Obtaining variance of gametic diversity with genomic models

D.J.A. Santos¹,⁴, J.B. Cole², P.M. VanRaden², T.J. Lawlor³, H. Tonhati¹ & L. Ma⁴
Breeding value inheritance components:

\[ a_i = \frac{1}{2} a_s + \frac{1}{2} a_d + m_i \]

### Mendelian sampling

**Sire 1**

<table>
<thead>
<tr>
<th>Value</th>
<th>AA</th>
<th>bb</th>
<th>cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Sire 2**

<table>
<thead>
<tr>
<th>Value</th>
<th>Aa</th>
<th>Bb</th>
<th>Cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

**Values:**

- A +5
- B +3
- C +2

**PTA expectation/average:**

- **+5**

**Possible values for the gametes:**

- **+5**
- **0,2,3,5,7,8,10**

**Probability (binomial distribution):**

- **1**
- **0.125,0.125,0.125,0.25,0.125,0.125,0.125**

**If linked (phased ABC|abc):**

- **Recombination rates:** 0.2 AB|BC
  - **0.32, 0.08, 0.02, 0.16, 0.02, 0.08, 0.32**

**Probability:**

- **> 0 (0.68)**
- **> 7 (0.40)**

**Heterozygosity**

- **> 0 (0.875)**
- **Exactly 5**

How about future progeny?
Statistics Background

Binomial Variances and Covariances

\[ Var(x) = \sum x^2 - \left( \frac{\sum x^2}{N} \right)^2 = Np_A - \frac{(Np_A)^2}{N} = N(p_A - p_A^2) = N(p_A(1 - p_A)) \]

\[ Var(y) = \sum y^2 - \left( \frac{\sum y^2}{N} \right)^2 = Np_B - \frac{(Np_B)^2}{N} = N(p_B - p_B^2) = N(p_B(1 - p_B)) \]

\[ Cov(x, y) = \sum xy - \frac{\sum x \sum y}{N} = Np_{AB} - \frac{Np_A Np_B}{N} = N(p_{AB} - p_A p_B) \]

\[ \sigma^2_{[A+B]} = (\sigma^2_A + \sigma^2_B + 2\sigma_{AB}) \]

\[ \sigma^2_{[A+B]} = (\sigma^2_A + \sigma^2_B) \quad \text{If independent !!!} \]

\[ \sigma^2_{\text{gamete}} = \sigma^2 \sum N\text{locus} \]

Solutions

Gametic phase
\( \frac{1}{2} \) centiMorgan (0.01 Morgan)

- \( p_{AB} = 0.25 \Rightarrow \text{cov}_{ab} = 0 \cdot S_A S_B \)
- \( p_{AB} = 0 \) or 0.50 \( \Rightarrow \text{cov}_{ab} = \pm 0.25 \cdot S_A S_B \)

Homozygous loci:
\[ N \cdot p \cdot (1-p) \cdot S^2 = 0 \]
Method

Methods for computing

\[ \sigma^2 \sum N_{\text{locus}} = \left[ s_1 \ldots s_n \right] \]

\[ \begin{bmatrix} 0.25 & \ldots & a l_{1n} \left( -\frac{cM_{1n}}{200} + 0.25 \right) & \ldots & 0.25 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a l_{1n} \left( -\frac{cM_{1n}}{200} + 0.25 \right) & \ldots & 0.25 \end{bmatrix} \]

\[ \sigma^2_{\text{gamete}} = \sigma^2 \sum N_{\text{locus}} \]

Independent \( \Rightarrow cM = 0.25 \) (25% for each gamete)

Reference allele

locus 1 = A
locus 2 = B

\[ a l_{nn} = 1 \]

\[ a l_{nn} = -1 \]

> 50 cM is considered as independent (that is 50 cM)
Application

Confidence intervals

Strategies of selection

Genetic Gain in Future

\[ RPTA_i = PTA_i + \sigma_{g\text{ametic}_i} \times i_f \]

\[ \Delta G = r \times i \times \sigma_a \]

\[ \Delta GR = r \times i \times \sqrt{\sigma_a^2 + 4 \times \text{var}(\sigma_{g\text{ametic}_i}) \times i_f^2} \]
In practice, how can we obtain the variance of gametic diversity?

