Next generation plant and animal breeding

John Hickey
Plant VS animal breeding

• Animal breeding
  – Population improvement
  – Quantitative traits
  – Objective economic selection index
  – Automated computer driven systems

• Plant breeding
  – Finding the best individual in current population
  – Fixing lines for specific “characters”
  – Major QTL
  – G*E is important
  – Pyramiding and stacking of QTL
  – “Population improvement” = biproduct
Traditional livestock breeding

• Recurrent selection

• Response to selection depends on
  – Time to evaluate MS term
  – Accuracy of MS term
  – Intensity (cost)
  – Variance available

• Traditionally done via progeny testing
  – 100 candidate bulls each mated to 100 cows
  – Breeding value of candidate ≈ mean milk yield of progeny
  – Expensive, takes 5 years and accuracy is not 1.0!

• Multiple trait economic index used for selection
  – e.g 20 traits
Why is genomic selection attractive?

• Directly addresses 3 of the 4 components of genetic gain
  – Time - via generation interval
  – Accuracy - via better analysis/use of data
  – Selection intensity - via cost

• Indirectly addresses diversity
In theory you can have sibs that are genetically unrelated. This is why I am different from my brother.

\[
A = \begin{bmatrix}
1.00 & 0.50 & 0.50 & 0.50 & 0.50 \\
0.50 & 1.00 & 0.25 & 0.25 & 0.25 \\
0.50 & 0.25 & 1.00 & 0.25 & 0.25 \\
0.50 & 0.25 & 0.25 & 1.00 & 0.25 \\
0.50 & 0.25 & 0.25 & 0.25 & 1.00 \\
\end{bmatrix}
\]

And “hidden” relationships.
Genomic selection

• GS0.0
  – The original model
  – Linkage disequilibrium based

• GS1.0
  – What has happened in practice
  – Linkage based

• GS2.0
  – The future
  – LD and QTN based
  – Requires lots of data
GS1.0 has been a major success

• Accurate breeding values possible

• Shorter generation interval

• Dynamics are now well understood

• Good systems in place

• Most importantly for the future
  – Breeding programs can generate lots of genomic data
What has been achieved

**Single-Step Genomic Evaluation**

- Phenotypes collected until **2009**
- Progeny genotyped (60k): animals born in 2009 and after (n=2,023)

<table>
<thead>
<tr>
<th>TRAIT</th>
<th>ACC EBV</th>
<th>ACC GEBV</th>
<th>increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number born</td>
<td>0.25</td>
<td>0.42</td>
<td>68%</td>
</tr>
<tr>
<td>Stillborn</td>
<td>0.26</td>
<td>0.43</td>
<td>65%</td>
</tr>
<tr>
<td>Survival birth - weaning</td>
<td>0.17</td>
<td>0.26</td>
<td>53%</td>
</tr>
<tr>
<td>Litter weaning weight</td>
<td>0.23</td>
<td>0.35</td>
<td>52%</td>
</tr>
<tr>
<td>Interval weaning - mate</td>
<td>0.17</td>
<td>0.30</td>
<td>76%</td>
</tr>
</tbody>
</table>

July 2012

- With small number of individuals (2012)
- Higher now

---

<table>
<thead>
<tr>
<th>Trait</th>
<th>Acc PA EBV</th>
<th>Acc GEBV</th>
<th>EBV Acc of MS term</th>
</tr>
</thead>
<tbody>
<tr>
<td>USDA Dairy Total Merit</td>
<td>0.55</td>
<td>0.87</td>
<td>?</td>
</tr>
</tbody>
</table>
What an animal pedigree looks like

- 6 generations random mating followed by 3 generations
- 200 individuals per generation
The imputation problem for a 2.5k low density chip in pigs
The cost and accuracy of “sensible” strategies

