View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects

Citation for published version:
Carré, C, Gamboa, F, Cros, D, Hickey, JM, Gorjanc, G & Manfredi, E 2013, 'Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects' Genetica, vol 141, pp. 239-246. DOI: 10.1007/s10709-013-9722-9

Digital Object Identifier (DOI):
10.1007/s10709-013-9722-9

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genetica

Publisher Rights Statement:
Available under Open Access

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 28. Apr. 2017

Genetica (2013) 141:239–246
DOI 10.1007/s10709-013-9722-9

Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects

Clément Carré · Fabrice Gamboa · David Cros · John Michael Hickey · Gregor Gorjanc · Eduardo Manfredi

Received: 16 January 2013 / Accepted: 20 May 2013 / Published online: 30 May 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract Genetic prediction for complex traits is usually based on models including individual (infinitesimal) or marker effects. Here, we concentrate on models including both the individual and the marker effects. In particular, we develop a "Mendelian segregation" model combining infinitesimal effects for base individuals and realized Mendelian sampling in descendants described by the available DNA data. The model is illustrated with an example and the analyses of a public simulated data file. Further, the potential contribution of such models is assessed by simulation. Accuracy, measured as the correlation between true (simulated) and predicted genetic values, was similar for all models compared under different genetic backgrounds. As expected, the segregation model is worthwhile when markers capture a low fraction of total genetic variance.

Keywords Genetic prediction · Genomic selection · SNP · Mendelian sampling

C. Carré · E. Manfredi (✉)
UR 631 SAGA, INRA Toulouse, B.P. 52627, Auzeville, 31326 Castanet-Tolosan Cedex, France
e-mail: Eduardo.Manfredi@toulouse.inra.fr

C. Carré · F. Gamboa
IMT, Université Paul Sabatier, Toulouse, France

D. Cros
AGAP, CIRAD, Montpellier, France

J. M. Hickey
Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), 06600 Mexico, D.F., Mexico

G. Gorjanc
Animal Science Department, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia

Introduction

In recent years, new knowledge on molecular genetics and the rapid evolution of sequencing and genotyping technology has renewed the interest on genetic prediction of complex traits. It should be recalled, however, that genetic prediction of complex traits has been a traditional field in animal and plant breeding since the 40's in the framework of the Selection Index (SI) theory (e.g., Hazel 1943), extended later to the "best linear unbiased prediction" (BLUP; Henderson 1975). These genetic prediction methods, without DNA data, were based on the "individual" model where covariances amongst phenotypes of related individuals are translated into unobserved covariances amongst genetic values, via theoretical relatedness coefficients amongst individuals. Anticipating the availability of low-cost whole genome DNA data, Meuwissen et al. (2001) proposed "marker" models where many markers' genotypes represent genetic effects, while the individuals are not explicitly specified in the model. We concentrate here on a third group of models including both "marker" and "individual" effects. We first recall the families of models proposed for genetic prediction and then we develop a novel model, which is illustrated with an example. Then, we assess the relative performance of the novel model in relation to the marker model for different genetic scenarios, and we report results of the analyses of a public simulated sample. Finally, originality, limits and possible extensions of the model are discussed.

Individual models for genetic prediction

Both SI and BLUP are applied to the "infinitesimal" (or polygenic) genetic model which in its simplest version is

"phenotype = mean + additive genetic value + residual". This model has been called "polygenic" or "infinitesimal" since the additive genetic value is the sum of the effects, assumed to be small and homogeneous, of numerous genes on the phenotype. In the statistical model, built from the genetic model, "individual effects" are used to represent additive genetic effects, and they are assumed random because genotype configurations of individuals arise through random processes:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{e} \quad (1)$$

$\boldsymbol{y}$ is a vector of phenotypes

$\boldsymbol{\mu}$ is a constant vector (assumed known in SI and estimated in BLUP)

$\boldsymbol{Z}$ is an incidence matrix of order N_y(phenotypes) × N_i(individuals), relating each of the N_y phenotypes to each of the measured individuals. For simplicity, we assume only one measure per individual. In standard BLUP technology $\mathbf{Z} = [\mathbf{0} \quad \mathbf{I} \quad \mathbf{0}]$, i.e., null columns for base individuals without phenotypes, the identity matrix for individuals with phenotypes (when there is a single measure for each individual), and null columns for descendants without phenotype, the usual target of prediction. In this context of genetic prediction, 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.

