J. Dairy Sci. 102:5279–5294
https://doi.org/10.3168/jds.2018-15971

Variance of gametic diversity and its application in selection programs

D. J. A. Santos,[1,2]* J. B. Cole,[3] T. J. Lawlor Jr.,[4] P. M. VanRaden,[3] H. Tonhati,[2] and L. Ma[1]*
[1]Department of Animal and Avian Sciences, University of Maryland, College Park 20742
[2]Departamento de Zootecinia, Universidade Estadual Paulista, Jaboticabal, 14884-900, Brazil
[3]Henry A. Wallace Beltsville Agricultural Research Center, Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350
[4]Holstein Association USA, Brattleboro, VT 05302-0808

ABSTRACT

The variance of gametic diversity $\left(\sigma^2_{\text{gamete}}\right)$ can be used to find individuals that more likely produce progeny with extreme breeding values. The aim of this study was to obtain this variance for individuals from routine genomic evaluations, and to apply gametic variance in a selection criterion in conjunction with breeding values to improve genetic progress. An analytical approach was developed to estimate σ^2_{gamete} by the sum of binomial variances of all individual quantitative trait loci across the genome. Simulation was used to verify the predictability of this variance in a range of scenarios. The accuracy of prediction ranged from 0.49 to 0.85, depending on the scenario and model used. Compared with sequence data, SNP data are sufficient for estimating σ^2_{gamete}. Results also suggested that markers with low minor allele frequency and the covariance between markers should be included in the estimation. To incorporate σ^2_{gamete} into selective breeding programs, we proposed a new index, relative predicted transmitting ability, which better utilizes the genetic potential of individuals than traditional predicted transmitting ability. Simulation with a small genome showed an additional genetic gain of up to 16% in 10 generations, depending on the number of quantitative trait loci and selection intensity. Finally, we applied σ^2_{gamete} to the US genomic evaluations for Holstein and Jersey cattle. As expected, the *DGAT1* gene had a strong effect on the estimation of σ^2_{gamete} for several production traits. However, inbreeding had a small impact on gametic variability, with greater effect for more polygenic traits. In conclusion, gametic variance, a potentially important parameter for selection programs, can be easily computed and is useful for improving genetic progress and controlling genetic diversity.

Key words: Mendelian sampling, gamete, heterozygosity, selective breeding, dairy cattle

INTRODUCTION

Since the introduction of marker-assisted selection and genomic selection, technological improvements have resulted in widespread incorporation of molecular information into genetic evaluations (Nejati-Javaremi et al., 1997; Meuwissen et al., 2001; Schaeffer, 2006). Increased prediction accuracy, along with reduced generation intervals, has made genomic selection an important tool for achieving fast progress in dairy selection programs (García-Ruiz et al., 2016). Despite concerns about inbreeding in selection and mating designs, most selection programs only consider breeding values when making selection decisions. Even with genomic selection models, genomic breeding value or PTA and evaluation of future progeny are mostly based on expected breeding values without consideration of the variability of those values due to Mendelian sampling.

In addition to breeding value or PTA, other selection strategies have been proposed to increase the rate of genetic progress. One idea was to select animals that will provide greater genetic gains in the future rather than choosing the best animals in the current population. Goiffon et al. (2017) showed improved genetic gains when selecting for the best gametes from a subset of individuals in a population. Segelke et al. (2014) discussed the potential use of the variation within groups of offspring, which allows the assignment of probabilities to obtain progeny with a breeding value over a given threshold, as well as the number of matings required. In a follow-up study, Bonk et al. (2016) showed how exact within-family genetic variation can be calculated using data from phased genotypes. Recently, Müller et al. (2018) proposed a new selection criterion based on the expected maximum haploid breeding value. Collectively, these studies suggest that the incorporation of variation of future gametic values into mating decisions

Received November 10, 2018.
Accepted February 27, 2019.
*Corresponding authors: daniel_jordan2008@hotmail.com and lima@umd.edu

can improve genetic progress on top of the selection on breeding values.

