GWAS based on multiple imputation with low-depth sequencing data

Application to biofuel traits in reed canarygrass

Reed canarygrass

<table>
<thead>
<tr>
<th>Type</th>
<th>Cool-season, C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breeding history</td>
<td>Bred for forage quality</td>
</tr>
<tr>
<td>Tolerances</td>
<td>- Poorly drained soils</td>
</tr>
<tr>
<td></td>
<td>- Soil acidity</td>
</tr>
<tr>
<td></td>
<td>- Drought</td>
</tr>
<tr>
<td>Reference genome</td>
<td>NONE</td>
</tr>
<tr>
<td></td>
<td>Closely related to Brachypodium</td>
</tr>
</tbody>
</table>
Experimental design

956 genotypes

Association panel

Linkage populations
Experimental design

956 genotypes

Panel
- AP
- LP1
- LP2

Location
- Europe (East)
- Europe (West)
- North America
Experimental design

- Phenotypic data
  - Field traits (9)
  - Biofuel quality traits (26)

Locations:
- Arlington (WI)
- Ithaca (NY)

Years:
2009 to 2011

35 traits
Experimental design

- **Genotypic data**
- DNA extraction → *ApeKI* restriction → Sequencing

6,138 markers

Single-Nucleotide Polymorphisms (SNP)

...ATTAGCTACGAA...

...ATTAGCTACGAA...

...ATTAGCTACGAA...

...ATTAGCTACGAA...

...ATTAGCTACGAA...

...ATTAGCTACGAA...
The missing-data problem

- Low sequencing depth
The missing-data problem

• Low sequencing depth

• No reference genome
  ➢ Use of classification trees (CART)
The missing-data problem

• Low sequencing depth

• No reference genome
  ➢ Use of classification trees (CART)

• General pattern of missing data at all markers
  ➢ Iterative sampling from CART until an equilibrium is reached
Multiple imputation
Multiple imputation

Significance test for marker-associated effect $\beta$

with respect to

$$\hat{Var}(\hat{\beta}) = \hat{Var}_{within}(\hat{\beta}) + \hat{Var}_{between}(\hat{\beta})$$

Multiple imputation

Significance test for marker-associated effect $\beta$

with respect to

$$\widehat{Var}(\hat{\beta}) = \widehat{Var}_{within}(\hat{\beta}) + \widehat{Var}_{between}(\hat{\beta})$$

From regression theory

From multiple sampled imputes

Association studies: Methods

- Model:
  
  Linear mixed model

Phenotype

- Marker effect
- Population structure
- Relatedness
Association studies: Methods

Data:

<table>
<thead>
<tr>
<th>Marker-effect</th>
<th>Structure &amp; Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete cases</td>
<td>Non-missing data</td>
</tr>
<tr>
<td></td>
<td>Average across imputes for complete cases at marker</td>
</tr>
<tr>
<td>Averaged genotypes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average across imputes</td>
</tr>
<tr>
<td>Multiple imputes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple imputed datasets</td>
</tr>
</tbody>
</table>
Association studies: Methods

- Data:

  - Complete cases
  - Averaged genotypes
  - Multiple imputes

Reference

Assumption: Marker data at a given site is missing completely at random
Association studies: Results

Concordance across analyses
Marker effect

- from ‘Complete cases’ to ‘Averaged genotypes’

- No systematic imputation bias
Association studies: Results

Concordance across analyses

Marker effect

- from ‘Complete cases’ to ‘Multiple imputes’

- Shrinkage due to imputation uncertainty
Association studies: Results

Concordance across analyses

Significance of associations

- Possible false positives

- from ‘Complete cases’ to ‘Averaged genotypes’
Association studies: Results

Concordance across analyses

Significance of associations

- from ‘Complete cases’ to ‘Multiple imputes’

➤ Possibly, more *true* positives
Association studies: Results

Associations

Total stem count

$-\log_{10}(p)$ - Observed quantile vs $-\log_{10}(p)$ - Expected quantile
Association studies: Results

