Obtaining variance of gametic diversity with genomic models

D.J.A. Santos[1], J.B. Cole[2], P.M. VanRaden[2], T.J. Lawlor[3], H. Tonhati[1] & L. Ma[4]

[1] Universidade Estadual Paulista –FCAV- Departamento de Zootecinia, Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900, Jaboticabal, Brazil
daniel_jordan2008@hotmail.com (Corresponding Author)
[2]Animal Genomics and Improvement Laboratory, ARS, USDA –10300 Baltimore Ave, 20705, Beltsville- MD, USA
[3] Holstein Association USA- Holstein Place PO Box 808, 05301, Brattleboro-VT, USA
[4] University of Maryland - 8127 Regents Drive, 20742, College Park-MD, USA

Summary

Variance of gametic diversity () may be a useful tool for identifying matings with an above-average likelihood of producing progeny with extreme breeding values. The aim of this study was to show how this variance can be obtained from a statistical model and to verify the estimates of this variance for an individual receiving a routine genomic evaluation. An approach to obtain a normally distributed variance of all gametic values from the sums of the binomial variances of QTLs was employed. A small simulated genome was used to verify the adequacy of estimates of . For genomic evaluation, GBLUP and BLASSO models were used. BLASSO had better performance for estimation of . The results showed that markers with low MAF should be considered in analyses, as well the covariance (dependence) between the markers. Finally, SNP marker panels with medium to high-density may be sufficient for estimation.

Keywords: Mendelian sampling, heterozygozity, mating

Introduction

The availability of genomic information has supported the development of a variety of useful resources for breeding programs, such as genomic evaluation and the assessment of individual homozygosity (Kim et al., 2013). Although there is great concern about inbreeding/homozygosity in mating designs, traditionally in most breeding programs, only the breeding value has been used as a selection criterion. Even with genomic models (GM), the evaluation of mating and their future progeny are based only on expected values (parent averages) and disregard the variability of those values. Segelke et al., (2014) discussed potential uses of the variance of gamete values, and Bonk et al., (2016) showed how the exact within-family genetic variation could be calculated using data from phased marker genotypes. In this paper, we build on those results and show how this gametic variance can be obtained in a simplified way from a statistical point of view, and how that variance can be used for both mating design and individual selection. We also demonstrate how this variance can be estimated for any individual using results from a routine genomic evaluation.

Methods

Obtaining the variance of gametic diversity

In the following discussion we refer to the variance of gametic diversity (), which is equivalent to Mendelian sampling variance, because it is calculated as a function of the probabilities of recombination for the transmission of all QTLs for the gametes from one parent, without the need for validation by sampling progeny data. Suppose there are four independent QTLs, with alleles A|a, B|b, C|c, D|d and effects of allele substitutions equal to A = + 5, B = + 3, C = + 2, and D = + 1. An example of a mating with different types of genotypes among these loci is shown in Table 1 in order to verify all combinations of the parental genotypes.

It is easy to verify that the additive variance of the values of future progeny will be given by the sum of the parent's independently. Since we know all the possible gametes of the parents, and all possible combinations in the offspring, the variance of the population should be (). However, for large numbers of QTLs observing all gametic values and writing all possible combinations of those gametes is a difficult task. However, only heterozygous loci contribute to gametic variability, and those loci can be identified by inspection of parental genotypes. The transmission variance of a biallelic heterozygous locus *i* with an allelic substitution effect α_i can be calculated from the variance of a binomial distribution as , with probabilities of transmission *p* and *q* equal to 0.5 and the sampling number (*n*) equal to one. Thus, in the previous example, the could be calculated for the sire considering only the loci with heterozygous genotypes Bb and Dd as $0.25*(3^2+1^2) = 2.5$, and for the dam using only the genotypes Aa and Dd as $0.25*(5^2+1^2) = 6.5$, producing a progeny value of 2.5+6.5=9.

