Methods to estimate erosion factors of genomic breeding values of candidates due to long-distance linkage disequilibrium

D. Boichard, S. Fritz[1], P. Croiseau, V. Ducrocq, T. Tribout, M. Barbat[2], and B. Cuyabano

INRAE science for people, life & earth

GABI, Jouy-en-Josas, France

[1] ELIANCE Des éleveurs. Une ambition.

[2] GenEval Evaluation génétique des animaux d'élevage

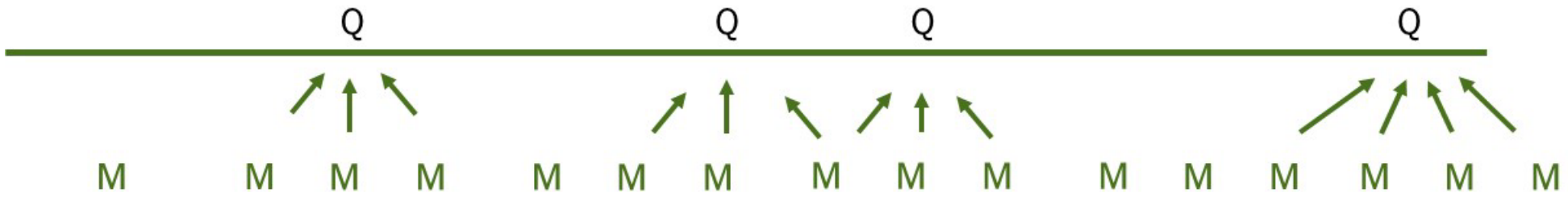
UMT eBIS

Lyon, August 26-27 2023

Introduction

In genomic evaluation (GE), SNP effects are
- estimated in the reference population
- applied to candidates for selection

A common interpretation:
Genetic markers are in close linkage disequilibrium (LD) with QTL and are good proxies of them.
This assumes that the estimates of the marker effects reflect the effects of the neighbouring QTLs only.



Under this assumption, associations observed in the reference are well conserved in the next generation as short distance LD is only slowly eroded by recombinations.

Introduction

BUT the reality is quite different:

- GE accuracy is highly dependent on the relationships between the candidates and the reference population (Habier et al, 2007, 2013; Legarra et al, 2008; Pszczola et al, 2012).

- Limited gain in accuracy in multi-breed evaluation (Erbe et al, 2012; Hozé et al, 2014)
⇒ distant reference populations are not informative.

- Decrease in accuracy with the number of generations when the reference population is not updated (Sonneson et al, 2009; Solberg et al, 2009).

- Impact of presence/absence of parents of candidates in the reference population

=> **Marker effects erode as the distance increases between the candidates and the reference population**

Introduction

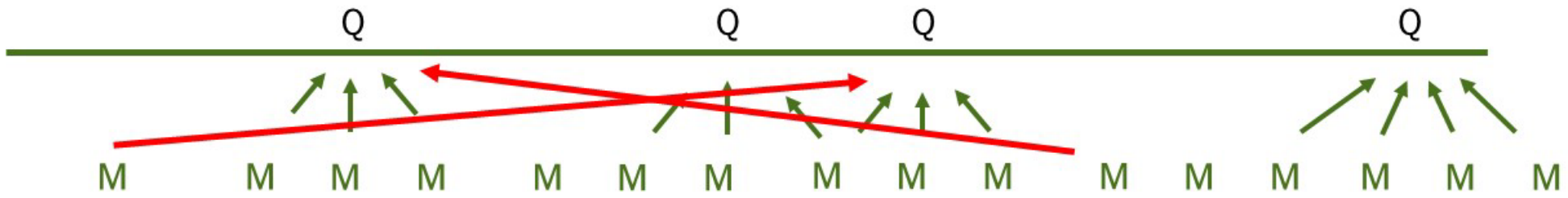
Validation studies of GE are frequently based on regression of later performances on early predictions

These studies nearly always observe an "inflation" pattern:
the regression coefficient is nearly systematically below 1 (frequently between 0.8 and 0.9 ?),
meaning that predictions are inflated
(i.e., too high for high predictions, too low for low predictions).

$\Rightarrow$ We will show that **the expectation is < 1**

Our hypothesis

- Long range LD exists, even between chromosomes
- Many markers capture some partial effects of physically unlinked QTL.
- Because this long range LD is not conserved over generations, these marker effects are rapidly eroded with generations



LD across chromosomes in French dairy cattle

1 every 20 SNP within chromosome
Statistics based on 2.8 – 3.2M r^2-values per breed (0.25%)

Breeds	# cows in reference population (2021)	Mean(r^2)	% |r| > 0.05
Vosgienne	2617	0.0029	33
Tarentaise	3788	0.0015	18
Abondance	7115	0.0012	15
Normande	69,220	0.00073	7
Montbeliarde	185,053	0.00053	4
Holstein	362,363	0.00038	1.5

Mean close to 0

… But many markers with some LD
(eg in Holstein, about 800 distant markers *vs* a few tens of close markers)

Note: **GE depends on r**, not on r^2

Questions

- Demonstrate the effect of long distance LD
- Measure the impact of a residual polygenic effect in the model
- How to measure erosion in practice ?

