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>
<thead>
<tr>
<th>Records</th>
<th>Selection Index Weights</th>
<th>Accuracy = ^TT^</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>h^2</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>(a)</td>
<td>na^2 {1 + (n-1)p}</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>Dam or sire or progeny</td>
<td>h^2/2</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>(a)</td>
<td>na^2 {1 + (n-1)p}{2}</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>Sire and dam</td>
<td>h^2/2; h^2/2</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>(a)</td>
<td>na^2 {1 + (n-1)p}{2}</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>One grandparent</td>
<td>h^2/4</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>Four grandparents</td>
<td>All h^2/4</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>One great-grandparent</td>
<td>h^2/8</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
<tr>
<td>Eight great-grandparents</td>
<td>All h^2/8</td>
<td>(\frac{h^2}{h^2})</td>
</tr>
</tbody>
</table>

BLUP

\[
\begin{bmatrix}
\mathbf{u} \\
\mathbf{v}
\end{bmatrix} = \begin{bmatrix}
\mathbf{0} \\
\mathbf{0}
\end{bmatrix} \begin{bmatrix}
\mathbf{0} \\
\mathbf{v}
\end{bmatrix}
\]

The solutions are:

\[
\begin{bmatrix}
\mathbf{u} \\
\mathbf{v}
\end{bmatrix} = \begin{bmatrix}
\mathbf{C}^{-1} \\
\mathbf{C}^{-1}
\end{bmatrix} \begin{bmatrix}
\mathbf{X}^T \\
\mathbf{Z}^T
\end{bmatrix}
\]

Pseudo-BLUP

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

1. phenotypic own performance (\(P_o\))
2. phenotypic information of full sibs (\(P_{ps}\))
3. phenotypic information of half sibs (\(P_{phs}\))
4. phenotypic information of progeny testing (\(P_{p\text{prop}}\))
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_{h-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, \( \alpha_{s1} \) casein...)
• Mutations are not the same across breeds
• Genes interact !!!!
• Genes mute
Eight known mutations of the BMP15 gene

<table>
<thead>
<tr>
<th>Exon 1</th>
<th>Intron</th>
<th>Exon 2</th>
<th>Mature protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
<td>S S O S O S</td>
<td></td>
</tr>
</tbody>
</table>

- Lacaune: C 321 Y
- Inverdale: V 299 D
- Hanna: Q 291 term
- Galway: T 239 term
- Belclare: S 367 I
- Olkuska: N 337 H
- Grivette: T 317 I
- Inverdale: V 299 D
- Hanna: Q 291 term
- Galway: T 239 term
- Rasa aragonesa: Δ6aa 154

**Homozygous status**

- S: sterile
- O: hyper-ovulating

C: Cysteine
Y: Tyrosine
V: Valine
I: Isoleucine
S: Serine
D: Aspartic Acid
Molecular characterization of the goat CSN1S1\textsuperscript{01} allele

Gianfranco Cosenza\textsuperscript{1}, Rosa Illario\textsuperscript{1}, Andrea Rando\textsuperscript{2}, Paola di Gregorio\textsuperscript{2}, Piero Masina\textsuperscript{2} and Luigi Ramunno\textsuperscript{3}

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 (Rumunno et al. 2001) loci; large DNA rearrangement (deletion/insertion) events of unknown origin and location characterize the two null alleles (CSN1S1\textsuperscript{01} and CSN1S1\textsuperscript{02}) of the CSN1S1 locus (Martin et al. 1999).

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

M. Streit,\textsuperscript{*} N. Neugebauer,\textsuperscript{*} T. H. E. Meuwissen,\textsuperscript{†} and J. Bennewitz\textsuperscript{‡}

Hashibe et al. (2008)
https://doi.org/10.1371/journal.pgen.1005765.g003

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

Legarra et al. 2014

<10 kb ~250 kb
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 $ZZ' \approx QQ'$?
- 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 \((\tilde{R})\)

**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).

All G-matrices are equal

Allele coding in genomic evaluation

Universidad Politécnica de Valencia, Spain
Relationships (2/2)

We advertise the unified theory of relationships based on metafounders

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

Other options?
Effective number of segments ($Me$) (1/3)

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

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

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

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

H. Wang$^1$, L. Mezaf$^2$ & A. Legarra$^3$
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