# Dr. Jianming Yu

## Iowa State University

### Professor of Agronomy

Plant Breeding Symposium
Texas A&M University
2015

Texas A&M Plant Breeding Symposium
February 19, 2015
10:00 – 11:00 am

# **Genomic Selection**: Historical Context, Technical Details, Empirical Findings, and Perspectives

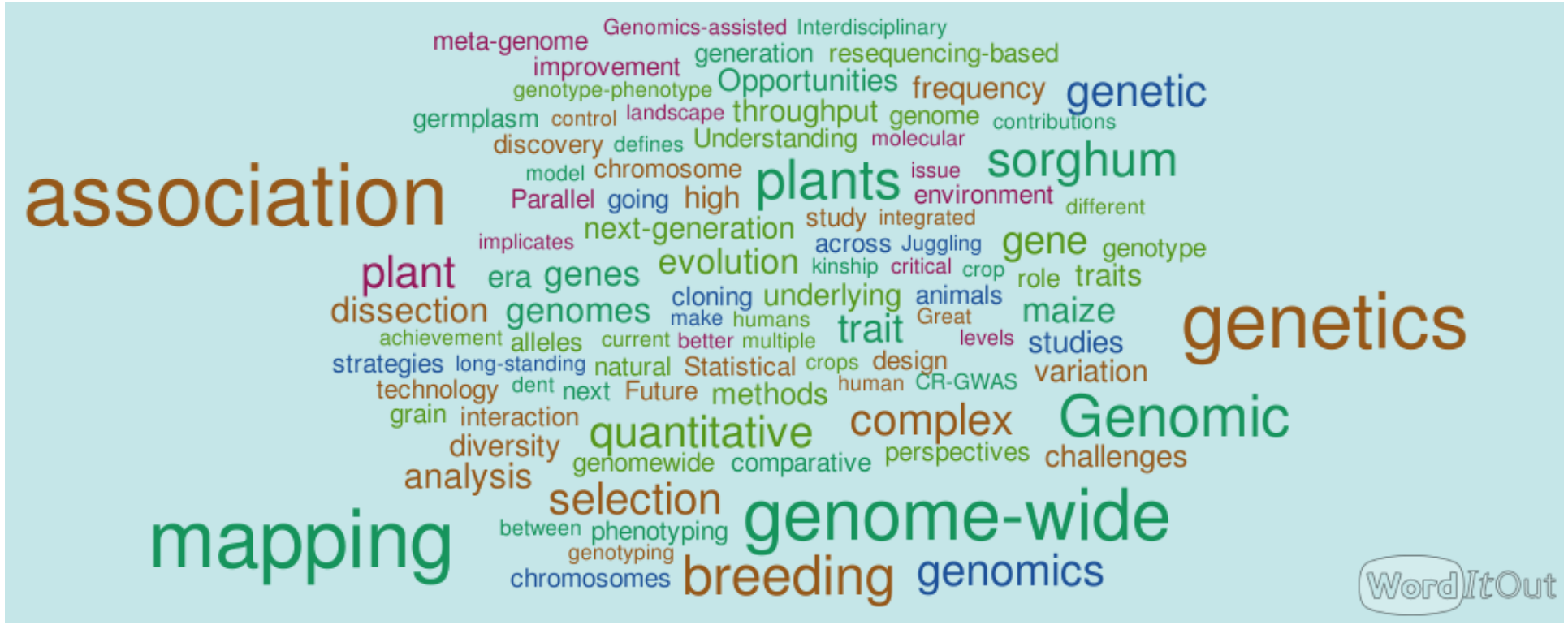
Jianming Yu

Department of Agronomy

Raymond F. Baker Center for Plant Breeding

# Introduction

From 56 presentation titles



USDA United States Department of Agriculture
National Institute of Food and Agriculture
USDA/DOE Plant Feedstock Genomics for Bioenergy

Targeted Excellence K-STATE Kansas State University

# 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)

Kansas GRAIN SORGHUM Commission

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)

Center for SORGHUM Improvement

**Xiaoqing Yu (ISU)**
Xianran Li (ISU)
Xin 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)

SAM Maize Shoot Apical Meristem
NSF
DUPONT PIONEER
IOWA STATE UNIVERSITY Department of Agronomy

