetic distance of markers, as well as the parental diplotypes. They allow the calculation of the within-family variation of the estimated genetic values based on exact considerations of the genotypes, degrees of homozygosity, and linkage phases in the parents. In the classical formula, $\frac{1}{4}\sigma^2_a(1 - F_{\male}) + \frac{1}{4}\sigma^2_a(1 - F_{\female})$, the loss of homozygosity in the parents is considered relative to the base population, where inbreeding coefficients are zero and the Mendelian variance is at maximum $(\frac{1}{2}\sigma^2_a)$ for all families. Our covariance matrices, in contrast, mirror the absolute level of marker homozygosity and the Mendelian variance reaches its maximum for fully heterozygous parents with all positive marker alleles in coupling phase. Different linkage phases can make a substantial difference in marker covariability, as shown by the example in Fig. 4, which presents the correlation matrices for two diplotypes with identical 16-marker genotypes but different linkage phases.

Fields of application and computational aspects

In general, the method described in this study can be applied to all diploid animals and plants. Of course, all relevant input parameters must be known, such as the marker maps, marker effects, and phased genotypes. Crosses of double haploid (i.e. fully inbred) lines occur as parents in breeding programs for plant species such as e.g. maize. An advantage of such parents is that they provide reliable diplotypes because of the complete homozygosity of genotyped grandparents, whereas the derivation of diplotypes is prone to some degree of phasing error in non-inbred populations [10]. In cases where the phase of some SNPs is only known at a probabilistic level, it

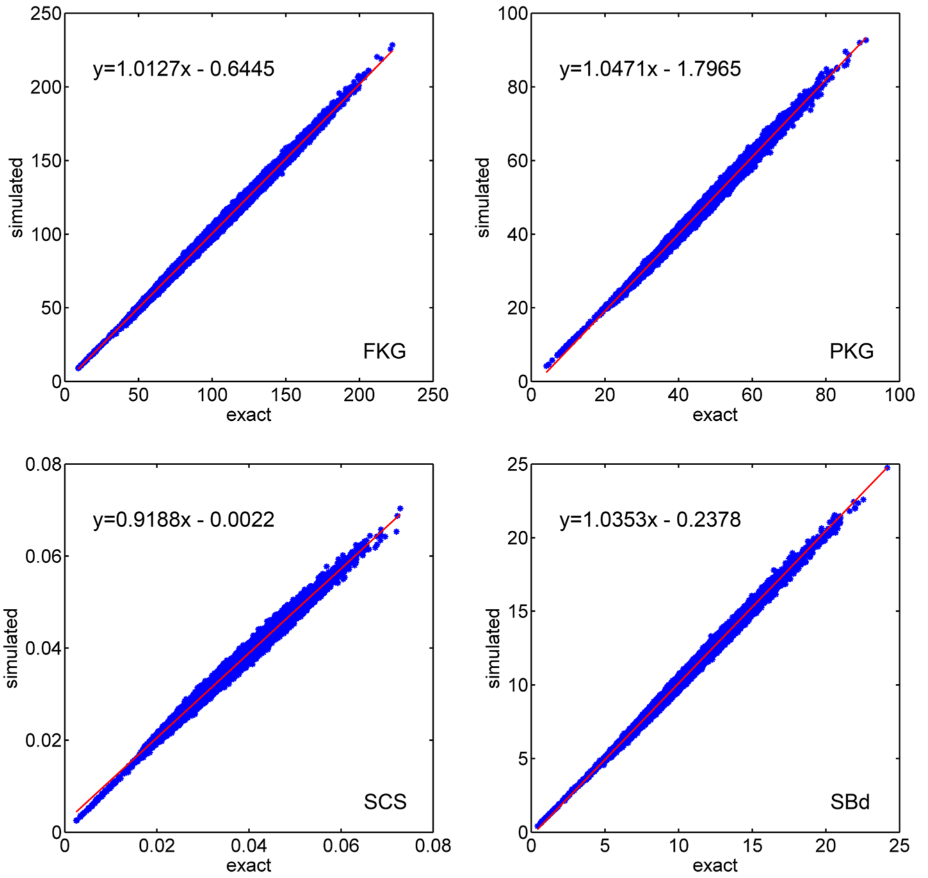

Fig. 2 *Scatter plots* of Mendelian gametic variances obtained from simulation versus the exact values. Differences between simulated values (each one derived from 100,000 simulated gametes per candidate) and exact values appear as vertical deviations from the linear regression lines (in *red*) and represent Monte Carlo errors. The traits include fat yield (FKG), protein yield (PKG), somatic cell score (SCS), and direct genetic effect on stillbirth (SBd)

