Accuracy of Genomic Prediction in Switchgrass Improved by Accounting for Linkage Disequilibrium

Guillaume P. Ramstein, Joseph Evans, Shawn M. Kaeppler, Robert B. Mitchell, Kenneth P. Vogel, C. Robin Buell, Michael D. Casler

University of Wisconsin-Madison, Michigan State University and University of Nebraska-Lincoln
Outline

- General approach of genomic selection
- Context of genomic prediction
- Optimization of genomic prediction
- Conclusions and perspectives
Outline

▪ General approach of genomic selection
▪ Context of genomic prediction
▪ Optimization of genomic prediction
▪ Conclusions and perspectives
Motivation for genomic selection

- Phenotypic selection
  - Field measurements
  - Estimated breeding values

- Genomic selection
  - Genomic marker features
  - Genomic estimated breeding values
Motivation for genomic selection

Phenotypic selection
- Field measurements
- Estimated breeding values

Genomic selection
- Genomic marker features
- Genomic estimated breeding values

Reduction in:
- Accuracy
- Cost?
Motivation for genomic selection

Phenotypic selection

Field measurements

Estimated breeding values

Genomic selection

Genomic marker features

Genomic estimated breeding values

Reduction in:
❌ Accuracy
✔ Cost?
✔ Time!
Genomic prediction procedures

- Genomic marker features
- Prediction procedure
- Genomic estimated breeding values
Genomic prediction procedures

Additional investment in genomic selection

Calibration set

Genomic marker features

Field measurements

Prediction procedure
Genomic prediction procedures

**Objective:** Developing good prediction procedures

**Instances**
What set of individuals?

**Features**
What variables from the marker data?

**Prediction model**
What types of marker-feature effects?

**Calibration set**
- Genomic marker features
- Field measurements

**Prediction procedure**

**Outcome**
How to measure the trait of interest?
Outline

- General approach of genomic selection
- Context of genomic prediction
- Optimization of genomic prediction
- Conclusions and perspectives
Populations: size and structure

Breeding panels

WS4U-C2 (137 Half-sib families)
Upland ecotype
Collection of diverse accessions

Liberty-C2 (110 Half-sib families)
Upland x Lowland
Cross between two cultivars
Populations: size and structure

Breeding panels

WS4U-C2 (137 Half-sib families)
Upland ecotype
Collection of diverse accessions

≠

Liberty-C2 (110 Half-sib families)
Upland x Lowland
Cross between two cultivars
Markers: type and distribution

**Single nucleotide polymorphisms (SNP)**

- Located around coding sequences
Markers: type and distribution

**Single nucleotide polymorphisms (SNP)**

- Located around coding sequences
- Unevenly distributed
Markers: type and distribution

**Single nucleotide polymorphisms (SNP)**

- Located around coding sequences
- Unevenly distributed
- Correlated
Phenotypes: traits and types of measurements

**Dry matter yield (DMY)**

- **Two locations**
  - Wisconsin
  - Nebraska

Phenotypes: traits and types of measurements

Dry matter yield (DMY)

• Two locations
  ⇒ Wisconsin
  ⇒ Nebraska

• Two years
  ⇒ 2012
  ⇒ 2013

Context for application of genomic selection

- Distantly related populations
- Dense and uneven marking
- Small population sizes / Many markers

- Calibration set

- Genomic marker features
- Field measurements

Prediction procedure

Dry matter yield in two diverse locations
Outline

- General approach of genomic selection
- Context of genomic prediction
- Optimization of genomic prediction
- Conclusions and perspectives
Optimization of prediction procedures by component

- Instances
- Outcome
- Features
- Prediction model

Calibration set
- Genomic marker features
- Field measurements

Prediction procedure
Optimization of prediction procedures by component

**Instances**
- Among/Within-population calibration

**Outcome**
- Among/Within-environment calibration

**Features**
- Marker-data transformations
  (redundancy in marker information)

**Prediction model**
- More complex models than standard
  (differential and/or non-linear effects)
Optimization of prediction procedures by component

