Validation of genomic selection in an outbred mice population

A Legarra, JM Elsen, E Manfredi, C Robert-Granié
andres.legarra@toulouse.inra.fr

UR631, INRA-SAGA, BP 52627
31326 Castanet Tolosan CEDEX

27 August 2007, EAAP, Session 18, abstract 1071

Introduction to Genomic selection

Let there be a "SNP" model of the breeding value: $BV = \sum(SNP_i)$.
Meuwissen et al 2001 showed by simulation:

- High predicting accuracies (up to 0.85).
- Overpasses practical problems in MAS (Boichard 2006).
- Very interesting breeding tool (Schaffer, 2006; Dekkers, 2007).

The idea is based on the existence of Linkage Disequilibrium between QTL and markers

Why to test genomic selection?

Why to test genomic selection?

- It is expensive (200€ per animal?)
- Restrictive assumptions (equilibrium mutation-drift, big population, no selection)
- Simple genetic model

What about an experiment?

- Slow and expensive
- Let use public data today

The data

Nature Genetics 38:879-887 (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice

W Valdar, LC Solberg, D Gauguier, S Burnett, P Klenerman, WO Cookson, MS Taylor, J Nicholas, P Rawlins, R Mott & J Flint

- http://gscan.well.ox.ac.uk
- Heterogeneous Stock Mice, 50 generations of random mating
- 13,459 SNPs, 1,904 fully phenotyped mice
- Weight at 6 weeks, highly heritable

How to test?

“Accuracy” of Classical BLUP *vs* genome-wide models by cross-validation.

1. Split the data into two at random : $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]$.
   $\mathbf{y}_1 \longrightarrow$ training; $\mathbf{y}_2 \longrightarrow$ validation.
2. Estimation
   - Estimate SNP effects $\hat{\mathbf{a}}$ from $\mathbf{y}_1$
   - Estimate Classical BLUP EBVs $\hat{\mathbf{u}}$ from $\mathbf{y}_1$
3. Validation
   - Estimate $\hat{\mathbf{y}}_2$ from SNP estimates $\hat{\mathbf{a}}$
   - Estimate $\hat{\mathbf{y}}_2$ from Classical BLUP EBVs $\hat{\mathbf{u}}$
4. Compute $r(\mathbf{y}_2, \hat{\mathbf{y}}_{2SNP})$, and $r(\mathbf{y}_2, \hat{\mathbf{y}}_{2BLUP})$.

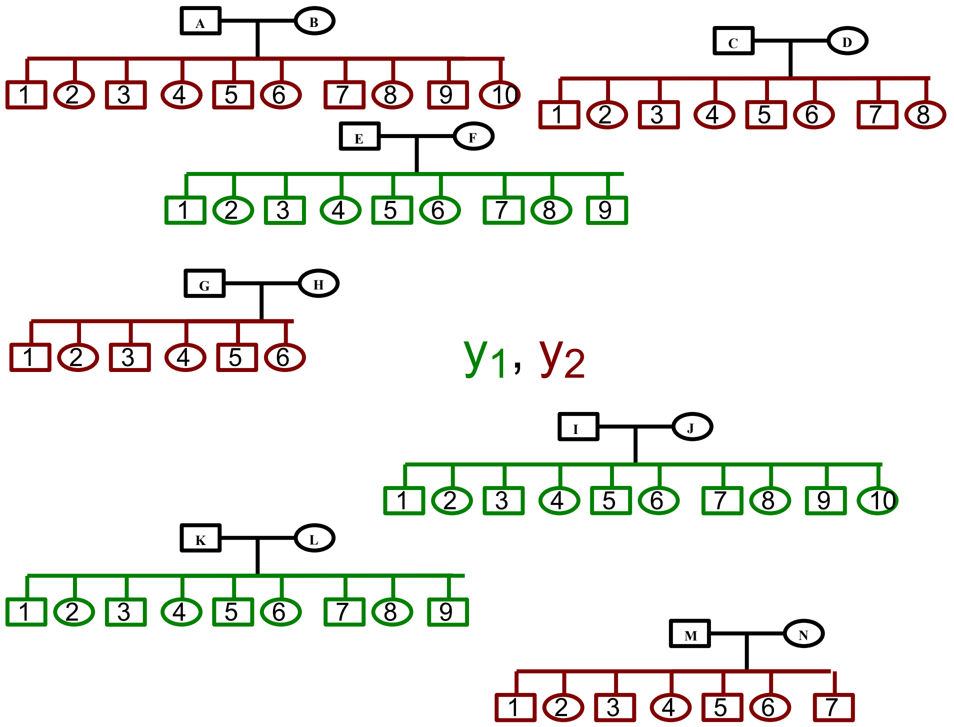
In a selection process: $\Delta G = i \cdot r(\mathbf{y}_2, \hat{\mathbf{y}}_2) \cdot \sigma_{y_2}$.

Cross-validation

How to split $\mathbf{y}$ in $[\mathbf{y}_1, \mathbf{y}_2]$?