$\mathbf{u}$ is a vector of additive genetic effects, with $\mathrm{Var}(\mathbf{u}) = \mathbf{A}\sigma_u^2$, with $\mathbf{A}$ being the relationship matrix amongst individuals.

$\mathbf{e}$ is a vector of residuals, with $\mathrm{Var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$, with $\mathbf{I}$ being an identity matrix

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

The only information available to distinguish genetic effects from residuals are the structures of the (co)variance matrices of $\mathbf{u}$ and $\mathbf{e}$. In other words, the model describes a network of phenotypic covariances (observed) which are translated into genetic covariances (unobserved) via the theoretical genetic model, in particular the relatedness coefficients in the relationship matrix $\mathbf{A}$.

Marker and individual models

With molecular data available, prediction models evolved to include this new information (e.g., Fernando and Grossman 1989; Meuwissen et al. 2001). Fernando and Grossman (1989) proposed a prediction model which included several genetic effects: an infinitesimal effect $\mathbf{u}$ plus haplotype effects of maternal and paternal origin at marked quantitative trait loci (QTL) positions. Their model was reasonably conservative, given the genomic tools available by that time (say, 500 microsatellites to cover the entire genome in farm animals). In this context, they assumed that a marker allele may mark different QTL alleles in different families. Later, with many more markers (10,000 multi-allelic markers), Meuwissen et al. (2001) switched from the previous conservative model to "marker" models exploiting linkage disequilibrium at the population level:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{ZWm} + \mathbf{e} \quad (2)$$

where:

$\mathbf{m}$ is a vector of marked genetic effects (usually termed "marker effects", although the usual hypothesis is that markers do not have a true effect *per se* on the phenotype)

$\mathbf{W}$ is a matrix of marker genotypes of order N_i(individuals) × N_m(markers). With biallelic markers such as SNP, usual elements of $\boldsymbol{W}$ are 0, 1 or 2, the number of, say, the allele "1" of the marker genotype.

Usually assumed (co)variances are:

$$\mathrm{Var}(\mathbf{m}) = \mathbf{I}_{N_m}\sigma_m^2$$
$$\mathrm{Cov}(\mathbf{e}, \mathbf{m}) = \mathbf{0}$$

where $\mathbf{I}_{N_m}$ is an identity matrix of order N_m.

If we further assume that $\mathbf{u} = \mathbf{Wm}$ and $\mathrm{Var}(\mathbf{u}) = \mathbf{WW}'\sigma_m^2$, it is possible to compute predictions for $\mathbf{u}$ 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).

Marker plus individual model

Alternative assumptions in an outbred population are $\mathbf{u} \neq \mathbf{Wm}$ and $\mathrm{Var}(\mathbf{u}) \neq \mathbf{WW}'\sigma_m^2$. There are theoretical reasons and experimental results to support this point of view. Theoretically, in a Bayesian context, Gianola et al. (2009) claimed that the functional relationship between σ_u^2 and σ_m^2 is elusive. They did propose simple approximations under Hardy–Weinberg and linkage equilibria (LE) to relate the marked genetic variance and the additive genetic variance as $\sigma_u^2 = 2\sum_{i=1}^{N_m} p_i q_i \sigma_m^2$, where p_i and q_i are the allelic frequencies for marker i. However, assuming LE is not compatible with the essential assumption of linkage disequilibrium in the context of genome-wide analysis. Furthermore, in most experimental studies, the sum of

variances due to marker associations does not add up to the additive genetic variance due to individual infinitesimal effects raising the problem of the "hidden heritability" (e.g., Yang et al. 2011).