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Other</th>
<th>MGS + PGS</th>
<th>MGD + PGD</th>
<th>Sire</th>
<th>Dam</th>
<th>Candidates</th>
<th>Individual cost</th>
<th>Accuracy of Imputation R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>60k</td>
<td>60k</td>
<td>0</td>
<td>60k</td>
<td>0</td>
<td>384</td>
<td>!</td>
<td>0.878</td>
</tr>
<tr>
<td>SC2</td>
<td>60k</td>
<td>60k</td>
<td>384</td>
<td>60k</td>
<td>384</td>
<td>384</td>
<td>$20.58</td>
<td>0.929</td>
</tr>
<tr>
<td>SC3</td>
<td>60k</td>
<td>60k</td>
<td>3k</td>
<td>60k</td>
<td>3k</td>
<td>384</td>
<td>$24.74</td>
<td>0.950</td>
</tr>
<tr>
<td>SC4</td>
<td>60k</td>
<td>60k</td>
<td>6k</td>
<td>60k</td>
<td>6k</td>
<td>384</td>
<td>$26.28</td>
<td>0.944</td>
</tr>
<tr>
<td>SC5</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>384</td>
<td>$34.84</td>
<td>0.964</td>
</tr>
<tr>
<td>SC6</td>
<td>60k</td>
<td>60k</td>
<td>0</td>
<td>60k</td>
<td>0</td>
<td>3k</td>
<td>!</td>
<td>0.968</td>
</tr>
<tr>
<td>SC7</td>
<td>60k</td>
<td>60k</td>
<td>384</td>
<td>60k</td>
<td>384</td>
<td>3k</td>
<td>!</td>
<td>0.972</td>
</tr>
<tr>
<td>SC8</td>
<td>60k</td>
<td>60k</td>
<td>3k</td>
<td>60k</td>
<td>3k</td>
<td>3k</td>
<td>$35.58</td>
<td>0.984</td>
</tr>
<tr>
<td>SC9</td>
<td>60k</td>
<td>60k</td>
<td>6k</td>
<td>60k</td>
<td>6k</td>
<td>3k</td>
<td>$41.28</td>
<td>0.983</td>
</tr>
<tr>
<td>SC10</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>3k</td>
<td>$49.84</td>
<td>0.993</td>
</tr>
<tr>
<td>SC11</td>
<td>60k</td>
<td>60k</td>
<td>0</td>
<td>60k</td>
<td>0</td>
<td>6k</td>
<td>!</td>
<td>0.982</td>
</tr>
<tr>
<td>SC12</td>
<td>60k</td>
<td>60k</td>
<td>384</td>
<td>60k</td>
<td>384</td>
<td>6k</td>
<td>!</td>
<td>0.983</td>
</tr>
<tr>
<td>SC13</td>
<td>60k</td>
<td>60k</td>
<td>3k</td>
<td>60k</td>
<td>3k</td>
<td>6k</td>
<td>!</td>
<td>0.986</td>
</tr>
<tr>
<td>SC14</td>
<td>60k</td>
<td>60k</td>
<td>6k</td>
<td>60k</td>
<td>6k</td>
<td>6k</td>
<td>$48.58</td>
<td>0.991</td>
</tr>
<tr>
<td>SC15</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>6k</td>
<td>$62.84</td>
<td>0.996</td>
</tr>
<tr>
<td>SC16</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>$120.00</td>
<td>1.000</td>
</tr>
</tbody>
</table>

nSires = 480
nDams = 11884
nCandidates = 100000

60k chip = $120
6k chip = $48
3k chip = $35
384 chip = $20 (This is now $15, we can make it $11)
What is the recipe?

- Clear goal for use of GS
- Do well designed expensive research and then make it cheap
- Operational simplicity

- Good IT infrastructure

- Simple statistical model

- Reliable, consistent, standardized, and cheap (@FEC) genotyping and imputation

Crops examples
- Early generation
- Intensity/time/environments/cost at later generations
- Robustness to G*E
- Hard to measure traits (broad breeding goals)
- Recurrent selection

Old

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + A^{-1}\lambda
\end{bmatrix}^{-1}
\begin{bmatrix}
X'y \\
Z'y
\end{bmatrix} = \hat{b} \quad \hat{u}
\]

