Dimensionality of genomic information and its impact on GWA and variant selection: A simulation study

Sungbong Jang

S. Tsuruta, N.G. Leite, I. Misztal, D. Lourenco

University of Georgia

2021 ASAS-CSAS-SSASAS
07/16/2021
GWA in AB&G

• GWA searches for major genes for traits of interest
• Sequence data is becoming available (> 30 M SNPs)
  • GWA narrows down to the most important sequence variants

• How many samples do we need?
  • Explore the limited dimensionality of genomic information
  • Chromosome segments segregating independently in the population (Me)

Ex) ↓ \( N_e \) → ↑ relationship among individuals and ↑ LD → ↓ \( M_e \)
Dimensionality of genomic information

**Independent chromosome segments** ($M_e$)
- $N_e$ and $M_e$ of livestock animals are small
- $N_e = 30 \sim 150$ and $M_e = 4000 \sim 15000$ (chicken, pig, dairy and beef cattle)
- $M_e = 4N_eL$ (Stam et al., 1980)

**If limited $M_e$ contains all additive information**
- 4000 $\sim$ 15000 of $M_e$ represents the dimensionality of genomic information
- How to approximate the dimensionality of genomic information?
- SVD of $Z \Rightarrow Z = U\Delta V$ ($U'U = I$ and $V'V = I$), where $\Delta$ = diagonal of singular values
- Eigen of $G \Rightarrow \text{var}(u) = \text{var}(U\Delta V a) \sim U\Delta\Delta'U'$, where $\Delta\Delta'$ = diagonal of eigenvalues
- Rank of $G \leq \min(N_{snp}, N_{ind}, M_e)$, limited dimensionality

Pocrnic et al., (2016)

Misztal et al., (2016)
Objectives

- Explore the limited dimensionality of genomic information for GWA
  - Establish the relationship between the number of animals representing
    the number of independent chromosome segments and the ability to
detect causative variants
Data: Simulation

• QMSim (Sargolzaei et al., 2009)
• Cattle population
• Random mating
  • Ne = 20: 5 males / 15,000 females
  • Ne = 200: 50 males / 15,000 females
• 20 generations – used from 11-20
  • 75k genotyped animals from gen. 16-20

• 29 Chromosomes (Total 23 Morgans)
• 500k SNPs
• 200, 2000 QTNs
• Heritability = 0.3, 0.9, 0.99
Step1: Genotype data scenarios

- withQTN: 500k SNPs + QTNs (sequence data)
- Different Ne: Ne = 20 and Ne = 200
- Different numbers of QTNs: QTN = 200 and QTN = 2000
- Different heritabilities to mimic different amounts of data:
  - $h^2 = 0.3, 0.9, 0.99$: low to very high accuracy of EBVs
### Step 2: Number of animals for discovery/training

<table>
<thead>
<tr>
<th>Generation</th>
<th>Number of Animals</th>
<th>Discovery-training candidates (N = 60,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen16 (N = 15k)</td>
<td>N = 530</td>
<td>Eig50</td>
</tr>
<tr>
<td>Gen17 (N = 15k)</td>
<td>N = 920</td>
<td>Eig60</td>
</tr>
<tr>
<td>Gen18 (N = 15k)</td>
<td>N = 1540</td>
<td>Eig70</td>
</tr>
<tr>
<td>Gen19 (N = 15k)</td>
<td>N = 2650</td>
<td>Eig80</td>
</tr>
<tr>
<td>Gen20 (N = 15k)</td>
<td>N = 5300</td>
<td>Eig90</td>
</tr>
<tr>
<td></td>
<td>N = 8800</td>
<td>Eig95</td>
</tr>
<tr>
<td></td>
<td>N = 15200</td>
<td>Eig98</td>
</tr>
<tr>
<td></td>
<td>N = 22000</td>
<td>Eig99</td>
</tr>
<tr>
<td></td>
<td>N = 30000</td>
<td>All</td>
</tr>
</tbody>
</table>

#### Process:
- Random sampling
- Discovery set
- Keep animals in previous EigXX
- Remaining 30,000

![Diagram showing the allocation of animals across different generations and the discovery/training process.](image-url)
Step 2: Number of animals for discovery/training

- Case of $N_e = 20$ and $N_e = 200$

<table>
<thead>
<tr>
<th>$N_e = 20$</th>
<th>$N_e = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTN = 2000</td>
<td>QTN = 2000</td>
</tr>
<tr>
<td>$h^2 = 0.3$</td>
<td>$h^2 = 0.3$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eig50</th>
<th>80</th>
<th>530</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eig60</td>
<td>140</td>
<td>920</td>
</tr>
<tr>
<td>Eig70</td>
<td>220</td>
<td>1540</td>
</tr>
<tr>
<td>Eig80</td>
<td>400</td>
<td>2650</td>
</tr>
<tr>
<td>Eig90</td>
<td>890</td>
<td>5300</td>
</tr>
<tr>
<td>Eig95</td>
<td>1800</td>
<td>8800</td>
</tr>
<tr>
<td>Eig98</td>
<td>4100</td>
<td>15200</td>
</tr>
<tr>
<td>Eig99</td>
<td>7100</td>
<td>22000</td>
</tr>
<tr>
<td>All</td>
<td>30000</td>
<td>30000</td>
</tr>
</tbody>
</table>

- $\downarrow N_e \rightarrow \downarrow M_e$
Step 3: GWA

- GEMMA (Zhou et al., 2012)
- Population structure was accounted for by G
- Total variance explained by significant QTNs
  - \( \% \text{Var} = \frac{2pq(\beta)^2}{\sigma_a^2} \)
GWA results: $N_e = 20$  $QTN = 2000$  $h^2 = 0.3$
GWA results: $Ne = 20$ \hspace{1em} QTN = 2000 \hspace{1em} h^2 = 0.99
GWA results: \( Ne = 200 \)  \( QTN = 2000 \)  \( h^2 = 0.3 \)
GWA results: $N_e = 200$  $QTN = 2000$  $h^2 = 0.99$
Total variance explained by QTNs

- $N_e = 20$

QTN = 200

QTN = 2000
Total variance explained by QTNs

- Ne = 200

QTN = 200

QTN = 2000
Conclusions

• The suitable sample size depends on the Ne and the number of QTNs
  • Smaller populations require more data to capture causative variants
  • Larger populations: Eigen98 – Eigen99 to capture most informative
  • More polygenic trait requires more data to identify causative variants

• Dimensionality of genomic information allows to approximate the suitable sample size for GWA

• In progress:
  • Deriving equations to relate sample size and amount of data with Ne and Me
  • Testing the impact of incorporating selected variants for GP
Thanks ☺

jsbng@uga.edu