**Instances**
- Among/Within-population calibration

**Outcome**
- Among/Within-environment calibration

**Features**
- Marker-data transformations (redundancy in marker information)

**Prediction model**
- More complex models than standard (differential and/or non-linear effects)

**Calibration set**
- Genomic marker features
- Field measurements

**Prediction procedure**
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights on marker features</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Differential weights</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on marker features</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights on marker features</td>
<td>Additive effects</td>
<td><strong>GBLUP</strong> (Standard)</td>
</tr>
<tr>
<td>Differential weights on marker features</td>
<td>Additive effects</td>
<td>Additive effects + Interactions</td>
</tr>
</tbody>
</table>

The table shows the comparison between Equal weights and Differential weights on marker features, with additive effects and additive effects plus interactions. **GBLUP** (Standard) is highlighted for Equal weights on marker features.
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights on marker features</td>
<td>GBLUP (Standard)</td>
<td>RKHS</td>
</tr>
<tr>
<td>Differential weights on marker features</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights on marker features</td>
<td>GBLUP (Standard)</td>
<td>RKHS</td>
</tr>
<tr>
<td>Differential weights on marker features</td>
<td>GBLUP + weights on markers</td>
<td>RKHS + weights on markers</td>
</tr>
</tbody>
</table>
Optimization of prediction procedures by component

**Prediction model**

<table>
<thead>
<tr>
<th>Equal weights on marker features</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GBLUP (Standard)</td>
<td>RKHS</td>
</tr>
<tr>
<td>Differential weights on marker features</td>
<td>BayesA</td>
<td>GBLUP + weights on markers</td>
</tr>
<tr>
<td></td>
<td>BayesB</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RKHS + weights on markers</td>
</tr>
</tbody>
</table>
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Additive effects</th>
<th>Additive effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights on marker features</td>
<td>GBLUP (Standard)</td>
<td>RKHS</td>
</tr>
<tr>
<td>Differential weights on marker features</td>
<td>BayesA</td>
<td>GBLUP + weights on markers</td>
</tr>
<tr>
<td></td>
<td>BayesB</td>
<td></td>
</tr>
</tbody>
</table>
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original features</td>
</tr>
<tr>
<td>Weighted features</td>
</tr>
<tr>
<td>Decorrelated features</td>
</tr>
<tr>
<td>Principal components</td>
</tr>
</tbody>
</table>

Accounting for correlation between markers (linkage disequilibrium)
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original features</td>
</tr>
<tr>
<td>Weighted features</td>
</tr>
<tr>
<td>Decorrelated features</td>
</tr>
<tr>
<td>Principal components</td>
</tr>
</tbody>
</table>

![Genome browser image](image.png)
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
<th>Centered numerical codes on every diallelic SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original features</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Weighted features</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Decorrelated features</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Principal components</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Base (Standard)</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Additive code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homozygous at allele 1</td>
<td>0</td>
</tr>
<tr>
<td>Heterozygous</td>
<td>1</td>
</tr>
<tr>
<td>Homozygous at allele 2</td>
<td>2</td>
</tr>
</tbody>
</table>

![Genotype codes example](image.png)
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
<th>Original features weighted by their level of redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original features</td>
<td>Base (Standard)</td>
</tr>
<tr>
<td>Weighted features</td>
<td>LD</td>
</tr>
<tr>
<td>Decorrelated features</td>
<td></td>
</tr>
<tr>
<td>Principal components</td>
<td></td>
</tr>
</tbody>
</table>

Sparsely tagged regions: High weights
Densely tagged regions: Low weights

---

Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
<th>Original features</th>
<th>Weighted features</th>
<th>Decorrelated features</th>
<th>Principal components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base (Standard)</td>
<td>LD</td>
<td>Cor</td>
<td></td>
</tr>
</tbody>
</table>

Original features scaled through a correlation matrix
Optimization of prediction procedures by component

<table>
<thead>
<tr>
<th>Features</th>
<th>Base (Standard)</th>
<th>LD</th>
<th>Cor</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decorrelated features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principal components</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Synthetic uncorrelated variables represented by different sets of markers