Using marker effects estimated from routine genomic evaluation!!

Subsequent questions about this approach:

1. Should the recombination rate also be considered (dependence) between the markers?
2. What should the density panel marker be?
3. Which models to use?
4. What is the MAF effect?

To answer these questions, simulation study was proposed!!!
**Simulation - Population**

**Phase 1** - 500 generations:
- Constant size:
  - 500 males
  - 500 females individuals

**Phase 2** - 500 generations
- Constant reduction:
  - from 1,000 to 200 individuals
  - equal proportion male/female
  - LD/drift-mutation balance

**Phase 3** - 10 generations
- Expansion:
  - from 200 to 3,000 individuals.
  - equal proportion male/female
  - 200 males and 800 females (last generation)

**Recent Generations**

**Historical Generations**

**9th Generation: Genomic evaluation**
- 9th and 10th:
  - Estimated $\sigma^2_{\text{gamete}}$ from the estimated marker effects;
  - True $\sigma^2_{\text{gamete}}$: effects of the QTLs and their genotypes ($\sigma^2_{\text{Nlocus}}$)

**Traditional evaluation and selection**
- 9 generations
- 5 progeny per dam
- Selection: Blup
- Mating: random
- Cutting: Blup
- Replacement rate: 20% dams and 60% for sires
Simulation – Genome and Traits

**Genome Size 200 cM**

**Others Genome Parameters**
- Mutation Rater QTL: $2.5 \times 10^{-5}$
- Mutation Rater Marker: $2.5 \times 10^{-3}$
- Marker positions in genome: Evenly spaced
- QTL position in genome: Random (uniform distribution)
- QTL allele effect: Gamma distribution ($\beta = 0.4$)

**Scenarios:** 4 traits (QTLs x $h^2$) x 2 SNPs panels

**Traits:**
- $\text{N}^o$ of QTL:
  - 20 (0.1 QTL/cM) (low density)
  - 200 (1 QTL/cM) (Meuwissen et al., 2001)
- $h^2$: 0.1 and 0.3
- $\sigma^2_{\text{phenotypic}} = 1$
- 4 replicates for each trait

**Markers and Panels:**
- 200,000 markers were simulated and randomly distributed
- HD => 10% of the polymorphic markers sampled each 0.5 cM
- SEQ => 20% of the markers also sampled every 0.5 cM and all QTLs

All simulations were performed QMSim version 1.10 (Sargolzaei & Schenkel, 2009)
Genomic Model

Depends on the effects of the markers:

\[ y = \mu + Ma + e \]

- **Marker** ~ N(0, \( \sigma_e^2 \))
- **Residual** ~ N(0, \( \sigma_e^2 \))

MAF ≥ 0.05 (to mimic a conventional genomic evaluation)

1 - Traditional (SNP-BLUP/GBLUP)

\[ a \sim N(0, \sigma_a^2) / u \sim N(0, G \sigma_a^2) \]

2 - Differential shrinkage (Improved LASSO)

\[ \Pr(a_i | \tau^2) = N(0, \tau_i^2) \]
\[ \Pr(\tau_i^2 | \lambda) = \lambda^2 \exp(-\lambda^2 \mid \tau_i^2 \mid) \]

The analyses were performed using GS3 v.3 software (Legarra et al., 2015)

**Variance components:**
- initial values = true values
- interactions: 20,000
- burn-in: 2,000.
Gametic Variance

\[
\begin{align*}
1 - \sigma_g^2 &= \text{All QTL} \\
2 - \sigma_{g\text{-maf}}^2 &= \text{QTL with } MAF \geq 0.05 \\
3 - \sigma_{\text{dia}}^2 &= \text{All QTL} \\
4 - \sigma_{\text{diamaf}}^2 &= \text{QTL with } MAF \geq 0.05
\end{align*}
\]

\[
\begin{bmatrix}
0.25 & \ldots & 0 \\
\vdots & \ddots & \vdots \\
0 & \ldots & 0.25
\end{bmatrix}
\]
Results
### Correlation of True Values

