


# Beyond genomic selection: The animal model strikes back (one generation)!

R.J.C. Cantet<sup>1,2</sup>  | C.A. García-Baccino<sup>1</sup> | A. Rogberg-Muñoz<sup>1,3</sup> | N.S. Forneris<sup>1</sup> | S. Munilla<sup>1</sup>

<sup>1</sup>Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina

<sup>2</sup>Instituto de Investigaciones en Producción Animal (INPA) UBA-CONICET, Buenos Aires, Argentina

<sup>3</sup>Instituto de Genética Veterinaria (IGEVEV), Facultad de Ciencias Veterinarias, Universidad Nacional de La Plata (UNLP) – CONICET, La Plata, Provincia de Buenos Aires, Argentina

## Correspondence

R.J.C. Cantet, Departamento de Producción Animal, Facultad de Agronomía, UBA, Buenos Aires, Argentina.

Email: rcantet@agro.uba.ar

## Funding information

FONCyT PICT, Grant/Award Number: 2013-1661; UBACyT, Grant/Award Number: 20020150100230B/2016; PIP CONICET, Grant/Award Number: 833/2013

## Summary

Genome inheritance is by segments of DNA rather than by independent loci. We introduce the *ancestral regression* (AR) as a recursive system of simultaneous equations, with grandparental path coefficients as novel parameters. The information given by the pedigree in the AR is complementary with that provided by a dense set of genomic markers, such that the resulting linear function of grandparental BV is uncorrelated to the average of parental BV in the absence of inbreeding. AR is then connected to segmental inheritance by a *causal* multivariate Gaussian density for BV. The resulting covariance structure ( $\Sigma$ ) is Markovian, meaning that conditional on the BV of parents and grandparents, the BV of the animal is independent of everything else. Thus, an algorithm is presented to invert the resulting covariance structure, with a computing effort that is linear in the number of animals as in the case of the inverse additive relationship matrix.

## KEYWORDS

breeding value, causal inference, Gaussian Markov density, genomic data, segmental inheritance

## 1 | INTRODUCTION

From 1986 to 1990, Professor Daniel Gianola supervised the PhD of the first author at the University of Illinois. His teaching was insightful and deep; his guidance was essential to transmit to his students and postdocs a rigorous and critical thinking. We wish him the best in the occasion of his retirement, which will probably mean a slow reduction in office hours but not in papers (as breeders need him to keep us going in the right direction!).

Speaking of that, we introduce here the ancestral regression (AR) model: a generalization of the “parental regression” from the animal model, with extra parameters to complement the information given by the pedigree with that provided by a dense set of genomic markers. The AR

is well suited to be fitted in large data sets where few animals are genotyped, because it does not rely on a reference population but on the classic idea of “identity by descent” (Malecot, 1969) at the level of genome segments. The resulting distribution of breeding values (BV) remains to be Gaussian, and the covariance structure (which will be denoted with the Greek letter  $\Sigma$  throughout the manuscript) is Markovian; that is conditional on the BV of parents and grandparents, the BV of the animal is independent of everything else. Moreover,  $\Sigma$  can be inverted by an extension of the rules of Henderson (1984) for the additive relationship matrix. For reasons of space, we defer to a further publication all formal arguments and detailed derivations of the covariance between relatives and asymptotic normality of BV under AR.

## 1.1 | Causal distribution for breeding values

We seek for a distribution of BV (or vector  $a$ , i.e., the sum of all additive effects across the genome of an individual) that accounts for the inheritance of the genome across generations by segments of random length rather than by independent loci. In the regular animal model (Henderson, 1984), the Gaussian density represents the Markov process by which the individual genome and its BV are generated by the addition of randomly formed half-parental BV plus a Mendelian segregation residual ( $\phi$ ), that is a “genetic error” term. Recent population genetics literature (Kelleher, Etheridge, Véber, & Barton, 2016; Matsen & Evans, 2008) differentiates between this *expected* process and the *realized* genetic process that results from sampling from a given (or fixed) pedigree (Wakeley et al., 2012). The specification of inheritance in the distribution of BV relates to the ancestral contributions from parents and further ancestors to the genome of the individual. In the classic setting, each genetic contribution of a parent to the genome of a progeny is one half, and each grandparental contribution is one quarter, and in general  $2^{-G}$  where  $G$  is the number of meioses from the ancestor to the descendant. However, there are ancestors that leave genealogical descendants but no genetic material (Chang, 1999; Kelleher et al., 2016; Matsen & Evans, 2008). This difference between ancestral contributions to the individual genome is the result of recombination, linkage and linkage disequilibrium (LD). For populations ranging in  $N_e$  (effective population size) from 50 to 1000, the fractions of genome shared IBD by the ancestor with each descendant could be close to zero (Chang, 1999) after 6–10 generations, respectively. Following this reasoning, the realized contribution of each grandparent to the genome of the individual could differ from one quarter, that is greater or smaller, thus being a random variable with expectation equal to 0.25. As a consequence, to model inheritance of genome segments across generations, the distribution must take into account both the expected relationship or genomic contribution as well as Mendelian segregation that makes grandparental contributions to differ across meioses.

