Haplotyping, imputation, and genomic selection

J.M. Hickey
Outline

• Three components to talk
  – Genomic selection - background
  – AlphaImpute
    • Method to impute genotypes in pedigreed populations
  – Harnessing the output of AlphaImpute to increase the accuracy of Genomic Selection
Genomic selection

- Meuwissen, Hayes, Goddard (2001) Genetics
- Complete coverage of genome with markers
- All QTL in linkage disequilibrium with at least 1 marker
- No QTL size thresholds needed
- Accurate breeding values of individuals at birth

Reference population
Known genotypes and phenotypes

Selection candidates
Marker genotypes

Prediction equation
Genomic breeding value = \( w_1x_1 + w_2x_2 + w \cdot x \)

Selected breeders
Using genomic breeding values

Michael E. Goddard & Ben J. Hayes
Nature Reviews Genetics 10, 381-391 (June 2009)
Typical data structure

- **Generation**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10

- **Information available**
  - Pedigree information on all animals
  - Genotype information

- **Validation population**
  - 500 selection candidates
  - 500 selection candidates
  - 500 selection candidates
Genomic selection

- Meuwissen, Hayes, Goddard (2001) Genetics
  - Complete coverage of genome with markers
  - Exploits linkage disequilibrium between markers and QTL
  - No QTL size thresholds needed
  - Accurate breeding values of individuals at birth

Was based on common QTL of large size
Relatively easy to find
Works across the population

BUT

Real data results indicate that the true model is polygenic
Lots of phenotypes of genotyped animals are needed

of individuals at birth
Size of training data set

As the number of animals in training population increases the accuracy increases.
Size of training data set

- Empirical results (VanRaden et al., 2009)
  - 3,500 Holstein bulls with 38,000 SNPs
    - Highly accurate phenotypes (EBVs with high accuracy)
  - Reducing marker number by 75%
    - Reduced accuracy for net merit from 0.53 to 0.50
  - Decreasing training population by 68%
    - Reduced accuracy from 0.53 to 0.35
The promise and problem of sequence data
So what do we need?

- Genotyping at high density is expensive
- Genotyping at low density is cheaper
- Imputation is free
- A genotype imputation method that allows us to get lots of genotyped and phenotyped animals at a low cost
What underlies a genotype?

Father

Mother

Proband

1111022211111111121021

1011112121121212211221121

10110122121110212101220022
What underlies a genotype?

- Father: 1010011101100111001110011
  - 010100111000110011001
  - 11110221111111111121021

- Mother: 001001111001010110011001
  - 01011101011111111111110
  - 101111212112122121121

- Proband: 1010011101100111001110011
  - 0001001111001010110011001
  - 10110122121110212101220022
The basic idea of imputation

General pedigree with its haplotypes represented
Segregation analysis and haplotype library imputation

- Individual’s are densely, sparsely, or not genotyped
- Pedigree information available
- Single locus segregation analysis for each SNP
- Match each pair of haplotypes with low density genotypes and genotype probabilities

Genotyping strategy in terms of high density, low density and not genotyped

Haplotype library for population
At the beginning the aim was

• An imputation method which would be:
  – Robust
    • Pedigree structure
    • Genotype density
    • Genotyped animals
  – Scalable
    • Large pedigrees
    • Large numbers of SNP
  – Fast
  – Suitability for routine/weekly use
  – Impute mapped and unmapped SNP
  – Impute ALL genotypes for ALL animals in the pedigree
The basic engine is long-range phasing

- Long-range phasing
  - A fast rule based phasing method
  - Basically it is a pedigree free linkage approach
  - Completely unrelated animals contribute phasing information
  - Even animals from a different breed can contribute
Phasing a Trio

Father

101001110111001111001110011
010101111001100110011010

Mother

00010011110010110011001110
10101101011011111111111110

Proband

10100111011100111001110011
00010111110100101100110011
Phasing a Trio

Father

010101101110011100111001

0101011000110011010101

Proband

10100110111001110011011011

00010011110011011011001110

Cannot phase this locus!!
Surrogate parents are the driver of long range phasing

Proband

Genotype:

Father:

Pat Hap: 1010011101110011100110011
Mat Hap: 0101011110001100110011010
Genotype: 1111022211111111111112101
Proband G: 1011022121110212101220022
Opp Homo: **********************************************

Mother:

Pat Hap: 000100111100101100110011
Mat Hap: 10101101011111111111111110
Genotype: 10111121211121212211221121
Proband G: 1011022121110212101220022
Opp Homo: **********************************************

Other:

Pat Hap: 10111101011111010101100110011
Mat Hap: 101011100101111111111111110
Genotype: 20212220111121221100222220
Proband G: 1011022121110212101220022
Opp Homo: **x*****x**x**x**x**x**x**x**x**x**x**x**x

Not a surrogate parent!
Surrogate parents are the driver of long range phasing

Proband:
- Genotype: 10100111011100111001110011
- Other: 10100111011100111001110011
- Father:
  - Pat Hap: 10100111011100111001110011
  - Mat Hap: 000100111100101100110011
  - Genotype: 01010111100011000110011010
  - Proband G: 10110122121110212101220022
  - Opp Homo: ***********************

Mother:
- Pat Hap: 10100111011100111001110011
- Mat Hap: 000100111100101100110011
- Genotype: 000100111100101100110011
- Proband G: 10110122121110212101220022
- Opp Homo: ***********************

Other:
- Pat Hap: 10100111011100111001110011
- Mat Hap: 0101011110001100110011010
- Genotype: 11110222111111111111121021
- Proband G: 10110122121110212101220022
- Opp Homo: ***********************

A surrogate parent! (Even without pedigree information)
Phasing a Trio

Could be a female
Could be a descendant
Could be many generations distant
Can be ‘unrelated’

Surrogate Father

Mother

Proband

Can now phase this locus!!
Long Range Phasing

• Erdös 1 surrogates are surrogates of the proband.
• Erdös n+1 surrogates are surrogates of Erdos n surrogates of the proband.
Haplotype library imputation

- Build library of all completely phased haplotypes
- Find haplotypes in the library which can explain an individual's genotype
- Low error rates
- Computationally fast
- Useful for extremely large data sets
  - Strategic use

[Binary sequences]
AlphaPhase

- Software implementing the phasing part
- Partitions surrogates
  - Pedigree
  - Pedigree free
- Storage of surrogate information
- Handles genotype error
- Cores and tails
- Iterative algorithm
- Multiple surrogates
Diversion to humans!

- Orkney Islands north of Scotland
- Population of about 20,000 people
- Traditionally isolated – closed population

- Orcades project  
  http://www.orcades.ed.ac.uk/
  - Intensively phenotyped and genotyped
  - Over 700 new causal mutations identified
  - Results published in Nature, Science, AJHG, Nature Genetics etc.

- Small collaboration on phasing and imputation
  - Phase genotype data
  - Impute exome sequence data

- Justification
  - Speed and accuracy of other methods not sufficient
Phasing results simulated data

Percentage of alleles correctly phased / incorrectly phased by the most optimal core and CplusT lengths\(^1\) for each pedigree and data scenario and for both effective population sizes.