Major contributions to component

![Genetic markers](image-url)
Protocol of optimization

1. **Evaluation of all possible prediction procedures** (combinations of prediction model and feature type for prediction accuracy)
   ⇒ Selected prediction procedure
Protocol of optimization

1. **Evaluation of all possible prediction procedures** (combinations of prediction model and feature type for prediction accuracy)
   ⇒ Selected prediction procedure

2. **Comparison of the selected prediction procedure to the standard** (GBLUP on original features) in *replicated* validations
   ⇒ Benefit from:
   - marker-data transformation?
   - more complex prediction model?
Evaluation of procedures (DMY in NE)
Comparison to standard (DMY in NE)

**WS4U-C2: DMY in NE**

- **Prediction accuracy**
  - Base - GBLUP
  - Cor - GBLUP
  - Cor - BayesA

- **Correlation**
  - $r = 0.14$ (p = 0.0063)
  - +0.095 (p = 0.87)
  - +0.01

**Liberty-C2: DMY in NE**

- **Prediction accuracy**
  - Base - GBLUP
  - Base - BayesB

- **Correlation**
  - $r = 0.5$ (p = 0.73)
  - +0.042
Outline

▪ General approach of genomic selection

▪ Context of genomic prediction

▪ Optimization of genomic prediction

▪ Conclusions and perspectives
Useful components of prediction procedures

<table>
<thead>
<tr>
<th>Instances</th>
<th>Among/Within-population calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>Among/Within-environment calibration</td>
</tr>
<tr>
<td>Features</td>
<td>Marker-data transformations</td>
</tr>
<tr>
<td></td>
<td>(redundancy in marker information)</td>
</tr>
<tr>
<td>Prediction model</td>
<td>More complex models than standard</td>
</tr>
<tr>
<td></td>
<td>(differential and/or non-linear effects)</td>
</tr>
</tbody>
</table>
Useful components of prediction procedures in our study

<table>
<thead>
<tr>
<th>Instances</th>
<th>Distantly related populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among/Within-population calibration</td>
<td>Two diverse locations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Dense and uneven marking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among/Within-environment calibration</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>Small population sizes / Many markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker-data transformations (redundancy in marker information)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prediction model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>More complex models than standard (differential and/or non-linear effects)</td>
<td></td>
</tr>
</tbody>
</table>
Continue selection on Liberty-C2

Phenotypic Selection

Genomic Selection

Cycle 1 PS

Cycle 1 GS
Cycle 2 GS
Cycle 3 GS

2016 - 2019
Acknowledgements

Joseph Evans
Robin Buell
Steve Masterson
Rob Mitchell
Ken Vogel

Nick Baker
Joseph Halinar
Shawn Kaeppler
Michael Casler
Jeremy Schmutz

U.S. Department of Energy Great Lakes Bioenergy Research Center, DOE BER Office of Science DE-FC02- 07ER64494 (GLBRC)

Agriculture and Food Research Initiative Competitive Grant No. 2011-68005-30411 from the USDA National Institute of Food and Agriculture (CenUSA)
Results: DMY in NE
(Evaluation of procedures by models only)

WS4U-C2: DMY in NE
- Base - GBLUP
- Base - RF

r = 0.14
+p = 0.99

Liberty-C2: DMY in NE
- Base - GBLUP
- Base - BayesB

r = 0.5
+p = 0.72

+0.042
Populations: size and structure

**Breeding panels**

- **Half-sib families**
  - WS4U-C2
    - Upland ecotype
    - Collection of diverse accessions
  - Liberty-C2
    - Upland x Lowland
    - Cross between two cultivars

- **Distantly related**

- **Small sizes**
Results

1. Evaluation of all possible prediction procedures

Gains in prediction accuracy from marker-data transformations in
- WS4U-C2: WI and NE
- Liberty-C2: WI
Results

2. Comparison of the selected prediction procedure to the standard

Gains in prediction accuracy from marker-data transformations in

• WS4U-C2: NE