Simulation study

Simulation study

In the real female 2021 Normande reference population: 69,220 genotyped cows with records

- **5 chromosomes** considered (**13,608 SNP**) .
- **nq=200 QTLs** simulated by randomly sampling 200 SNPs with MAF>0.02, then excluded from the model
- QTL effects were independently drawn in a normal distribution (h^2=0.3).

- With / without polygenic effect (20% σ_g^2 when present)
  - the cows were born from 1108 sires

Contribution of QTL to SNP effects

The SNP random part of the MME is

$$(\mathbf{M'M} + \lambda\,\mathbf{I})\,\hat{\mathbf{s}} = \mathbf{M'}\,\mathbf{y}$$

with **M** the matrix of standardized genotypes, **s** the SNP effects, λ the variance ratio

y is replaced by its value $\mathbf{y} = \mathbf{Pq} + \mathbf{e}$,

with **P** the matrix of QTL genotypes and **q** the QTL effects

$$\Rightarrow \hat{\mathbf{s}} = (\mathbf{M'M} + \lambda\,\mathbf{I})^{-1}\,\mathbf{M'(Pq + e)}.$$

If $\mathbf{c}_i$ is line i of $\mathbf{C} = (\mathbf{M'M} + \lambda\,\mathbf{I})^{-1}$, the contribution of QTL j to effect of SNP i is : $\mathbf{c}_i\,\mathbf{M'}\,\mathbf{P}_j\,q_j$.

(200 x 13,408 = 2,681,600 contributions)

From these contributions, 4 partial DGV were computed for each cow, as well as their variance and correlations

(1) distance (QTL, SNP) < 5 Mb
(2) 5 Mb < distance (QTL, SNP) < 20 Mb
(3) distance (QTL, SNP) > 20 Mb
(4) the QTL and the SNP are on different chromosomes

Accounting for the polygenic effect

$$\begin{bmatrix} \mathbf{Z'Z} + \kappa\mathbf{A^{-1}} & \mathbf{Z'M} \\ \mathbf{M'Z} & \mathbf{M'M} + \lambda\,\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'y} \\ \mathbf{M'y} \end{bmatrix}$$

The polygenic effects **u** can be absorbed into the SNP effects

$$\left[\mathbf{M'}\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z'Z} + \kappa\mathbf{A^{-1}}\right)^{-1}\mathbf{Z'}\right)\mathbf{M} + \lambda\,\mathbf{I}\right]\hat{\mathbf{s}} = \mathbf{M'}\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z'Z} + \kappa\mathbf{A^{-1}}\right)^{-1}\mathbf{Z'}\right)(\mathbf{Pq} + \mathbf{e})$$

Again, if $\mathbf{C} = \left[\mathbf{M'}\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z'Z} + \kappa\mathbf{A^{-1}}\right)^{-1}\mathbf{Z'}\right)\mathbf{M} + \lambda\,\mathbf{I}\right]^{-1}$

$$\mathbf{M^{*\prime}} = \mathbf{M'}\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z'Z} + \kappa\mathbf{A^{-1}}\right)^{-1}\mathbf{Z'}\right)$$

Again, if $\mathbf{c}_i$ is line i of **C,** the contribution of QTL j to the effect of SNP i is : $\mathbf{c}_i\ \mathbf{M^{*\prime}}\ \mathbf{P}_j\ q_j$.

Contribution of partial DGV variances to total DGV variance

Results on 30 replicates.

Classes of Partial DGV defined according to QTL-SNP relationship	Relative contribution (%)	
	Model without polygenic effect	Model with polygenic effect
1 : distance(QTL,SNP) < 5 Mb	**73.5**	**68.9**
2 : 5 Mb < distance < 20 Mb	**13.7**	**14.3**
3 : distance(QTL,SNP) > 20 Mb	**4.9**	**5.2**
4 : QTL and SNP on different chromosomes	**8.0**	**11.5**

⇒ **Inclusion of a polygenic effect seems to reinforce erosion**
Probably because SNP effects are more regressed

Correlations between partial DGV

Results on 30 replicates.