The unknown vector **m** represents the effects of unobserved genes that should be marked by observed markers. This model should fit all genome-wide additive effects simultaneously. However, it is not warranted that all the actual additive genetic effects in the studied genome will be effectively traced by the available markers (Yang et al. 2011). Potential problems are poor marker coverage (low density but also insufficient representation of independent DNA segments), rare alleles, small (infinitesimal) gene effects, multi-allelic genes having additive effects that are poorly traced by bi-allelic markers, or other molecular genetics mechanisms. The main assumption is that each marker allele or haplotype is associated with each unobserved QTL allele in identical way for each individual in the studied population. This may be true in some cases but it is not true in general. While an association between a marker and the QTL may be stable within parents and progeny, open populations over several generations are built up by subpopulations, each one with its own QTL allele-marker allele association. Reintroduction of infinitesimal effects in the prediction model is one of the recommended ways to control partially the lack of perfect association between marker alleles and causative alleles (Goddard and Hayes 2009). The model becomes:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e} \tag{3}$$

with additional assumptions:

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

where $\mathbf{R}\sigma_u^2$ is the symmetric (co)-variance matrix of individual effects of order N_i. 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 $\mathbf{u} = \mathbf{Wm}$.

The idea in model (3) is to include residual genetic values not taken into account by the marked effects **m.** In applications, this model gave better predictions than the marker model (2) (e.g., De los Campos et al. 2009; Duchemin et al. 2012).

Mendelian segregation model

Here, we develop a model where the genetic value of an individual is a function of infinitesimal effects of ancestors (individuals in the base, with unknown parents) and Mendelian sampling which can be traced by DNA data. In the following it is assumed that all individuals have complete genotype data and all descendants have known parents. We then discuss the departures from this complete data situation.

The model starts as in (3):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e}$$

It is convenient to separate individuals in two groups: the base ancestors with unknown parents (indexed by b) and the descendants (indexed by d). We can now expand and decompose the vector of infinitesimal values **u** as:

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_b \\ \mathbf{u}_d \end{bmatrix}$$

Let **P** be a $N_i \times N_i$ matrix with two 1's in each row, indicating the parents of each individual (rows of **P** for base individuals are null).

We define the matrix **M** as:

$$\mathbf{M} = \left(\mathbf{I} - \frac{1}{2}\mathbf{P}\right)$$

The matrix **M** is interpretable in biology (each row of **M** represents the individual minus half the sum of parents) and in mathematics since **M** has the form of a Laplacian matrix, representing the pedigree graph, with **P** being the adjacency matrix with elements equal to 1 at the intersection of adjacent nodes (parent and progeny nodes) or 0 otherwise.

Let $\boldsymbol{\phi}$ be a vector of infinitesimal mendelian sampling effects which are deviations of individual genetic values from their respective parental averages. Then, the matrix operator $\mathbf{M^{-1}}$ can be used to construct additive genetic values u as linear combinations of ancestor genetic values $\mathbf{u}_b$ and mendelian sampling $\boldsymbol{\phi}$ of their descendants, as illustrated in part (a) of Fig. 1, so we can write:

$$\mathbf{u} = \mathbf{M}^{-1}\begin{bmatrix} \mathbf{u}_b \\ \phi \end{bmatrix}$$