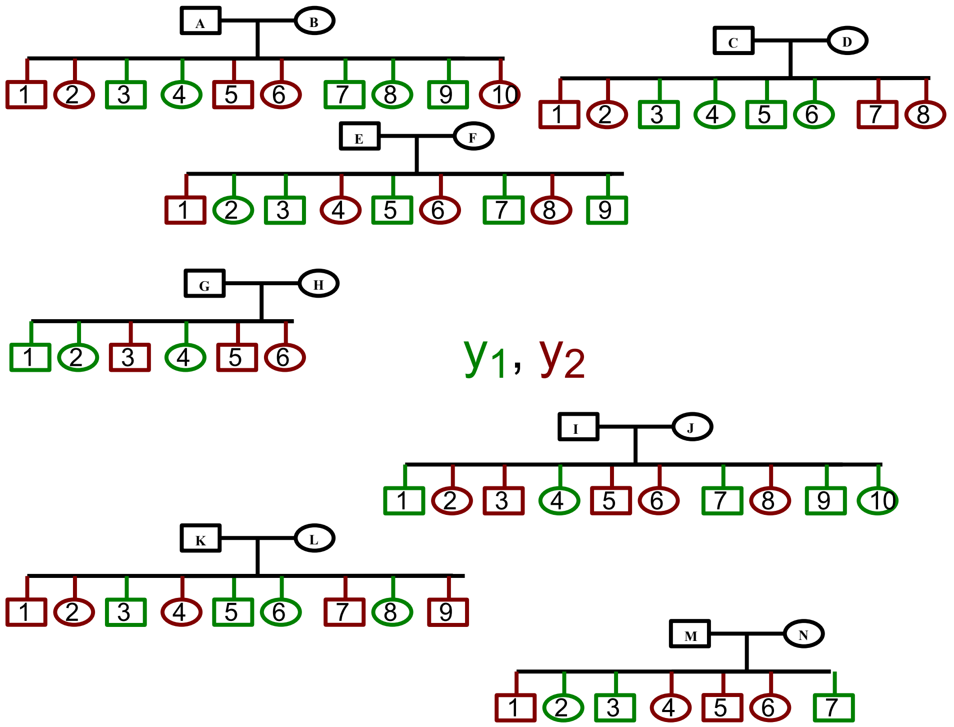
- Sampling families: Most LD is only at the population level, less powerful. BLUP does not give information in this case (no known relatives).
- Splitting families in two. High LD because there is a family structure and we use full-brothers to predict full-brothers. Comparable to a two-generations (dairy cattle) design.

Sampling families



y_1, y_2

Splitting families



y_1, y_2

Models

1. Classical BLUP $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$
2. SNP $\mathbf{y} = \mathbf{X}\beta + \mathbf{Wa} + \mathbf{e}$
3. Mixture allows for SNPs without any effect.
   - $a_i \sim N(0, \sigma_a^2)$ with probability p_a
   - $a_i = 0$ with probability $1 - p_a$
4. Classical+SNP $\mathbf{y} = \mathbf{X}\beta + \mathbf{Wa} + \mathbf{Zu} + \mathbf{e}$
5. ...and combinations of the above
6. ...and we tried different priors (including Meuwissen et al. 2001)

We used MCMC for everything.

Genomic selection ≈ Classical BLUP?

Genomic selection ≈ Classical BLUP?

Look at model 2; define a pseudo-overall breeding value $v, v_i = \sum a_{ij}$.
Then:
$\mathbf{y} = \ldots + \mathbf{Wa} = \ldots + \mathbf{Zv}$ where
$\mathbf{v} = \mathbf{Wa}$, $\mathbf{v} \sim N(\mathbf{0}, \mathbf{WW}'\sigma_a^2)$.

Genomic selection is akin to Classical BLUP where $\mathbf{WW}'$ is an IBS pseudo-relationship matrix. For the mixture approach, some row/cols in $\mathbf{W}$ are nullified.

Results (10 replicates), sampling families

Table: Correlations $r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$, sampling families

Method	$r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$
Classical BLUP	0
SNP	0.21
Mixture	0.21
Classical BLUP + SNP	0.19
...	
Others	≤ 0.21

Results (10 replicates), splitting families

Table: Correlations $r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$, splitting families

Method	$r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$
Classical BLUP	0.59
SNP	0.49
Mixture	0.49
Classical BLUP + SNP	0.60
...	
Others	≤ 0.49

The end

Conclusions:

1. The genomic model performs
   - *better* than classical BLUP when there is no information from relatives
   - *worse* when there is family information (real-life situations)
2. The simplest "SNP" model performs better than more complex ones
3. Historical LD can be used but is less powerful than close LD due to family relationships
4. The genomic model implicitely assumes a pseudo-relationship matrix based on identity by state among markers. Sometimes this information might be better than pedigree.

Why?

- Are different loci segregating in different families?
- How many QTLs around?

The end

Homework assignment (for us)

- Analyze more traits
- More models? Non parametric?

Take-home message

1. (We have) reasonable doubts whether genomic selection will work immediately.
2. More testing has to be done in real-life data (e.g. Sölkner, this conference). Cross-validation is a good tool.
3. We need a better modeling of marker locus effects allowing for population *and* familiar LD *and* LA.

Thank you

Extended results (10 replicates), sampling families

Table: Correlations $r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$, sampling families

Method	Mean	S.D.	var($\hat{\mathbf{y}}_2$)
Classical BLUP + SNP	0.19	0.03	0.26
SNP	0.21	0.04	1.33
SNP - prior	0.17	0.04	4.14
Mixture	0.21	0.05	1.32
Classical BLUP	0	0	0
...			
Others	$\leq$ 0.21		

Extended results (10 replicates), splitting families

Table: Correlations $r(\mathbf{y}_2, \hat{\mathbf{y}}_2)$, splitting families

Method	Mean	S.D.	var($\hat{\mathbf{y}}_2$)
Classical BLUP + SNP	0.60	0.01	2.26
SNP	0.49	0.01	2.35
SNP - prior	0.43	0.02	4.16
Mixture	0.49	0.02	1.28
Classical BLUP	0.59	0.01	2.28
...			
Others	$\leq$ 0.49		