When two loci *i* and *j* are dependent the resulting variance can be obtained as , where , and p_i and p_j are always 0.5, and p_{ij} is the probability that two alleles of two loci are inherited together. That probability can be calculated knowing the linkage phase between the alleles and the recombination rate between these loci. The total variance can be obtained as the sum for all *n* heterozygous QTLs in the genome of an individual, where and . The value of can be easily computed as:

where *MP* is the (co)variance matrix of Mendelian probabilities among *n* loci. The diagonal of *MP* matrix is composed of elements equal to, and off-diagonal elements by). al_n is the result of multiplication of the alleles related to the p_{ij} frequency of one of the phases (maternal or paternal), encoding as -1 for allele 1, and as 1 for allele 2. Loci with genetic distances greater than 50 cM on the same chromosome, or between loci on different chromosomes, are considered to be independent. For genomic models, can be easily obtained by formula [1], where is the solution for the marker effect. Thus, for the GM the expression [1] will be close to that described by Bonk et al., (2016), to obtain the Mendelian variance for the additive effect.

Simulation and Genomic analysis

Simulations were performed using QMSim version 1.10 (Sargolzaei & Schenkel, 2009) and included three phases, with the first consisting of 500 generations

(constant size of 1,000 individuals), the second of an additional 500 generations (constant reduction from 1,000 to 200 individuals), and the third (expansion) phase included 10 generations (increasing from 200 to 3,000 individuals). 200 males and 800 females from the last generation were randomly selected as founders of the contemporary population, which consisted of 9 generations with 5 progeny per dam and a replacement rate of 20% for dams and 60% for sires. In the ninth generation, the empirical for all individuals was obtained from the estimated marker effects. The true was calculated from the effects of the simulated QTLs and their genotypes, as presented in equation [1]. Four traits were simulated with heritability of 0.1 and 0.3 combined with a number of QTLs of 20 and 200. The phenotypic variance was assumed to be 1 for all traits. Four replicates were run for each trait.

A small genome with four autosomal chromosomes of 50 cM was simulated. The QTL effects were based on a Gamma distribution (parameter $\beta = 0.4$). The mutation rate for marker and QTL were fixed at $2.5x10^{-5}$. The number of crossovers was sampled from a Poisson distribution. Originally, 200,000 markers were simulated and randomly distributed along the genome. However, a panel formed with 10% of the polymorphic markers sampled each 0.5 cM (HD panel), and another panel with 20% of the markers also sampled every 0.5 cM and all QTLs (SEQ panel) were used.

Since depends on the effects of the markers, we used genomic evaluation models without and with differential shrinkage (GBLUP and BLASSO, respectively). The analyses were performed using GS3 v.3 software (Legarra et al., 2015). In order to mimic a conventional genomic evaluation, only markers with MAF greater than 0.05 were considered. The model included the population mean, the markers effect, and the residual.

Results and discussion

was calculated considering dependence and independence (MP diagonal) between the loci, for all QTLs, for QTLs with MAF ≥5%, and for HD and SEQ panels using solutions obtained with GM. The Pearson correlation between the true and estimated ranged from medium to high (Table 2), indicating that estimates from GM may be useful tools to enhance selection programs. In general, correlations increased as h^2 increased, while such a relationship was not apparent for the number of QTL number. BLASSO had higher correlations among true and predicted than did GBLUP (Table 2) for all scenarios simulated. This result is expected and can be attributed to the more accurate estimation of QTLs effects by this model. In addition, GBLUP showed higher predicted bias (Table 3), with the less-desirable (higher) values of MSE and linear regression coefficients (b) farther from 1. The overestimation by GBLUP (values for b much less than 1) revealed the desirability of differential shrinkage estimators for the effect of many markers. However, the large differences observed between the models may be caused by the very small genome size simulated.