As the actual number and the location of the genes controlling the variability of a complex trait is uncertain, identifiability of BV in the animal model (Henderson, 1984) historically rested on information from the covariance structure between related individuals (Kempthorne, 1954). Let the covariance matrix of  $a$  be equal to  $\sum \sigma_A^2$ , where  $\Sigma$  is the covariance structure, and  $\sigma_A^2$  is the additive variance. We will refer to element  $\sum_{XY}$  of  $\Sigma$  as the genome-wide relationship (GWR) between animals X and Y. Formally

$$\sum_{XY} = P_C(g_X = g_Y | g_X \leftarrow g_C \rightarrow g_Y) = P(g_X \equiv g_Y) \quad (1)$$

where  $g_X$  and  $g_Y$  are random variables representing the genomes of X and Y, and  $g_C$  is the genome of all common ancestors of both animals. The arrows indicate the passing of genome material from C to X and Y. Hence,  $P(g_X \equiv g_Y)$  is the fraction of genome shared IBD between X and Y (Guo, 1995) and symbolized as ‘ $\equiv$ ’. Indeed, GWR is the “realized or observed” relationship that results from the combined action of the process that produces the “expected relationship”  $A_{XY}$  (or additive relationship, Henderson, 1984), plus the Mendelian segregation. Overall, for individuals X and Y with BV in  $a$ , the covariance between their BV is equal to

$$\begin{aligned} \text{cov}(a_X, a_Y) &= \text{cov}(a_X, a_Y | g_X \equiv g_Y) P(g_X \equiv g_Y) \\ &\quad + \text{cov}(a_X, a_Y | g_X \not\equiv g_Y) [1 - P(g_X \equiv g_Y)] \\ &= \text{Var}(a_i) P(g_X \equiv g_Y) + 0 [1 - P(g_X \equiv g_Y)] \\ &= \sigma_A^2 P(g_X \equiv g_Y) \end{aligned} \quad (2)$$

The value  $P(g_X \equiv g_Y)$  is the joint probability:

$$P(g_X \equiv g_Y) = P(X_1 \equiv Y_1, X_2 \equiv Y_2, \dots, X_n \equiv Y_n) \quad (3)$$

where  $X_i$  and  $Y_i$  are random variables representing the event that the additive effect from gene variant  $i$  is carried by animals X and Y, respectively. In general,  $n$  is large and unknown. Calculating the joint density (3) is a daunting task, so that (3) is approximated by relaxing some assumptions. For example, all models employed so far to predict BV assume *independent additive effects* such that

$$P(g_X \equiv g_Y) = P(X_i \equiv Y_i) \quad (4)$$

The equality in (4) indicates that the fraction of genome shared IBD has the same probability at any causal loci. After scaling the variances of individual additive effects that give rise to  $\sigma_A^2$ , the probability in (3) is equal to the “expected relationship”  $2P(X_i \equiv Y_i) = 2P(X \equiv Y) = A_{XY}$ . Hence, identifiability of BV under non-independent additive effects requires approximating the Mendelian segregation process such that  $\sum_{XY} \approx P(g_X \equiv g_Y)$ . For pairs of linked loci, Weir and Cockerham (1969) attributed to J.B.S. Haldane the inequality

$$P(X_1 \equiv Y_1, X_2 \equiv Y_2) > P(X_1 \equiv Y_1) P(X_2 \equiv Y_2) \quad (5)$$

The meaning of (5) is that the effect of linkage is to increase the probabilities of IBD at both loci. Under segmental inheritance and the Mendelian genetic process, the conditions under which the orthogonal decomposition of genetic effects rests are violated. Under LD, additive effects from genes located in the same segment will be correlated and will pick variability connected with additive-by-additive effects (Mäki-Tanila & Hill, 2014). This suggests that relationships calculated with  $P(X_1 \equiv Y_1, X_2 \equiv Y_2)$  would recover a larger fraction of  $\sigma_A^2$  in a segmental (or non-independent) inheritance model. Fisher

(1949) developed a theory of “junctions”: sites in the genome that separate segments originated in different ancestors; or sites where recombination took place. The process models inheritance of DNA encompassing non-independent causal variants. In sum, the distribution of additive effects is undefined by number and location of additive effects (identification problem) and the complex random behaviour introduced by the processes of Mendelian segregation under linkage and LD.

A well-defined point of departure is to model the inheritance distribution by conditioning on the BV of the immediate ancestors, that is parents and grandparents, and departing from Hardy–Weinberg-random mating assumptions. A *causal distribution* (Pearl, 2000) provides a structure and a set of assumptions from which a joint distribution can be computed. In our terms, the goal is to predict the value of a vector of BV  $a_t$  at time  $t$ , given the BV of ancestors ( $a_{t-1}$ ) measured on a previous time ( $t - 1$ ), by means of the conditional probability distribution  $P(a_t|a_{t-1})$ . Once the pedigree is observed (i.e., fixed), mating is not random and the resulting distribution is  $P(a_t|a_{R,t-1})$ ;  $R$  is the set of ancestral BV responsible for the make-up of the genomes at time  $t$ . The parameters of the conditional distribution  $P(a_t|a_{t-1})$  are identifiable at the *individual* level and reflect the “expected relationship”: the regression coefficients of the parental BV on the BV of offspring are equal to 0.5 (Bulmer, 1985; Henderson, 1984). However, parameters of  $P(a_t|a_{R,t-1})$  are not only the two halves associated with parental contributions, but also the differential contributions of grandparental genomes whose coefficients differ from 0.25. This new set of parameters is *identifiable only* with information from a dense set of SNPs, CNVs or DNA sequences, as detection of segments requires locating recombination sites. If we take a multivariate Gaussian density as the causal distribution of  $a$  (a fact that is described in detail elsewhere),  $P(a_t|a_{R,t-1})$  represents a *structural regression* model (Kiiveri, Speed, & Carlin, 1984; Wermuth, 1980) for the vector  $a$  of order  $q$  (the number of animals), which includes the BV of all individuals from all generations. Parameters of  $P(a_t|a_{R,t-1})$  are *directional* or *causal* path coefficients (Wright, 1934). The direction is related to the fact that additive effects from genes that affect BV flow always from ancestors to descendants and never the opposite direction. Hence, ancestral BV “cause” descendants BV in causal inference terms (Pearl, 2000).

