Dr. Jianming Yu
IOWA STATE UNIVERSITY
PROFESSOR OF AGRONOMY
Genomic Selection: Historical Context, Technical Details, Empirical Findings, and Perspectives

Jianming Yu
Department of Agronomy
Raymond F. Baker Center for Plant Breeding
Introduction

From 56 presentation titles
Acknowledgements

Kraig Roozeboom (KSU)
Tesfaye Tesso (KSU)
Donghai Wang (KSU)
Vara Prasad (KSU)
Ramasamy Perumal (KSU)
Scott Bean (USDA-ARS)
Frank White (KSU)
Clare Nelson (KSU)
Harold Trick (KSU)
Kerry Mayfield (Chromatin)
Larry Lambright (Chromatin)
Scott Staggenborg (Chromatin)

Mike Scanlon (Cornell)
Patrick Schnable (ISU)
Gary Muehlbauer (U of Minn.)
Marja Timmermans (CSHL)

MingLi Wang (USDA-ARS)
Gary Pederson (USDA-ARS)
Rex Bernardo (U of Minn.)
Sharon Mitchell (Cornell)
Guihua Bai (USDA-ARS)
Mitch Tuinstra (Purdue)
Ismail Dweikat (UNL)
Tom Clemente (UNL)
Min Zhang (Purdue)
H-P Piepho (Hohenheim)
Xiaoming Liu (U. Texas)

Xiaoqing Yu (ISU)
Xianran Li (ISU)
Tingting Guo (ISU)
Matt Dzievit (ISU)
Jinyu Wang (ISU)
Paul White (ISU)

Jode Edwards (ISU)
Fernando E. Miguez (ISU)
Sotiris Archontoulis (ISU)
Liang Dong (ISU)
Kevin Price (AgPixel)
Outline

• Genomic Selection
  – Historical Context & Technical Details
  – Breeder’s 10 Tenets

• Empirical Findings
  – 44,733, 1,000, 300, 200, 100, 50, 50

• Perspectives
  – Optimizing GS following Breeder’s Tenets
What is “Selection”?

- Differential viability (survival) and fertility (reproduction) among genotypes. Natural selection versus artificial selection.
- In plant breeding, selection changes allele frequencies and genotype frequencies.
- Selection happens at different stages of the breeding cycle.
What is not “Genomic” Selection?

- Selection based on observations (from the unreplicated trial)
- Selection based on rankings from the replicated, multi-environment trials
- Selection on best linear unbiased prediction of the phenotypes on the individual and its relatives
What is not “Genomic” Selection?

• Genotyping and selection based on known markers (e.g., disease resistance, insect resistance, or major QTLs)
  – Assume we have identified the informative markers

• Genotyping and selection based on sum of estimated genotypic values across multiple QTLs
  – Assume we have detected multiple QTLs
  – Assume the summation across individual QTL effects provides a good estimate of the overall genotypic value
  – $G_Q = \sum_{k=1}^{q}(\alpha_i^k + \alpha_j^k)$, where $q$ is the number of declared QTLs
What is Genomic Selection?

• Genotyping and selection based on Genomic Estimated Breeding Value (GEBV)
  - $G_{EBV} = \sum_{k=1}^{m} (\alpha^k_i + \alpha^k_j)$, where $m$ is the number of genome-wide markers
  - Assume we can have accurate estimates of marker effects
  - Assume $G_{EBV}$ is a better estimate of the overall genotypic value than $G_Q$

• A procedure where genetic merits are predicted for unphenotyped individuals so that selection can be made
  - **Model Training** Genome-wide marker information across a group of phenotyped individuals is used in developing the prediction model
  - **Prediction** Prediction is made for a different group of unphenotyped individuals but with genome-wide marker information
  - **Selection** Either directly advance the breeding generation, or significantly narrow down the candidate pool before field evaluation
What is Genomic Selection?

At the start, should you be thinking about how close your $r$ is to “1”?
or how different your $r$ is from “0”?

$r = 0.77$ in this case
What is Genomic Selection?

It is all about having a step to improve the plant breeding process by better utilizing resources: time and cost.

What is Genomic Selection?