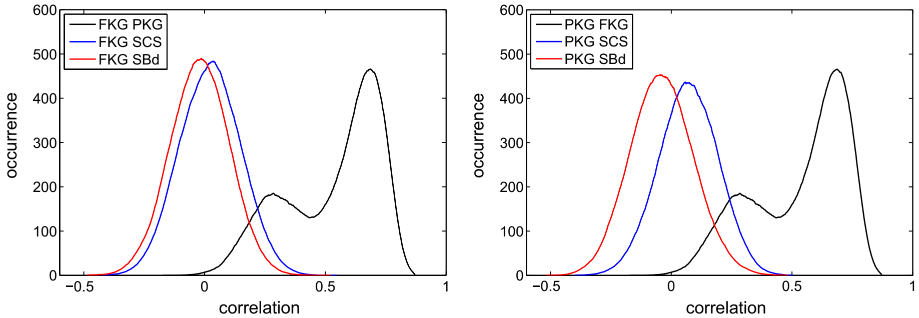

Fig. 3 Frequency distributions of the Mendelian correlations. The frequency distributions of the correlations between FKG and all the other traits (PKG, SCS, and SBd) are shown on the *left*, and those between PKG and all the other traits are shown on the *right*. *FKG* fat yield, *PKG* protein yield, *SCS* somatic cell score, and *SBd* direct genetic effect on stillbirth

may be an option to average the Mendelian sampling variances over all possible linkage phases. Simplifications may be possible, e.g. by taking only the most probable diplotypes into account. However, we did not investigate this question in detail.

For humans and mice, it has been found that marker maps generally differ for male and female parents [11]. All the covariances can be adjusted easily for sex-specific recombination rates, which is achieved most easily in the pure additive case by applying the male and female

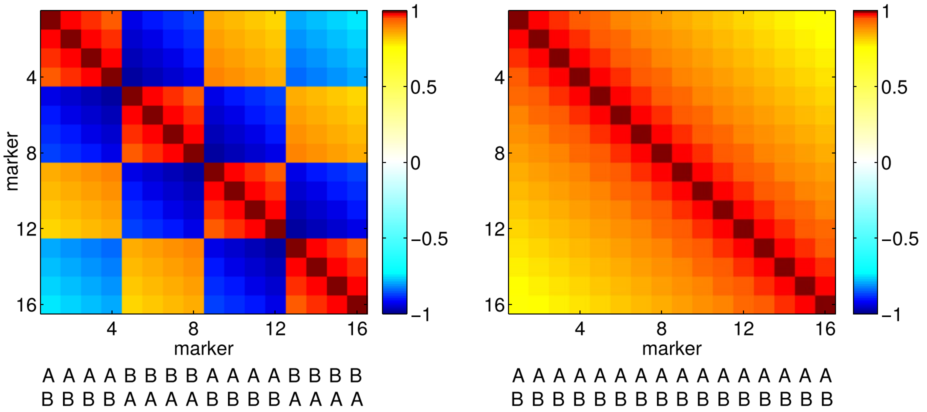

Fig. 4 Parent-specific correlation matrix $\boldsymbol{\Omega}^{♂}$ of the gametes. The respective phased genotypes are listed below. All markers are heterozygous, with a genetic distance of 1 cM between markers. For the first animal (*left*), markers 1–4 and 9–12 are in coupling phase, whereas markers 5–8 and 13–16 are in repulsion phase with respect to the first marker. For the second animal (*right*), all markers are in coupling phase

recombination rates to set up $\boldsymbol{\Omega}^{♂}$ and $\boldsymbol{\Omega}^{♀}$. In the general case, the LD measures for both the paternal and maternal gametes $D^{♂}$ and $D^{♀}$ must be adjusted in order to obtain the adjusted covariances.