However, a few questions need to be answered before the application of gametic variance to breeding programs, such as how to assess the variation of future gametic values of an individual, how large is the gametic variance, how to use this information for selection, and how to estimate the variance of gametic diversity and use it in existing genomic evaluations. In this study, we aimed to address these questions from a statistical point of view, demonstrating the equivalence between gametic variance and Mendelian sampling variance in the classical BLUP (pedigree) model. We also sought to explore how this variance can be used as a selection criterion in conjunction with breeding values, with the goal of maximizing future genetic gains. We propose an approach for estimating this variance from routine genomic evaluations, verifying the adequacy of the estimates for individuals with and without progeny, and estimating the variance of breeding values of future progeny for a given mating. Finally, we evaluate the application of gametic variance to improve the selection of dairy traits in the US Holstein and Jersey populations.

MATERIALS AND METHODS

Estimation of the Variance of Gametic Diversity

We refer to the variance of gametic diversity as σ^2_{gamete}, which is equivalent to half of the Mendelian sampling variance (Appendix A1). σ^2_{gamete} measures the deviation of progeny breeding values from parent average and can be calculated using the probabilities of transmission of alleles at all QTL from an individual to its gametes. Gametic variance represents the variability of all possible gametic values generated by the permutation and recombination of each parental chromosome. In fact, only the heterozygous loci of an individual contribute to σ^2_{gamete}, so we only consider heterozygous loci in the following text.

Let's first consider one locus. For a biallelic locus j of an individual i with allele substitution effect α_j, σ^2_{gamete} can be calculated from a binomial variance of $\sigma^2_{[j]} = np(1-p)\alpha_j^2$, where the probability of transmission of a reference allele to a gamete $p = 0.5$ and the number of alleles transmitted to a gamete $n = 1$. When 2 loci, j and k, are considered for an individual i, the resulting variance can be obtained as

$$\sigma^2_{[j+k]} = \sigma^2_{[j]} + \sigma^2_{[k]} + 2\sigma_{jk}$$

$$\text{and } \sigma_{jk} = \left(p_{jk} - p_j p_k\right)\alpha_j\alpha_k, \quad [1]$$

where $p_j = p_k = 0.5$, and p_{jk} is the probability that the 2 reference alleles of the 2 loci are transmitted together; p_{jk} can be obtained from the linkage phase and recombination rate between the 2 loci. For example, $p_{jk} = 0.25$ and $\sigma_{jk} = 0$ when the loci are in linkage equilibrium; $p_{jk} = 0.5$ and $\sigma_{jk} = 0.25\alpha_j\alpha_k$ when the 2 reference alleles are on the same chromosome and the loci are in complete linkage.

Extending this calculation from 2 loci to all QTL on the genome, the σ^2_{gamete} of individual i can be obtained as the sum across all N heterozygous QTL:

$$\sigma^2_{\text{gamete}} = \sum_{j=1}^{N}\sigma^2_{[j]} + 2\sum_{j=1}^{N}\sum_{k=j+1}^{N}\sigma_{jk}.$$

This can be represented in matrix format as follows:

$$\sigma^2_{\text{gamete}} = [\alpha_1 \ldots \alpha_N]\mathbf{M}[\alpha_1 \ldots \alpha_N]', \quad [2]$$

where $\alpha_j\,(j = 1,\ldots,N)$ are the allele substitution effects, and $\mathbf{M}$ is the (co)variance matrix of the Mendelian transmission probabilities for the N heterozygous loci:

$$\mathbf{M} = \begin{bmatrix} 0.25 & \ldots & al_{1,N}\left(-\frac{cM_{1,N}}{200} + 0.25\right) \\ \vdots & \ddots & \vdots \\ al_{N,1}\left(-\frac{cM_{N,1}}{200} + 0.25\right) & \ldots & 0.25 \end{bmatrix},$$

where al_{jk} is a phase indicator for loci j and k, with value 1 when both loci have the reference allele on the same chromosome and -1 otherwise; cM_{jk} is the genetic distance between the 2 loci (in centimorgans). Any 2 loci with genetic distance >50 cM on the same chromosome, or on different chromosomes, are assumed to be independent and thus have zero values for the corresponding elements of $\mathbf{M}$. When all the loci are independent,

$$\mathbf{M} = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 0.25 \end{bmatrix}.$$

Instead of using genetic distances, $\mathbf{M}$ can be set up when direct recombination rates are available.