Individuals
\[ y = X\beta + Ig + e \]
\[ V = \text{Var}(y) = A\sigma_g^2 + I\sigma_e^2 \]

Markers
\[ y = X\beta + Wu + e \]
\[ V = \text{Var}(y) = WW'\sigma_u^2 + I\sigma_e^2 \]

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{\gamma}
\end{bmatrix} = \left[ X'X + \frac{X'\sigma_u^2}{\sigma_e^2} I \right]^{-1} \begin{bmatrix}
X' \ y \\
\end{bmatrix}
\]

Connection
\[ A\sigma_g^2 = WW'\sigma_u^2 \]

\[ \sigma_u^2 = \sigma_e^2 / m, A = WW'/m \]

\[ \hat{\gamma} = A \frac{\sigma_u^2}{\sigma_e^2} \hat{\gamma}^{-1} (y - X\hat{\beta}) \]
\[ \hat{u} = W'A^{-1} \hat{\gamma}/m = W'(WW')^{-1} \hat{\gamma} \]

Genomic Estimated Breeding Value (GEBV) of the individuals with only marker data
\[ \hat{g}_{new} = W_{new} \hat{u} \quad G_{EBV} = \sum_{k=1}^{m} (\alpha_i^k + \alpha_j^k) \]
What is Genomic Selection?

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{\mu}
\end{bmatrix} = \left[ X'X & X'W \\
W'X & W'W + \frac{\sigma_e^2}{\sigma_u^2} \mathbf{I}
\right]^{-1} \begin{bmatrix} X'y \\ W'y \end{bmatrix}
\]

\[
V = \text{Var}(y) = WW'\sigma_u^2 + I\sigma_e^2
\]

\[
\lambda = \frac{\sigma_e^2}{\sigma_u^2}
\]

Marker Effect Estimates

\(\hat{\mu}\) estimated from Training Population, and used for Prediction/Validation Population

Prediction Accuracy

\[
\begin{align*}
r_1 &= \text{Corr} (y_{new}, \hat{y}_{new}) \\
r_1 &= \text{Corr} (y_{new}, \hat{y}_{new}) / \sqrt{h^2}
\end{align*}
\]

• What is the distribution of quantitative trait variants?
  
  – **GBLUP**, Single distribution \(u \sim N(0, \sigma_u^2)\)
  
  – **BayesA**, Different distributions \(u_k \sim N(0, \sigma_{u_k}^2), \sigma_{u_k}^2 \sim \chi^{-2}(v, s)\) with probability of 1-\(\pi\)
  
  – **BayesB**, \(\sigma_{u_k}^2 = 0\) with probability of \(\pi\) AND \(u \sim N(0, \sigma_u^2)\) with probability of 1-\(\pi\); \(\pi\) estimated from the data
  
  – **Kernel-based methods**, capturing non-additive effects
What Can We Do with Genomic Selection?

• Genomic selection has been extensively studied in recent years
  – Prediction accuracy, Different models, Different genetic architectures
  – Resources allocation: “Plant breeding is a number’s game.” “Genetic gain per unit time and cost.”

• Genomic selection has been well integrated into many plant breeding programs
  – Inbred development
  – Testcross evaluation
  – Parental selection for generating breeding populations
  – Hybrid prediction
Emerging Areas of Research

- Genomic selection under G x E
  - Environmental covariates
  - Modeling G x E prior to genomic selection
  - Reaction norm model
  - Integration of genomic selection and crop model

- Genomic selection for exotic germplasm introgression

- Genomic selection for germplasm enhancement?
Ten Things a Successful Plant Breeder Understands - Roger Boerma

1. **Definition and exploitation of base population of environments and genotypes.**

3. **Uniform plots strengthen the association between phenotype and genotype;**

4. **Use of heritability of a trait to achieve maximum gain per year.**

6. **There is a large amount of additive genetic variance for most quantitative traits.**

7. **Marker-assisted selection is a tool, not a discipline; and genome-wide selection requires experimental verification.**

8. **Include check genotypes in breeding experiments and know how you will analyze an experiment before you plant it.**

10. **Time and cost are the adversaries of plant breeders.**
Bioenergy Crop Breeding and Genomics

• New crops that not much breeding and genetics have been conducted in the past. But, many genomics and phenomics resources are available.

• **Genomics-Assisted Breeding**
  
  - How do we tap into the vast plant germplasm collections for biomass crop improvement?
  - How do we increase the information contained in genotypic and phenotypic data for the selected germplasm so that this information can generate maximum knowledge to enrich our understanding of the germplasm and genotype-phenotype relationship?
Start Right with Available Germplasm Information

• Selection of the initial biomass sorghum accessions (base population)
  
  – Step 1: Flowering = 6, photoperiod sensitive; Plant height (280 ~ 400cm)
  – Step 2: Seed availability at GRIN for distribution
  – Step 3: Germplasm information: collection site, race, height uniformity, disease, lodging, etc.

