A Mendelian Sampling Model for Genetic Prediction

Clément Carré [1,2], Fabrice Gamboa[2], David Cros[3], Gregor Gorjanc[4] and Eduardo Manfredi[1]
[1] *INRA, UR 631 SAGA, F-31326 Castanet-Tolosan, France*
[2] *IMT - Université Paul Sabatier, Toulouse, France*
[3] *AGAP - CIRAD, Montpellier, France*
[4] *Animal Science Department, Biotechnical Faculty, University of Ljubljana, Slovenia*

Abstract

Genetic prediction for complex traits is usually based on models including individual or marker effects. Alternatively, models can include both the individual and the marker effects. In particular, we studied a model combining effects for base individuals, realized Mendelian sampling in descendants and marker effects. The predictive ability of this model, measured as the correlation between true (simulated) and predicted genetic values, was similar to that of the marker model. As expected, the Mendelian sampling model was worthwhile when markers captured a low fraction of total genetic variance.

Key words: Genetic prediction, Genomic selection, SNP, Mendelian sampling

Introduction

Most models used for genetic prediction and genetic evaluation include "individual" or "marker" data to represent genetic effects. Carré *et al.* (2013) studied a third group of models including both "marker" and "individual" effects. We recall this model and we discuss its predictive ability, relative to the usual marker model. Finally, originality, limits and possible extensions of the model are discussed.

Standard models for genetic prediction

Without DNA data, "individual effects" are used to represent additive genetic effects (e.g., Henderson, 1975):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{e} \qquad [1]$$

$\boldsymbol{y}$ is a vector of phenotypes
$\boldsymbol{\mu}$ is a constant vector
$\boldsymbol{Z}$ is an incidence matrix
$\mathbf{u}$ is a vector of individual effects, with $\mathrm{Var}(\mathbf{u}) = \mathbf{A}\sigma^2_u$, with $\mathbf{A}$ being the relationship matrix amongst individuals.
$\mathbf{e}$ is a vector of residuals, with $\mathrm{Var}(\mathbf{e}) = \mathbf{I}\sigma^2_e$, with $\mathbf{I}$ being an identity matrix

A further usual assumption is $\mathrm{Cov}(\mathbf{e}, \mathbf{u}) = \mathbf{0}$.

When there is only one phenotype per individual $\mathbf{Z}=\mathbf{I}$ and the only information available to distinguish individual effects from residuals is given by the coefficients in the relationship matrix $\mathbf{A}$ which depend on genetic transmission data (diploids originate from two parental gametes) and the assumed known variance of the unobserved Mendelian sampling effects (Quaas, 1976; Henderson, 1976).

With molecular data available, prediction models evolved to include markers in the model (Meuwissen *et al.*, 2001):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{ZWm} + \mathbf{e} \qquad [2]$$

where:

$\mathbf{m}$ is a vector of "marker effects"
$\mathbf{W}$ is a matrix of marker genotypes. With biallelic markers such as SNP, usual elements of $\mathbf{W}$ are 0, 1 or 2, the number of, say, the allele "1" of the marker genotype.

In genomic selection models, the simplest assumed (co)variances are:

$\mathrm{Var}(\mathbf{m}) = \mathbf{I}_{N_m}\sigma^2_m$, with N_m being the number of markers, and $\mathrm{Cov}(\mathbf{e}, \mathbf{m}) = \mathbf{0}$

If it is further assumed that $\mathbf{u} = \mathbf{Wm}$ and $\mathrm{Var}(\mathbf{u}) = \mathbf{WW}'k\sigma^2_m$, k being a scaling

factor, it is possible to compute predictions with the individual model [1], amended such that the relationship matrix $\mathbf{A}$ is replaced by the realized "genomic relationship" matrix $\mathbf{G} = \mathbf{WW}'$ (VanRaden, 2008; Goddard, 2009). Application of BLUP to this model has been termed "genomic BLUP" and improvements have been proposed to make assumptions more realistic (departures from the homogeneous variances for marked effects in model [2]) and practical implementations when only part of the individuals are genotyped making necessary to mix the $\mathbf{A}$ and the $\mathbf{G}$ matrices for the combined analyses of individuals with or without genotypes (e.g. Aguilar *et al.*, 2011).

