A Mendelian Sampling model for genetic prediction

Clement Carre, Fabrice Gamboa, David Cros, Gregor Gorjanc, Eduardo Manfredi

INRA
Génétique Animale

E. Manfredi

CIRAD (France)
IMT (France)
University of Ljubljana

Additive models for prediction

Individual:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{u} + \mathbf{e}\,, \qquad \mathbf{var}(\mathbf{u}) = \mathbf{A}\sigma_u^2 \text{ and } \mathbf{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$$

Marker:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{W}\mathbf{m} + \mathbf{e}\,, \qquad \mathbf{var}(\mathbf{m}) = \mathbf{I}\sigma_m^2 \text{ and } \mathbf{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$$

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{u}^* + \mathbf{e}\,, \qquad \mathbf{var}(\boldsymbol{u}^*) = \mathbf{W}\mathbf{W}'\mathbf{f}(\sigma_m^2) \text{ and } \mathbf{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$$

Individual + marker: $\mathbf{y} = \boldsymbol{\mu} + \mathbf{u} + \mathbf{W}\mathbf{m} + \mathbf{e}$

Individual + marker models: when?

When the individual variance σ_u is larger than the variance due to markers f(σ_m)

- Theory: elusive relationship between σ_u and σ_m (Gianola et al., 2009)

- Reports on missing heritability (Yang et al., 2011)

- Improvement of prediction accuracy in some applications (de los Campos et al., 2009; Duchemin et al., 2012)

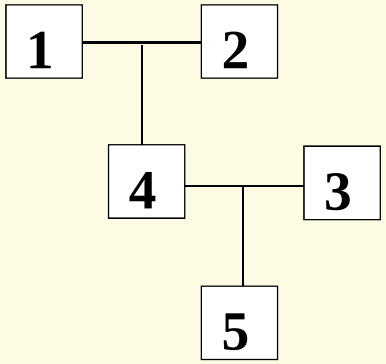
- Genome complexities: LD is subjected to noise and to data conditions (marker coverage, reference and target populations), no-nucleotide polymorphism

So, revisit the model: $y = \mu + u + Wm + e$

A Mendelian Sampling Model (1/4)

Individual effects of descendants are a fonction of effects of base individuals and Mendelian Sampling (MS) (Quaas, 1976):

$$\mathbf{u_d} = \mathbf{M}^{-1} \begin{bmatrix} \mathbf{u_b} \\ \mathbf{s} \end{bmatrix}, \text{ with } \mathbf{M}=\mathbf{I}-1/2\ \mathbf{P}$$



$$M^{-1} =$$

	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

$$u_1 = u_1$$

$$u_5 = \left(\frac{u_1}{4} + \frac{u_2}{4} + \frac{s_4}{2}\right) + \frac{u_3}{2} + s_5$$

A Mendelian Sampling Model (2/4)

Replace random Mendelian Sampling s by realized Mendelian Sampling r:

$$\mathbf{r} = [\mathbf{M_{db}} \quad \mathbf{M_{dd}}] \begin{bmatrix} \mathbf{W_b} \\ \mathbf{W_d} \end{bmatrix} \mathbf{m}$$

	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

$$\mathbf{r_4} = [1 - \tfrac{0+2}{2}]\,\mathbf{m_1} + [2 - \tfrac{1+1}{2}]\,\mathbf{m_2} + [0 - \tfrac{1+0}{2}]\,\mathbf{m_3}$$

INRA
Génétique Animale
E. Manfredi

Mendelian sampling model (3/4)

Using parts (1) and (2) write individual effects of descendants as a fonction of effects of base individuals and MS:

$$\mathbf{u_d = Du_b + (W_d - DW_b)m}$$

with $\mathbf{D = -M^{-1}_{dd} M_{db}}$

Mendelian Sampling model (4/4)

Then, several models for phenotypes of descendants can be proposed:

Disjoint model:

$$y_d = \mu + Du_b + (W_{1d}\text{-}DW_{1b})m_1 + W_{2d}m_2 + e$$

For m1 and m2, subsets of m (e.g. markers of QTL and rest of markers)

Embedded model:

$$y_d = \mu + Du_b + (W_d - DW_b)m + W_d m + e$$

$$= \mu + Du_b + (2W_d - DW_b)m + e$$

Testing the embedded model

1. Simulated data (QMSim **;** Sargolzaei and Frenkel, 2009)
Demography : G0 (n=25), G1 to G3 (training; n=120), G4 (target ; n=40)
Random mating; family size=1
Genome: 2 chromosomes (2M) with 10 biallelic QTL each
Trait: phenotypic variance=1 and genetic variance (QTL + infinitesimal effects)=0.4
Conditions studied: 200 vs 2000 SNP/chromosome ; variance explained by qtl: 10, 50 and 90% (200 replicates/condition)

Predictive ability: correlation between known simulated values (phenotypes and genetic values) and their corresponding predictions

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

Discussion

First test of the embedded model suggests advantages (NS) of the MS model when markers capture a small proportion of genetic variance. On the other hand, the embedded model did not improve the accuracy of the reference model when markers do capture a high proportion of genetic variance (NS).

Some simplifying assumptions that we used at this stage may limit the potential contribution of the embedded model: the assumption of uncorrelated effects of base individuals, and also the use of known variances.

The MS model gave an extra accuracy of +16% over the marker model in the analysis of a QTLMAS simulated data set (Lund et al., 2009), with a complex genetic model and large family size (n=10)

Conclusions

The decomposition of individual effects into base individual effects and mendelian sampling can be revisited to integrate realized MS instead of random MS.

This open new possibilities for modelling: by splitting markers into groups or by embedding marker and individual effects. Also, the model allows the easy integration of genetic groups.

Ongoing work is on variance components estimation of the embedded and disjoint models, including the incomplete marker data situation, and the analyses of real data.

INRA
Génétique Animale

E. Manfredi

CIRAD (France)
IMT (France)
University of Ljubljana