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Ne 100 With pedigree</th>
<th>Ne 100 Without pedigree</th>
<th>Ne 1000 With pedigree</th>
<th>Ne 1000 Without pedigree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedigree 1</td>
<td>99.11 / 0.30</td>
<td>NR</td>
<td>99.11 / 0.30</td>
<td>NR</td>
</tr>
<tr>
<td>Pedigree 2 without parents genotyped</td>
<td>97.85 / 0.43</td>
<td>98.49 / 0.63</td>
<td>97.73 / 0.29</td>
<td>97.88 / 0.39</td>
</tr>
<tr>
<td>Pedigree 2 with parents genotyped</td>
<td>98.85 / 0.49</td>
<td>99.03 / 0.42</td>
<td>99.70 / 0.17</td>
<td>99.48 / 0.17</td>
</tr>
<tr>
<td>Pedigree 3 without parents genotyped</td>
<td>98.35 / 0.38</td>
<td>98.61 / 0.63</td>
<td>99.23 / 0.14</td>
<td>99.14 / 0.27</td>
</tr>
<tr>
<td>Pedigree 3 with parents genotyped</td>
<td>99.23 / 0.37</td>
<td>99.05 / 0.41</td>
<td>99.76 / 0.16</td>
<td>99.58 / 0.13</td>
</tr>
<tr>
<td>Pedigree 4 without parents genotyped</td>
<td>98.20 / 0.41</td>
<td>98.61 / 0.63</td>
<td>98.19 / 0.41</td>
<td>98.61 / 0.63</td>
</tr>
<tr>
<td>Pedigree 4 with parents genotyped</td>
<td>99.35 / 0.31</td>
<td>99.29 / 0.32</td>
<td>99.74 / 0.20</td>
<td>99.59 / 0.15</td>
</tr>
<tr>
<td>Pedigree 5</td>
<td>97.59 / 0.42</td>
<td>98.28 / 0.60</td>
<td>99.30 / 0.30</td>
<td>99.31 / 0.22</td>
</tr>
<tr>
<td>Pedigree 6 sires</td>
<td>97.05 / 0.45</td>
<td>98.40 / 0.62</td>
<td>99.05 / 0.17</td>
<td>99.25 / 0.20</td>
</tr>
<tr>
<td>Pedigree 6 last 2000</td>
<td>98.24 / 0.39</td>
<td>98.24 / 0.39</td>
<td>99.34 / 0.20</td>
<td>99.42 / 0.26</td>
</tr>
<tr>
<td>Pedigree 7 sires</td>
<td>97.56 / 0.40</td>
<td>98.71 / 0.50</td>
<td>98.98 / 0.20</td>
<td>99.15 / 0.29</td>
</tr>
<tr>
<td>Pedigree 7 last 2000</td>
<td>96.86 / 0.46</td>
<td>98.40 / 0.66</td>
<td>98.85 / 0.20</td>
<td>99.34 / 0.26</td>
</tr>
<tr>
<td>Pedigree 8</td>
<td>95.01 / 1.10</td>
<td>96.67 / 1.39</td>
<td>96.02 / 0.57</td>
<td>96.36 / 1.01</td>
</tr>
</tbody>
</table>

\(^1\)Core length was 100 SNPs, CplusT length varied between 300 and 500 SNPs
# Phasing results real data

Table 2. Numbers of individuals in the data set, numbers of SNPs to be phased, core and CplusT length and missing genotype / genotype error % (M/E%) error threshold parameters, computation time, and percentage of alleles phased for 6 real data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th># Individuals</th>
<th># SNPs</th>
<th>Core / CplusT length</th>
<th>M/E%</th>
<th>(^1)Time</th>
<th>% Phased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheep Chr. 4</td>
<td>1019</td>
<td>2278</td>
<td>100 / 300</td>
<td>1.00</td>
<td>3 min. 39 sec.</td>
<td>98.17</td>
</tr>
<tr>
<td>Sheep Chr. 5</td>
<td>1016</td>
<td>1927</td>
<td>100 / 400</td>
<td>1.00</td>
<td>5 min. 1 sec.</td>
<td>97.62</td>
</tr>
<tr>
<td>Pig Chr. 1</td>
<td>2723</td>
<td>3999</td>
<td>100 / 500</td>
<td>0.00</td>
<td>364 min.</td>
<td>96.87</td>
</tr>
<tr>
<td>Beef Chr. 24</td>
<td>2171</td>
<td>874</td>
<td>100 / 300</td>
<td>0.00</td>
<td>17 min. 8 sec.</td>
<td>98.42</td>
</tr>
<tr>
<td>Dairy Chr. 1</td>
<td>5057</td>
<td>2296</td>
<td>100 / 400</td>
<td>0.00</td>
<td>456 min.</td>
<td>97.99</td>
</tr>
<tr>
<td>Human Chr. 1</td>
<td>879</td>
<td>4472</td>
<td>100 / 300</td>
<td>1.00</td>
<td>3 min. 29 sec.</td>
<td>93.73</td>
</tr>
</tbody>
</table>

\(^1\)A 64 bit desktop with an Intel i7 3.07 GHz quad core processor running Linux was used to measure computation time. Computation time includes time required to parse and summarise the data and write out the results.
Segregation analysis and haplotype library imputation

- Individual’s are densely, sparsely, or not genotyped
- Pedigree information available
- Single locus segregation analysis for each SNP
- Match each pair of haplotypes with low density genotypes and genotype probabilities

Genotyping strategy in terms of high density, low density and not genotyped

Haplotype library for population
Example of haplotypes
The imputation problem for a 2.5k low density chip in pigs
What information do we have to solve this problem?

• Knowledge

• Low density genotypes

• Pedigree information

• Linkage

• Linkage disequilibrium
What knowledge do we have?