Based on analytical developments (Gianola *et al.* (2009) and experimental evidence (Yang *et al.*, 2011; De los Campos *et al.*, 2009; Duchemin *et al.*, 2012), Carré *et al.* (2013) claimed that the alternative assumptions $\mathbf{u} \neq \mathbf{Wm}$ and $\mathrm{Var}(\mathbf{u}) \neq \mathbf{WW}'\mathrm{k}\sigma^2_{\mathrm{m}}$ are likely in an outbred population, and they studied a model including individual and marker effects:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e} \qquad [3]$$

with assumptions:

$\mathrm{Var}(\mathbf{u}) = \mathbf{R}\sigma^2_{\mathrm{u}}$, and
$\mathrm{Cov}(\mathbf{e}, \mathbf{u}) = \mathrm{Cov}(\mathbf{u}, \mathbf{m}) = \mathbf{0}$,

where $\mathbf{R}\sigma^2_u$ is the (co)-variance matrix of individual effects. Usually, as in model [1], $\mathbf{R} = \mathbf{A}$, the additive relationship matrix computed theoretically from genealogy data. Note that the terms in model [3] are redundant if it is assumed that $\mathbf{u} = \mathbf{Wm}$.

Mendelian segregation model

Starting from model [3], Carré *et al.* (2013) developed a model where the individual effect of a descendant is a function of individual effects of its ancestors (individuals in the base) and Mendelian sampling which can be traced by DNA data:

$$\mathbf{u_d} = \mathbf{Du_b} + (\mathbf{W_d} - \mathbf{DW_b})\mathbf{m} \qquad \mathbf{[4]}$$

where:

$\mathbf{u_d}$ and $\mathbf{u_b}$ represent the effects of descendants and base individuals
$\mathbf{W_d}$ and $\mathbf{W_b}$ represent the genotypes of descendants and base individuals
$\mathbf{D}$ is a genetic transmission matrix

Base individuals are defined for a given genealogy as the most distant known ancestors of individuals with recorded phenotypes, i.e., they do not have phenotypes and their parents are unknown.

Using [4] in model [3] gives:

$$\mathbf{y_d} = \boldsymbol{\mu} + \mathbf{Z}_{\mathrm{d}}\mathbf{Du}_{\mathrm{b}} + \mathbf{Z}_{\mathrm{d}}\,(\mathbf{W}_{\mathrm{d}} - \mathbf{DW}_{\mathrm{b}})\mathbf{m} + \mathbf{Z}_{\mathrm{d}}\,\mathbf{W}_{\mathrm{d}}\mathbf{m} + \mathbf{e} \qquad \mathbf{[5]}$$

where the phenotypes of descendants $\mathbf{y_d}$ are the sum of the effects of base individuals (second term), realized Mendelian sampling effects (third term) and marker effects (fourth term). In the term $\mathbf{Z}_{\mathrm{d}}\mathbf{Du}_{\mathrm{b}}$, $\mathbf{Z}_{\mathrm{d}}$ relates records to individuals (descendants d) and $\mathbf{D}$ relates individuals to ancestor effects $\mathbf{u}_{\mathrm{b}}$ via simple coefficients of genetic transmission (including consanguinity, i.e., multiple contributions of an ancestor to an individual).

This approach opens further possibilities for modeling, according to the context of prediction. If previous knowledge on marker effects is available a possible "disjoint model" is:

$$\mathbf{y_d} = \boldsymbol{\mu} + \mathbf{Z}_{\mathrm{d}}\mathbf{Du}_{\mathrm{b}} + \mathbf{Z}_{\mathrm{d}}\,(\mathbf{W}_{\mathrm{d1}} - \mathbf{DW}_{\mathrm{b1}})\mathbf{m1} + \mathbf{Z}_{\mathrm{d}}\,\mathbf{W}_{\mathrm{d2}}\mathbf{m2} + \mathbf{e} \qquad \mathbf{[6]}$$

where **m2** represents markers of QTL and **m1** the rest of the markers.

In the absence of previous knowledge, an alternative is the "embedded model":

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_{\mathrm{d}}\mathbf{Du}_{\mathrm{b}} + \mathbf{Z}_{\mathrm{d}}\,(2\mathbf{W}_{\mathrm{d}} - \mathbf{DW}_{\mathrm{b}})\mathbf{m} + \mathbf{e} \qquad \mathbf{[7]}$$

The term $\mathbf{Z}_{\mathrm{d}}\,(2\mathbf{W}_{\mathrm{d}}\text{-}\mathbf{DW}_{\mathrm{b}})\,\mathbf{m}$ in [7] groups two parts: $\mathbf{Z_d}\,\mathbf{(W_d\text{-}DW_b)}\,\mathbf{m}$, the realized mendelian sampling effects, and $\mathbf{Z_dW_dm}$ which represents the direct relations between markers and phenotypes.