To estimate gametic variance in real data where genomic evaluation is available, we proposed to use the estimated SNP effects to replace true QTL effects in Equation [2]. This approximation of QTL with SNP marker effects is similar to that described by Bonk et al. (2016). Note that using estimated SNP effects in [2] may bias the estimation due to the covariance between estimated effects of SNP in linkage disequilibrium (LD) and potential biases from shrunken estimators of SNP effects, which warrants further investigation.

Application of Gametic Variance in Selection Programs

A new selection strategy using σ^2_{gamete} can be proposed, focusing on the future genetic progress (Bijma et al., 2018). When a small proportion of animals are selected for breeding, σ^2_{gamete} can help identify those that are most likely to produce progeny with extreme breeding values. Assuming selection intensity is maintained across generations, the average genetic value of the animals selected in the future will be related to the variance of gametes of the selected animals in the current generation. The average breeding value transmitted to future progeny can be calculated by summing the expected value and i times the standard deviation of gametic diversity $\left(i\sigma_{gamete}\right)$. The selection intensity (i) represents the number of standard deviations between the population average and the average of selected individuals. The same intensity can be applied when using PTA as the expected value and σ_{gamete} as standard deviation, to obtain the mean breeding value transmitted to the selected individuals in the next generation. Similar approaches have been proposed by Lehermeier et al. (2017) via a usefulness criterion (**UC**) with genomic EBV (**GEBV**) and the standard deviation of a given mating.

Here, we propose a new selection criterion relative to the intensity of selection applied in the next generation (i_f) for an individual i (unknowing mating),

$$RPTA_i = PTA_i + \sigma_{gamete\_i} \times i_f, \quad [3]$$

where $RPTA_i$ (relative PTA) is the average of the genetic values relative to the group of progeny that will be selected in the future (see Appendix A2). In addition, we introduce a new concept of coefficient of relative variation (**CRV**) as a measure of the variability of the additive genetic values (u) transmitted from an individual to its progeny (Appendix A3). The CRV of an individual i is defined as follows (where E indicates expected value):

$$CRV_i = \frac{\sigma_{gamete}}{0.5\sqrt{E\left(u_i^2\right)}}. \quad [4]$$

Simulation

To verify the estimation of σ^2_{gamete} by genomic models and the use of this new parameter to aid selection, we simulated different scenarios with various QTL, genotype, and phenotype data using the QMSim version 1.10 software (Sargolzaei and Schenkel, 2009). In brief, we simulated a historical population, a 10-generation recent population, and a 10-generation future population (Table 1).

To mimic real populations, a historical population was simulated with the same proportion of males and females that were mated randomly. This population was generated in 3 phases: the first phase consisted of 500 generations with a constant population size of 1,000 individuals; the second phase had 500 generations with a constant reduction of population size from 1,000 to 200 to generate LD and establish drift-mutation balance; and the third phase included 10 generations of expansion, where the population size increased from 200 to 3,000. From the last generation of this historical population, 200 males and 800 females were randomly selected as founders to generate the study population, which consisted of 10 generations with 5 progeny per dam and a ratio of 50% males in the offspring. We simulated a selection for breeding values estimated by the classical BLUP (Henderson, 1975). The replacement ratio was set at 20% for dams and 60% for sires (Brito et al., 2011), and mating was random among selected individuals. The replacement ratio is the proportion of animals to be culled and replaced in each generation.