The difference between a regression coefficient  $\beta_{XY}$  and a causal path coefficient  $\beta_{XY.O}$  from a multivariate normal distribution of BV is seen as follows:

$$\beta_{XY} = \frac{\text{cov}(a_Y, a_X)}{\text{Var}(a_X)} \quad \beta_{XY.O} = \frac{\sum_{YX} - \sum_{YO} \sum_{OO}^{-1} \sum_{OX}}{\sum_{XX} - \sum_{XO} \sum_{OO}^{-1} \sum_{OX}} \quad (6)$$

The subscripts in  $\Sigma$  relate to individual  $X$  (parent),  $Y$  (progeny), and a set of individuals ( $O$ ) related to  $Y$  and  $X$ . The additive variance  $\sigma_A^2$  cancels out from the parameters in (6) because it appears in the numerator and the denominator:

$$\begin{aligned} \beta_{XY.O} &= \frac{\sum_{YX} \sigma_A^2 - \sigma_A^2 \sum_{YO} \sum_{OO}^{-1} (\sigma_A^2)^{-1} \sum_{OX} \sigma_A^2}{\sum_{XX} \sigma_A^2 - \sigma_A^2 \sum_{XO} \sum_{OO}^{-1} (\sigma_A^2)^{-1} \sum_{OX} \sigma_A^2} \\ &= \frac{\sum_{YX} - \sum_{YO} \sum_{OO}^{-1} \sum_{OX}}{\sum_{XX} - \sum_{XO} \sum_{OO}^{-1} \sum_{OX}} \end{aligned}$$

The covariance in the numerator and the variance in the denominator are adjusted for everything else such that the set  $O$  of BV is marginalized in the estimation process. The matrix representation of the Gaussian recursive linear system is the following:

$$a = Ba + \phi \quad (7)$$

where  $\phi$  is the vector of Mendelian residuals (Bulmer, 1985). Matrix  $B$  is lower triangular with nonzero elements being path coefficients relating the BV of ancestors to descendants. Solution of (7) is equal to

$$a = (I - B)^{-1} \phi \quad (8)$$

BV in (8) have expected value and variance, respectively, equal to

$$\begin{aligned} E(a) &= (I - B)^{-1} E(\phi) = 0 \\ \text{Var}(a) &= (I - B)^{-1} D (I - B')^{-1} \sigma_A^2 \end{aligned} \quad (9)$$

We are interested in the covariance structure of  $a$ :

$$\Sigma = (I - B)^{-1} D (I - B')^{-1} \quad (10)$$

The covariance structure of Mendelian residuals is the diagonal matrix  $D$ . The structure  $\Sigma$  is positive definite as 1)  $(I - B)$  is positive definite (see Lemma 2.1, chapter 6 of Berman & Plemmons, 1994), and 2) the diagonals of  $D$  are all positive real numbers. Then,  $(I - B)^{-1}$  exists and it is equal to (Quaas, 1988)

$$(I - B)^{-1} = I + \sum_{k=1}^{\infty} B^k = I + B + B^2 + B^3 + \dots \quad (11)$$

As  $(I - B)^{-1}$  exists, the unique solution to (7) is (8).

A remarkable property of the Gaussian causal distribution  $P(a_t|a_{R,t-1})$  is the Markov recursion to parental and/or ancestral BV. Moreover, in conditional Gaussian densities, zero elements in  $\Sigma^{-1}$  indicate conditional independence (Wermuth, 1980) and enable the joint density function to be decomposed into the product of conditional distributions (Lauritzen & Wermuth, 1989; Speed & Kiiveri, 1986). The result facilitates computation of  $\Sigma^{-1}$  with large number of animals for prediction purposes. The linear system that results from this setting is referred to as the Bartlett decomposition (Munilla & Cantet, 2012; expression (4)). It has

been employed to represent inheritance of BV across generations by Cantet, Schaeffer, and Smith (1992), expression [A9]. In the current framework, the Bartlett decomposition of the covariance matrix of BV is as follows:

$$\begin{aligned} \text{Var}(a) &= \sum \sigma_A^2 = \begin{bmatrix} \sum_{t=1}^{t-1} & \sum_{t=1}^{t-1,t} \\ \sum_{t=1}^{t-1,t'} & \sum_t \end{bmatrix} \sigma_A^2 \\ &= \begin{bmatrix} \sum_{t=1}^{t-1} & \sum_{t=1}^{t-1} B_t' \\ B_t \sum_{t=1}^{t-1} & B_t \sum_{t=1}^{t-1} B_t' + \text{Var}(\phi_t) \end{bmatrix} \sigma_A^2 \end{aligned} \quad (12)$$

where the subscripts  $t-1$  and  $t$  indicate the covariance structures of ancestral and offspring generations, respectively.

