A general approach for calculation of genomic relationship matrices for epistatic effects.

L. Varona[1], **Z.G. Vitezica**[2], **S. Munilla**[1,3], **A. Legarra**[2]
[1]Universidad de Zaragoza, Zaragoza, España. [2] INRA, UMR1388, Toulouse, France. [3] Universidad de Buenos Aires, Buenos Aires, Argentina.

ABSTRACT: The aim of this study is to present a general procedure to calculate, from SNP markers, the covariances between individuals due to additive, dominant and epistatic effects, e.g. "additive x dominant genomic relationships". The proposed method expands the orthogonal approach called NOIA and does not assume Hardy-Weinberg equilibrium. It is thus applicable to, e.g., crosses. A real mice data set was used to illustrate its implementation. Estimated variance components show that epistatic interactions may explain an important portion of the overall genetic variability for some traits of interest, such as growth speed. Some of the potential applications of the procedure within the genomic selection scope are briefly discussed.
Keywords: Dominance; Epistasis; Genomic Selection

Introduction

Since the advent of massive genotyping platforms, genomic evaluation models usually fit marker additive effects, either explicitly (Meuwissen et al., 2001; Van Raden, 2008; De los Campos et al., 2009) or implicitly through the "genomic" relationship matrix (Van Raden, 2008; Goddard, 2009). However, it is possible that dominance or higher order interaction terms play an important role in the genetic determinism of some traits of economic interest in livestock or plants. The existence of non-negligible interactions is supported by the wide application of crossbreeding as a breeding strategy. Further, it is known that the use of assortative mating can improve the performance of livestock and crop traits when dominance or epistasis is present (Toro and Varona, 2010).

In livestock populations, one of the main reasons why dominance or higher order interaction terms have not been estimated is that pedigree relationships are not enough informative. Recently, genomic data have renewed the interest in the prediction of non-additive genetic effects (Su et al., 2012; Vitezica et al., 2013; Nishio and Satoh, 2014), because it is much easier to work with dominance in view of heterozygote genotypes at the individuals.

Higher order interaction terms can be also modeled by using the "genomic" relationship approach, but a general framework for calculation of relationships matrices is needed. In particular, cited developments apply to populations in Hardy-Weinberg equilibrium but not to populations like F1, backcrosses or three-way crosses. In this study, we develop a completely general procedure to estimate "genomic" relationship matrices for interactions terms of any order expanding the natural and orthogonal (NOIA) approach (Álvarez-Castro and Carlborg, (2007)) within the scope of the covariance between individuals and we discuss some of the potential applications of the procedure.

Materials and Methods

Theory. A model including additive, dominant and higher order interaction terms can be written as:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ta} + \mathbf{Xd} + \sum_{i=A,D}\sum_{j=A,D}\mathbf{K}_{ij}\mathbf{o}_{ij} + \sum_{k=A,D}\sum_{i=A,D}\sum_{j=A,D}\mathbf{M}_{ij}\mathbf{p}_{ij} + \ldots + \mathbf{e}$$

where **y** is the vector of phenotypic records, **a** and **d** are the vector of additive and dominant effects, $\mathbf{o}_{ij}$ is vector the *ijth* second order epistatic effect and $\mathbf{p}_{ijk}$ is the *ijkth* third order epistatic effect and so on. The epistatic effects can be subdivided according to the interaction involving breeding values (A) or dominance deviations (D). For instance, there are three types of the *ij* second order epistatic effect: AxA, AxD and DxD. Further, **T**, **X**, $\mathbf{K}_{ij}$ and $\mathbf{M}_{ijk}$ are incidence matrices that include the covariates that link the genetic effects with the phenotypic records. In fact, there are several available parameterizations for these covariates. For instance, for the additive and dominant effects and following Su et al. (2012) the elements of **T** and **X** can be defined as 1, 0, -1 and 0, 1, 0 for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. An alternative parameterization was proposed by Vitezica et al. (2013) that define the elements of **T** and **X** as (2-2p), (1-2p), -2p and $-2q^2$, 2pq and $-2p^2$ for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Further, Vitezica et al. (2013) proved that both approaches are equivalent and identified the effects of the first approach as the "biological" additive and dominant effects and, for the second, as the "statistical" substitution effects and dominant deviations.