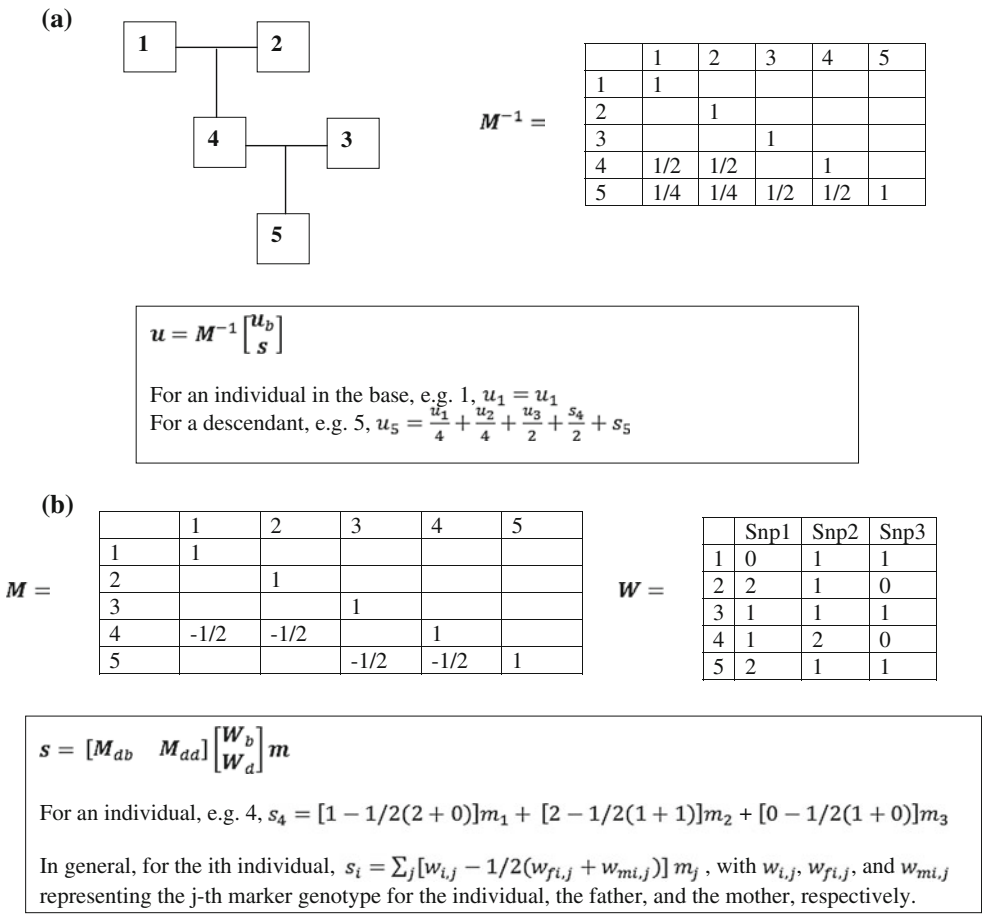
where **u** can be found by partitioning the **M** matrix in $\mathbf{M}_{bb}$, $\mathbf{M}_{dd}$, $\mathbf{M}_{db}$ and $\mathbf{M}_{bd}$ blocks, as:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{bb} & \mathbf{M}_{bd} \\ \mathbf{M}_{db} & \mathbf{M}_{dd} \end{bmatrix}, \text{with} \mathbf{M}_{bb} = \mathbf{I}, \text{ and } \mathbf{M}_{bd} = 0.$$

Using known results about the inverse of a lower triangular matrix, we obtain:

$$\mathbf{u} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{M}_{dd}^{-1}\mathbf{M}_{db} & \mathbf{M}_{dd}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{u}_b \\ \boldsymbol{\phi} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_b \\ -\mathbf{M}_{dd}^{-1}\mathbf{M}_{db}\mathbf{u}_b + \mathbf{M}_{dd}^{-1}\boldsymbol{\phi} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_b \\ \mathbf{u}_d \end{bmatrix} \tag{4}$$

Equation (4) uses standard results under infinitesimal models developed when it was impossible to observe DNA, and a theoretical distribution was assigned to the unknown $\boldsymbol{\phi}$ (see Quaas 1976). Availability of genotypes for progeny and parents gives a realized "molecular" mendelian



	1	2	3	4	5
1	1				
2		1			
3			1		
4	1/2	1/2		1	
5	1/4	1/4	1/2	1/2	1

	1	2	3	4	5
1	1				
2		1			
3			1		
4	-1/2	-1/2		1	
5			-1/2	-1/2	1

	Snp1	Snp2	Snp3
1	0	1	1
2	2	1	0
3	1	1	1
4	1	2	0
5	2	1	1

Fig. 1 Genetic transmission and Mendelian sampling effects in the prediction model. **a** Transmission: genetic values of descendants are a function of genetic values of base individuals $\boldsymbol{u}_b$ and Mendelian sampling effects predicted by $\boldsymbol{s}$. **b** Observed Mendelian sampling effects

sampling s, a predictor of $\boldsymbol{\phi}$ which can be approached as a function of marked gene effects $\mathbf{m}$:

$$\mathbf{s} = [\,\mathbf{M}_{db} \quad \mathbf{M}_{dd}\,]\begin{bmatrix}\mathbf{W}_b \\ \mathbf{W}_d\end{bmatrix}\mathbf{m} \tag{5}$$

where matrices $\mathbf{W}_b$ and $\mathbf{W}_d$ contain the marker genotypes of base and descendant individuals, respectively. Figure 1b illustrates how expression (5) represents individual deviations from parental means, in terms of marked genetic effects, for a hypothetical genealogy of 5 individuals and 3 markers.