	Without polygenic effect			With Polygenic effect		
	<5	5-20	>20	<5	5-20	>20
5-20	0.30			0.54		
>20	0.17	0.15		0.26	0.29	
Other Chrom	0.11	0.05	0.12	0.25	0.22	0.27

⇒ **All correlations are slightly to moderately positive,**
illustrating that distant QTL contribute to many marker effects

⇒ **Inclusion of a polygenic effect increases these correlations**
showing that long distance effects are reinforced

Factors affecting erosion of SNP effects

Erosion has two components

- Long distance LD in the reference population (halved at each generation)
  - Effective size of the population (non-zero LD baseline when Ne is small)
  - **Relationship within the reference population** (LD higher than baseline)

- Number of generations between the candidate and the reference population
  - On both sire and dam pathways

How to estimate the erosion factor ρ ?

Additional work needed for a theoretical derivation

But again, **simulation** is a practical solution

Two methods are proposed

Method 1: Estimation of the erosion factor:
by simulating a new generation

- Nq QTL are simulated in the reference population (G_r)
- SNP effects are estimated by $\widehat{\mathbf{s_r}} = (\mathbf{M'M} + \lambda\mathbf{I})^{-1}\,\mathbf{M'}\,\mathbf{P}\,\mathbf{q}$
- A next generation (G_n) is simulated, with parents sampled in the reference population
- Their DGV obtained from reference population SNP estimates are $\mathbf{DGV_r} = \mathbf{M_n}\,\widehat{\mathbf{s_r}}$
- New SNP effects can be obtained from generation n only $\widehat{\mathbf{s_n}} = (\mathbf{M_n'}\,\mathbf{M_n} + \lambda\mathbf{I})^{-1}\;\mathbf{M_n'}\mathbf{P_n}\mathbf{q}$
- New DGV are obtained from new SNP estimates $\mathbf{DGV_n} = \mathbf{M_n}\,\widehat{\mathbf{s_n}}$
- An estimate of ρ is the regression coefficient of $\mathbf{DGV_n} = \mu + \rho\,\mathbf{DGV_r} + \varepsilon$

Method 2: Estimation of the erosion factor:
By regressing each QTL to SNP contribution according to genetic map

- Nq QTL are simulated in the reference population
- All contributions f_{ij} of QTL j to effect of SNP i are computed $f_{ij} = \mathbf{c}_i \mathbf{M'} \mathbf{P}_j q_j$.
- $\mathbf{DGV_r}$ in the reference population is the sum of all contributions $\mathbf{DGV_r} = \mathbf{M f 1_q}$
- All f_{ij} are regressed according to the genetic map, with coefficients r varying from 1 to 0.5: $h_{ij} = r_{ij} f_{ij}$
- Eroded DGV ($\mathbf{DGV_e}$) is the sum of all regressed contributions $\mathbf{DGV_e} = \mathbf{M h 1_q}$
- An estimate of $\sqrt{\rho}$ is the regression coefficient of $\mathbf{DGV_e} = \mu + \sqrt{\rho}\, \mathbf{DGV_r} + \varepsilon$

Example of the Normande breed

G0 : the actual (2021) reference population of 69,220 genotyped cows with records
5 chromosomes, 13608 markers, 200 QTL simulated

Method 1 :
G1 : A new generation of 50,000 cows is simulated, born from 1000 sires and 50,000 dams sampled at random in G0

Regression of $\mathbf{DGV_n = \alpha + \rho\ DGV_r + \varepsilon}$
ρ Mean = 0.87, std = 0.015 over 30 replicates

Method 2 :

Regression of $\mathbf{DGV_e = \alpha + \sqrt{\rho}\ DGV_r + \varepsilon}$
ρ Mean = 0.84, std = 0.01 over 10 replicates

NB : both methods are not fully equivalent, as method 1 depends on the new generation sampled

Of course, various simulations scenarios (eg, #qtl) can be tested

Conclusion

Erosion exists and is due to long range LD over the genome.
Application of erosion of SNP estimates is required to obtain unbiased genomic prediction of candidates

Including a residual polygenic effect does not solve the problem

Erosion factor is specific to each reference population and can be estimated by simulation.
Deterministic formulae would be desirable

Breeding schemes with accelerated generations without updating reference populations accumulate more erosion, and are less attractive than generally believed

Erosion has been implemented in the French Single Step bovine evaluation since 2022

Thank you for your attention !

How to apply erosion ?

We assume the erosion factor per generation (ρ) is known

1-ρ is similar to a generalized recombination rate (Dekkers et al, GSE, 2021)

Adjusted (eroded) direct genomic values (DGV) of candidate i, with parents s and d, is

$$DGV_i = 0.5\ (DGV_s + DGV_d) \quad + \quad \rho^{k/2} \quad \phi_i$$

with ϕ the Mendelian sampling component

k the number of generations between i and the reference population summed over paternal and maternal pathways (Dekkers et al, 2021)

– up to the reference population

This formula is recursive: sire and/or dam's DGV should be eroded first if they do not belong to the reference.