• Knowledge
  – Inheritance is “chunkular”
  – Chunks of DNA are inherited together
  – Recombination events breaks these chunks up
  – Approximately 70 recombination events during meiosis
  – Thus approximately 150 chunks per animal
  – WE CALL THESE CHUNKS HAPLOTYPES

• Pedigree information

• Linkage
  – Family statistic
  – Correlation between adjacent markers within a family
  – Long haplotype information

• Linkage disequilibrium
  – Population statistic
  – Correlation between adjacent markers within a population
  – Short haplotype information
How can we use low density information?

- We are trying to impute this individual

```
..........................
    0..............1......
Low density genotype
```

```
10110122121110212101220022
True genotype
```
How can we use pedigree information?

**Low density genotype**

```
....0..............1......
```

**Father**

```
111102221111111111121021
```

**Mother**

```
10111121211112121221121121
```

**True genotype**

```
10110122121110212101220022
```
How can we use pedigree information?

Low density genotype

```
....0.2............1.2..2.
```

Three offspring

```
10120122121120212101121022
10220122121120212101120122
21110122121110212101220022
```

- We know that at this locus the animal has been mated to three dams who are homozygous 2
- Therefore the animal must be heterozygous
- The capturing of pedigree information is automated in GeneProb

True genotype

```
10110122121110212101220022
```

```
....0.2......1.......1.2..2.
```
How can we use linkage information?

• Two step procedure
  – Phase
  – Choose which parental haplotype was passed to offspring

• What is phasing?
  – Phasing strips a string of SNP into its two gametes/haplotypes
  – Genotypes represented as 0, 1, or 2
  – Each parent contributes a 0 or a 1
  – We use long range phasing to do this

True phase

<table>
<thead>
<tr>
<th></th>
<th>Paternal gamete</th>
<th>Maternal gamete</th>
</tr>
</thead>
<tbody>
<tr>
<td>10100111011100111001110011011</td>
<td>000100111100101100110011011</td>
<td>000100111100101100110011011</td>
</tr>
</tbody>
</table>

True genotype

<table>
<thead>
<tr>
<th>Genotype is the sum of the gametes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011012212111021210122022</td>
</tr>
</tbody>
</table>
How can we use linkage information?

Low density genotype

Father Phased

Mother Phased

Individual Phased

True phase

True genotype

- Work at the level of phase
- Can fill in alleles where parent is homozygous
- Crucially fill in one paternal allele where animal is known to be heterozygous because mother is homozygous
- Can use this to chose between fathers haplotypes
How can we use pedigree information?

Low density genotype

Father Phased

10100111011100111001110011

01010111100011000110011010

Mother Phased

000100111100101101100110011

10101101011111111111111110

Individual Phased

10100111011100111001110011

000100111100101101100110011

10101110101111111111111110

True phase

10100111011100111001110011

000100111100101101100110011

10101110101111111111111110

True genotype

1011012121110212101220022

- Work at the level of phase
- Can fill in alleles where parent is homozygous
- Crucially fill in one paternal allele where animal is known to be heterozygous because mother is homozygous
- Can use this to chose between fathers haplotypes
How can we use pedigree information?

Low density genotype

Father Phased

Mother Phased

Individual Phased

True phase

• Impute alleles in maternal gamete via the complement of the paternal gamete and the genotype
How can we use pedigree information?

Low density genotype

Father Phased

10100111011100111001110011
010101111100011001100111010

Mother Phased

000100111100101101100110011
101011101011111111111111110

Individual Phased

10100111011100111001110011
000100111100101101100110011

• Impute alleles in maternal gamete via the complement of the paternal gamete and the genotype

• Use this allele to chose between mothers haplotypes

1010122121110212101220022
Segregation analysis and haplotype library imputation

- Individual’s are densely, sparsely, or not genotyped
- Pedigree information available
- Single locus segregation analysis for each SNP
- Match each pair of haplotypes with low density genotypes and genotype probabilities

Genotyping strategy in terms of high density, low density and not genotyped

Haplotype library for population
It all looks easy......

• But it is completely dependent on ability to determine parental phase

• Or

• Our ability to have the haplotypes the animal carries stored in a library
Linkage disequilibrium – the information we are ignoring

- The correlation between adjacent SNP
  - Population average parameter

- fastPHASE, Beagle, IMPUTE2 are popular software

- But how useful are population correlations at explaining within family Mendelian segregation anyway?

- Can we model recombination statistically?