### 1.1.1 | Phasing and probabilities of IBD under linkage

We assume that the four genomes or haplotypes from two individuals (X and Y) are phased, so they can be expressed as a mosaic of segments originated in their eight grandparents. Aligning the four parental haplotypes in X and Y in the four possible pairings ( $i = 1$ , father of X – father of Y;  $i = 2$ , father of X – mother of Y;  $i = 3$ , mother of X – father of Y and  $i = 4$ , mother of X – mother of Y) allows us to estimate marginal values of  $\sum_{XY}$ , which we denote as  $r_{XY}$ .

For the causal distribution to account for linkage and LD, IBD probabilities have to be computed by segments rather than for independent loci. As a first step, we propose to detect IBD segments using software such as Beagle (Browning & Browning, 2013). Let  $l_{ij}$  be the length of segment  $ij$  shared IBD between X and Y and measured in cM or Megabases. Subscript  $j$  runs from 1 to  $N_{IBD_i}$ , the last segment detected IBD for pair  $i$  of haplotypes,  $i = 1, \dots, 4$  as in the previous paragraph. At each value of  $j$ , there may be subsegments of IBD from two individuals. Once segments are detected to be identical, we calculate the probability of IBD between segments,  $P(S_X \equiv S_Y)$ , by multiplying  $l_{ij}$  by the probability of IBD sharing between one grandparent of X and one grandparent of Y, and then summing up to the last IBD segment. The linkage distribution is tractable if the dependence structure of the segments is modelled using probabilities of IBD at two loci. Weir and Cockerham (1969) and Cockerham and Weir (1973) defined three different probabilities for linked genes at two loci according to the number of gametes involved (2, 3, or 4), as follows:

Digametic:  $P(S_X \equiv S_Y) = P(X_1 \equiv Y_1, X_2 \equiv Y_2)$

Trigametic:  $P(S_X \equiv S_Y, S_X \equiv S_W) = P(X_1 \equiv Y_1, X_2 \equiv W_2)$

Tetragametic:  $P(S_X \equiv S_Y, S_W \equiv S_Z) = P(X_1 \equiv Y_1, W_2 \equiv Z_2)$

The capital letters to the left indicate the individuals the gametes belong. Digametic probabilities require expansion

to di-, tri- and tetragametic probabilities, whereas trigametic probabilities expand in tri- and tetragametic probabilities. Finally, tetragametic probabilities are functions only of tetragametic probabilities. All in all

$$P(g_X \equiv g_Y) = \sum_{i=1}^4 \sum_{j=1}^{N_{IBD_i}} l_{ij} P(X_{i1} \equiv Y_{i1}, W_{i2} \equiv Z_{i2}) \quad (13)$$

where  $P(X_{i1} \equiv Y_{i1}, W_{i2} \equiv Z_{i2})$  is one of the three gametic probabilities previously described. On dividing (13) by the total length of genome ( $L$ ), we obtain

$$r_{XY} = \frac{\sum_{i=1}^4 \sum_{j=1}^{N_{IBD_i}} l_{ij} P(X_{i1} \equiv Y_{i1}, X_{i2} \equiv Y_{i2})}{L} \quad (14)$$

## 1.2 | The ancestral regression

Our goal is to provide for a genetic model where BV of parents and grandparents are uncorrelated. As a result, the covariance structure of the process is Markovian and direct inversion of  $\Sigma$  is possible using an algorithm that is linear in the number of animals ( $O(q)$ ), as Henderson (1984) did with the animal model or “parental regression.” By conditioning on the number of generations ( $G_{MRCA}$ ) to the most recent common ancestor (MRCA) of a pair of animals, Palamara, Lencz, Darvasi, and Pe’er (2012) obtained the distribution of the length of a non-recombinant segment shared IBD between two individuals. If the pedigree is known,  $G_{MRCA} = G$  or the exact number of generations between the animals. Under this coalescent specification, Palamara et al. (2012) observed that “individual segments carry little information about specific sites unless the common ancestor is extremely recent (e.g., <10 generations before present)” that is  $G \leq 10$ . This figure agrees with the theoretical results derived by Chang (1999) who observed that at approximately  $1.77 \log_2(N_e)$  generations all partial ancestry ends. Consequently, for populations with  $N_e$  ranging from 50 to 1,000 (the range of most domestic animal populations), fractions of genome shared IBD in descendants induce the covariance between BV to be practically null in 6 to 10 generations. We compared estimation of GWR with simulated and real data using genomic information, without or with the aid of pedigree records. Whereas both methods produced estimates of GWR that were almost unbiased, the IBD based method (pedigree and markers) had higher correlation (0.956) with true or realized GWR than the method that does not employ the pedigree (0.678) (Forneris et al., 2016). In a related research, García-Baccino et al. (2016) found that the method that does not use pedigree information had larger variance of estimation of half-sib GWR than the method that employs genomic and pedigree data. Additionally, Kumar, Feldman, Rehkopf, and Tuljapurkar (2016) observed that GWR estimated