Empirical Findings

44,733 accessions
16 species, from 116 countries
34,844 (78%) photoperiod sensitive
Selective Phenotyping

- Trait evaluation of only those individuals that contain the maximum genetic information
  - Genotyping a large set of accessions
  - Select a subset for a given sample size to maximize genetic information content (allele number, gene diversity, recombination break point, etc.)
- Mapping power and resolution
- Prediction accuracy
Selective Phenotyping

Genotyping-By-Sequencing (GBS) of all

- Choose most representative accessions & Phenotype
- Build and evaluate genomic prediction models

Predict genomic estimated breeding values of the rest
Empirical Findings

Genetic Relationship

1000

300
Two replications in each environment
Prediction Potential
Evaluated by Cross Validation

Empirical Findings

Training Population Size
Model Performance
Evaluated by Cross Validation

Empirical Findings

- Biomass
- Height

Prediction Accuracy vs. Training Population Size

- Models: Kinship, EXP, GAUSS, BayesA, BayesB

300
Genome-wide Prediction of Unphenotyped Individuals

Empirical Findings

Total for empirical validation: 200
Biomass

Empirical Findings

Biomass Yield (Mt/ha) vs. Estimated Breeding Value

- r = 0.42
- r = 0.67
Empirical Findings

Plant Height

Observed Height vs. Genomic Estimated Breeding Value

$r = 0.70$

$r = 0.71$
Empirical Findings

Root Lodging

Genomic Estimated Breeding Value

$r = 0.42$

$r = 0.67$

200

Empirical Findings

Root Lodging

Genomic Estimated Breeding Value

$r = 0.42$

$r = 0.67$

200
Bioenergy crops provide a great opportunity to implement and test new strategies of genomics-assisted breeding
Why did Things Work out Nicely?

44,733 accessions
16 species, from 116 countries
34,844 (78%) photoperiod sensitive

Breeder’s Tenet 6: There is a large amount of **additive genetic variance** for most quantitative traits
Base Population

Breeder’s Tenet 1: Definition and exploitation of **base population** of **environments** and **genotypes**

- Is it possible to define the base population of genotypes at the germplasm collection level?

- Is it the right time to GBS all accessions of a species in a gene bank?

- Are we saying that one can generate a potentially functional GS model?
Phenotyping

• Breeder’s Tenet 3: Uniform plots strengthen the association between phenotype and genotype

• Breeder’s Tenet 4. Use of heritability of a trait to achieve maximum gain per year
  
  – Is it important to make sure our phenotypic data for training population reliable? Validation population?

  – Can we achieve higher accuracy with High Throughput Phenotyping?

  – Will there be many startup companies that are focused on generating high accuracy field-level phenotypes?
High Throughput Phenotyping

Sorghum Canopy Upper Tier Classification of Leaf Area Contiguity for Select Plots

Area (m²)

- Less Contiguity
- More Contiguity
- Most Contiguity
Environment

• Breeder’s Tenet 8: Include **check genotypes** in breeding experiments and know how you will **analyze** an experiment before you plant it.

  – *Should we rely on environment variables to generate the kernel (variance-covariance) among different environments, particularly training and validation stage?*

  – *Should we include a small subset of training population to mark up the testing environments at validation stage?*
Optimization

• Breeder’s Tenet 1: Definition and exploitation of base population of environments and genotypes
Optimization, Optimization

• Are there better algorithms to “design” the training population? Testing sites?

• Should we consider this design under a “progressive” context so that the validation information can be used to update the prediction model?

• Are there options to “optimize” the validation set?

• Should we “embed” checks that capture the GxE dynamics so we are well prepared to deal with uncertainties?
– Is it possible to define the base population of genotypes at the germplasm collection level?
– Is it the right time to GBS all accessions of a species in a gene bank?
– Are we saying that one can generate a potentially functional GS model?
– Is it important to make sure our phenotypic data for training population reliable? Validation population?
– Can we achieve higher accuracy with High Throughput Phenotyping?
– Will there be many startup companies that are focused on generating high accuracy field-level phenotypes?
– Should we rely on environment variables to generate the kernel (variance-covariance) among different environments, particularly training and validation stage?
– Should we include a small subset of training population to mark up the testing environments at validation stage?
– Are there better algorithms to “design” the training population? Testing sites?
– Should we consider this design under a “progressive” context so that the validation information can be used to update the prediction model?
– Are there options to “optimize” the validation set?
– Should we “embed” checks that capture the GxE dynamics so we are well prepared to deal with uncertainties?

Yes to all these questions!
Take-home Message

• Genomic Selection is more than a difficult-to-understand equation.

What is Genomic Selection?

• Genomic Selection is a tool/strategy in plant breeding. Itself and the whole process of plant breeding is being, and can be further optimized.
QUESTION & ANSWER SESSION
Thank you to our sponsors: