

Andres Legarra, andres.legarra@inra.fr



INRA
SCIENCE & IMPACT

Genomic accuracy depends on... *what?*

- Starting points for the discussion diverge among people
 - Simulations, N_e , M_e , LD, relationships, n , h^2 , ...
- Historically:
 - Forefathers of animal breeding assumed large populations and infinitesimal genomes:
 - Selection index on “unrelated” candidates to selection
 - Relationship matrix
 - BLUP
 - This leads to meaningful estimates of accuracy from a few parameters.
- Can we reach a similar consensus?

What you can achieve with theory

Selection index

TABLE 8.1. WEIGHTS AND ACCURACY VALUES FOR PREDICTING ADDITIVE GENETIC VALUE FROM RECORDS OF VARIOUS RELATIVES. (h^2 IS HERITABILITY; r IS REPEATABILITY).

Records		Selection Index Weights	Accuracy = $r\Pi$
Individual	(1)	h^2	$\sqrt{h^2}$
	(n)	$nh^2/[1 + (n-1)r]$	$\sqrt{nh^2/[1 + (n-1)r]}$
Dam or sire or progeny	(1)	$h^2/2$	$.50\sqrt{h^2}$
	(n)	$nh^2/[1 + (n-1)r](2)$	$.50\sqrt{nh^2/[1 + (n-1)r]}$
Sire and dam	(1)	$h^2/2; h^2/2$	$.71\sqrt{h^2}$
	(n)	$.5nh^2/[1 + (n-1)r];$ $.5nh^2/[1 + (n-1)r]$	$.71\sqrt{nh^2/[1 + (n-1)r]}$
One grandparent		$h^2/4$	$.25\sqrt{h^2}$
Four grandparents		All $h^2/4$	$.50\sqrt{h^2}$
One great-grandparent		$h^2/8$	$.125\sqrt{h^2}$
Eight great-grandparents		All $h^2/8$	$.35\sqrt{h^2}$

BLUP

$$\begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} e \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(Z'X \quad Z'Z + \sigma^2 G^{-1}) \begin{pmatrix} \hat{\beta} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} Z'y \\ 0 \end{pmatrix}$$

The solutions are:

$$\begin{pmatrix} \hat{\beta} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} C^{xx} & C^{xz} \\ C^{zx} & C^{zz+} \end{pmatrix} \begin{pmatrix} X'y \\ Z'y \end{pmatrix} \quad \text{where}$$

$$\begin{pmatrix} C^{xx} & C^{xz} \\ C^{zx} & C^{zz+} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma^2 G^{-1} \end{pmatrix}^{-1}$$

The inverse of a non-full rank coefficient matrix is generalized inverse without affecting the PEV.

Pseudo-BLUP

the current generation. Such an index is called a pseudo-BLUP index. Thus the information sources are:

1. phenotypic own performance (P_i)
2. phenotypic information of full sibs (P_{FS})
3. phenotypic information of half sibs (P_{hs})
4. phenotypic information of progeny testing (P_{prog})
5. estimated breeding value of the sire (EBV_s)
6. estimated breeding value of the dam (EBV_d)
7. average estimated breeding values of the dams of the half sibs ($EBV_{hs-dams}$)

Four “horsemen” that “ride” genomic selection

- Simulations
- Linkage disequilibrium
- Relationships
- Effective number of segments



Everyone agrees that these are important notions



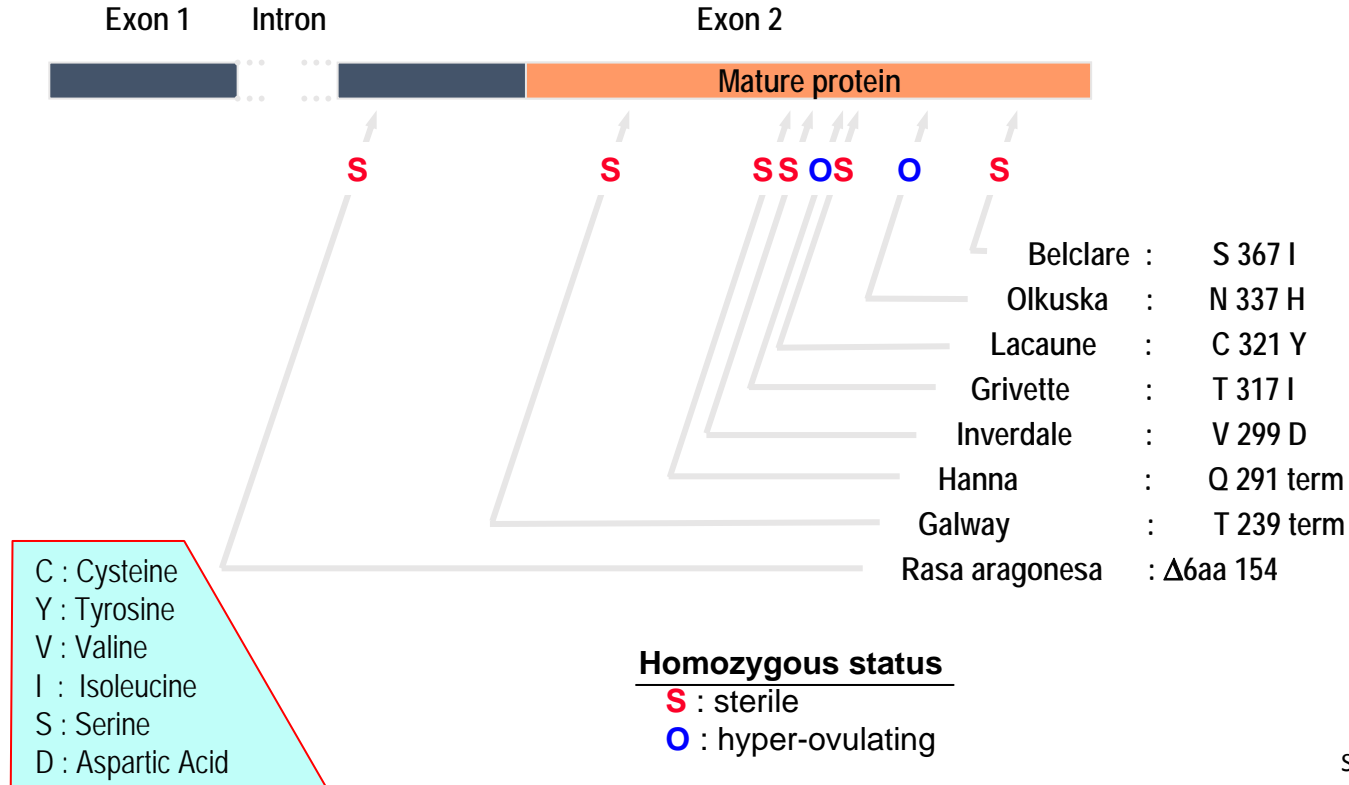
Simulations (1/2)

We rely too much on simulations as substitute for theory ...and we do very poor simulations

- Genes are not QTN: biallelic, single nucleotide polymorphisms
- Genes have coding parts, deletions, enhancers, promoters
- Genes are multiallelic with “fuzzy” locations (PRNP, α_{s1} casein...)
- Mutations are not the same across breeds
- Genes interact !!!!
- Genes mute



Eight known mutations of the BMP15 gene



Molecular characterization of the goat *CSN1S1*⁰¹ allele

Gianfranco Cosenza¹, Rosa Illario¹, Andrea Rando², Paola di Gregorio², Piero Masina² and Luigi Ramunno^{1*}

Mahè & Grosclaude, 1993). Such alleles are characterized by different mutations: single point mutations, responsible for premature stop codons, characterize null alleles of the *CSN2* (Rando et al. 1996; Persuy et al. 2000) and *CSN1S2* (Ramunno et al. 2001) *loci*; large DNA rearrangement (deletion/insertion) events of unknown origin and location characterize the two null alleles (*CSN1S1*⁰¹ and *CSN1S1*⁰²) of the *CSN1S1* locus (Martin et al. 1999).

More
than
QTN

Short communication: Evidence for a major gene by polygene interaction for milk production traits in German Holstein dairy cattle

M. Streit,* N. Neugebauer,* T. H. E. Meuwissen,† and J. Bennewitz*¹

GxG

rs1229984 (*ADH1B*)

	OR	95% CI
Overall	0.56	0.47–0.66 ($P = 4 \times 10^{-11}$)

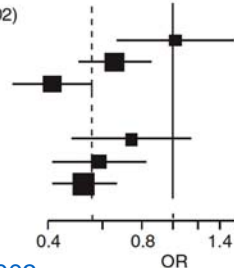
By drinking intensity ($\chi^2_1 = 14.0$; $P_{\text{trend}} = 0.0002$)

Never drinkers	1.02	0.66–1.56
≤Med	0.65	0.50–0.85
>Med	0.42	0.31–0.56

By smoking status ($\chi^2_1 = 3.45$; $P_{\text{trend}} = 0.063$)

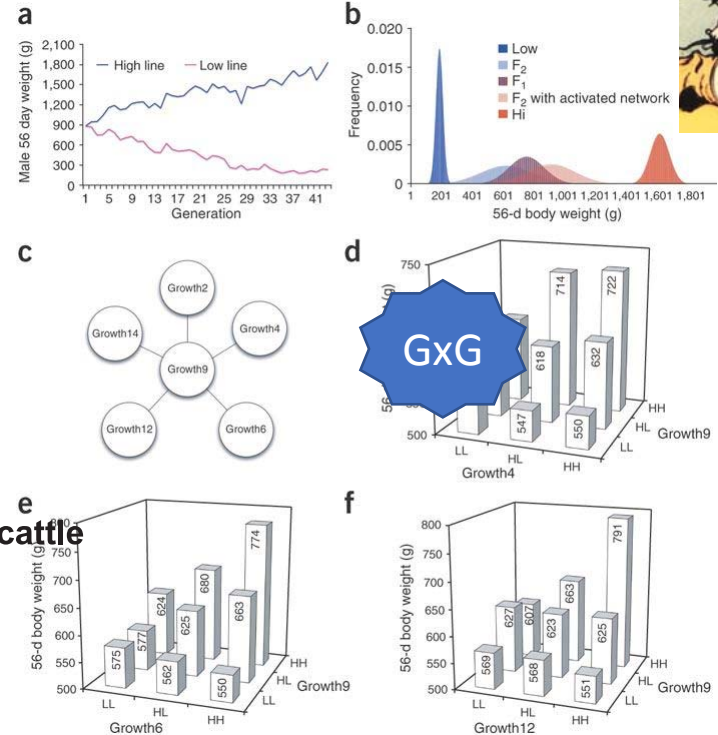
Never smokers	0.74	0.48–1.14
Former smokers	0.58	0.41–0.82
Current smokers	0.53	0.41–0.66

GxE



Hashibe et al. (2008)

<https://doi.org/10.1371/journal.pgen.1005765.g003>



Carlborg, Örfan, et al. "Epistasis and the release of genetic variation during long-term selection." *Nature genetics* 38.4 (2006): 418.



Simulations (2/2)

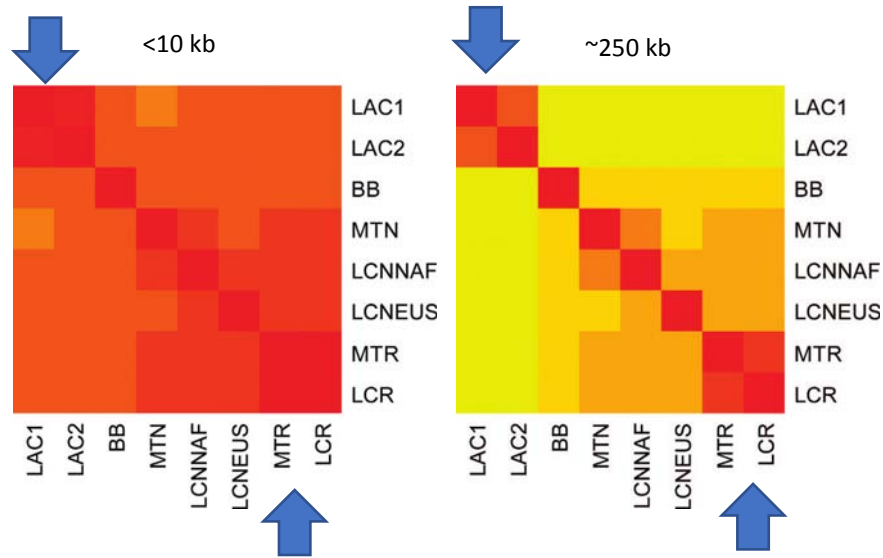
From simulations, we had the following “fake news”