With higher order epistatic effects, the coefficients included in matrices $\mathbf{K}_{ij}$ and $\mathbf{M}_{ijk}$ can be also defined under several approaches (Cockerham, 1954). Furthermore, Álvarez-Castro and Carlborg (2007) defined a general framework to obtain orthogonal estimates of genetic effects using different reference points (e.g. a particular genotype). Under a simple additive and dominant model for one locus A, the coefficients of **T** and **X** can be calculated by the following locus genetic-effect design matrix:

$$S_A = \begin{bmatrix} 1 & -p12_A - 2p22_A & \frac{-2p12_A p22_A}{p11_A + p22_A - (p11_A - p22_A)^2} \\ 1 & 1 - p12_A - 2p22_A & \frac{4p11_A p22_A}{p11_A + p22_A - (p11_A - p22_A)^2} \\ 1 & 2 - p12_A - 2p22_A & \frac{-2p11_A p12_A}{p11_A + p22_A - (p11_A - p22_A)^2} \end{bmatrix}$$

where $p11_A$, $p12_A$ and $p22_A$ are the genotypic frequencies for the genotypes A_1A_1, A_1A_2 and A_2A_2 for the locus A. Note that under the assumption of Hardy-Weinberg equilibrium this design matrix reduces to the "statistical" approach of Vitezica et al. (2013), but it also applies to populations *not* in such equilibrium. Following Álvarez-Castro et al. (2007), the coefficients for second order epistatic effects between locus A and B can be calculated by the Kronecker product between design matrices S_A and S_B:

$$S_{A\times B} = S_A \otimes S_B$$

and, subsequently, for third and higher order epistatic effects as:

$$S_{A\times B\times C} = S_A \otimes S_B \otimes S_C \ldots$$

However, it should be noted that the definition of a model to estimate second or higher order epistatic effects involves a quadratic or cubic increase of the number of effects to be estimated. To solve that problem, the "genomic" relationship matrix approach (VanRaden, 2008; Goddard, 2009) allows to develop a model that include only a genetic effect of each additive, dominant or epistatic term for each individual, after the definition of an appropriate "genomic" relationship matrix for each one. From the output of the S_{AxB} design matrix, the "genomic" (co) variance relationship matrices will be computed as:

$$\mathrm{cov}(X) = \frac{\mathbf{hh'}}{\mathrm{var}(h)}\sigma_X^2 = \mathbf{G}_X\sigma_X^2$$

where X take the values A, D, AxA, AxD and DxD in the case of second order epistatic effects. In the case of the additive effects for the locus A, the elements of the **h** vector for the *ith* individual are:

$$h_i = \begin{cases} -p12_A - 2p22_A & for \quad A_1A_1 \\ 1 - p12_A - 2p22_A & for \quad A_1A_2 \\ 2 - p12_A - 2p22_A & for \quad A_2A_2 \end{cases}$$

And for the dominance matrix are:

$$h_d = \begin{cases} \dfrac{-2p12_A p22_A}{p11_A + p22_A - (p11_A - p22_A)^2} & for \quad A_1A_1 \\ \dfrac{4p11_A p22_A}{p11_A + p22_A - (p11_A - p22_A)^2} & for \quad A_1A_2 \\ \dfrac{-2p11_A p12_A}{p11_A + p22_A - (p11_A - p22_A)^2} & for \quad A_2A_2 \end{cases}$$