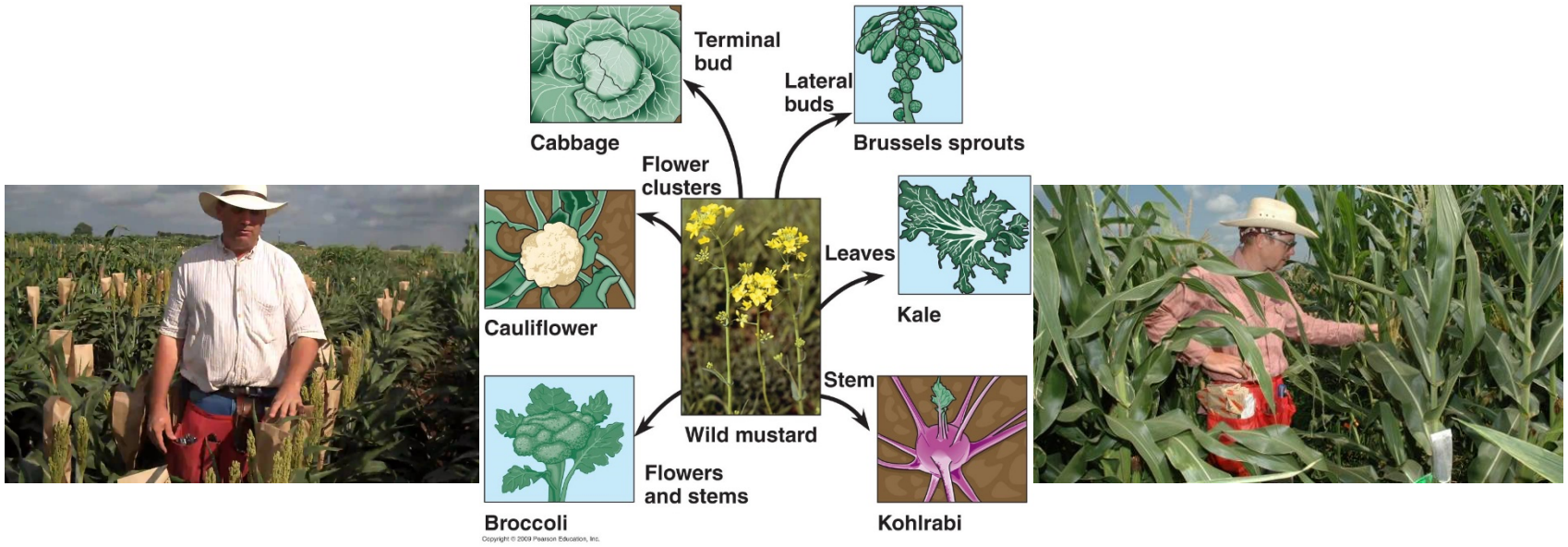
# Outline

- 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

Historical Context

# *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.



Historical Context

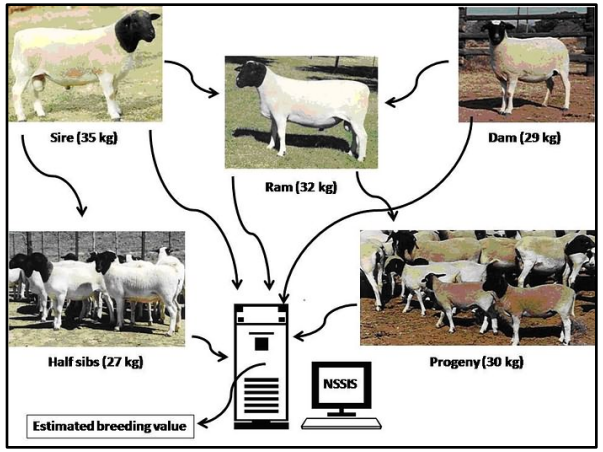
# *What is not "Genomic" Selection?*

- 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

Parallel selection for non-shattering
Nature Genetics 44:720–724

Texas A&M AgriLife Extension
Service photo by Dr. Gaylon Morgan

Breeding value for weaning weight of a ram




GROOTFONTEIN
Agricultural Development Institute

Historical Context

# *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
  - $\boldsymbol{G_Q} = \sum_{k=1}^{q} (\alpha_i^k + \alpha_j^k)$, where $\boldsymbol{q}$ is the number of declared QTLs

Historical Context

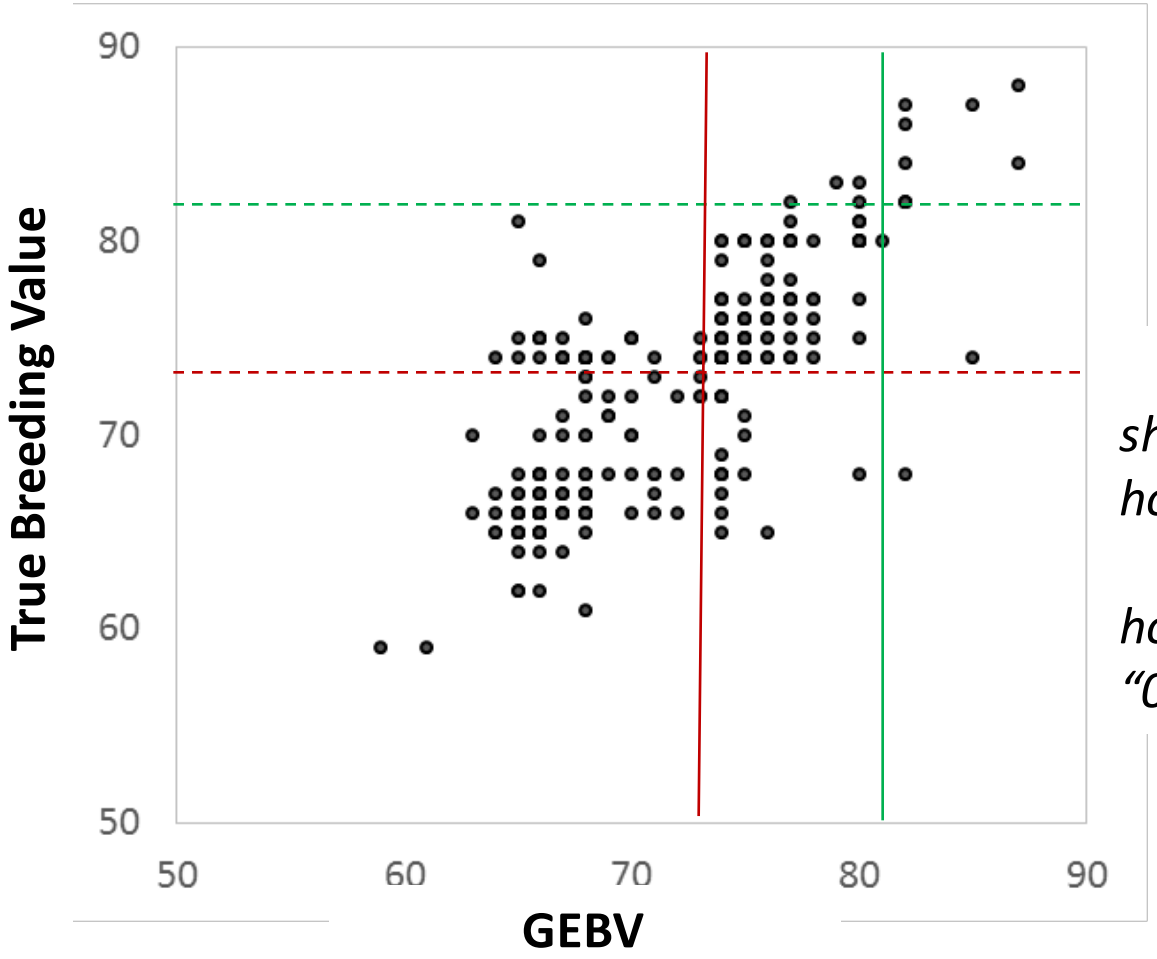
# *What is Genomic Selection?*

- Genotyping and selection based on Genomic Estimated Breeding Value (GEBV)
  - $G_{EBV} = \sum_{k=1}^{m} (\alpha_i^k + \alpha_j^k)$, 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