For the trait with $h^2 = 0.10$ and 20 QTL (Table 2), the correlations between obtained with all QTLs and with QTLs with MAF ≥5% were of medium-high magnitude, lower than that of other traits (high magnitude), resulting in lower correlations with the estimated by GMs. While this result may be attributable to allele frequency fluctuations in the historical population, it also implies that QTLs with low

MAF are important for obtaining accurate estimates of . This variance does not depend directly on population allele frequencies, but only on the individual's heterozygous state (allele carrier). Although MAF control (≥5%) is used to improve the prediction of GEBV (Uemoto et al., 2015), markers with low MAF may have greater linkage disequilibrium with low MAF QTLs, providing better predictions of progeny performance.

In order to facilitate the process of obtaining of in routine evaluations, the covariance (dependence) between markers was ignored and compared with the true . The correlations ranged in magnitude from medium to high using estimates obtained from QTLs, and from low to high magnitude when obtained by the GM (Table 2). However, the high correlation observed for one of the scenarios ($h^2 = 0.30$ and QTL = 20) can be attributed to the randomness of the QTLs distribution in the genome. Thus, the covariance between the markers should be considered for calculation of .

No differences in the correlations of the obtained with BLASSO were observed between the HD and SEQ scenarios (Table 2). This result shows that there is no need for QTL genotypes in the analyses and that panels with lower marker densities are sufficient. However, a decrease in correlation was observed for estimates obtained with GBLUP when the SEQ panel was used, regardless of the quantity of simulated QTLs. This, together with the increase in overestimation due to the increase in the number of markers (Table 3), confirms the preference for shrinkage model for estimation of .

In conclusion, this study verified the feasibility of obtaining by GM using HD panels without the need to use sequencing data. For improving the accuracy of the estimations, differential shrinkage models are preferred and markers with low MAF should be used, and the covariance (dependence) among markers should be considered.

List of References

Bonk S., M. Reichelt, F. Teuscher, D. Segelke & N. Reinsch, 2016. Mendelian sampling covariability of marker effects and genetic values. Genet. Sel. Evol., 48(46):1-11.

Kim E.S., J.B. Cole, H. Huson, G.R. Wiggans, C. P. Van Tassell, B.A. Crooker, G. Liu, Y. Da & T.S. Sonstegard, 2013. Effect of Artificial Selection on Runs of Homozygosity in U.S. Holstein Cattle. PLoS One, v. 8, e80813.

Legarra, A., A. Ricard & O. Filangi, 2015. GS3 Genomic Selection — Gibbs Sampling — Gauss Seidel (and BayesCπ).

Sargolzaei, M. & F. S. Schenkel, 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5): 680–681.

Segelke D., F. Reinhardt, Z. Liu & G. Thaller, 2014. Prediction of expected genetic variation within groups of offspring for innovative mating schemes. Genet. Sel. Evol., 46(42):1-10.

Uemoto Y., S. Sasaki, T. Kojima, Y. Sugimoto & T. Watanabe, 2015. Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese Black cattle. BMC Genet., 16(1):134.

List of Tables

Table 1. Scheme of a mating, with their genotypes for four independent QTL, additive values, mean and variance of the values of the gametes and of future offspring.

	Sire	Dam

Mating	Genotypes	AABbccDd	AaBBCCDd
	Value	14	16
Gametes	Genotypes	ABcD,ABcd,AbcD,Abcd	ABCD, ABCd, aBCD, aBCd
	Values	9,8,6,5	11,10,6,5
	Mean	7	8
	Variance	2.5	6.5
Offspring	Genotypes	ABcDABCD,ABcDABCd,ABcDaBCD,ABcDaBCd, ABcdABCD,ABcdABCd,ABcdaBCD,ABcdaBCd, AbcDABCD,AbcDABCd,AbcDaBCD,AbcDaBCd, AbcdABCD,AbcdABCd,AbcdaBCD,AbcdaBCd	
	Values	20, 19, 15, 14, 19, 18, 14, 13, 17, 16, 12, 11, 16, 15, 11, 10	
	Mean	15	
	Variance	9	

Here is considered A = + 5, B = + 3, C = + 2, and D = + 1 and a=b=c=d=0.