using solely marker information is a function of the random matrix of sampled SNP that has large numbers of very small singular values. They also found that estimates of heritability using such GWR were biased and had large variance. Beside the sampling errors to detect relationship, methods to estimate GWR using information from markers *unconstrained* by the pedigree (identity in state) tend to pick up distant (6–10 meiosis or more) relationships, that is  $GWR \approx 0.01\text{--}0.001$ . These convey very little information on the BV of the distant relative and increase prediction error variance. Therefore, a possible prediction model in which pedigree and genomic information are complementary would be based on pedigree—to fix the parental path coefficients, which are exactly equal to 0.5—and on genomic data to carry information on the Mendelian process through the path coefficients going from the BV of the grandparents to the BV of the animal. Within the framework of  $P(a_i|a_{i-1})$ , these additional “regressors” should reduce the variance of the Mendelian residual ( $\text{Var}(\phi_X)$ ) only if the average of the BV of the parents is uncorrelated to a linear function of grandparental BV. Cantet and Vitezica (2014) proved that the  $\text{Var}(\phi_X)$  arising from the prediction of BV with methods that combine pedigree and genomic information is always less than (or at best equal to)  $\text{Var}(\phi_X)$  with  $P(a_i|a_{i-1})$ .

The problem is then to find a set of conditions such that the grandparental path coefficients are *identifiable* (Kenny, 1979). This occurs when the number of parameters is less or equal to the number of distinct elements of  $\Sigma$  or correlations among pairs of individuals:  $0.5 q(q-1)$ . More formally, the recursive Gaussian model is as follows:

$$a_X = 0.5 a_S + 0.5 a_D + \beta_{X,PGS,R} a_{PGS} + \beta_{X,PGD,R} a_{PGD} + \beta_{X,MGS,R} a_{MGS} + \beta_{X,MGD,R} a_{MGD} + \phi_X \quad (15)$$

To estimate  $\beta_{X,GP|P}$  where  $GP = GS, GD$ , and  $P = S, D$ , we will take advantage of a theorem by Cochran (1938), expressing the total regression coefficient from  $a_{GP}$  to  $a_X$  as

$$\begin{aligned} \beta_{X,GP} &= \beta_{X,GP|P} + \beta_{X,P|GP} \beta_{P,GP} = \beta_{X,GP|P} + (0.5)(0.5) \\ &= \beta_{X,GP|P} + 0.25 \end{aligned}$$

A similar expression is obtained from the other grandparent of X. If we add both expressions and equate the result to the parental contribution (equal to 0.5), we obtain

$$0.50 = \beta_{X,GS|P} + \beta_{X,GD|P} + 0.50$$

Consequently, the following *over-identification restrictions* (Kenny, 1979, pp. 42–49) are needed on the grandparental path coefficients:

$$\beta_{X,PGS|R} = -\beta_{X,PGD|R} = \beta_S \quad \beta_{X,MGS|R} = -\beta_{X,MGD|R} = \beta_D \quad (16)$$

Positive values of  $\beta_S$  and  $\beta_D$  in (16) mean that the paternal and maternal grandsires, respectively, are in excess, whereas negative values indicate that the granddams are in excess. To obtain the “AR”, we have to include in (15) the path coefficients under restrictions (16) as follows:

$$a_X = 0.5 a_S + 0.5 a_D + \beta_S (a_{PGS} - a_{PGD}) + \beta_D (a_{MGS} - a_{MGD}) + \phi_X \quad (17)$$

The path coefficients for the grandparents are defined as

$$\begin{aligned} \beta_S &= \frac{\sum_{X,PGP} - \sum_{PGP,R} \sum_{RR}^{-1} \sum_{R,PGP}}{\sum_{PGP,PGP} - \sum_{PGP,R} \sum_{RR}^{-1} \sum_{R,PGP}} \\ \beta_D &= \frac{\sum_{X,MGP} - \sum_{MGP,R} \sum_{RR}^{-1} \sum_{R,MGP}}{\sum_{MGP,MGP} - \sum_{MGP,R} \sum_{RR}^{-1} \sum_{R,MGP}} \end{aligned} \quad (18)$$

with

$$\begin{aligned} a_{PGP} &= [0.5 (a_{PGS} - a_{PGD}) | 0.5 a_S \rightarrow a_X] \\ a_{MGP} &= [0.5 (a_{MGS} - a_{MGD}) | 0.5 a_D \rightarrow a_X] \end{aligned}$$

Thus,  $a_{PGP}$  and  $a_{MGP}$  are half the difference between grandsire and granddam from each parent given the passing of 0.5 parental BV to X.

### 1.2.1 | Relationship with the covariance structure

Consider calculation of the row of  $\Sigma$  related to animal X. Let R be a set comprised by individuals S, D, PGS, PGD, MGS and MGD. Any row of matrix  $B$  has at most six elements different from zero and  $q-6$  zeros. Depending on date of birth, BV of grandparents and parents are ordered in  $a$  such that the row of  $B$  for X may look like

$$B_X = [0 \quad \beta_S \quad 0 \dots 0 \quad -\beta_S \quad \dots 0 \dots 0 \quad \beta_D \quad -\beta_D \quad 0.5 \quad 0 \quad \dots 0 \quad 0.5 \quad \dots 0] \quad (19)$$