Technical Details

# *What is Genomic Selection?*

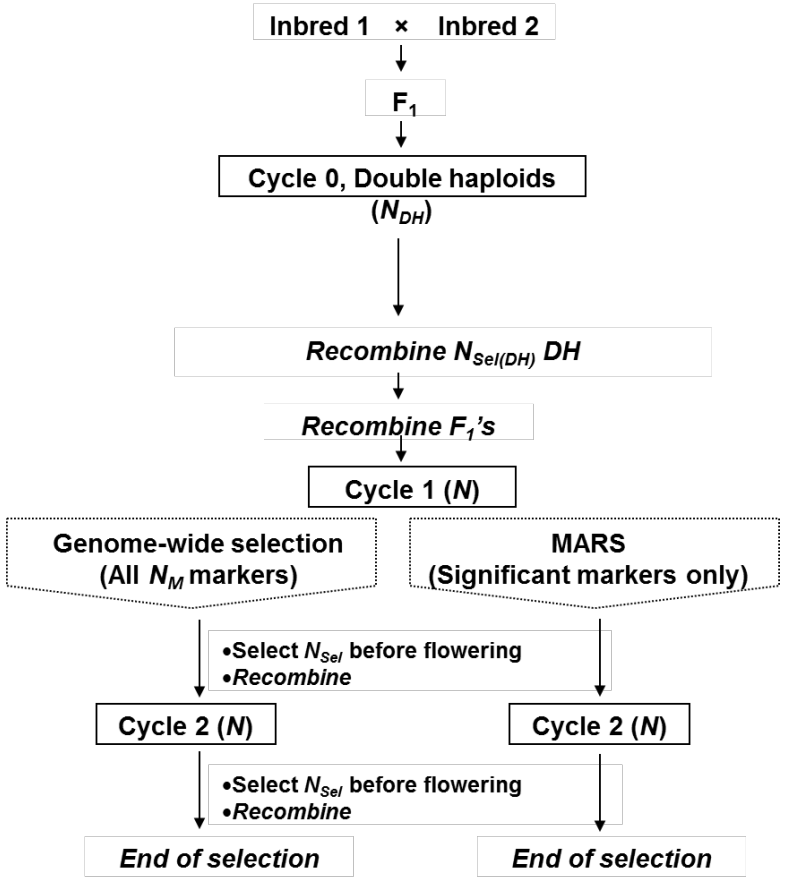



*r* = 0.77 in this case

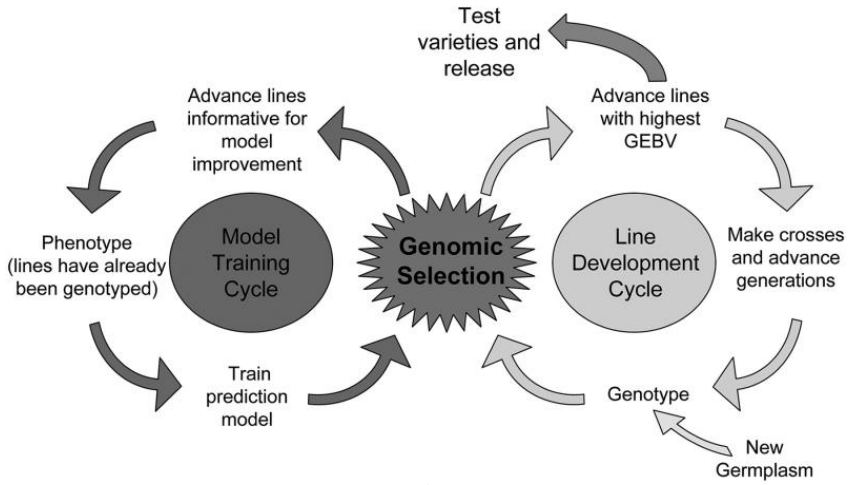
At the start,
*should you be thinking about how close your r is to "1"?*
or
*how different your r is from "0"?*

Technical Details

# *What is Genomic Selection?*




Bernardo and Yu, 2007. Crop Science 47:1082-1090.




Heffner *et al.* 2009. Crop Science 49:1-12.

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

Technical Details

# *What is Genomic Selection?*

Individuals

$$y = X\beta + Ig + e$$

$$V = Var(y) = A\sigma_g^2 + I\sigma_e^2$$

Markers

$$y = X\beta + Wu + e$$

$$V = Var(y) = WW'\sigma_u^2 + I\sigma_e^2$$

$W_{ij}=(M_{ij}-2p_i)/\mathrm{Sqrt}(2p_i(1-p_i))$

$$\begin{bmatrix}\hat{\beta}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}X'X & X'\\ X & I+\frac{\hat{\sigma}_e^2}{\hat{\sigma}_g^2}A^{-1}\end{bmatrix}^{-1}\begin{bmatrix}X'y\\ y\end{bmatrix}$$

Connection

$$A\sigma_g^2 = WW'\sigma_u^2$$

$$\sigma_u^2 = \sigma_g^2/m,\ A = WW'/m$$

$$\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix} = \begin{bmatrix}X'X & X'W\\ W'X & W'W+\frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}I\end{bmatrix}^{-1}\begin{bmatrix}X'y\\ W'y\end{bmatrix}$$

$$\hat{g} = A\frac{\hat{\sigma}_g^2}{\hat{\sigma}_e^2}\hat{V}^{-1}(y - X\hat{\beta})$$

$$\hat{u} = W'A^{-1}\hat{g}/m = W'(WW')^{-1}\hat{g}$$

$$\hat{\lambda}=\frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}$$

Genomic Estimated Breeding Value (GEBV) of the individuals with only marker data

$$\hat{g}_{new} = W_{new}\hat{u} \qquad \boldsymbol{G_{EBV}}=\sum_{k=1}^{m}(\alpha_i^k + \alpha_j^k)$$