New

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + G^{-1}\lambda
\end{bmatrix}^{-1}
\begin{bmatrix}
X'y \\
Z'y
\end{bmatrix} = \hat{b} \quad \hat{u}
\]

60,000 SNP on males
384 SNP on candidates
$15.50$ FEC
Implementation in pigs

- Largest breeding program globally
  - 100,000 candidates per year

- Start 2009 - Four key projects identified
  - Simple statistical algorithm (One post-doc)
    - Needed to minimize changes to routine
  - Genotype imputation and genotyping strategy (One post-doc)
    - Needed to be less than $22
  - Experiment to proof the principle (One post-doc)
    - Used historical data – because good databases existed
  - Infrastructure team (4 people)

- Routine testing as of 2012
- Routine use September 2013
- Accuracy doubled (Validation confirmed!)
- Total cost – GUESS ≈ Data $600,000 + 7 FTE
What are we doing next?

• Sequencing millions of animals for Genomic Selection 2.0

  – Agri-Tech Catalyst funding (applied for)
Overall hypothesis

- Sequence data has huge potential in livestock breeding

- Huge volumes of sequence are needed to realize this potential

- Individual breeding programs with 1 million animals with sequence information:
  - Will be normal “by the end of the decade”
In the future …

- Huge data sets
- Sequenced
- Phenotyped
- Industrial scale fine mapping
  - Perhaps 50% of total genetic variance mapped to its causal variants
How might we benefit?

• We can do better what we do today
  – Operational simplicity
  – Persistent predictions
    • (e.g. multiplier layers, across breeds, train with commercial data)
  – Cheaper

• Or we could be bolder
  – Explicit utilization of de-novo mutations
  – Higher recombination rate
  – Genome editing for complex traits
  – Rapid response to some disease outbreaks
  – Much greater biological understanding
  – Better monitoring and utilization of variation
  – ??????
Genetic gain – edits per sire

Editing all 25 selected sires

1000 QTN

10,000 QTN

20 edits

No editing
Plant applications
What we have learnt (conjecture!)

• Accuracy affected by
  – Relationships between the training and prediction population
  – Size of the prediction set
  – Heritability of trait
  – Random sampling
  – Interactions between these factors

• Accuracy NOT affected by
  – Marker density beyond 10000 markers
  – Statistical model (except the Legarra model and MT models)
  – Level of linkage disequilibrium (Sort of!!!!!)

• LINKAGE information drives predictions

• How to validate
Relationships between T & P

$R^2 = 0.962$

Accuracy of GEBV

- Mean of the Top Ten Relationships
- Training set size = constant
- Ne/LD = constant
How to validate

- If it is about population improvement it is about the MS term

- If accuracy reflects structure population replacement will result

- Predicting mid-parent with 100% accuracy will not affect response to selection

- Response to selection is only affected by the accuracy of the MS term
Relationships between T & P

• Simulation in crops

• Target is prediction of breeding value of $F_2$ in BP-X

• Simulated 121 connected bi-parental families with different degrees of relatedness to BP-X

• Polygenic trait

• Large range of phenotype numbers

• Large range in marker density

• Accuracy is for MS term in BP-X
Pedigree design

Coalescent simulator to generate 1000 founder haplotypes according to historical effective population size of Maize

200 doubled-haploid parents for the population created by randomly sampling 200 gametes and doubling

40 bi-parental families that are unrelated to BP-X (BP-U)

Grandparent 1
- 10 BP families with Grandparent 1 in common (BP-G)
  - Parent 1
    - 20 BP families with Parent 1 in common (BP-P)
  - 10 BP families with Grandparent 2 in common (BP-G)
    - Parent 2
      - 20 BP families with Parent 2 in common (BP-P)

Grandparent 2
- 10 BP families with Grandparent 2 in common (BP-G)
  - Parent 1

Grandparent 3
- 10 BP families with Grandparent 3 in common (BP-G)
  - Parent 2

Grandparent 4
- 10 BP families with Grandparent 4 in common (BP-G)
  - 40 bi-parental families that are unrelated to BP-X (BP-U)
    - BP-X
Train and predict in BP-X