The diagonal element in (19) is always zero because it corresponds to the BV of animal X. Because of the values of the path coefficients in (17) or (19), parental BV are uncorrelated with the linear functions in (18),  $(I-B)$  is non-singular and of rank  $q$ , such that  $\Sigma = (I-B)^{-1} D (I-B')^{-1}$  is positive definite. The diagonal element of the Mendelian covariance matrix for animal X is equal to  $D_X = \text{Var}(\phi_X) (\sigma_A^2)^{-1}$ . Let  $\sum_R$  be the causal covariance structure among the parents and grandparents of X, thus a square matrix of order 6. Then, using the Bartlett decomposition of the covariance structure (12) and the following “reduced” version  $B_{X(r)} = [\beta_S \quad -\beta_S \quad \beta_D \quad -\beta_D \quad 0.5 \quad 0.5]$ , we can calculate the relationships between X and its parents and grandparents as follows:

$$\begin{bmatrix} \sum_R & \sum_R B_{X(r)}' \\ B_{X(r)} \sum_R & B_{X(r)} \sum_R B_{X(r)}' + D_X \end{bmatrix} \quad (20)$$

The variance of the Mendelian residual of  $X$  is equal to

$$\begin{aligned}\text{Var}(\phi_X) &= \text{Var}(a_X - B_{X(r)} a_R) \\ &= [(1 + F_X) + B_{X(r)} \sum_R B'_{X(r)} \\ &\quad - 2B_{X(r)} \sum_R B'_{X(r)}] \sigma_A^2\end{aligned}$$

where  $F_X$  is the inbreeding coefficient of  $X$  calculated as  $0.5 r_{SD}$ . Thus

$$\text{Var}(\phi_X) = [(1 + F_X) - B'_{X(r)} \sum_R B_{X(r)}] \sigma_A^2 = D_X \sigma_A^2 \quad (21)$$

### 1.2.2 | Estimation of $\beta_S$ and $\beta_D$

In the framework of the Gaussian causal model of inheritance discussed so far, it is possible to use the Bartlett decomposition of  $\Sigma$  to write the multivariate normal distribution of the vector of BV  $a$  as follows:

$$f(a) = f(a_R, a_O, a_X) = f(a_O, a_X | a_R) f(a_R) \quad (22)$$

where  $a_X$  is the BV of  $X$ ,  $a_R$  are the BV of the grandparents and parents of  $X$  and  $a_O$  are the BV of all other individuals which are neither  $X$  nor the parents and grandparents of  $X$ . Now, we take advantage of the Markov property (Speed & Kiiveri, 1986) of this Gaussian density and, conditional on  $a_R$ ,  $a_X$  is independent of  $a_O$  and (22) is written as:

$$f(a) = f(a_O | a_R) f(a_X | a_R) f(a_R) \quad (23)$$

Hence, all statistical information on  $\beta_S$  and  $\beta_D$  is contained on the normal density  $f(a_X | a_R)$ , and sufficient statistics for  $\beta_S$  and  $\beta_D$  belong to this distribution. A closer look to (23) suggests that the conditional covariances

$$\begin{aligned}\sum_{X(PGS-PGD)|R} &= \sum_{X,PGP} - \sum_{PGP,R} \sum_{RR}^{-1} \sum_{R,PGP} \\ \sum_{X(MGS-MGD)|R} &= \sum_{X,MGP} - \sum_{MGP,R} \sum_{RR}^{-1} \sum_{R,MGP}\end{aligned}$$

are sufficient for  $\beta_S$  and  $\beta_D$  (Lauritzen & Wermuth, 1989). Hence, using the difference of GWR between  $X$  and each grandparent calculated with (14) produces estimates of the conditional covariances of the differences:

$$\begin{aligned}\hat{\sum}_{X(PGS-PGD)|R} &= r_{XPGS|R} - r_{XPGD|R} \\ \hat{\sum}_{X(MGS-MGD)|R} &= r_{XMGS|R} - r_{XMGD|R}\end{aligned} \quad (24)$$

Under the causal distribution, the expected value of both estimators are

$$E\left(\hat{\sum}_{X(GS-GD)|R}\right) = 0.25 + \beta_P - 0.25 - (-\beta_P) = 2\beta_P$$

Then, moment estimators of  $\beta_S$  and  $\beta_D$  respectively are equal to

$$\begin{aligned}\hat{\beta}_S &= 0.5[r_{XPGS|R} - r_{XPGD|R}] \\ \hat{\beta}_D &= 0.5[r_{XMGS|R} - r_{XMGD|R}]\end{aligned} \quad (25)$$

Once the values of  $\beta_S$  and  $\beta_D$  are estimated, the row  $B_X$  associated with animal  $X$  is formed as in (19) by setting to the following: (i) zero: all elements in the columns pertaining to individuals in  $a$  that do not belong to the set  $R$ ; the last element or position of  $X$  is also equal to zero; (ii)  $\beta_S$ ,  $\beta_D$ ,  $-\beta_S$  and  $-\beta_D$ : in the columns of the grandparents; 3) 0.5: in both parental columns.