Technical Details

# *What is Genomic Selection?*

| Markers $y = X\beta + Wu + e$ | Marker Effect Estimates |
|---|---|
| $\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix} = \begin{bmatrix}X'X & X'W\\ W'X & W'W + \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}I\end{bmatrix}^{-1}\begin{bmatrix}X'y\\ W'y\end{bmatrix}$<br>$V = Var(y) = WW'\sigma_u^2 + I\sigma_e^2$<br>$\hat{\lambda} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}$ | $\hat{u}$ estimated from Training Population, and used for Prediction/Validation Population<br>Prediction Accuracy<br>$r_1 = Corr\,(y_{new}, \hat{g}_{new})$<br>$r_1 = Corr\,(y_{new}, \hat{g}_{new})/\sqrt{h^2}$ |

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

Technical Details

# *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

Technical Details

# 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?*

Breeders' Tenets

# Ten Things a Successful Plant Breeder Understands - Roger Boerma

10 Tenets

1. De ments
and
3. Ur type
and
4. Us year
6. Th ost
quar
7. M
geno
8. In ow
how
10. T

## Ten Things a Successful Plant Breeder Understands

H. Roger Boerma, Executive Director, Georgia Seed Development & Emeritus Professor, University of Georgia

4th Annual Meeting of the Nationa
8th Annual Meeting of the Plant Br
Minneapolis, MN - 6 August 2014

GEORGIA SEED DEVELOPMENT

# Bioenergy Crop Breeding and Genomics

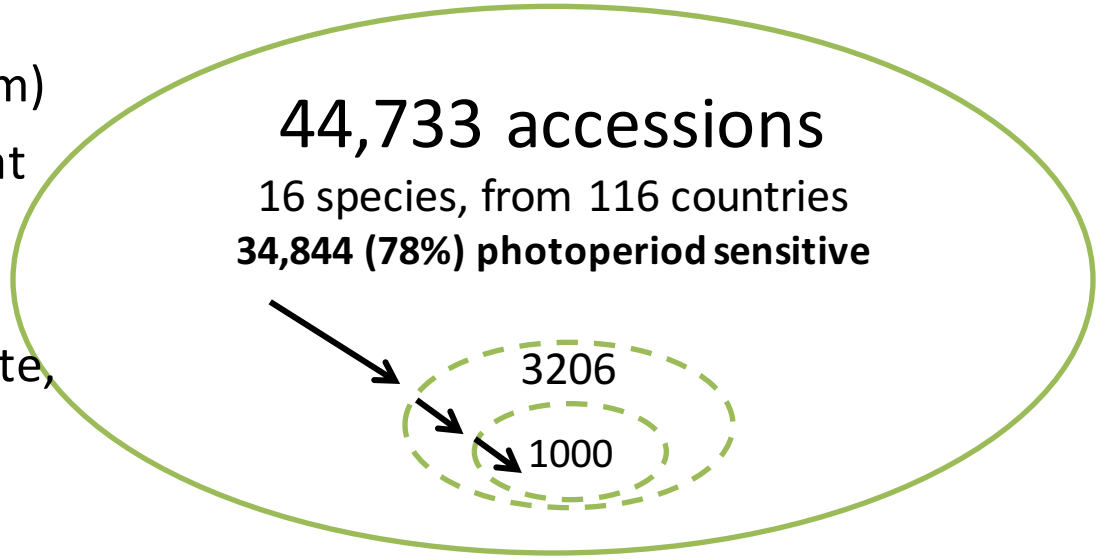
- New crops that not much breeding and genetics have been conducted in the past. But, many genomics and phenomics resources are available.
- **Genomics-Assisted Breeding**

➢ *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**?*

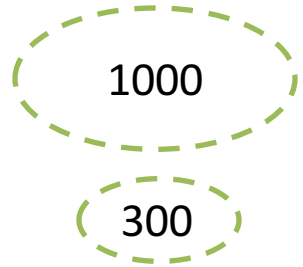
# Start Right with Available Germplasm Information

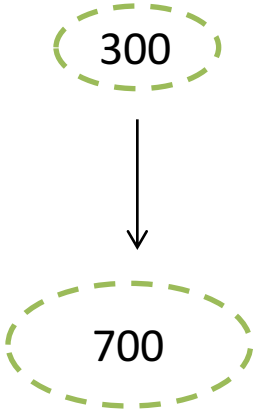
- Selection of the initial biomass sorghum accessions (base population)
  - Step 1: Flowering = 6, photoperiod sensitive; Plant height (280 ~ 400cm)
  - Step 2: Seed availability at GRIN for distribution
  - Step 3: Germplasm information: collection site, race, height uniformity, disease, lodging, etc.



# Selective Phenotyping

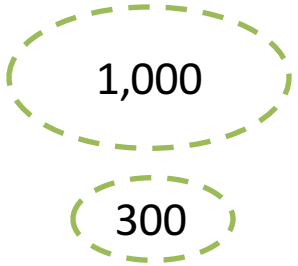
- Trait evaluation of only those individuals that contain the maximum genetic information
  - Genotyping a large set of accessions
  - Select a subset for a given sample size to maximize genetic information content (allele number, gene diversity, recombination break point, etc.)
  - Mapping power and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



