Identification of causal variants using one million individuals with whole-genome sequence information

Janez Jenko, Andrew Whalen, R. Chris Gaynor, Christos Dadousis, Gregor Gorjanc and John M. Hickey
PAGE at WCGALP 2014

Genetic Gain (since generation -20)

Genomic selection + PAGE genetic gain (since generation 0)

Genomic selection only

Generations

PAGE can potentially double genetic gain!
How can we achieve this?

• Over the 20 years we edited ~300 distinct causal variants
  – They explain 36% of genic variance
  – 3% of all the causal variants
  – 15 variants per year

• Old approach to variant discovery will not work

• Allele testing approach
Allele testing scheme

A process to game the odds
~25 million segregating sites
~10,000 affect the trait
~1,000 work in a simple additive way How to find 15 of these?

1. GWAS
2. eQTL
3. Functional
4. Cell lines
5. In vivo editing
6. PAGE
7. Outcome

Increase ratio of causal variants in subset

Increase probability that variant is NOT highly deleterious

| 0.0000006 | 0.000015 | 0.00015 | 0.0015 | 0.015 | 0.03 | 0.15 | 1.0 |
| 0.0 | ? | ? | 0.015 | ? | ? | 0.1 | 1.0 |

How to find 15 of these?
Aim of current study

Million animals

WGS

Change in the ratio of causal variants in subset
Simulating 1 million animals

• Historical sequences for 10 related populations

• 1 million animals (10 populations with 10 generations)

• Polygenic trait with 10,000 causal variants

• Phenotype with 0.3 heritability
Facilitating simulations

• 9 chromosomes with SNP information

• 1 chromosome with WGS information

1% of genome

91 causal variants and 100,000 neutral variants

Ranked 44th, 420th, 574th... across the whole on the effect size
Single SNP regression model

\[ y = \mu + X\beta + g + e \]

- \( y \) - vector of phenotypes
- \( \mu \) - mean
- \( X \) - incidence matrix
- \( \beta \) - fixed effects
- \( g \) - random genetic effect \( N(0, G\sigma_g^2) \)
- \( e \) - residual \( N(0, I\sigma_e^2) \)
Analysed scenarios
Causal and neutral variants

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of causal variants</th>
<th>Number of neutral variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Analysed region</td>
<td>Whole genome approximation</td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>6,900</td>
</tr>
<tr>
<td></td>
<td>84</td>
<td>8,400</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>7,900</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>6,700</td>
</tr>
<tr>
<td></td>
<td>84</td>
<td>8,400</td>
</tr>
</tbody>
</table>
Manhattan plots

-log_{10} P

Segregating variant position (cM)
# Significant variants statistics I.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of causal variants</th>
<th></th>
<th>Number of neutral variants</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Analysed region</td>
<td>Whole genome approximation</td>
<td>Analysed region</td>
<td>Whole genome approximation</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>200</td>
<td>176</td>
<td>17,600</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>400</td>
<td>256</td>
<td>25,600</td>
</tr>
</tbody>
</table>
Significant variants statistics II.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Genetic variance explained (%)</th>
<th>Correlations between the causal variant effect and $-\log_{10} P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>21.3</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>22.9</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Change in the ratio of causal variants in the subset

- Before GWAS: 1 causal variant out of 1018 variants (84/85,519)
- After GWAS: 1 causal variant out of 64 variants (4/260)

GWAS increased the ratio of causal variants in the subset for ~16 times
Conclusions

• GWAS is effective first step in allele testing scheme

• GWAS discovered ~400 causal variants

• ~25,000 false positives

• The next steps in allele testing will be to reduce these false positives to 3000
Acknowledgements