Efficient use of historical data for genomic selection: a case study in wheat

Jessica Rutkoski

Faster gains in crop production are needed

Adapted from Tester and Langridge Science 2010
Genetic gain per unit time

\[ G_t = \frac{i \cdot r \cdot \sigma_a}{T} \]
Using genome-wide markers to increase genetic gain

1) More accurate breeding value estimations

2) ‘Marker-only’ breeding value prediction (Genomic selection)

\[ G_t = i \times r \times \sigma_a \]

- **Increase** $r$
- **Decrease** $T$
Genomic selection (GS)

Advantage: Identify superior individuals before phenotypic evaluation

Process:

1. Individuals with phenotypes + genotypes
2. Calibrate Prediction Model
3. Predict Individuals with genotypes only
Training populations (TP)s: Two ideas

Population specific TP
Use subset of the population of interest

Historical TP
Use pre-existing data from the breeding program

\[ P_1 \times P_2 \rightarrow F_1 \rightarrow F_2 \text{ pop.} \]

Training set

Predict
Ongoing project: Implement GS for quantitative stem rust resistance

- Historical data
- Selection candidates
- Breeding pipeline

Train model and predict
Objective: Use recurrent selection case study to...

1) Compare accuracies of a population specific and historical TP

2) Evaluate potential of TP optimization

3) Determine if historical data can augment a population specific TP
Historical and selection candidate populations

![Histogram and selection candidate populations](image)
LD decay in historical and selection candidate populations
Training population accuracy comparison

Historical TP

Population Specific TP

Selection candidates

Validation set

Within population prediction

Across population prediction
Model: Genomic BLUP (G-BLUP)

\[ Y = X\beta + Zu + \varepsilon \]

\[ V(u) = A*V_A \]

Phenotypes (some are missing) → Fixed effects → Random effects, breeding values → Covariance of breeding values → Marker based relationship matrix (A)
Prediction accuracy comparison

![Graph showing prediction accuracy comparison for different training set sizes. The x-axis represents the training set size (73, 146, 219, 292, 365), and the y-axis represents prediction accuracy. Two training sets are compared: Historical and Pop. Specific. The bars indicate the accuracy with error bars showing the variability.]
Training population optimization
Criteria

Maximize ‘connectedness’
Rincent et al. 2012

- Generalized coefficient of determination \((CD_{mean})\)
- Prediction error variance \((PEV_{mean})\)

Minimize genetic differentiation

\(F_{st}\)
Optimization algorithm

1. Select random sample
2. Remove individual & add new
3. Keep change if more optimal
4. Save TP at convergence

Repeat steps 1-4 100 times, determine overall best TP

Method adapted from Rincent et al. 2012
Training population optimization results

Population Specific TP

Historical TP

Accuracy

TP Size
Can historical data augment a population specific TP?

\[
\text{Population specific TP} + \text{Sample historical TP} = \text{Combined TP}
\]

Manipulate $H^2$

Calculate accuracy

Repeat 1000 times & compute mean accuracy
Effect of adding historical individuals to a population specific training set

Starting TP $h^2 = 0.2$

Starting TP $h^2 = 0.6$

$H^2$ of Individuals Added
- 0.2
- 0.3
- 0.4
- 0.6

No. of historical lines added
Effect of adding historical individuals to a population specific training set

![Graphs showing the effect of adding historical lines to a population specific training set.](image)

- **Starting TP $h^2 = 0.2$**
  - No. of historical lines added vs. Accuracy
  - Small benefit when pop specific TP $h^2$ is low and historical $h^2$ is high

- **Starting TP $h^2 = 0.6$**
  - No. of historical lines added vs. Accuracy
  - Detrimental when pop specific TP $h^2$ is high and historical $h^2$ is low

- $H^2$ of Individuals Added
  - • 0.2
  - □ 0.3
  - ▲ 0.4
  - ▲ 0.6
Populations can be weighted according to heritability

Mixed model:

\[ y = X\beta + Zu + \varepsilon \]

Covariance of residuals:

\[ V(e) = R*V_R \]

Covariance of breeding values:

\[ V(u) = A*V_A \]

\[
R = \begin{bmatrix}
1 - H^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 - H^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 - H^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 - H^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 - H^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 - H^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 - H^2
\end{bmatrix}
\]
Weighting each population by its heritability leads to improved accuracy.

<table>
<thead>
<tr>
<th>No. of historical lines added</th>
<th>Result without weighting</th>
<th>Result with weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Starting $TP \ h^2 = 0.6$

- $H^2$ of Individuals Added
  - 0.2
  - 0.3
  - 0.4
  - 0.6
Weighting each population by its heritability leads to improved accuracy.
Summary

1) Population specific TP was 4.4x to 1.5x more accurate than a historical TP depending on the TP size

2) Accuracy from the historical TP could be improved by selecting an optimal subset

3) When a population specific TP is available, its best to not include historical data in the TP
Acknowledgements

Funding

• Supported in part by a grant from the Bill & Melinda Gates Foundation to Cornell University for the Borlaug Global Rust Initiative Durable Rust Resistance in Wheat Project.

• USDA Cooperative State Research, Education and Extension Service, Coordinated Agricultural Project 2005-05130-Wheat Applied Genomics

• ASPB Pioneer Hi-Bred Fellowship

• USDA National Needs Fellowship Grant 2008-38420-04755: Provided Fellowship for Jessica Rutkoski

Collaborators

• Ravi Singh, CIMMYT
• Julio Huerta-Espino, INIFAP
• Sridhar Bhavani, CIMMYT
• Jesse Poland, KSU & USDA-ARS

Committee

• Mark Sorrells
• Rebecca Nelson
• Jean-Luc Jannink
• Ronnie Coffman