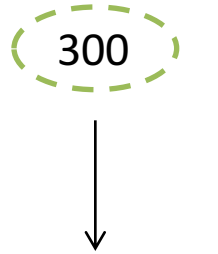
# Selective Phenotyping

Genotyping-By-Sequencing (GBS) of all




Choose most representative accessions & Phenotype

Build and evaluate genomic prediction models

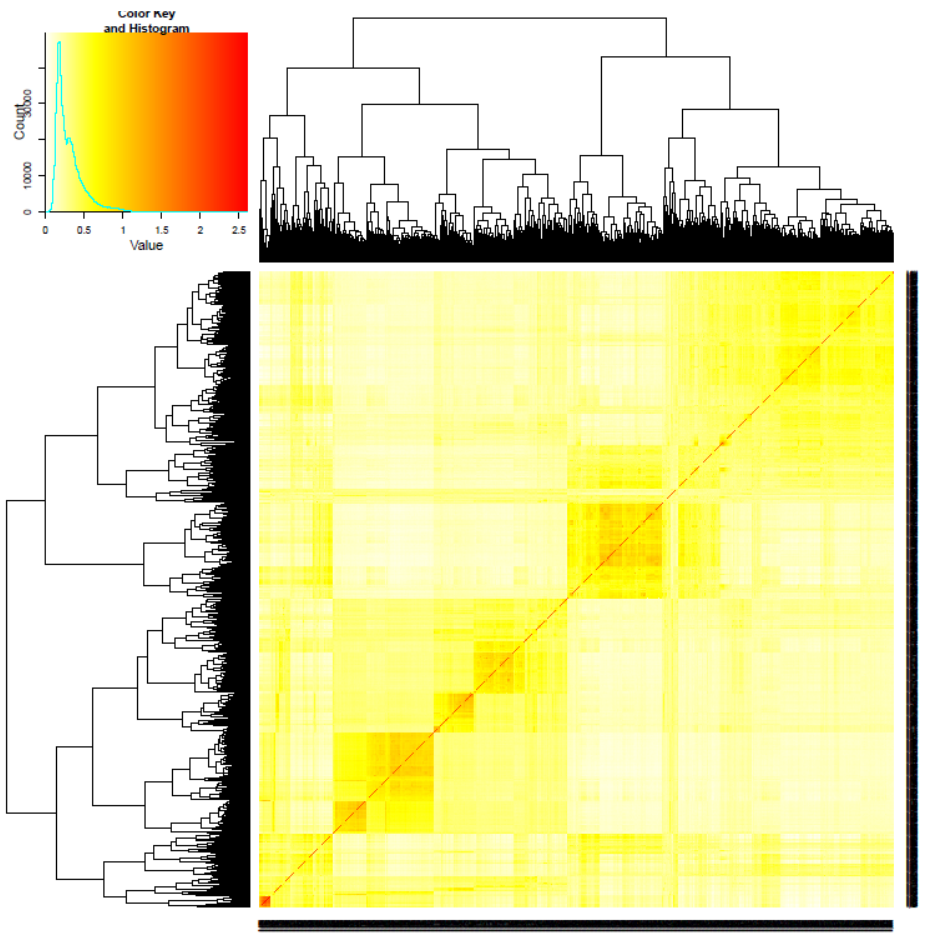



Predict genomic estimated breeding values of the rest

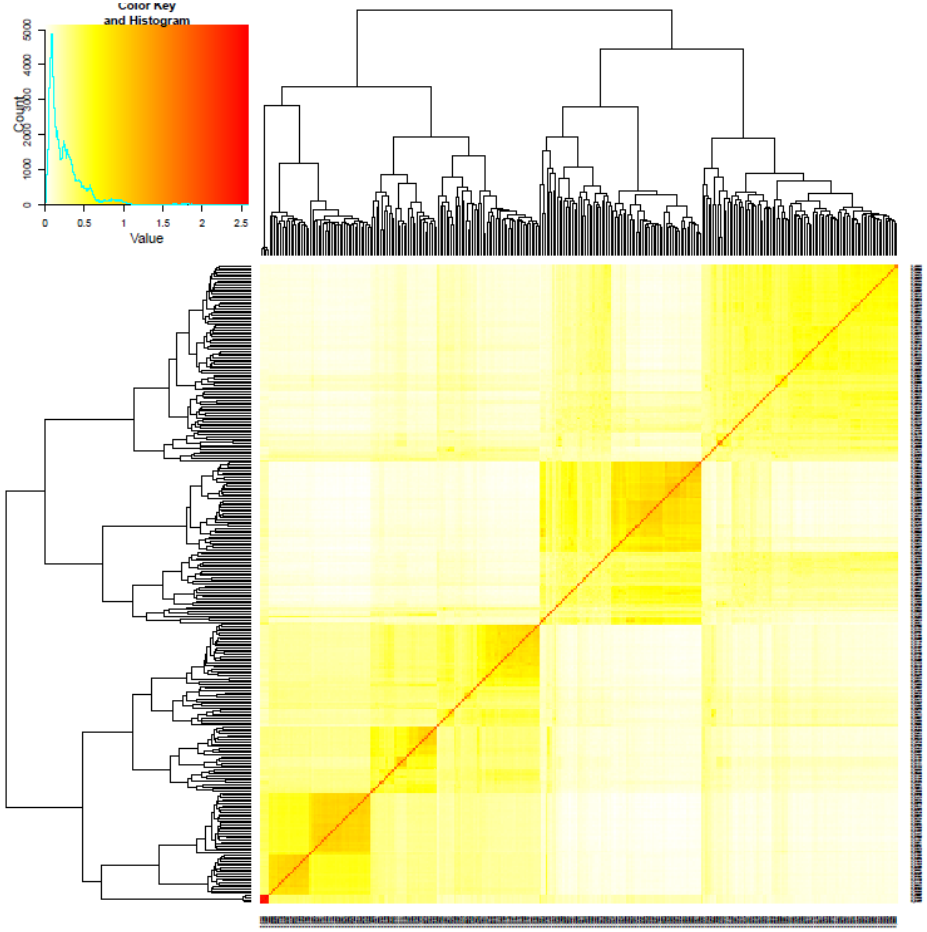
Empirical Findings

# Genetic Relationship

1000




300

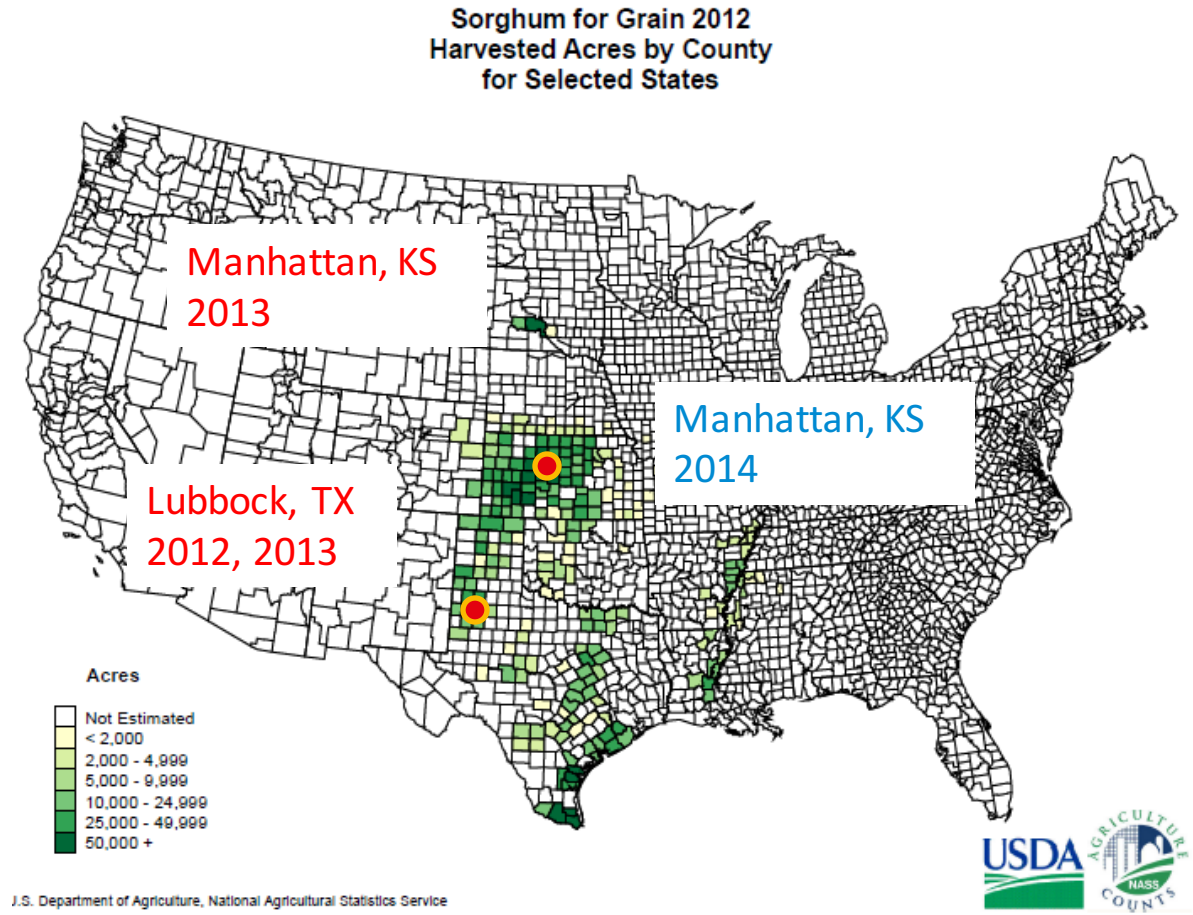


# Experiment Location

300

700

200



