Sequencing millions of animals for GS2.0
John Hickey
John Woolliams, Janez Jenko, Gábor Mészáros
Miguel Perez Enciso, Mara Battagin, Andreas Kranis
Alan Mileham, Matthew Cleveland, Gregor Gorjanc

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"

Sequencing hypothesis
- Target is the population not the individual
- Think in terms of sequencing an individual as part of a population and not in isolation
- Individuals share haplotypes
  - Why sequence the same haplotype 30 times in 30 individuals?
- Cost assumptions!
  - Library = $3/$15/$60
  - 1x sequencing = $20/$120
  - Thus 0.1x sequence for $5!!! Or $30!!!

Overview
- Review of GS1.0
- Properties and promises of GS2.0
- GS2.0 – how do we get the data?
- Example benefits of GS2.0
  - That go beyond the accuracy!!!

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
GS1.0 depends on relationships

\[ R^2 = 0.962 \]

How to pivot line on this point and move it up?

We have so little information for imputation!

- Not all low-density markers are informative
  - At all positions
  - For both gametes
- Example data
  - 23 markers on one chromosome
    - (circa 700 markers genome wide)
    - On average only 2.45 are informative
    - Still achieve an 0.97 correlation!

In the future …..

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

Imputation has been central to GS1.0

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Other</th>
<th>nSires</th>
<th>nDams</th>
<th>nCandidates</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>60k</td>
<td>0</td>
<td>60k</td>
<td>0.878</td>
</tr>
<tr>
<td>SC2</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>384</td>
<td>60k</td>
<td>0.929</td>
</tr>
<tr>
<td>SC3</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.964</td>
</tr>
<tr>
<td>SC4</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.972</td>
</tr>
<tr>
<td>SC5</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.994</td>
</tr>
<tr>
<td>SC6</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.997</td>
</tr>
<tr>
<td>SC7</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.998</td>
</tr>
<tr>
<td>SC8</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>0.999</td>
</tr>
<tr>
<td>SC9</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>60k</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Time to change our thinking?

- Sequence information is falling in price
- Can we sequence everybody?

How will we sequence everybody?

- How will we benefit?
- At next WCGALP results of studies with >1 million sequenced individuals will be presented!!

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
  - ??????
How will we generate sequence on everybody?

- 1 million individuals at $1000 = $1 billion!!!
- Imputation will be central
- Could use a standard approach

Some definitions!!!

- \( x \) = the number of reads at a position
  - \( 1^x \) = 1 read, \( 2^x \) = 2 reads, etc.
  - \( 30^x \) is typical

Everybody at low-coverage

- Rather than sequence the individual at high \( x \) we try to sequence haplotypes at high \( x \)
  - Many individuals share long haplotypes
  - Sequence data can be used on both sides of imputation task

- Akin to a “Big Data” approach
  - Lots of cheap data
  - Each data point is unreliable
  - Algorithms can recover quality
  - Total volume overcomes the individual weakness

Also different ways to reduce representation

Little example

True haplotypes

8x sequencing of individuals
Prototype algorithm

- 158,281 GBS markers (1.1x)
- 504 DH lines
- 44% of markers missing
  - Low-coverage
  - Mutations in restriction site
- After running imputation algorithm only 20% of markers missing

Everybody at low X

1.000 x of sequence in total
1. Sequence 25 sires at 40x and genotype progeny with 200 SNP
2. Sequence 1000 progeny at 1x

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Imputation accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.752</td>
</tr>
<tr>
<td>2</td>
<td>0.921</td>
</tr>
<tr>
<td>*</td>
<td>0.887</td>
</tr>
</tbody>
</table>

*Infinite x sequence on sires and genotype progeny with 200 SNP
Sequence sperm cells

Sperm cells are haploid and thus large pieces of the sire are phased 1500x in total

1. Sequence sires at 30x and genotype progeny with 200 SNP

2. Do not sequence sires
   Sequence 1500 sperm cells (30 per sire) at 1x
   Genotype progeny with 200 low-density SNP markers

3. Sequence sires at 10x
   Sequence 1000 sperm cells (20 per sire) at 1x
   Genotype progeny with 200 low-density SNP markers

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Imputation accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.756</td>
</tr>
<tr>
<td>2</td>
<td>0.933</td>
</tr>
<tr>
<td>3</td>
<td>0.967</td>
</tr>
</tbody>
</table>

What is the power of such data for achieving what we currently do?

Expand training set

Genome editing

GE = the process of precise editing genome

Nucleotides can be added deleted replaced

How might we benefit?

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

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

Genome editing

• Focus to date has been on simple traits

• Most traits of economic importance are complex

• When it will be cost effective to edit lots of alleles in sires …..

• … what will the benefit of GE be for complex traits?

• GE = controlled recombination
  – Use existing variation faster
Genetic gain – edits per sire

Recombination is the rate limiting factor in all of genetics!

- If the ideotype is Q/Q and A/A
  - Q/A
  - Q/A

- If we start with
  - Q/a
  - q/A

- Without recombination selection can never deliver the ideotype!!

Manipulating recombination

- Most standing genetic variance is "trapped"
- Could higher recombination "release" more variance and thus enable greater genetic progress?
- Recombination is under genetic and environmental control
  - Heritability of 0.1 to 0.3
  - Causal variant that increases it by 15%
  - Included in selection index
  - Genome editing
  - Environmental manipulation

Exhausting variance at the same rate

What needs to happen next?

- Huge data sets
- Phenotyped
- Sequenced
- Current biggest research question is:
  - How much variance can our industrial fine mapping explain?
- Dairy has been the poster boy for GS1.0
  - Pigs and poultry will be for GS2.0!!!

Acknowledgements

- My Co-authors
- Genus/PIC
  - Dave McLaren
  - Alan Mileham
- Aviagen