- Benefit from allele frequency
  - This is why the correlation is the key statistic
Alphalmpute compared to Impute2.0

- **Alphalmpute**
  - Uses pedigree and linkage information

- **Impute2.0**
  - Pedigree free imputation which uses linkage disequilibrium
  - Similar algorithm to Beagle / fastPHASE

- **Data set for comparison**
  - PIC pig data set
  - LIC multiple breed cattle data set
AlphalImpute compared to Impute2.0

• Data set for comparison
  – PIC single line test set
  – 6473 animals in pedigree file
  – 3200 genotyped at high density
  – 509 genotyped at low density
    • These are animals in the current generation
    • We know their high density genotypes
    • Different categories of animals (which ancestors genotyped)

  – High density chip contains 4221 SNP on chromosome 1

  – Low density
    • 15% of SNP (approx. 7500 SNP density)
    • 10% of SNP (approx. 5000 SNP density)
    • 5% of SNP (approx. 2500 SNP density)
    • 1% of SNP (approx. 500 SNP density)
AlphaImpute compared to Impute2.0

• Data set for comparison
  – LIC multiple breed test set
  – 24,017 animals in pedigree file
  – 4431 genotyped at high density
  – 626 genotyped at low density
    • These are randomly selected from across all generations
    • Restriction that the sire and dam identified
    • We know their high density genotypes
    • Different categories of animals (which ancestors genotyped)

  – High density chip contains 2297 SNP on chromosome 1

  – Low density
    • 15% of SNP (approx. 7500 SNP density)
    • 10% of SNP (approx. 5000 SNP density)
    • 5% of SNP (approx. 2500 SNP density)
    • 1% of SNP (approx. 500 SNP density)
## Results PIC pig data set

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>0.5k LD</th>
<th></th>
<th>2.5k LD</th>
<th></th>
<th>5k LD</th>
<th></th>
<th>7.5k LD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AlphaImpute</td>
<td>IMPUTE2</td>
<td>AlphaImpute</td>
<td>IMPUTE2</td>
<td>AlphaImpute</td>
<td>IMPUTE2</td>
<td>AlphaImpute</td>
<td>IMPUTE2</td>
</tr>
<tr>
<td>BothParents</td>
<td>51</td>
<td>0.98</td>
<td>0.77</td>
<td>0.99</td>
<td>0.92</td>
<td>1.00</td>
<td>0.96</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>SireMGS</td>
<td>62</td>
<td>0.93</td>
<td>0.80</td>
<td>0.98</td>
<td>0.92</td>
<td>0.99</td>
<td>0.94</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>DamPGS</td>
<td>47</td>
<td>0.96</td>
<td>0.79</td>
<td>0.98</td>
<td>0.92</td>
<td>0.99</td>
<td>0.95</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>Sire</td>
<td>45</td>
<td>0.89</td>
<td>0.78</td>
<td>0.97</td>
<td>0.92</td>
<td>0.99</td>
<td>0.95</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>Dam</td>
<td>13</td>
<td>0.90</td>
<td>0.76</td>
<td>0.96</td>
<td>0.89</td>
<td>0.98</td>
<td>0.93</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>Other</td>
<td>291</td>
<td>0.86</td>
<td>0.79</td>
<td>0.94</td>
<td>0.91</td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Correlation is the statistic that matters
## Results LIC cattle data set

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>0.5k LD AlphaImpute</th>
<th>0.5k LD IMPUTE2</th>
<th>2.5k LD AlphaImpute</th>
<th>2.5k LD IMPUTE2</th>
<th>5k LD AlphaImpute</th>
<th>5k LD IMPUTE2</th>
<th>7.5k LD AlphaImpute</th>
<th>7.5k LD IMPUTE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BothParents</td>
<td>28</td>
<td>0.97</td>
<td>0.64</td>
<td>0.99</td>
<td>0.92</td>
<td>0.99</td>
<td>0.94</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>SireMGS</td>
<td>224</td>
<td>0.87</td>
<td>0.60</td>
<td>0.97</td>
<td>0.91</td>
<td>0.98</td>
<td>0.95</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td>DamPGS</td>
<td>7</td>
<td>0.92</td>
<td>0.63</td>
<td>0.97</td>
<td>0.87</td>
<td>0.98</td>
<td>0.91</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>Sire</td>
<td>144</td>
<td>0.86</td>
<td>0.60</td>
<td>0.96</td>
<td>0.90</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Dam</td>
<td>4</td>
<td>0.95</td>
<td>0.63</td>
<td>0.98</td>
<td>0.90</td>
<td>0.99</td>
<td>0.97</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>Other</td>
<td>219</td>
<td>0.84</td>
<td>0.58</td>
<td>0.94</td>
<td>0.90</td>
<td>0.96</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Robust across data sets