Table 2. Pearson correlation between gametic dispersion variance for all QTL (), for QTLs with maf≥0.05 () and disregarding the dependency for all QTL (), and QTLs with maf≥0.05 (), and their estimations using a high-density marker panel and sequencing data by genomic model GBLUP (bp) and BLASSO (ls), considering (and) and disregarding(and) the dependency of the markers.

		High-sensity panel				Sequencing data				QTLs data			
QTLs													
0.1	20	0.49	**0.56**	0.17	0.39	0.46	**0.57**	0.20	0.40	-	0.75	0.96	0.69
		0.53	**0.74**	0.21	0.54	0.48	**0.75**	0.25	0.55	0.75	-	0.66	0.93
		0.45	**0.53**	0.15	0.43	0.43	**0.53**	0.19	0.43	0.96	0.66	-	0.71
		0.50	**0.74**	0.18	0.61	0.45	**0.73**	0.24	0.61	0.69	0.93	0.71	-
	200	0.50	**0.60**	0.29	0.37	0.46	**0.61**	0.29	0.40	-	0.96	0.50	0.48
		0.48	**0.61**	0.29	0.39	0.45	**0.63**	0.30	0.41	0.96	-	0.46	0.49
		0.29	0.28	**0.51**	0.30	0.28	0.27	**0.48**	0.31	0.50	0.46	-	0.97
		0.27	0.29	**0.52**	0.32	0.26	0.29	**0.49**	0.33	0.48	0.49	0.97	-
0.3	20	0.64	**0.83**	0.28	0.66	0.59	**0.83**	0.07	0.65	-	0.94	0.95	0.90
		0.65	**0.87**	0.28	0.68	0.59	**0.87**	0.07	0.68	0.94	-	0.90	0.95
		0.60	**0.81**	0.30	0.69	0.54	**0.81**	0.07	0.68	0.95	0.90	-	0.95
		0.60	**0.85**	0.30	0.71	0.55	**0.85**	0.07	0.70	0.90	0.95	0.95	-
	200	0.63	**0.77**	0.25	0.49	0.59	**0.77**	0.29	0.48	-	0.95	0.55	0.52
		0.62	**0.78**	0.25	0.51	0.57	**0.78**	0.29	0.49	0.95	-	0.53	0.53
		0.42	0.48	0.52	**0.63**	0.40	0.49	0.54	**0.62**	0.55	0.53	-	0.99
		0.41	0.48	0.52	**0.63**	0.39	0.48	0.54	**0.63**	0.52	0.53	0.99	-

Values in bold represent the most accurate estimates using high-density marker panel and sequencing data.

Table 3. Mean squared prediction (MSE), intercept (a) and coefficient of the linear regression (b) between the gametic dispersion variance for QTL with maf≥0.05 and its estimation using a high density marker panel (HD) and sequencing data (SEQ) by genomic models (GBLUP and BLASSO).

Trait	Model	HD			SEQ		
QTLs		MSE	a	b	MSE	A	B

0.1	20	GBLUP	0.0014	**-0.0010**	0.27	0.0022	**-0.00033**	0.20
		LASSO	**8e-05**	0.0027	**1.20**	**8e-05**	0.00185	**1.26**
	200	GBLUP	0.0010	**0.0058**	0.23	0.0016	**0.00637**	0.18
		LASSO	**0.0001**	0.0074	**1.01**	**0.0001**	0.00681	**1.03**
0.3	20	GBLUP	0.0017	**-0.00697**	0.43	0.0028	**-0.00625**	0.35
		LASSO	**0.0002**	0.00282	**1.46**	**0.0002**	0.00247	**1.41**
	200	GBLUP	0.0021	0.00979	0.40	0.0035	0.01123	0.33
		LASSO	**0.0004**	**0.00945**	**1.14**	**0.0004**	**0.00950**	**1.13**

Values in bold represent the least biased estimates