<table>
<thead>
<tr>
<th>Scenario</th>
<th>QTLs data</th>
<th>$\sigma^2_g$</th>
<th>$\sigma^2_{g,\text{maf}}$</th>
<th>$\sigma^2_{\text{dia}}$</th>
<th>$\sigma^2_{\text{dia,maf}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h^2$</td>
<td>QTL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td></td>
<td>0.75</td>
<td>0.96</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.96</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>0.3</td>
<td>20</td>
<td></td>
<td>0.94</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td>0.95</td>
<td>0.55</td>
<td>0.52</td>
</tr>
</tbody>
</table>

- **Medium magnitude**: $h^2 = 0.1$
- **High magnitude**: $h^2 = 0.3$

It implies that QTLs with low MAF are important for obtaining accurate estimates of $\sigma^2_{\text{gamete}}$

$\sigma^2_{\text{gamete}}$ does not depend directly on population allele frequencies but on the individual's heterozygous state (allele carrier).
### Correlation between True and Estimated $\sigma^2_{\text{gamete}}$

<table>
<thead>
<tr>
<th>Scenario</th>
<th>High-sensitivity panel</th>
<th>Sequencing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_{\text{gbup}}$</td>
<td>$\sigma^2_{\text{glas}}$</td>
</tr>
<tr>
<td>$h^2$</td>
<td>QTL</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.50</td>
</tr>
<tr>
<td>0.3</td>
<td>20</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**Best accuracy!**

**Worst accuracy!**

**Similar accuracy!**
### Bias

<table>
<thead>
<tr>
<th>Trait</th>
<th>Model</th>
<th>HD</th>
<th>SEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0.1</td>
<td>GBLUP</td>
<td>0.0014</td>
<td>-0.0010</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>8e-05</td>
<td>0.0027</td>
</tr>
<tr>
<td></td>
<td>GBLUP</td>
<td>0.0010</td>
<td>0.0058</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.0001</td>
<td>0.0074</td>
</tr>
<tr>
<td>0.3</td>
<td>GBLUP</td>
<td>0.0017</td>
<td>-0.00697</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.0002</td>
<td>0.00282</td>
</tr>
<tr>
<td></td>
<td>GBLUP</td>
<td>0.0021</td>
<td>0.00979</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.0004</td>
<td>0.00945</td>
</tr>
</tbody>
</table>

- **Mean squared prediction (MSE):** ↓values
- **Coefficient of the linear regression (b):** close to one
- **GBLUP - higher predicted bias (overestimation)**
- **HD X SEQ - Similar Bias**
Conclusions

1 - The $\sigma^2_{\text{gamete}}$ can be obtained by GM using HD panels without the need to use sequencing data.

2 - Differential shrinkage models are preferred;

3 - Markers with low MAF should be also used;

4 - The covariance (dependence) among markers should be considered.

For improving the accuracy of the estimations
Financial Support

- BARD Research Project US-4997-17
- USDA-NIFA Foundational Grant 2016-67015-24886
- FAPESP 2017/00462-5

Thank you!!!
Real Data: USDA/Jersey

Biased distribution among chromosomes

Even distribution among chromosomes
Distribution of $\sigma^2_{\text{gamete}}$ for Production Traits

Atypical Gaussian curve

Close to typical Gaussian curve
Applied example: USDA/Jersey

Correlation (r) between $\sigma^2_{\text{gamete}}$ and variance of progeny GEBV for different traits per minimum number of offspring per sire.

<table>
<thead>
<tr>
<th>Minimum no of offspring</th>
<th>No Sires</th>
<th>$r_{\text{Milk Yield}}$</th>
<th>$r_{\text{Fat Yield}}$</th>
<th>$r_{\text{Protein Yield}}$</th>
<th>$r_{\text{Fat %}}$</th>
<th>$r_{\text{Protein %}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1109</td>
<td>0.24</td>
<td>0.20</td>
<td>0.16</td>
<td>0.58</td>
<td>0.30</td>
</tr>
<tr>
<td>50</td>
<td>451</td>
<td>0.40</td>
<td>0.46</td>
<td>0.33</td>
<td>0.75</td>
<td>0.50</td>
</tr>
<tr>
<td>100</td>
<td>311</td>
<td>0.53</td>
<td>0.47</td>
<td>0.34</td>
<td>0.85</td>
<td>0.60</td>
</tr>
<tr>
<td>200</td>
<td>183</td>
<td>0.64</td>
<td>0.49</td>
<td>0.31</td>
<td>0.95</td>
<td>0.77</td>
</tr>
<tr>
<td>300</td>
<td>128</td>
<td>0.68</td>
<td>0.55</td>
<td>0.40</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>400</td>
<td>97</td>
<td>0.66</td>
<td>0.61</td>
<td>0.43</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td>500</td>
<td>77</td>
<td>0.66</td>
<td>0.62</td>
<td>0.51</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td>600</td>
<td>66</td>
<td>0.69</td>
<td>0.66</td>
<td>0.54</td>
<td>0.97</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Motivating Results – TRUE RPTA / PTA