### 1.3 | Algorithms for calculating $\Sigma^{-1}$

The algorithm is a variant of the one given by Henderson (1984) to calculate  $A^{-1}$ . Let  $(I - B)_X$  be the row of  $(I - B)$  related to  $X$ , and  $D_X^{-1}$  the inverse of the corresponding diagonal element of  $D$ . As noted by Quaas (1988) for  $A^{-1}$ , we can compute  $\Sigma^{-1}$  as follows:

$$\Sigma^{-1} = \sum_{i=1}^q (I - B)'_i (I - B)_i (D_{ii}^{-1}) \quad (26)$$

The first step is to set up the pedigree in an array composed by the individual, both parents and the four grandparents; ordered by date of birth. Then, do the following:

- (i) calculate the GWR,  $r_{ij}$ , using (14);
- (ii) compute  $\Sigma^{-1}$  on a row by column basis starting from the first individual (oldest) and ending with the last one (youngest);
  - a estimate  $\beta_S$  and  $\beta_D$  for the grandparents and set up  $B_X$  as in (19);
  - b calculate  $D_X$  using (21);
  - c update the “reduced”  $\Sigma$  (order  $7 \times 7$ : grandparents, parents and the individual being processed) using (20);
  - d expand  $B_X$  with zeros in all remaining elements in the row and obtain

$$\begin{aligned}(I - B)_X &= [0 \quad -\beta_S \quad \beta_S \quad \dots 0 \dots 0 \quad -\beta_D \\ \beta_D - 0.5 \quad \dots -0.5 \quad 0 \dots 0 \quad 1 \quad 0 \dots 0];\end{aligned}$$

- e add the contributions of the matrix in (26), that is

$$(I - B)'_X (I - B)_X (D_X^{-1});$$

to the corresponding elements of  $\Sigma^{-1}$ ;

- f go back to 2)a).

A small data set with 10 animals is presented in supplementary materials to exemplify calculating  $\Sigma^{-1}$ .

## 2 | DISCUSSION

The contribution of this research is to obtain a quantitative genetic model to predict BV, the AR (17), that accounts for the expected relationship, and part of the Mendelian segregation process (recombination, linkage and LD) when BV are formed. The AR reduces Mendelian residual variance (Cantet & Vitezica, 2014) and increases the individual accuracy of BV, when compared to the regular animal model. If there is no inbreeding, the reduction of Mendelian variance in outbreds would be bounded by 0.125 for either  $\beta_S$  or  $\beta_D$  to a minimum of  $D_X = 0.25$ , but somewhat more under inbreeding. The AR provides for a unified framework where the contributions of pedigree and genomic information become complementary; that is, BV of parents and the function of grandparental BV defined in (18) are uncorrelated. The model is then Gaussian Markovian (Lauritzen & Wermuth, 1989; Speed & Kiiveri, 1986), which means that conditional on the BV of parents and grandparents, the individual BV is independent of every other BV in  $a$ . In the regular infinitesimal model where the “expected GWR” is calculated with the pedigree, the covariance between the BV of relatives goes to zero when the number of generations between the two individuals ( $G$ ) increases, such that the speed of convergence to normality is a function of  $2^{-G}$ . In the segmental inheritance model, the genetic process induces the covariance structure of BV to go to zero at the faster rate  $2^{-G}Q_G$ , where  $Q_G$  is the expected fraction the genome from the ancestor present in the descendant (after Lemma 8 in Matsen & Evans, 2008).

The utility of AR goes beyond segmental inheritance (non-independent loci), where it is most advantageous in terms of recovering a larger fraction of  $\sigma_A^2$  and can also be employed with independent loci using other estimators of GWR than (14). A distinctive advantage of the AR is that  $\Sigma^{-1}$  is computed in a linear fashion  $O(q)$  as  $A^{-1}$  by a simple extension of the rules of Henderson (1984), and direct inversion is not required. A less advantageous aspect of AR relates to the observation that prediction of BV with an IBD derived covariance structure did not gain much more accuracy, with respect to a method that employs only marker data to estimate the covariance structure (0.02–0.03, Forneris et al., 2016). Additionally, two grandparental parameters enter into the model for each animal and have to be estimated with all implications of such procedures.

Research is underway to estimate grandparental path coefficients when grandparents are not genotyped but related individual, non-ancestors and contemporaneous are. It would be tempting to include path coefficients related to great-grandparents or further ancestral generations to reduce the Mendelian variance of the parental regression. However, adding the BV from great-grandparents or from ancestors further back will most likely induce correlations

among the estimated path coefficients and overfitting of BV. This is troublesome as  $D_X$  has to be smaller than one for  $\text{Var}(\phi_X) > 0$  such that  $\Sigma > 0$ . When discussing the effect of multicollinearity in models used for prediction purposes, Shmueli (2010) quoted the following comment in a textbook by S.G. Makradakis, S.C. Wheelwright and R. J. Hyndman, R. J., for prediction purposes “multicollinearity is not a problem unless either (i) the individual regression coefficients are of interest, or (ii) attempts are made to isolate the contribution of one explanatory variable to  $Y$ , without the influence of the other explanatory variables.” Assumptions (i) and (ii) are applicable to prediction of BV using genomic information because 1) matrices  $\Sigma$  and  $G$  (VanRaden, 2008) are estimates of the true covariance structure of the inheritance process. Therefore, those estimates will have variance around the true value and the individual “regression coefficients” (i.e., the ancestral  $\beta$ ) will have an effect on prediction error variance of the individual, thus affecting the accuracy of different animals and, consequently, selection response with those predictions; 2) the AR accounts for the contributions of the different “explanatory variables,” in our context the BV of ancestral and related animals in the equations of relatives, as long as the model remains causal and Markovian. That is not the case with matrix  $G$  that generally has neither properties. The Markov property in AR is due to orthogonalizing the grandparental from the parental contributions. Under the absence of inbreeding, the orthogonalization produces that the grandparental extra contributions do not inflate the variance (Höskuldsson, 1994) of the  $\beta$  coefficients even though BV of parents and grandparents are correlated. Hence, the orthogonalization in the AR allows for a “balance between the improvement in fit, and the increase in the model uncertainty” (Höskuldsson, 1994). As a consequence, AR is more suited to be fitted into a continuous and traditional evaluation because of the complementarity between the pedigree and genomic information that generalize the regular animal model. Moreover, a large “reference population” of genotyped animals is not needed for a single individual to gain in accuracy, but rather a group of ancestors or related individuals.