• PIC data
  – Animals come from 1 line
  – Highly related
  – Many females genotyped at high density

• LIC data
  – Multiple breed
  – Within each breed groups of highly related animals
  – Not many females genotyped
  – Likely that Jersey information contributes to Holstein imputation
Diversion to humans!

- Veterans Affairs Hospitals
  - Largest hospital network in the United States
  - Excellent electronic records

- Pilot project about to begin to test this method
  - Framingham data set

- Potential to create extremely large data set cheaply

- Collaborator – University of Alabama (Birmingham)
Diversion to plants!

- CIMMYT – Plant breeding for the developing world

- Once a haplotype has been genotyped at high density and identified it does not need to be re-genotyped at high density

- Suppose we have 30,000 haplotypes floating around the CIMMYT maize populations

- Find them, store them, and impute them

- Global haplotype library
Haplotype library imputation

Central Library
Genotyped at high density
Key ancestral lines
Lines which are represented in two or more sub breeding programs
Parents

Satellite breeding program
e.g. Drought resistance in India
Genotype at low density

Satellite breeding program

Satellite breeding program

Satellite breeding program

Satellite breeding program

Pedigree information
Haplotype library imputation

Central Library

Genotyped at high density

Thousands of individuals genotyped at high density
represented in two or more sub breeding programs
Parents

Satellite breeding program

Hundreds of individuals genotyped at low density

Satellite breeding program

Hundreds of individuals genotyped at low density

Satellite breeding program

Hundreds of individuals genotyped at low density

Satellite breeding program

Hundreds of individuals genotyped at low density

Pedigree information
Haplotype library imputation

Central Library
- Thousands of individuals genotyped at high density
  - represented in two or more sub breeding programs
  - Parents

Satellite breeding program
- Hundreds of individuals genotyped at low density

Genotype information on selected individuals

Pedigree information
AlphaImpute – the output

• Among the output:

  – Imputed genotypes or sum of the allele probabilities for ALL SNP for ALL animals in the pedigree

  – Can be thought of as sum of the allele probabilities enhanced by linkage information via haplotypes or recombination modelling
Simulated data

<table>
<thead>
<tr>
<th>Generation</th>
<th>Information available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Genotype information on sires only</td>
</tr>
<tr>
<td>2</td>
<td>Genotype information on sires only</td>
</tr>
<tr>
<td>3</td>
<td>Genotype information on all animals</td>
</tr>
<tr>
<td>4</td>
<td>Pedigree information</td>
</tr>
<tr>
<td>5</td>
<td>Phenotype information</td>
</tr>
</tbody>
</table>

500 selection candidates

Validation population
Alternative single stage genomic evaluation?

- Scenario 1 - Train using the 3200 HD animals
- Scenario 2 - Train using the 3200 HD and 2764 completely ungenotyped animals
- Predict in the 509 testing animals
- Growth trait with h² of 0.61

Results - correlation with progeny test EBV from BLUP

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Unrestricted</th>
<th>Restricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>0.42</td>
<td>0.51</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>0.49</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Availability

- [http://sites.google.com/site/hickeyjohn/](http://sites.google.com/site/hickeyjohn/)
  - AlphaBayes
    - Bayesian GWAS
  - AlphaPhase
    - Pedigree and pedigree free phasing
  - AlphaImpute
    - Phasing and imputation in pedigreed populations
  - AlphaDrop
    - Simulation
Acknowledgements

• **Supervision and mentoring**
  - Julius van der Werf
  - Brian Kinghorn
  - Bruce Tier

• **Collaborators**
  - Gregor Gorjanc
  - Matt Cleveland
  - Matt Kelly
  - Gustavo de los Campos
  - Jose Crossa
  - Brigit Gredler

• **Funding**
  - Australian Research Council
  - Genus PTY
  - Aviagen LTD
  - Pfizer Animal Genetics
  - Sheep CRC
  - CIMMYT
“Genetics thrives in Armidale!”