Simulation: Future generations; sires (i=1.75) and Dam (i=0.97).

TRUE RPTAs were corrected for number of offspring;

- 0.1 QTL/cM
  - $\sigma^2_a = 0.3(h^2=0.3)$
  - $\Delta G = 0.33\%$
  - 7 generations

- 1 QTL/cM
  - $\sigma^2_a = 0.3(h^2=0.3)$
  - $\Delta G = 54.75\%$
  - 25 generations
## Genetic summary for Top 10 Sires for Milk Yield.

<table>
<thead>
<tr>
<th>Sire_ID</th>
<th>Year</th>
<th>$\sigma_g^2$</th>
<th>CRV</th>
<th>N</th>
<th>PTA</th>
<th>rankPTA</th>
<th>RPTA_1.5</th>
<th>rankRPTA</th>
<th>Pr&gt;1,100</th>
</tr>
</thead>
<tbody>
<tr>
<td>59250449</td>
<td>2010</td>
<td>27.905</td>
<td>0.50</td>
<td>96</td>
<td>1,057</td>
<td>1</td>
<td>1.308</td>
<td>1</td>
<td>0.40</td>
</tr>
<tr>
<td>62902902</td>
<td>2012</td>
<td>23.724</td>
<td>0.47</td>
<td>83</td>
<td>1,027</td>
<td>2</td>
<td>1.259</td>
<td>4</td>
<td>0.32</td>
</tr>
<tr>
<td>56893061</td>
<td>2009</td>
<td>23.526</td>
<td>0.47</td>
<td>85</td>
<td>1,021</td>
<td>3</td>
<td>1.251</td>
<td>5</td>
<td>0.30</td>
</tr>
<tr>
<td>63345061</td>
<td>2012</td>
<td>30.756</td>
<td>0.53</td>
<td>107</td>
<td>983</td>
<td>5</td>
<td>1.241</td>
<td>6</td>
<td>0.24</td>
</tr>
<tr>
<td>54319065</td>
<td>2008</td>
<td>29.600</td>
<td>0.52</td>
<td>103</td>
<td>973</td>
<td>6</td>
<td>1.272</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>63561482</td>
<td>2012</td>
<td>39.800</td>
<td>0.65</td>
<td>164</td>
<td>963</td>
<td>7</td>
<td>1.204</td>
<td>8</td>
<td>0.20</td>
</tr>
<tr>
<td>68432385</td>
<td>2014</td>
<td>25.722</td>
<td>0.50</td>
<td>95</td>
<td>958</td>
<td>8</td>
<td>1.203</td>
<td>9</td>
<td>0.19</td>
</tr>
<tr>
<td>66011155</td>
<td>2013</td>
<td>26.721</td>
<td>0.50</td>
<td>97</td>
<td>928</td>
<td>9</td>
<td>1.142</td>
<td>25</td>
<td>0.11</td>
</tr>
<tr>
<td>65096622</td>
<td>2013</td>
<td>26.532</td>
<td>0.45</td>
<td>78</td>
<td>927</td>
<td>10</td>
<td>1.171</td>
<td>14</td>
<td>0.14</td>
</tr>
</tbody>
</table>

$CRV = \frac{\sigma_g^2}{0.5 \sqrt{E[\mu^2]}}$; $N = \frac{(1.96)^2 \cdot (CRV)^2}{(0.1)^2}$; $RPTA_{1.5} = PTA + \sigma_g \cdot 1.5$