From the study population (last 10 generations of the simulation), genotype and QTL data were obtained for the 9th generation (treated as a reference population) and the 10th generation (the validation population). The marker effects were first estimated in the reference generation. The σ^2_{gamete} values for all individuals were estimated for both the reference and validation populations using the marker effects estimated in the reference generation. For comparison, true gametic variance was also calculated using the QTL effects and their genotype data from the simulation.

To reduce computational load, a small genome, with 4 autosomal chromosomes of 50 cM each, was simulated. The mutation rate was fixed at 2.5×10^{-5} in the historical population. The number of crossovers was sampled from a Poisson distribution. A total of 200,000 markers and different sets of QTL were simulated to be randomly distributed along the genome. After the ge-

nome was simulated, a panel with 10% of the polymorphic markers was sampled every 0.5 cM and another panel with 20% of the markers was sampled every 0.5 cM. The first panel was chosen to mimic a high-density SNP panel and the second for sequence data. A detailed description of the parameters is reported in Table 1.

Six traits were simulated with heritabilities of 0.1, 0.3, and 0.5 and 20 QTL (i.e., 0.1 QTL per cM) or 200 QTL (i.e., 1 QTL per cM), respectively. We used 2 QTL densities similar to those used by Meuwissen et al. (2001). The QTL effects were generated based on a gamma distribution with parameter $\beta = 0.4$ (Hayes and Goddard, 2001). The phenotypic variance was assumed to be 1 for all traits. Four replicates were used for each trait. In addition, 10 future generations were simulated where the individuals were selected either by the true breeding value (T_PTA) or by true RPTA (T_RPTA) to verify and compare the genetic gains obtained using these criteria. To assess the effect of these indices on selection in the future generations, the replacement ra-

Table 1. Summary of simulation parameters

Parameter	Value
Genome parameter	
Genome size	200 cM
Number of chromosomes	4
Number of QTL	20 and 200
Number of markers	10,000 (high-density panel) and 20,000+ QTL (sequence data)
Mutation rate, QTL	2.5×10^{-5}
Mutation rate, marker	2.5×10^{-3}
Marker positions in genome	Evenly spaced
QTL position in genome	Random (uniform distribution)
QTL allele effect	Gamma distribution ($\beta = 0.4$)
Trait parameters	
Number of traits	6
Heritability	0.10, 0.30, 0.50
Phenotypic variance	1
Sex-limited trait	No
Population structure parameters	
Historical generation	
Phase 1	
Number of generations	500
Number of animals	Constant (500 males and 500 females)
Mating	Random
Phase 2	
Number of generations	500
Initial number of animals	1,000
Final number of animals	200 (100 males and 100 females)
Mating	Random
Phase 3	
Number of generations	10
Initial number	200 (100 males and 100 females)
Final number	3,000 (1,500 males and 1,500 females)
Mating	Random
Recent generation	
Number of generations	10
Reference population	9th
Validation population	9th and 10th
Number of offspring per dam	5
Founders	1,000 (200 males 800 females)
Mating	Random
Selection	BLUP
Cutting	BLUP
Replacement	20% females and 60% males
Overlapping generation	Yes
Generation 9–10 (predictability)	Correlation $\left(\sigma^2_{\text{gamete}}, \sigma^2_{\text{gamete\_estimated}}\right)$
Future generation	
Number of generations	10
Criterion of selection[1]	T_PTA = TRUE/2 or T_RPTA (TRUE/2) + σ^2_{gamete}
Number of offspring per dam	5 or 10
Replacement	100% females and 100% males
Better criterion	Genetic gain per generation

[1]T_PTA = true PTA; T_RPTA = true relative PTA; σ^2_{gamete} = variance of gametic diversity.

tio was maintained at 100% and the number of offspring per dam was 5 (corresponding to a selection intensity of 0.996 for females and 1.76 for males) or 10 (corresponding selection intensities of 1.4 for females and 2.06 for males). As the predicted σ^2_{gamete} is a latent variance, its realized value depends on the number of progeny of an individual. Any inference using this variance should be regarded as a bet (probability of an event considering the number of attempts). Therefore, the selection intensity applied to RPTA (i_f) may need to be adjusted accordingly, and 3 values of i_f (0.5, 0.8, and 1) were tested in this study.