Then, replacing $\boldsymbol{\phi}$ by s in (4), and using (5) in (4), with $\mathbf{D} = -\mathbf{M}_{dd}^{-1}\mathbf{M}_{db}$, we get:

$$\mathbf{u}_d = \mathbf{D}\mathbf{u}_b - \mathbf{D}\mathbf{W}_b\mathbf{m} + \mathbf{W}_d\mathbf{m} = \mathbf{D}\mathbf{u}_b + (\mathbf{W}_d - \mathbf{D}\mathbf{W}_b)\mathbf{m} \tag{6}$$

And the model for phenotypes is then:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_d\mathbf{D}\mathbf{u}_b + \mathbf{Z}_d(2\mathbf{W}_d - \mathbf{D}\mathbf{W}_b)\mathbf{m} + \mathbf{e} \tag{7}$$

In the term $\mathbf{Z}_d\mathbf{D}\mathbf{u}_b$, $\mathbf{Z}_d$ (of order Ny × Nd) relates records to individuals (descendants d) and $\mathbf{D}$ relates individual genetic values to ancestors' genetic values $\mathbf{u}_b$ via simple coefficients of genome sharing (including consanguinity, i.e., multiple contributions of an ancestor to an individual). So this term in (7) concentrates all phenotype information of descendants to estimate the ancestors' infinitesimal values. The term $\mathbf{Z}_d\,(2\mathbf{W}_d - \mathbf{D}\mathbf{W}_b)\,\mathbf{m}$ in (7) groups two parts: $\mathbf{Z_d}\ \mathbf{(W_d - DW_b)}\ \mathbf{m}$, the "molecular" mendelian sampling effects where individual marked effects deviate from ancestors' marked effects, and $\mathbf{Z_dW_d}\ \mathbf{m}$ which represents the direct relations between markers and phenotypes.

Assumptions of the model

A set of possible assumptions is:

$$\mathbf{u_b} \sim \mathbf{N}\left(\mathbf{0},\ \mathbf{I}\sigma_u^2\right)$$
$$\mathbf{m} \sim \mathbf{N}\left(\mathbf{0},\ \mathbf{I}\sigma_m^2\right)$$
$$\mathbf{Cov}(\mathbf{u_b}, \mathbf{m}) = \mathbf{0}$$

The assumption of independent base individuals is usual in quantitative genetics. 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 $\mathbf{H}$ represents a genomic matrix, thus recognizing that individuals in the base populations may share genes. Again, the model is redundant if it is assumed that $\mathbf{u_b} = \mathbf{W_b m}$ and $\mathbf{H} = \mathbf{W_b W_b'}\sigma_m^2$. Alternatively, model (7) can also accommodate fixed genetic values for individuals in the base population.

Distribution of marked effects $\mathbf{m}$ is assumed normal but other distributions such as the Gamma may be chosen, to take into account experimental results indicating few loci with large effects and many more loci with small effects (Goddard and Hayes 2009).

Analyses of data

Firstly, repeated simulations were conducted to assess the predictive ability of the Mendelian segregation model MS (Eq. 7) relative to the marker model M (Eq. 2). Then, we analyzed a public sample simulated for the 12th European QTLMAS workshop by Lund et al. (2009), using several models including individual and marked genetic effects.

We preferred to use simulated data at this exploratory stage to understand the behavior of the compared models. Also, to simplify interpretation at this stage, estimation and prediction were limited to the unknowns in the models ($\boldsymbol{\mu}$, the vector of marked effects $\mathbf{m}$ and the vector of individual genetic values $\mathbf{u}$) by applying known variances used to simulate the data.

We used the same statistical method BLUP to all models compared, which have either one (Eqs. 1 and 2) or two (Eqs. 3 and 7) random effects in addition to random residuals. BLUP of random effects were computed as detailed in the "Appendix".

Relative predictive performance of the Mendelian segregation (MS) model

Data were simulated using the QMSim software (Sargolzaei and Schenkel 2009). The simulated population had 1 base generation (25 individuals), 3 training generations (120 individuals) and the last generation (40 individuals) taken as prediction target. Mating was at random and the family size was 1. The simulated genome had 2 chromosomes of 1 Morgan each and 10 biallelic QTL/chromosome were responsible for the QTL fraction of genetic variance. Number of SNP markers used was either 2,000 or 200 per chromosome. Phenotypes in the base and target generations were simulated but not used to predict genetic values of the target generation. The phenotypes had variance 1 and overall heritability (infinitesimal + QTL effects) was 0.4. Three genetic scenarios were replicated 200 times: high (90 %), intermediate (50 %), or low (10 %) proportion of genetic variance explained by QTL.