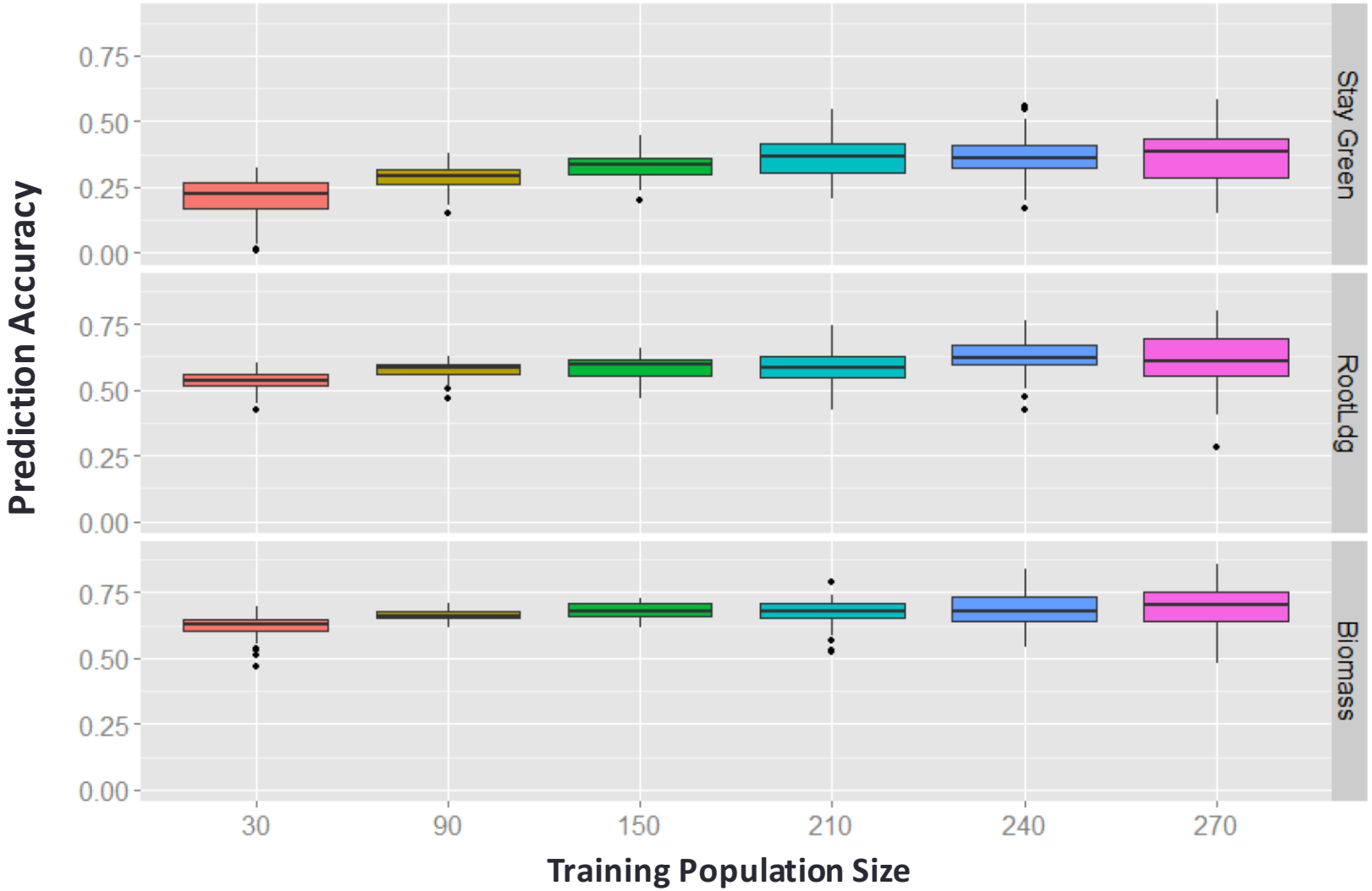

Two replications in each environment

Empirical Findings

300

# Prediction Potential

## Evaluated by Cross Validation

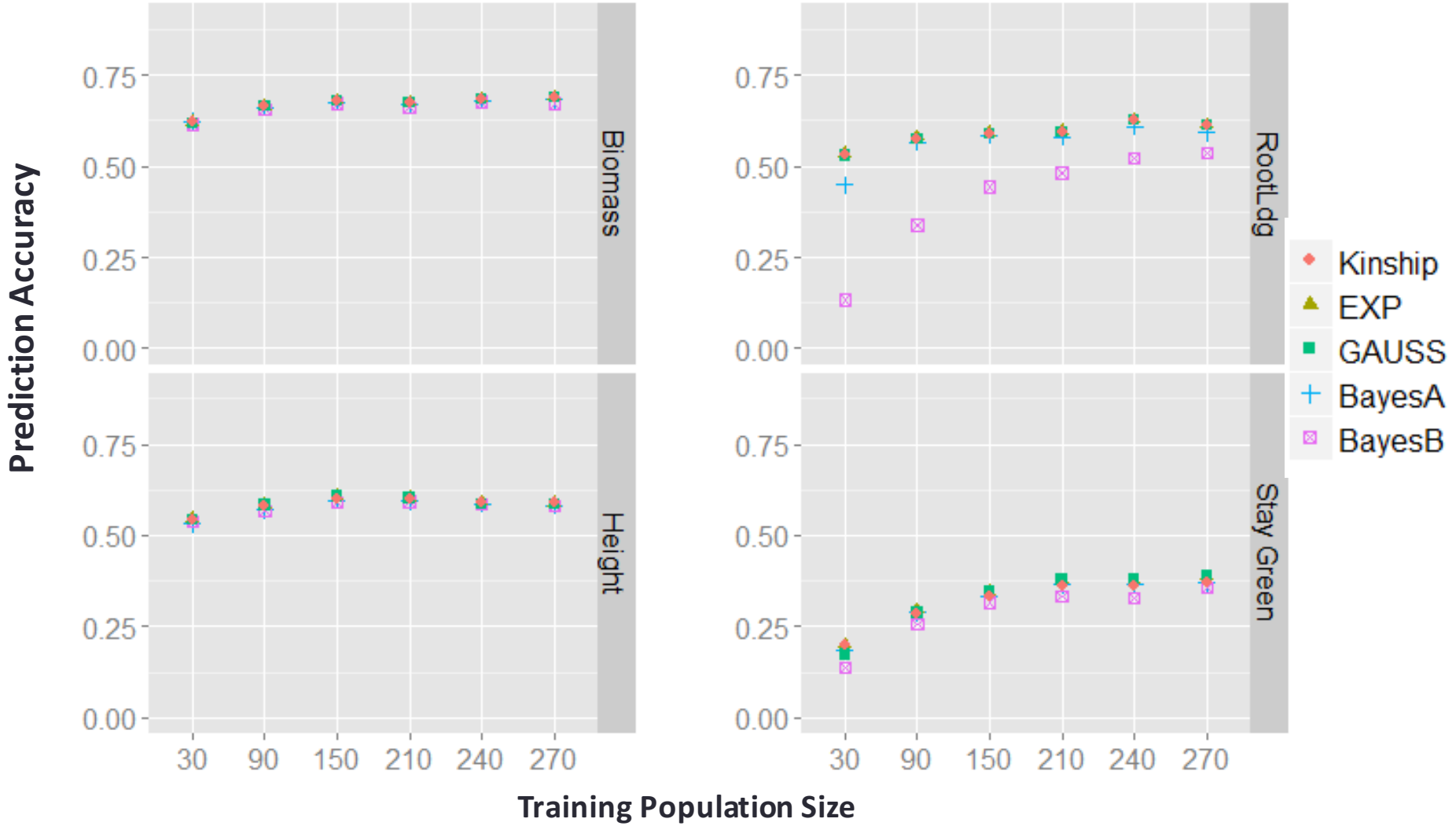


Empirical Findings

# Model Performance

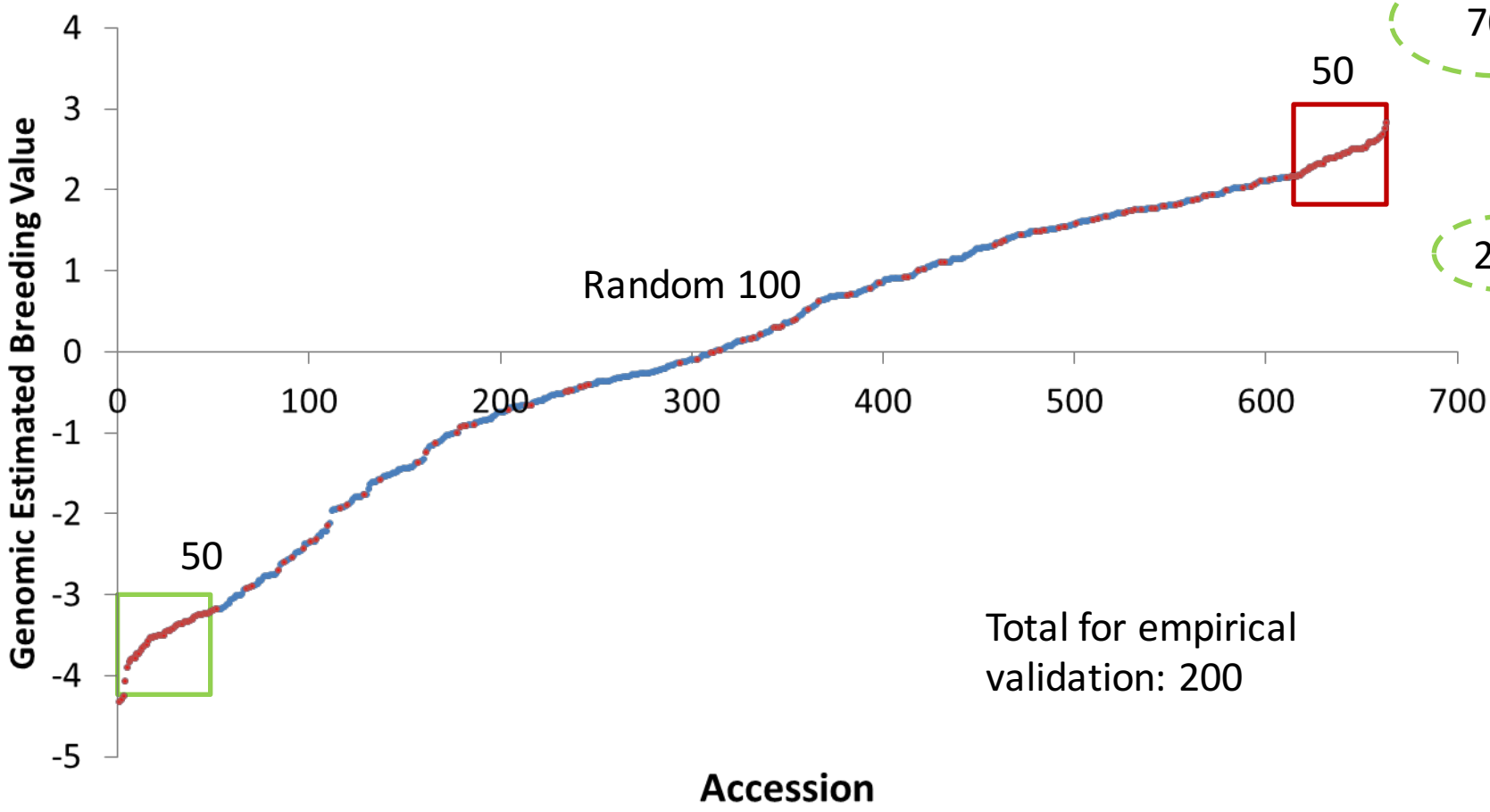