Genomic Analysis

Because σ^2_{gamete} depends on the marker effects in genomic models, we used a model that assumed homogeneity of variance of marker effects, **GBLUP** (SNP-BLUP), and another model that allowed heterogeneity of marker effects with differential shrinkage through the improved Bayesian LASSO (**BLASSO**; Legarra et al., 2011). The analyses were performed using the GS3 v.3 software (Legarra et al., 2015). The model included the population mean, marker effects, and residual. Only markers with minor allele frequency (**MAF**) >0.05 were considered. For estimation of additive and residual variances, the simulated true values were used as initial values to reduce computational complexity, followed by 20,000 iterations with the burn-in of 2,000 initial chains.

Application of Gametic Variance to Real Data

The data used were part of the 2017 US genomic evaluations from the Council on Dairy Cattle Breeding (CDCB, Bowie, MD), consisting of 1,364,278 Holstein and 164,278 Jersey cattle from the national dairy cattle database. Five dairy traits based on up to 5 lactations were analyzed: milk (**MY**), fat (**FY**) and protein (**PY**) yields, and fat (**F%**) and protein (**P%**) percentages. The genotype data were generated from different SNP arrays with the number of SNP ranging from 7K to 50K. All individuals were imputed to a common panel of 60,671 SNP and their linkage phase were determined by FindHap version 3 (VanRaden et al., 2011). The σ^2_{gamete} was calculated using Equation [2] with estimated SNP effects $(\hat{\alpha}_1)$. The marker effects were derived from the PTA obtained from the genomic evaluation. Sex-specific recombination rates between SNP markers in Holstein and Jersey cattle were directly used in this study (Ma et al., 2015; Shen et al., 2018). Thus, a modification to the off-diagonal elements of the **M** matrix in Equation [2] was applied to incorporate recombination rate

$$\mathbf{M}_{jk} = al_{jk}\left(-\frac{rate_{jk}}{2} + 0.25\right),$$

when the recombination rate is <0.5; and $\mathbf{M}_{jk} = 0$ when the rate ≥0.5.

RESULTS AND DISCUSSION

Estimation of Gametic Variance with Genomic Models

The variance of progeny breeding values has been investigated in previous studies (Cole and VanRaden, 2011; Segelke et al., 2014; Bonk et al., 2016). Here, we sought to use simulation to evaluate the predictability of gametic variance as a parameter for selection. To evaluate the predictability, a comparison with classical simulation studies with genomic prediction was adopted. The variance of gametic diversity $\left(\sigma^2_{\text{gamete}}\right)$ was calculated considering both dependence and independence between loci, using all QTL and QTL with MAF ≥5%, and utilizing high-density SNP and sequence data with marker effects obtained from genomic models. The Pearson correlation between the true and estimated σ^2_{gamete} ranged from medium to high (Table 2), similar to those studies on breeding values (Meuwissen et al., 2001; Daetwyler et al., 2010; Clark et al., 2011). In general, the correlation increased when the heritability (h^2) of traits increased, whereas the same relation was not apparent when the number of QTL was large. Differently, for the GEBV prediction, the increase in accuracy has been reported with increased h^2 and for scenarios with a small number of QTL, particularly when these were estimated by differential shrinkage models (Daetwyler et al., 2010; Clark et al., 2011).

We observed higher correlations between the true and predicted σ^2_{gamete} using BLASSO compared with GBLUP in all scenarios (Table 2). These results were partly due to the small genome and large QTL effects simulated. Although GBLUP can have a similar or slightly better performance for prediction of GEBV than differential shrinkage models for scenarios with a large number of QTL (Daetwyler et al., 2010), the accuracy of the estimated marker effects, mainly for QTL regions, is greater from differential shrinkage models (Meuwissen et al., 2001; Shepherd et al., 2010; Legarra et al., 2011). For estimating σ^2_{gamete}, the marker effect has a greater impact than for GEBV prediction because

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.