A direction for future research is on the search for efficient algorithms to calculate the GWR in (14), as well as those for estimating  $\beta_S$  and  $\beta_D$ . The conditions under which these latter parameters are identifiable when grandparents are not genotyped are another issue for future research.

## ACKNOWLEDGEMENTS

We are grateful to Andrés Legarra, Zulma Vitezica, Graciela Boente and Matías Schrauf for their helpful comments. This research was funded by grants FONCyT PICT

2013-1661, UBACyT 20020150100230B/2016 and PIP CONICET 833/2013.

## REFERENCES

- Berman, A., & Plemmons, R. J. (1994). *Non-negative matrices in the mathematical sciences*. Philadelphia, USA: SIAM.
- Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194, 459–471.
- Bulmer, M. G. (1985). *The mathematical theory of quantitative genetics*. Oxford: Clarendon Press.
- Cantet, R. J. C., Schaeffer, L. R., & Smith, C. (1992). Reduced animal model with differential genetic grouping for direct and maternal effects. *Journal of Animal Science*, 70, 1652–1660.
- Cantet, R. J. C., Vitezica, Z. G. (2014). Properties of Mendelian residuals when regressing breeding values using a genomic covariance matrix. 10th WCGALP. Vancouver.
- Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Advances in Applied Probability*, 31, 1002–1026.
- Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Journal of the Royal Statistical Society*, 5, 171–176.
- Cockerham, C. C., & Weir, B. S. (1973). Descent measures for two loci with some applications. *Theoretical Population Biology*, 4, 300–330.
- Fisher, R.A. (1949). *The theory of inbreeding*. Edinburgh: Oliver and Boyd.
- Forneris N. S., Steibel J. P., Legarra A., Vitezica Z. G., Bates R. O., Ernst C. W., Basso A. L., & Cantet R.J. (2016). A comparison of methods to estimate genomic relationships using pedigree and markers in livestock populations. *Journal of Animal Breeding and Genetics*, 133, 452–462.
- García-Baccino C. A., Munilla S., Legarra A., Vitezica Z. G., Forneris N. S., Bates R. O., Ernst C. W., Raney N. E., Steibel J. P., & Cantet R. J. C. (2016). Estimates of the actual relationship between half-sibs in a pig population. *Journal of Animal Breeding and Genetics*, 134(2), 109–118.
- Guo, S. W. (1995). Proportion of genome shared identical by descent by relatives: Concept, computation, and applications. *American Journal of Human Genetics*, 56, 1468–1476.
- Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Guelph: University of Guelph.
- Höskuldsson, A. (1994). The H-principle: Tutorial. New ideas, algorithms and methods in applied mathematics and statistics. *Chemometrics and Intelligent Laboratory Systems*, 23, 1–28.
- Kelleher, J., Etheridge, A. M., Véber, A., & Barton, N. H. (2016). Spread of pedigree versus genetic ancestry in spatially distributed populations. *Theoretical Population Biology*, 108, 1–12.
- Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London. Series B*, 143, 102–113.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience.
- Kiiveri, H., Speed, T. P., & Carlin, J. B. (1984). Recursive causal models. *Australian Mathematical Society. Journal. Series A*, 36, 30–52.
- Kumar, S. K., Feldman, M. W., Rehkopf, D. H., & Tuljapurkar, S. (2016). Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences of the United States of America*, 113, E61–E70.
- Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17, 31–57.
- Mäki-Tanila, A., & Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198, 355–367.
- Malecot, G. (1969). *The mathematics of heredity*. San Francisco: Freeman.
- Matsen, F. A., & Evans, S. N. (2008). To what extent does genealogical ancestry imply genetic ancestry? *Theoretical Population Biology*, 74, 182–190.
- Munilla, S., & Cantet, R. J. C. (2012). Bayesian conjugate analysis using a generalized inverted Wishart distribution accounts for differential uncertainty among the genetic parameters – an application to the maternal animal model. *Journal of Animal Breeding and Genetics*, 129, 173–187.
- Palamara, P. F., Lencz, T., Darvasi, A., & Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91, 809–822.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. NY, USA: Cambridge University Press.
- Quaas, R. L. (1988). Additive genetic model with groups and relationships. *Journal of Dairy Science*, 71, 1310–1318.
- Shmueli, G. (2010). To Explain or to Predict? *Statistical Sciences*, 25, 289–310.
- Speed, T. P., & Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14, 138–150.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91, 4414–4423.
- Wakeley J., King L., Low B. S., & Ramachandran S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, 190, 1433–1445.
- Weir, B. S., & Cockerham, C. C. (1969). Pedigree mating with two linked loci. *Genetics*, 61, 923–940.
- Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *The Journal of the American Statistical Association*, 75, 963–972.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Cantet RJC, García-Baccino CA, Rogberg-Muñoz A, Forneris NS, Munilla S. Beyond genomic selection: The animal model strikes back (one generation)!. *J Anim Breed Genet*. 2017;134:224–231. <https://doi.org/10.1111/jbg.12271>