Relative predictive ability of the embedded Mendelian sampling (MS) model

Carré *et al.* (2013) studied by simulation the predictive ability of the embedded MS model (MS model as in [7]), relative to that of the marker model (M model as in [2]). Predictive ability was the correlation between true (simulated) and predicted individual effects.

Six contexts of prediction were studied: high and low marker density (200 and 2000 markers per chromosome of 1 Morgan) combined with the fraction of genetic variance explained by markers (90, 50 and 10%). Prediction was for a trait with overall heritability of 0.4. The studied population is described by Carré *et al.* (2013). Each prediction context was replicated 200 times using the QMSim software (Sargolzaei and Frenkel, 2009). Unknowns of the compared models, marker (M) and Mendelian sampling (MS), were obtained with the BLUP method, under the following assumptions for model MS:

$$\mathbf{u_b} \sim N(0, I\,\sigma_u^2)$$
$$\mathbf{m} \sim N(0, I\,\sigma_m^2)$$
$$\mathbf{Cov(u_b, m)=0} \qquad [8]$$

Mean accuracies over 200 replicates when using 2000 SNP markers are presented in Figure 1 for 10, 50 and 90% of total genetic variance explained by QTL. Accuracies were highest (0.76 for model M and 0.74 for model MS) in the training data when the genetic variance explained by QTL was high (90%). The lowest correlations occurred for the test data under scenario 10% (0.36 for M vs. 0.40 for MS). The MS model gave the best predictions when the infinitesimal effects were important (scenario 10%) and model M gave the best predictions when QTL effects represented 90% of genetic variance. Differences between mean accuracies of two models were small and non-significant ($P<0.05$).

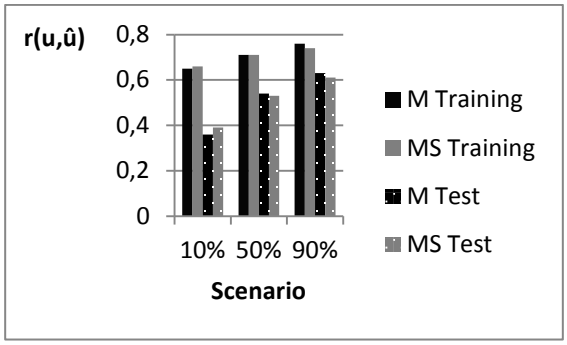

Figure 1. Accuracy of the Marker (M) and Mendelian segregation (MS) models for the three simulation scenarios with 10, 50, or 90% of the total genetic variance explained by QTL. From Carré *et al.* (2013)

When fewer markers were used (200 SNP per chromosome), all accuracies were lowered but the methods ranked as when using more markers (2000 SNP per chromosome; Table 1). The accuracy of the MS model was 12% higher than that of the M model for the scenario with the 10% of genetic variance explained by QTL and 5% lower when the QTL explained the 90% of total variance.

Table 1. Performance of the Mendelian sampling model: relative accuracies in the training and the test data*.

Simulated scenario	Training data (%)[a]	Test data (%)[a]
QTL variance 10%		
200 SNP markers	103	112
2000 SNP markers	102	108
QTL variance 50%		
200 SNP markers	100	100
2000 SNP markers	100	98
QTL variance 90%		
200 SNP markers	99	95
2000 SNP markers	97	97

[a](%) is 100 times the ratio between the average accuracy under the Mendelian segregation model and the average accuracy under the marker model

* From Carré *et al.* (2013)

Discussion

There are plausible arguments to combine marker effects with individual effects when analyzing complex traits. To do so, the strategy used in the MS models [5], [6] and [7] is to decompose the individual genetic value into two terms: a contribution from base individuals, weighted by the transmission matrix **D**, and a contribution from Mendelian sampling occurring at several meiosis from base individuals to their descendants. This modeling approach has two consequences. On one hand, we keep the genetic transmission information of the infinitesimal model, i.e., the equal contribution of parental gametes to progeny. On the other hand, the unobserved random MS is replaced by an estimate of the realized MS.