Further, the elements of the vector **h** for the *ijth* additive x additive combination of loci A and B is calculated from the elements the corresponding column of S_{AxB} matrix as:

$$h_{ij} = \begin{cases} (-p12_A - 2p22_A)(-p12_B - 2p22_B) & for \quad A_1A_1B_1B_1 \\ (-p12_A - 2p22_A)(1 - p12_B - 2p22_B) & for \quad A_1A_1B_1B_2 \\ (-p12_A - 2p22_A)(2 - p12_B - 2p22_B) & for \quad A_1A_1B_2B_2 \\ (1 - p12_A - 2p22_A)(-p12_B - 2p22_B) & for \quad A_1A_2B_1B_1 \\ (1 - p12_A - 2p22_A)(1 - p12_B - 2p22_B) & for \quad A_1A_2B_1B_2 \\ (1 - p12_A - 2p22_A)(2 - p12_B - 2p22_B) & for \quad A_1A_2B_2B_2 \\ (2 - p12_A - 2p22_A)(-p12_B - 2p22_B) & for \quad A_2A_2B_1B_1 \\ (2 - p12_A - 2p22_A)(1 - p12_B - 2p22_B) & for \quad A_2A_2B_1B_2 \\ (2 - p12_A - 2p22_A)(2 - p12_B - 2p22_B) & for \quad A_2A_2B_2B_2 \end{cases}$$

where $p11_B$, $p12_B$ and $p22_B$ are the genotypic frequencies for the genotypes B_1B_1, B_1B_2 and B_2B_2 for the locus B. Similarly, the elements of the **h** vector for additive x dominant and dominant x dominant matrices can be calculated from the elements of the appropriate column of S_{AxB}.

Given the availability of the "genomic" relationship matrices, the following linear model can be assumed:

$$\mathbf{y} = \mathbf{Wb} + \mathbf{Zg}_A + \mathbf{Zg}_D + \mathbf{Zg}_{AA} + \mathbf{Zg}_{AD} + \mathbf{Zg}_{DD} + \mathbf{Sc} + \mathbf{e}$$

where **b** is the vector of systematic effects, $\mathbf{g}_X$ is the vector of the X genetic effect (A, D, AA, AD and DD), **c** is the vector of any other random environmental effect and **e** is the vector of the residuals. Further, **W, S** and **Z** are the incidence matrices. The $\mathbf{g}_X$ and **c** vectors are assumed to follow a multivariate Gaussian distribution with the appropriate covariance matrix, calculated as described:

$$\mathbf{g}_X = N(\mathbf{0}, \mathbf{G}_X\sigma_X^2) \qquad \mathbf{c} = N(\mathbf{0}, \mathbf{I}\sigma_c^2)$$

Mice Data analysis: Legarra et al. (2008) analyzed phenotypes of mice data (http://mus.well.ox.ac.uk/mouse/HS/), composed by 1884 phenotypic records for different traits (such as growth speed and body length) and including 10,946 markers. We have reanalyzed the data for the two traits, with the same model of estimation as Legarra et al. (2008) that included a general mean and a random cage effect. Estimation was performed by a Bayesian approach using flat priors for the variance components and a Gibbs sampling algorithm. A single long chain of 100,000 iterations was run, after discarding a burn-in period of 25,000 iterations.

Results and Discussion

The results of the ratios of variance explained by the additive, dominant, additive x additive, additive x dominant and dominant x dominant effects are presented in Table 1.

Table 1. Ratios of additive, dominant, additive x additive, additive x dominant and dominant x dominant variances.

	h_a^2	h_d^2	h_{axa}^2	h_{axd}^2	h_{dxd}^2
Growth Speed	0.022 (0.014)	0.012 (0.008)	0.037 (0.030)	0.058 (0.049)	0.060 (0.047)
Body Length	0.109 (0.031)	0.008 (0.006)	0.038 (0.032)	0.046 (0.038)	0.061 (0.048)

Additive genetic variance explains most of the genetic variation for body length, whereas the epistatic genetic effects represent a relevant percentage of variation for growth speed. However, the posterior standard deviation for epistatic variance components is much higher, as reflected in Figure 1. The additive and dominance genetic variances were well estimated and results agree with Vitezica et al. (2013).

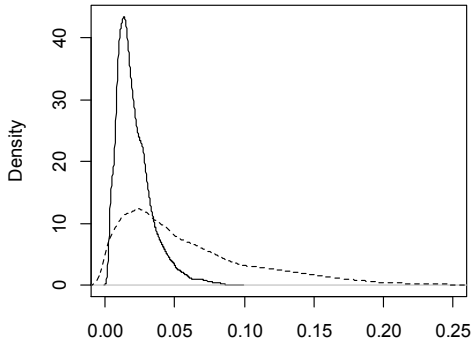

Figure 1. Posterior densities ratios for the additive (solid line) and additive x dominant (dotted line) variance for growth speed.