- Additive variance diminishes quickly (but mutation, dominance, epistasis refill)
- Across-breed predictions are possible (but gene substitution effects depend on background, environment)
- Sequence is more accurate than SNP chips (but it has high redundancy and genes are not QTN)
- Bayesian regressions are better than GBLUP (most often they're not)



Linkage disequilibrium (1/2)

- We don't have consensual global statistics to describe
 - the relationship between LD and accuracy in a population
 - Reduction of genetic variance due to LD (i.e. Bulmer effect)
- All that we have is those pairwise r^2
- Do we need n-loci statistics or higher moments?
- Can we correlate LD measures with genomic accuracy?
 - Maybe not



- High LD phase agreement...
- But it does not result in higher accuracy



Linkage disequilibrium (2/2)

- Mental model of Bayesian regression: there will be at least one SNP in complete LD with the QTL
 - Maybe, but then there will be *many* SNP in almost-complete LD
- Mental model of GBLUP: does $\mathbf{ZZ}' \approx \mathbf{QQ}'$?
- Is any of these models correct? To what extent?



Relationships (1/2)

Several definitions not easy to conciliate

Probabilistic: assuming an unrelated base population (which one ?)

- Expected IBD relationships *conditional* on the pedigree (**A**)
- Real unobserved IBD relationships ($\tilde{\mathbf{R}}$)

Statistical: using cross-products

- VanRaden's **G** (base population is whatever we use in p)

Pedigrees go back in time “forever”



A closed rabbit line of 45 discrete generations:
934 sires (yellow) with 1,950 dams (green) and
3,492 progeny (red).

Universidad Politécnica de Valencia, Spain

All G-matrices are equal



Allele coding in genomic evaluation

Ismo Strandén^{1*} and Ole F. Christensen²

On curious properties of genomic relationship matrices in mixed models

Bruce Tier and Karin Meyer

Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia

Which GRM?

► GRM that look very different ...

→ different allele coding, centering, scaling, etc.

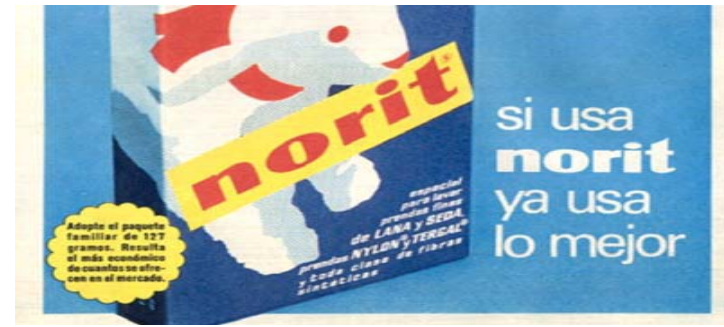
... give 'equivalent' predictions → shifted breeding values

... but not necessarily the same prediction error variances

Strandén, I., Christensen, O. F. 2011. Allele coding
in genomic evaluation. Genet. Sel. Evol. 43:25.

► IMPLICATIONS?

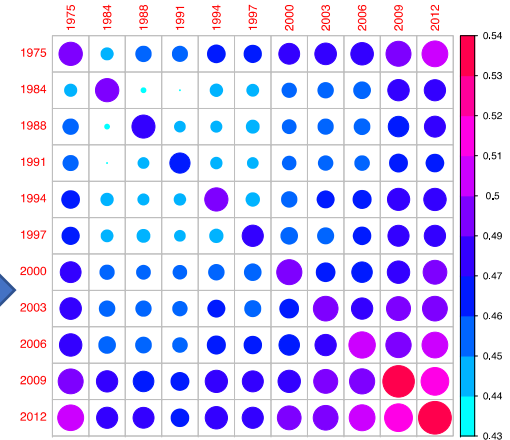
Relationships (2/2)



We advertise the unified theory of relationships based on metafounders

- \mathbf{G} = crossproduct of $\mathbf{Z} = \{-1,0,1\}$ is the absolute reference (Christensen, 2012)
- As a byproduct, pedigree base populations are related
- Other options?