Total stem count

- $\log_{10}(p)$ - Observed quantile
- $\log_{10}(p)$ - Expected quantile

Complete cases

Associations
Association studies: Results

Associations
Association studies: Results

Associations
Association studies: Results

Associations

Crude protein content
Association studies: Results

Crude protein content

Homology with PAL gene in Brachypodium
Association studies: Results

- Gains of power from imputation uncertainty less than 0.2-0.4
- “Good” imputation uncertainty from less than 25% missing values
Conclusions

✓ **Good consistency** of multiple imputation with complete-case analyses
  - No imputation bias

✓ **Some gain in significance** of marker-trait associations
  - Proof of concept for the usefulness of multiple imputation in GWAS

✗ Generally **large imputation uncertainty**
  - Importance of imputation accuracy
Conclusions

✓ **Good consistency** of multiple imputation with complete-case analyses
   ➢ No imputation bias

✓ **Some gain in significance** of marker-trait associations
   ➢ Proof of concept for the usefulness of multiple imputation in GWAS

✗ Generally **large imputation uncertainty**
   ➢ Importance of imputation accuracy

- Marker-trait associations may be **useful for further genetic studies**, in reed canarygrass, Brachypodium and other biofuel crops (switchgrass)
Thank you for your attention
## Marker data processing

<table>
<thead>
<tr>
<th>Filtering method</th>
<th>Number of markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>29,313</td>
</tr>
<tr>
<td>PMV ≤ 0.9 by panel</td>
<td>18,818</td>
</tr>
<tr>
<td>PMV ≤ 0.9 by panel + Non-collinearity</td>
<td>6,138</td>
</tr>
<tr>
<td>PMV ≤ 0.8</td>
<td>3,419</td>
</tr>
<tr>
<td>PMV ≤ 0.5</td>
<td>176</td>
</tr>
<tr>
<td>PMV ≤ 0.2</td>
<td>51</td>
</tr>
</tbody>
</table>
Significance test in multiple imputation

\[ \bar{\beta} = \frac{1}{m} \sum_{r=1}^{m} \hat{\beta}^{(r)} \]

\[ \bar{W} = \frac{1}{m} \sum_{r=1}^{m} \text{Var} \left( \hat{\beta}^{(r)} \right); \quad B^* = \frac{m+1}{m} \text{Var}_r \left( \hat{\beta}^{(r)} \right) \rightarrow T = \bar{W} + B^* \]

\[ p = \Pr \left( F_{1, \nu} > \frac{\bar{\beta}^2}{T} \right) \]

\[ \lambda = \frac{B}{T} \]

\[ \nu_{com} = n - (2 + t) \quad [t: \text{number of principal components}] \]

\[ \nu_{obs} = \frac{\nu_{com} + 1}{\nu_{com} + 3} \nu_{com} (1 - \lambda); \quad \nu_{m} = \frac{m-1}{\lambda^2} \]

\[ \nu = \frac{\nu_{m} \nu_{obs}}{\nu_{m} + \nu_{obs}} \]

Regression of $\hat{\beta}_{CC}$ on $\hat{\beta}_{MI} \times \gamma$

($\gamma$: Imputation uncertainty)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate (S.E.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.005 (0.004)</td>
<td>0.22485</td>
</tr>
<tr>
<td>$\hat{\beta}_{MI}$</td>
<td>0.890 (0.015)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.010 (0.009)</td>
<td>0.29103</td>
</tr>
<tr>
<td>$\hat{\beta}_{MI} \times \gamma$</td>
<td>6.063 (0.035)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.005 (0.007)</td>
<td>0.44761</td>
</tr>
<tr>
<td>$\hat{\beta}_{MI}$</td>
<td>4.706 (0.029)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.012 (0.014)</td>
<td>0.39762</td>
</tr>
<tr>
<td>$\hat{\beta}_{MI} \times \gamma$</td>
<td>-0.640 (0.060)</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Marker data processing