The sex of the full sibs matters for the inclusion of sex chromosomes because when all considered progeny are female, the sire can be treated as homozygous at all X-chromosomal loci and the calculation can proceed as usual. When the focus is on male progeny, such as young bulls obtained from elite matings, there is no X-chromosomal paternal contribution to the Mendelian sampling variance. Of course, dominance has no effect on the X-chromosomal Mendelian variance in males, unlike for females.

From a computational viewpoint, the purely additive case is most convenient because the parental contributions to the Mendelian variances and covariances can be calculated for a large list of potential parents and all traits by setting up the parent-specific covariance matrix. Subsequently, the parental contributions only have to be added to the total within-family variance for each mating considered. The computational time required by the exact method was roughly the same as that for the simulation approach. However, the computational demand would increase for the latter case if the Monte Carlo error needs to be reduced further.

Population-averaged Mendelian sampling variances for single traits were previously derived from large numbers of phased genotypes and available estimates of additive marker effects by Cole and VanRaden [6]. Neither simulation of gametes nor covariance matrices were used in this study since loci on the same chromosome were either assumed to be perfectly linked or fully independent. Consequently, their respective results can only be interpreted as upper and lower limits. In another study, Segelke et al. [5] took recombination within chromosomes into account by simulating gametes of individuals with known diplotypes. Parental contributions to the within-family additive genetic variance were expressed as standard deviations of gamete breeding values in a family-specific manner.

Consideration of the aggregate genotype calls for a full Mendelian sampling covariance matrix across traits, which, in the additive case, can also be derived by simulation, but this has not yet been reported in the literature. This requires that genomic breeding values are estimated for each trait of interest and each single simulated gamete and then averages of squares and cross-products are calculated over gametes. If dominance effects are to be included, pairs of paternal and maternal gametes have to be simulated. The simulation-inherent Monte Carlo errors of all single-trait variances and all pair-wise covariances will, of course, propagate and induce a joint Monte Carlo error of the resulting variability in the aggregate genotype.

From a producer's perspective, phenotypic uniformity of a population of plants or animals is desirable because it facilitates management. Matings with high additive genetic merit and low within-family genetic variance [6] may be attractive to achieve that goal. Dominance—if of some importance for the traits under consideration—could be included for the same purpose. Breeding organizations, in contrast, are probably more interested in offspring with exceptionally high breeding values [6], since e.g. in dairy cattle, semen prices are non-linearly related to the genetic merit of bulls. For a particular mating, the probability that the estimated breeding value of offspring will be greater than a certain threshold can be determined from a normal distribution with family-specific mean and variance. The opportunity to breed the desired animals of top genetic merit can then be maximized by choosing the matings with the highest probabilities among all possible matings, possibly by taking some constraints such as inbreeding into account.

The average observed degree of homozygosity in the German Holstein dataset was 65.3 %, with a range from 25.2 to 88.0 %. These high degrees of homozygosity were exploited for computational speed by deleting the rows and columns for homozygous markers from the covariance matrix and the respective marker effects from $\hat{\mathbf{m}}$. Therefore, the dimensions of the remaining vector of marker effects and the remaining covariance matrix were reduced greatly, leading to considerable computational time savings. Note that both parents had to be homozygous in the dominance case in order to reduce matrix dimension in a similar way. Clearly, computational time can be decreased by implementing parallel calculation of individuals and chromosomes. However, in the presence of dominance, each considered mating must be computed with its own covariance matrix, and thus only matings can be parallelized (the chromosomes are unaffected).

Choosing alternative genotype indicators

The formulae for the covariances and correlations are functions of the chosen genotype indicators, for which different options exist [12]. In the present study, we used (1, 0, −1) (Eq. 3) for additive effects and (−1, 1,−1) (Eq. 4) for dominance effects, but other possible indicators include (0, 1, 2) for additive effects and (0, 1, 0) or $\left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)$ for dominance effects. All these indicators can be transformed into each other by a shift and/or a multiplication by a constant. Simply shifting the indicators does not influence the (co)variance, so it is also possible to use the formulae described in this study when additive marker effects have been estimated via the (0, 1, 2) coding.