Results of simulations indicate that the predictive ability of the MS embedded model is comparable to that of the marker model. On one hand, the accuracies obtained in different genetic scenarios suggest that the MS model might be useful when markers do not fully explain the genetic background (low QTL variances with high individual variance, or low marker density). On the other hand, the marker model [2] yielded slightly higher predictive ability than MS when QTL were important and marker density was high. This result reflects sub-optimality of the MS model to exploit favorable situations where markers do effectively capture much of total genetic variance. This might be explained by the simple distributional assumptions that we assumed at this exploratory stage for the base individuals and the marker effects of model MS in [7] and accompanying assumptions [8]. In particular, the assumption of independent base individuals chosen in [8] is usual in quantitative genetics, but, with DNA information and complete data it would be possible to make more general assumptions like $\mathbf{u_b} \sim N(\boldsymbol{\mu_u}, \mathbf{H}\sigma_u^2)$, where **H** represents a genomic matrix, thus recognizing that individuals in the base populations may share genes. Besides, model [7] can accommodate fixed genetic values for individuals or for groups of individuals in the base population.

Further investigation is needed on variance component estimation of models including marker and individual effects. Duchemin *et al.* (2012) were able to estimate both components of variance from real data using model [3]. We are currently studying variance components estimation for model [7].

Also, at this stage of model development, we are assuming complete data, in particular genotypes of base individuals. In some situations, it may possible to impute missing data. Also, if genealogy is unknown and if all individuals are in the genotyped sample, parent-progeny pairs can be easily identified using DNA data (Rohlfs *et al.*, 2012). However, to cover many variable situations in real life, it should be necessary to expand model [7] to include heterogeneous variances where Mendelian sampling is observed for some individuals but it remains a random value for individuals without genotyped parents.

The MS model [7] is compatible with other representations of marker effects: haplotypes can be used instead of single non-phased SNP. The MS model is also compatible with approaches where some QTL are known (model [6]), markers are preselected or markers are weighted by their effects during prediction (e.g. Zhang *et al.*, 2011).

Conclusions

According to the literature on prediction of complex traits, it is justified to keep, both, individual (infinitesimal) and marked gene effects in the statistical predictive model. We gave a formal derivation of a Mendelian sampling MS model where individual effects are a function of infinitesimal effects of base individuals and Mendelian sampling in descendants, traced using available DNA data. At this stage of research, we are assuming complete data, simple distributional assumptions for individual and marker effects, and known variances. First simulation results suggest that these simplifying assumptions should be extended to show general advantages of the MS model.

Acknowledgements

The first author benefits from financial support from INRA (Animal Genetics Department) and the Midi-Pyrenees' Region (France).

We thank the computing support of the bio-informatics platform Genotoul (http://bioinfo.genotoul.fr/; Toulouse, France) and financial support from the Cost Action TD1101 of the European Union

Literature cited

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics 128,* 422-428.

Carré, C., Gamboa, F., Cros, D., Hickey, J.M., Gorjanc, G. & Manfredi, E. 2013. Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects. *Genetica 141,* 239-246. doi: 10.1007/s10709-013-9722-9.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K & Cotes, J.M. 2009. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. *Genetics 182,* 375-385.

Duchemin, S.I., Colombani, C., Legarra, A., Baloche, G., Larroque, H., Astruc, J.-M., Barillet, F., Robert-Granié, C & Manfredi, E. 2012. Genomic selection in the French Lacaune dairy sheep breed. *Journal of Dairy Science 95,* 2723-2733.

Gianola, D., De Los Campos, G., Hill, W.G., Manfredi, E. & Fernando, R. 2009. Additive Genetic Variability and the Bayesian Alphabet. *Genetics 183,* 347-363.

Goddard, M.E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica 136,* 245-257.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics 31,* 423-447.

Henderson, C.R. 1976. Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics 32,* 69-83.

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157,* 1819-1829.

Quaas, R.L. 1976. Computing diagonal elements and inverse of a large numerator relationship matrix. *Biometrics 32,* 949-953.

Rohlfs, R.V., Fullerton, S.M. & Weir, B.S. 2012. Familial identification: population structure and relationship distinguishability. *PLoS Genet 8: e1002469-e1002469.*

Sargolzaei, M. & Schenkel, F.S. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics, 25,* 680-681. First published January 28, 2009, doi:10.1093/bioinformatics/btp045.

VanRaden, P. 2008. Efficient method to compute genomic predictions. *J. Dairy Sci. 91,* 4414-4123.

Yang, J.T., Manolio, A., Pasquale, L.R., Boerwinkle, E., Caporaso, N. *et al.* 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics 43,* 519-U544.

Zhang, Z., Ding, X., Liu, J., De Koning, D.J. & Zhang, Q. 2011. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. *BMC Proc 5 Suppl 3,* S15-S15.