Training inside BP-X
50 individuals + 200 markers = accuracy of MS term of 0.60
Large sampling variance

Training in 40 BP with parent in common with BP-X
800 individuals
400 markers

Training in 40 BP unrelated to BP-X
6,000-20,000 individuals
10,000 markers
ROBUST
LD-LA conflict = model
Example breeding program

Population improvement

Product development

<table>
<thead>
<tr>
<th>Generation</th>
<th>Number of plants</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td></td>
<td>Select parents and cross</td>
</tr>
<tr>
<td>Year 2</td>
<td>F₁</td>
<td>Bulk seed; space plant for higher yield</td>
</tr>
<tr>
<td>Year 3</td>
<td>F₂</td>
<td>Space plant for easy visual selection</td>
</tr>
<tr>
<td>Year 4</td>
<td>F₃</td>
<td>Select and plant in spaced rows</td>
</tr>
<tr>
<td>Year 5</td>
<td>F₄</td>
<td>Identity superior rows; select 3–5 plants to establish family in progeny rows</td>
</tr>
<tr>
<td>Years 6–7</td>
<td>F₅–F₁₀</td>
<td>Establish family progeny rows; select individual plants to advance each generation</td>
</tr>
<tr>
<td>Year 8</td>
<td>F₇</td>
<td>Conduct preliminary yield trials; select individual plants to advance</td>
</tr>
<tr>
<td>Years 9–11</td>
<td>F₈–F₁₀</td>
<td>Conduct advanced yield trials with more replications and over locations and years</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cultivar release</td>
</tr>
</tbody>
</table>
One way to use GS

- Do a few rounds of recurrent selection at $F_2$ stage within each bi-parental

- How many rounds?
  - Diversity
  - Accuracy

- What is the accuracy trajectory?

- What training population?

- What genotyping platform?

- GplusE proposal is building towards this

Diagram:

- Year 1: $P_1 \times P_2$
- Year 2: $F_1$
- Cycle 1: select best $F_2$'s on GEBV
- Cycle 2: select best $F_2$'s on GEBV
- Cycle 3: select best $F_2$'s on GEBV
- Year 3: $F_2$
- Year 4: $F_3$
- Year 5: $F_4$
- Years 6–7: $F_5$–$F_6$
- Year 8: $F_7$
- Years 9–11: $F_8$–$F_{10}$
Breeding program design

Population improvement

- Recurrent selection
  - Select best F1’s on GEBV
  - Select best F1’s on GEBV
  - Select best F1’s on GEBV
  - Select best F1’s on GEBV
  - Select best F1’s on GEBV

Product development

- Select best F1’s on GEBV to pass to product development

Population improvement

- Operated by a quantitative geneticist
  (Bonus on basis of program wide ΔG per year!!)
- Multiple trait and environment selection index
- Manage utilization of diversity
- Entirely genomic selection
- Winter nurseries

Product development

- Operated by a traditional breeder
  (Bonus on basis of new lines released!!)
- Traditional method supplemented by GS

Could pepper with genomic information
Coalesce?

- **Methods**
  - Statistical methods are the same
  - Genotyping

- **Quantitative genetics of traits**

- **Plant breeders can bring**
  - Phenotyping expertise
  - Trial designs
  - Biological knowledge and appreciation!

- **Animal breeders**
  - Systematic objectivity!!
  - Quantitative methods
Acknowledgements

• Genus PLC
• Aviagen LTD
• CIMMYT
  – CRP Maize/Wheat
  – Seeds of Discovery
• ICRISAT
• GplusE project with NIAB, KWS, RAGT, Limagrain, Elsoms and BBSRC

• G. Gorjanc
• S. Dreisigacker
• S. Hearne
• B. Prasanna
• J. Crossa
• J. Jenko
• G. Meszaros
• J. Woolliams
• M. Cleveland
• A. Kranis
• D. McLaren
• I. Mackay
• E. Ober
• A. Bentley