However, multiplication of genotype codes by a constant factor does affect the (co)variance. For example, let $\mathbf{m}_{\tilde{d}}$ be the dominance marker effect when the indicator $c_{\tilde{d},i} = (0,1,0)$ is used for its estimation. The Mendelian sampling variance is then calculated by $\mathbf{m}'_{a,\tilde{d}}\boldsymbol{\Omega}_{a,\tilde{d}}\mathbf{m}_{a,\tilde{d}}$, where $\mathbf{m}'_{a,\tilde{d}}$ is known, but $\boldsymbol{\Omega}_{a,\tilde{d}}$ is not known. The type of parameterization does not affect the Mendelian sampling variance, so:

$$\mathbf{m}'_{a,\tilde{d}}\boldsymbol{\Omega}_{a,\tilde{d}}\mathbf{m}_{a,\tilde{d}} = \mathbf{m}'_{a,d}\boldsymbol{\Omega}_{a,d}\mathbf{m}_{a,d} \tag{33}$$

must hold, where the terms on the right-hand side are known. The relationship between indicators $c_{\tilde{d}}$ and c_d is $c_{\tilde{d},i} = \frac{c_{d,i}+1}{2}$; hence,

$$\begin{aligned}\operatorname{cov}\left(c_{\tilde{d},i}, c_{\tilde{d},j}\right) &= \operatorname{cov}\left(\frac{c_{d,i}+1}{2}, \frac{c_{d,j}+1}{2}\right) \\ &= \frac{1}{4}\operatorname{cov}\left(c_{d,i}, c_{d,j}\right),\end{aligned} \tag{34}$$

$$\begin{aligned}\operatorname{cov}\left(c_{\tilde{d},i}, c_{a,j}\right) &= \operatorname{cov}\left(\frac{c_{d,i}+1}{2}, c_{a,j}\right) \\ &= \frac{1}{2}\operatorname{cov}\left(c_{d,i}, c_{a,j}\right),\end{aligned} \tag{35}$$

and

$$\begin{aligned}\operatorname{cov}\left(c_{a,i}, c_{\tilde{d},j}\right) &= \operatorname{cov}\left(c_{a,i}, \frac{c_{d,j}+1}{2}\right) \\ &= \frac{1}{2}\operatorname{cov}\left(c_{a,i}, c_{d,j}\right),\end{aligned} \tag{36}$$

and thus the Hadamard product

$$\boldsymbol{\Omega}_{a,\tilde{d}} = \mathbf{U} \odot \boldsymbol{\Omega}_{a,d} \tag{37}$$

with

$$\mathbf{U} = \begin{bmatrix} \mathbf{1} & \frac{1}{2}\mathbf{1} \\ \frac{1}{2}\mathbf{1} & \frac{1}{4}\mathbf{1} \end{bmatrix} \tag{38}$$

transforms the $\boldsymbol{\Omega}_{a,d}$ matrix into $\boldsymbol{\Omega}_{a,\tilde{d}}$. Instead of transforming the covariance matrix $\boldsymbol{\Omega}$, the estimated marker effects can also be transformed, which can be implemented easily by multiplying the marker effects $\mathbf{m}_{\tilde{d}}$ by $\frac{1}{2}$.

Genetic differences can be parameterized in terms of substitution effects and dominance deviations, or in terms of additive and dominance genotype effects, as discussed in detail by Vitezica et al. [12]. For the former, the $\hat{\mathbf{m}}$-vector comprises estimates of the allele substitution effects $\mathbf{m}_{\alpha}$ and dominance effects $\mathbf{m}_{\tilde{d}}$. Allele substitutions effects are defined as (e.g. [13]):

$$m_{\alpha,i} = m_{a,i} + m_{\tilde{d},i}(v_i - u_i), \tag{39}$$

where u_i and $v_i = 1 - u_i$ are the population frequencies for alleles A and B, respectively. When only allele substitution effects $\mathbf{m}_{\alpha}$ are considered and dominance effects are ignored (e.g. [5]), it is possible to use the covariance matrix described in this study without changes.