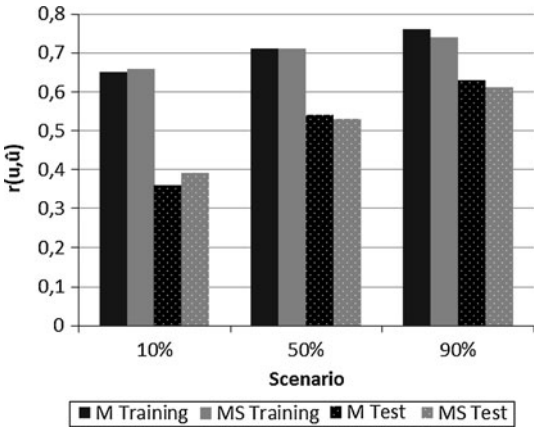


Fig. 2 Accuracy of the marker (M) and Mendelian segregation (MS) models for the three simulation scenarios with 10, 50, or 90 % of the total genetic variance explained by QTL

Mean accuracies over 200 replicates when using 2,000 SNP markers are presented in Fig. 2 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 ($P < 0.05$).

When fewer markers were used (200 SNP per chromosome), all accuracies were lower but the methods ranked as when using more (2,000 SNP per chromosome) markers (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.

Analyses of a public simulated sample

In the data simulated for the 12th European QTLMAS workshop (Lund et al. 2009), the simulated phenotypes were influenced by 50 loci, including 15 major effect loci and 35 minor effect loci with a total heritability of 0.3. Marker information was available for 6,000 SNP (only 5,925 were polymorphic and used in our analyses) on 6 chromosomes. The population was simulated under random mating and the absence of selection. Each male was mated to 10 females and each mating pair produced 10 offspring. A data set of 4,665 individuals was split into a training set (3,165 individuals) and a test set (1,500). In the

Table 1 Performance of the Mendelian segregation 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
2,000 SNP markers	102	108
QTL variance 50 %		
200 SNP markers	100	100
2,000 SNP markers	100	98
QTL variance 90 %		
200 SNP markers	99	95
2,000 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

training set, the base population (generation 0) included 165 individuals with unknown parents. The remaining 3,000 individuals had known parents and were born in generations 1 and 2. The test set had 1,500 individuals born in generation 3 with complete genealogy. The targets of prediction were the simulated genetic values and phenotypes of the test individuals. The data used in prediction were the phenotypes of 3,000 individuals of generations 1 and 2, and the marker genotypes of all individuals.

Four models were compared using the known variances used for the simulation: the marker model (M) as in (2), the marker plus individual model (MI) as in (3), the marker plus mendelian effects model (MS) given in (7), and the individual model where the (co)-variance matrix of individual effects was the additive relationship **A** (individual infinitesimal model; II). The method to estimate the unknowns of all the models was BLUP. The known variances were given by Lund et al. (2009): $\sigma_e^2 = 3.15$ and $\sigma_u^2 = 1.35$. The variance of marker effects was computed as $\sigma_m^2 = \sigma_u^2 / 2\sum_j p_j(1-p_j)$. Correlations between predicted values and simulated genetic values and phenotypes for the training and test populations are given in Table 2. The goodness of fit of model (7) for the training data was moderate $r((\widehat{\mathbf{u}}, \mathbf{y}) = 0.53)$ but it yielded the best predictions for genetic values $r((\widehat{\mathbf{u}}, \mathbf{u}) = 0.94)$ and phenotypes $r(\widehat{\mathbf{u}}, \mathbf{y}) = 0.55$ in the test sample. Model [7] was also the best to estimate the marked effects **m**: the correlations between estimates of **m** and the simulated allele substitution effects, in absolute values, were 0.69 for Model [7] and 0.56 for both the marker model and the "marker + individual" model.