The use of non-additive genetic effects in animal breeding is mainly devoted for the exploitation of heterosis in cross-breeding schemes. In this context, several approaches have been proposed to obtain prediction of the performance of candidates to selection, mimicking the classical reciprocal recurrent selection approach (Kinghorn et al., 2010; Zeng et al., 2013).

The distribution of the genetic variance into additive, dominant and interaction terms is strongly determined by the allelic or genotypic frequencies, and redistribution of variance between them is expected when alternative genotypic frequencies are used. In this study, "genomic" variance component were calculated by using the observed population genotypic frequencies as reference for the calculation of the **S** design matrix for each locus. However, the NOIA model proposed by Alvarez-Castro and Carlborg (2007) allow for an easy transformation of those matrices into alternative reference points R_2, by the application of the following equation:

$$\mathbf{S}_{R2} = \mathbf{S}_{R1} - \mathbf{P}_{R2}\mathbf{S}_{R1}\mathbf{I}^*$$

where, $\mathbf{S}_{R2}$ is the design matrix under the new reference point for every locus, $\mathbf{S}_{R1}$ is the previous design matrix, I* is the identity matrix with the first scalar replaced by zero, $\mathbf{P}_{R2}$ is the change of reference matrix for the new reference point, which takes the form:

$$\mathbf{P}_{R2} = \begin{pmatrix} p^*_{11} & p^*_{12} & p^*_{22} \\ p^*_{11} & p^*_{12} & p^*_{22} \\ p^*_{11} & p^*_{12} & p^*_{22} \end{pmatrix}$$

where p^*_{11}, p^*_{12} and p^*_{22} are the genotypic frequencies for the new reference point. Further research must be done to explore the possibilities of transformation for the "genomic" (co) variance matrices into alternative reference points, and, thus, to obtain predictions of the additive and non-additive genetic effects under the scope of a cross-breeding scheme. Also, the accuracy of these predictions must be investigated.

Conclusion

A procedure for calculation of "genomic" relationship matrices for second and higher order epistatic effects is presented. The procedure was applied to a mice data set, where a relevant percentage of variance was assigned to epistatic genetic effects for growth speed.

Acknowledgements

Luis Varona and Sebastian Munilla acknowledge the financial support of Spanish AGL2010-15903 and UE FP7-289592-GENE2FARM grants. Zulma Vitezica and Andrés Legarra acknowledge the financial support of X-Gen and GenSSeq actions from SelGen metaprogram (INRA).

Literature Cited

Álvarez-Castro, J. M. and Carlborg, O. (2007). *Genetics*. 176:1151-1167.

Cockerham, C. C. (1954). *Genetics* 39:859-882.

De los Campos, G., Naya, H., Gianola, D. et al. (2009). *Genetics* 182:375-385.

Goddard, M. E. (2009). *Genetica* 136:245-257.

Kinghorn, B. P., Hickey, J. M, Van der Werf, J.H.J. (2010). 9th WCGALP.

Legarra, A., Robert-Granié, C., Manfredi, E. et al. (2008). *Genetics* 180:611-618

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. (2001). *Genetics* 157:1819-1829.

Nishio, M. and Satoh, M. (2014). *Plos ONE* 9(1): e85792.

Su, G. Christensen, O. F., Ostersen, T. et al. (2012). *Plos ONE* 7(9):e45293.

Toro, M. and Varona, L. (2010). *Genet. Sel. Evol.* 42:33.

Van Raden, P. M. (2008). *J. Dairy Sci.* 91:4414-4423.

Vitezica, Z. G., Varona, L. and Legarra. A. (2013). *Genetics* 195:1223-1230.

Zeng, J., Toosi, A., Fernando, R. L. et al. (2013). *Genet. Sel. Evol.* 45:11