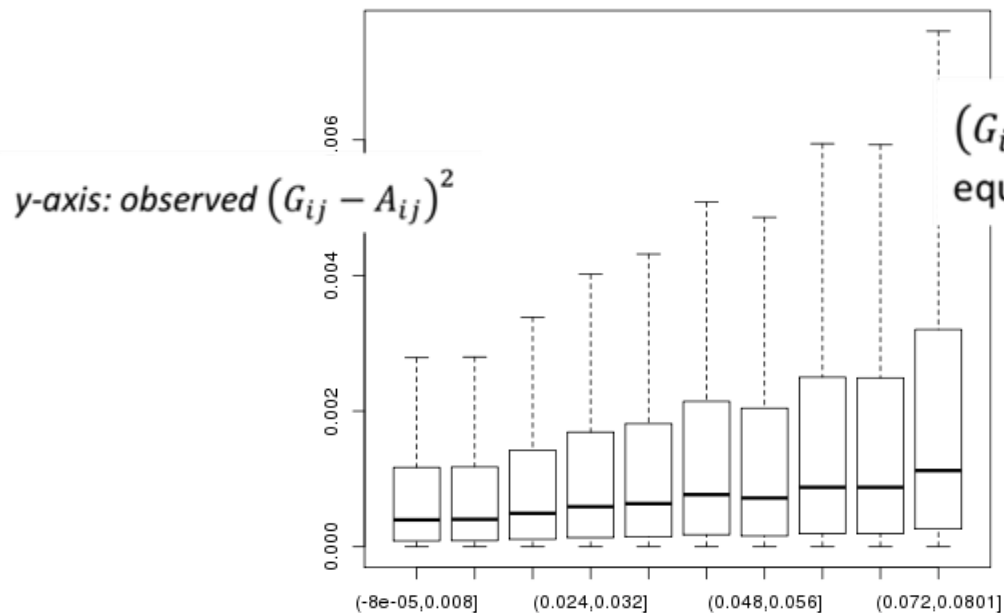
Relationships within/across Genetic Groups ,
Manech Tete Rousse





Effective number of segments (Me) (1/3)

- Me describes the “non infinitesimality” of the genome
 - If $Me = \infty$ (infinitesimal) then $\tilde{R}_{ij} = A_{ij}$ and $Var(\tilde{R}_{ij} - A_{ij}) = 0$
 - If $Me = 1$ (single locus) then $Var(\tilde{R}_{ij} - A_{ij}) = 4(\phi_{ij,ij} - \phi_{ij}\phi_{ij})$
- To me, Me is a parameter of the population like h^2
- To other people (Lee, Wientjes) this is data specific: an empirical quantity $\frac{1}{var(G_{ij}-A_{ij})}$ or $\frac{1}{r^2}$



$(G_{ij} - A_{ij})^2$ increases according to theoretical equation $4(\phi_{ij,ij} - \phi_{ij}\phi_{ij})$ based on pedigree

x-axis: expected $Var(\tilde{R}_{ij} - A_{ij})$ for 1 locus

Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals

H. Wang¹, I. Misztal² & A. Legarra³



Effective number of segments (Me) (2/3)

Paradoxes of data specific Me ; for 2 generations (Hill and Weir 2011) :

- $Me = \infty$ between father and offspring
- $Me = 636$ for fullsibs,
- $Me = 318$ for halfsibs and
- $Me = 503$ for cousins

I'd rather prefer a population parameter from which to deduce these values...



Effective number of segments (Me) (3/3)

Can it be a population parameter?

- The distribution of segments from an ideal infinite base population is described by the theory of junctions, too complicated ☹
- Segments should be created by meiosis and disappear by drift
- Is there an equilibrium?

An attempt to conclude

- Simulations are misleading
- LD is not well quantified
- What do we mean by relationship?
- Can we better define *Me*?



- We animal breeders should make an effort to clearly define concepts
- Lack of formalization leads to improvisation and misunderstanding
- Lack of agreement leads to disparate conclusions

Acknowledgments

- Poctefa funding project “ARDI”
- INRA SelGen metaprogram, project EpiSel



INRA
SCIENCE & IMPACT