Table 2 Correlations between the predicted genetic values ($\widehat{\mathbf{u}}$), simulated genetic values (**u**), and simulated phenotypes (**y**) in the training and test data

Model[a]	M	MI	MS	II
Training data				
$r(\widehat{\mathbf{u}}, \mathbf{u})$	0.87	0.84	0.94	0.69
$r(\widehat{\mathbf{u}}, \mathbf{y})$	0.59	0.77	0.53	0.74
Test data				
$r(\widehat{\mathbf{u}}, \mathbf{u})$	0.81	0.77	0.94	0.43
$r(\widehat{\mathbf{u}}, \mathbf{y})$	0.46	0.46	0.55	0.27

[a] Models. M: marker model (Eq. 2); MI: marker plus individual effect model (Eq. 3); MS: Mendelian segregation model (Eq. 7); II: individual infinitesimal model based on pedigree (Eq. 1)

Discussion

As reviewed in the Introduction, there are plausible arguments to combine marked effects models with other individual effects when analyzing complex traits. To do so, the strategy used in the MS model [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, instead of attempting to fit twice the additive genetic value of an individual as in model [3]. In traditional infinitesimal models, mendelian sampling is an unknown theoretical random term, so predictions of future phenotypes (of future progeny) are based on ancestor phenotypes and random terms. At present, with the availability of numerous markers, mendelian sampling is realized for each individual and it can be used to improve predictions.

Model [7] builds on very well-known results in quantitative genetics. Early work described how genetic transmission operates in the additive relationship matrix **A** (e.g., Quaas 1976 and Henderson 1976, who presented detailed factorizations of the **A** matrix). Subsequent models included genetic transmission at unobserved segregating QTL (e.g., Fernando and Grossman 1989; Meuwissen and Goddard 2000; Legarra and Fernando 2009) and combined within family and between family marker effects in the context of methodology for QTL search (e.g., Abecasis et al. 2000). In animal breeding, efforts have focused on combining genotype data with genealogy data in individual genomic models, as reviewed by Meuwissen et al. (2011). The model [7] developed here builds on previous work by the simultaneous inclusion of infinitesimal and marked genetic effects. In this way the model might capitalize on two advantages of molecular information: the improvement of the infinitesimal prediction by the estimation of realized mendelian sampling in descendant individuals, and by capturing marked gene effects without bias due to family structure, i.e., to predict marked effects and infinitesimal effects simultaneously and without redundancy. Here, marked effects are estimated at the level of the population (marked effects **m** in model MS

[7] are not defined within family) but the family structure is taken into account in the estimation model.

Results of simulations indicate that the predictive ability of the MS 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 are not adequate to fully explain the genetic background (low QTL variances with high infinitesimal variance, or low marker density).

On the other hand, the marker model M yielded slightly higher predictive ability than MS when QTL were important and marker density was high. This result might reflect 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 marked effects of model MS in [7] and accompanying assumptions. In particular, the marker model [2], and, more explicitly, its equivalent model "Genomic BLUP", capitalizes the complete data setting studied here by estimating covariances among base individuals, and covariances between base individuals and descendants. So, for the MS model to be fully competitive, its distributional assumptions should be extended to take into account those relationships.

Results for the QTLMAS example are encouraging but unique and different from those of replicated simulations. At least two reasons may be advanced to explain these different results: the more complicated genetic background and the large family size, a full-sib design, simulated in the QTLMAS data set. But the impact of such factors on predictive ability needs further investigation.

Further investigation is also 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], i.e., the variance of individual effects and the variance of marker effects. We are currently studying variance components estimation for model [7], with infinitesimal effects defined only for the base individuals and variance structure designed to avoid identifiability problems.

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.

Another potential improvement of the MS model in [7] is the representation of genetic transmission (as in expression [5]) and marked genetic effects (as in [2] and [7]) which may be certainly improved. Haplotypes can be used instead of single non-phased SNP. The model is also compatible with approaches where some QTL are known, 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 marked genetic effects, and known variances. First simulation results suggest that these simplifying assumptions should be extended to render the MS model fully competitive.