If dominance effects are not ignored, the allele substitution effects must be transformed into additive marker effects in order to allow the use of the derived covariance matrix. This transformation is achieved by:

$$\begin{aligned}m_{a,i} &= m_{\alpha,i} - m_{\tilde{d},i}(v_i - u_i) \\ &= m_{\alpha,i} - \frac{1}{2}m_{d,i}(v_i - u_i).\end{aligned} \tag{40}$$

The allele frequencies u and v, which are required for the transformation, are already available because they are required for routine estimation.

Conclusions

In this study, we proposed a new method for the exact calculation of Mendelian sampling (co-)variances based on knowledge of phased marker genotypes and marker effect estimates and we derived all the requisite formulae. The method considers inbreeding but also the absolute level of homozygosity, as indicated by the marker genotypes, while it also considers the linkage phase of the markers in both parents.

We demonstrated the applicability of our method by comparing its results with results produced by an established simulation method using a large dairy cattle dataset. We found that both approaches agreed within the range of the Monte Carlo error, which is inherent in the simulation, but which can be fully avoided because the derived covariance matrices represent an infinitely large number of progeny.

Authors' contributions

SB and FT derived the formulae for all the covariances, with the help of NR. FT introduced the use of LD parameters and their clear representation. The theory was implemented in Fortran programs by MR and SB, with some help from NR. DS ran the simulations and the exact calculations at VIT, while SB performed the comparative analysis. The manuscript was drafted by SB, with contributions from all the coauthors. NR was responsible for the general concept and supervised all of the work. All authors read and approved the final manuscript.

Author details

[1] Institute of Genetics and Biometry, Leibniz-Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany. [2] Vereinigte Informationssysteme Tierhaltung w.V., Heideweg 1, 27283 Verden, Germany.

Acknowledgements

We thank Friedrich Reinhard (VIT Verden) for fruitful discussions. We thank three anonymous reviewers for their valuable comments. Financial support for SB and MR from BMBF via Phänomics under Grant 0315536G is gratefully acknowledged. The publication of this article was funded by the Open Access fund of the Leibniz Association.

Competing interests

The authors declare that they have no competing interests.

Received: 10 June 2015 Accepted: 7 April 2016
Published online: 23 April 2016

References

1. Dempfle L. Problems in the use of the relationship matrix in animal breeding. In: Gianola D, Hammond K, editors. Advances in statistical methods for genetic improvement of livestock, vol. 18. 1st ed. New York: Springer; 1990.
2. Quaas RL, Pollak EJ. Mixed model methodology for farm and ranch beef cattle testing programs. J Anim Sci. 1980;51:1277–87.
3. Neugebauer N, Räder I, Schild HJ, Zimmer D, Reinsch N. Evidence for parent-of-origin effects on genetic variability of beef traits. J Anim Sci. 2010;88:523–32.
4. Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976;32:69–83.
5. Segelke D, Reinhardt F, Liu Z, Thaller G. Prediction of expected genetic variation within groups of offspring for innovative mating schemes. Genet Sel Evol. 2014;46:42.
6. Cole JB, VanRaden PM. Use of haplotypes to estimate Mendelian sampling effects and selection limits. J Anim Breed Genet. 2011;128:446–55.
7. Wittenburg D, Guiard V, Teuscher F, Reinsch N. Comparison of statistical models to analyse the genetic effect on within-litter variance in pigs. Animal. 2008;2:1559–68.
8. Wolfram S. The MATHEMATICA ® book, version 4. Cambridge: Cambridge University Press; 1999.
9. Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet. 1919;8:299–309.
10. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12:703–14.
11. Kong X, Murphy K, Raj T, He C, White PS, Matise TC. A combined linkage-physical map of the human genome. Am J Hum Genet. 2004;75:1143–8.
12. Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195:1223–30.
13. Falconer D, Mackay T. Introduction to quantitative genetics. 4th ed. Harlow: Longman Group Ltd; 1996.