Acknowledgments 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.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

Appendix: Computation of individual and marked genetic effects using BLUP

Let σ_i^2, σ_M^2, and σ_e^2 be the variance of infinitesimal effects, the genetic variance due to all QTL, and the residual variance, respectively. Also the variance of individual markers is $\sigma_m^2 = \sigma_M^2/k$, with $k = 2\sum_j p_j\left(1 - p_j\right)$. Then:

$$\alpha_u = \sigma_e^2/\left(\sigma_i^2 + \sigma_M^2\right)$$
$$\alpha_m = \sigma_e^2/\sigma_m^2$$
$$\alpha_i = \sigma_e^2/\sigma_i^2$$

Solutions for models compared were:

For model [1]:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 1'1 & 1'\mathbf{Z} \\ \mathbf{Z}'1 & \mathbf{Z}'\mathbf{Z} + \alpha_u \mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} 1'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where 1 is a vector of 1 and $\mathbf{Z}$ is the incidence matrix. $\widehat{\mu}$ is the BLUE (best linear unbiased estimator) of the general mean, and $\widehat{\mathbf{u}}$ is the solution for individual effects.

For model [2]:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} 1'1 & 1'\mathbf{X} \\ \mathbf{X}'1 & \mathbf{X}'\mathbf{X} + \alpha_m \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} 1'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

where X = $\boldsymbol{ZW}$, i.e., the incidence matrix times the matrix of genotypes, centered by column. $\widehat{\mathbf{m}}$ is the solution for marked effects.

Predictions from model [2] can be also obtained with the individual model:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} 1'1 & 1'\mathbf{Z} \\ \mathbf{Z}'1 & \mathbf{Z}'\mathbf{Z} + \alpha_u \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} 1'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where $\mathbf{G} = \mathbf{W}\mathbf{W}'/k$

For model [3]:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \\ \widehat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} 1'1 & 1'\mathbf{X}_1 & 1'\mathbf{X}_2 \\ \mathbf{X}_1'1 & \mathbf{X}_1'\mathbf{X}_1 + \alpha_u \mathbf{A}^{-1} & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'1 & \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 + \alpha_m \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} 1'\mathbf{y} \\ \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}$$

where $X_1 = Z$ and $X_2 = ZW$. $\widehat{\mathbf{u}}$ is the solution for individual effects and $\widehat{\mathbf{m}}$ is the solution for marked effects.

For model [7]:

$$\begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{u_b}} \\ \widehat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} 1'1 & 1'\mathbf{X}_1 & 1'\mathbf{X}_2 \\ \mathbf{X}_1'1 & \mathbf{X}_1'\mathbf{X}_1 + \alpha_i \mathbf{I} & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'1 & \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 + \alpha_m \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} 1'\mathbf{y} \\ \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}$$

where $\mathbf{X}_1 = \mathbf{Z}_d\mathbf{D}$ and $\mathbf{X}_2 = \mathbf{Z}_d 2(\mathbf{W}_d - \mathbf{D}\mathbf{W}_b)$, and $\widehat{u_b}$ is the solution for base individuals and $\widehat{\mathbf{m}}$ is the solution for marked effects.

References

Abecasis GR, Cardon RL, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292

Aguilar I, Misztal I, Legarra A, Tsuruta S (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J Animal Breed Genet 128:422–428

De Los campos G, Naya H, Gianola D, Crossa J, Legarra A et al (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

Duchemin SI, Colombani C, Legarra A, Baloche G, Larroque H et al (2012) Genomic selection in the French Lacaune dairy sheep breed. J Dairy Sci 95:2723–2733

Fernando RL, Grossman M (1989) Marker assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467–477

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363

Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–257

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

Hazel LN (1943) The genetic basis for constructing selection indexes. Genetics 28:476–490

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Henderson CR (1976) Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83

Legarra A, Fernando RL (2009) Linear models for joint association and linkage QTL mapping. Genet Sel Evol 41:43

Lund MS, Sahana G, De Koning DJ, Su G, Carlborg O (2009) Comparison of analyses of the QTLMAS XII common dataset. I: genomic selection. BMC Proc 3 (Suppl 1):S1

Meuwissen TH, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421–430

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Meuwissen TH, Luan T, Woolliams JA (2011) The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet 128:429–439

Quaas RL (1976) Computing diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32:949–953

Rohlfs RV, Fullerton SM, Weir BS (2012) Familial identification: population structure and relationship distinguishability. PLoS Genet 8:e1002469

Sargolzaei M, Schenkel FS (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 91:4414–4423

Yang JT, Manolio A, Pasquale LR, Boerwinkle E, Caporaso N et al (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43:519–544

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