## Evaluated by Cross Validation

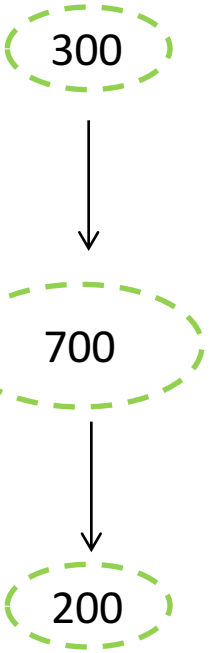
300



Empirical Findings

# Genome-wide Prediction of Unphenotyped Individuals

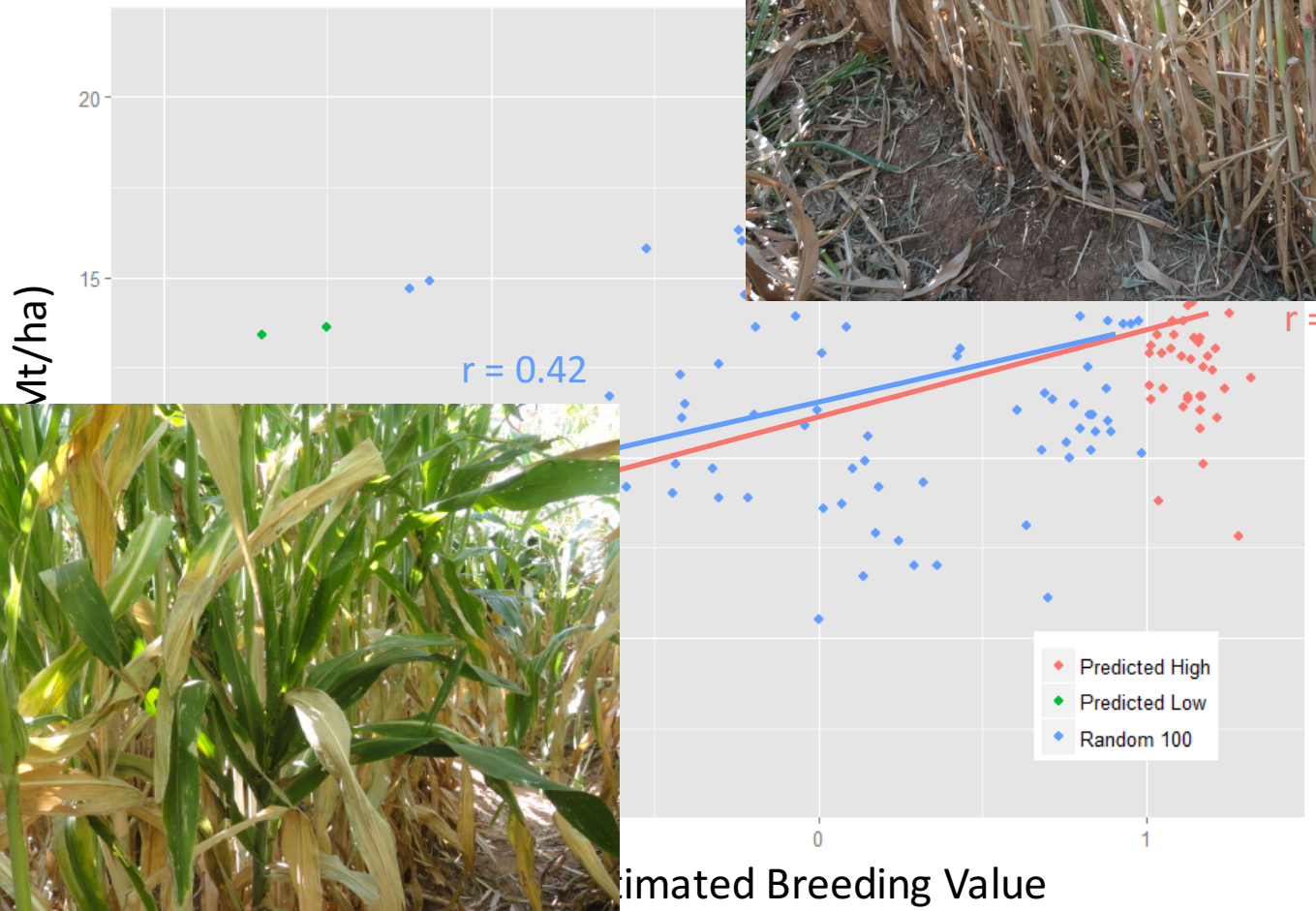



Total for empirical validation: 200

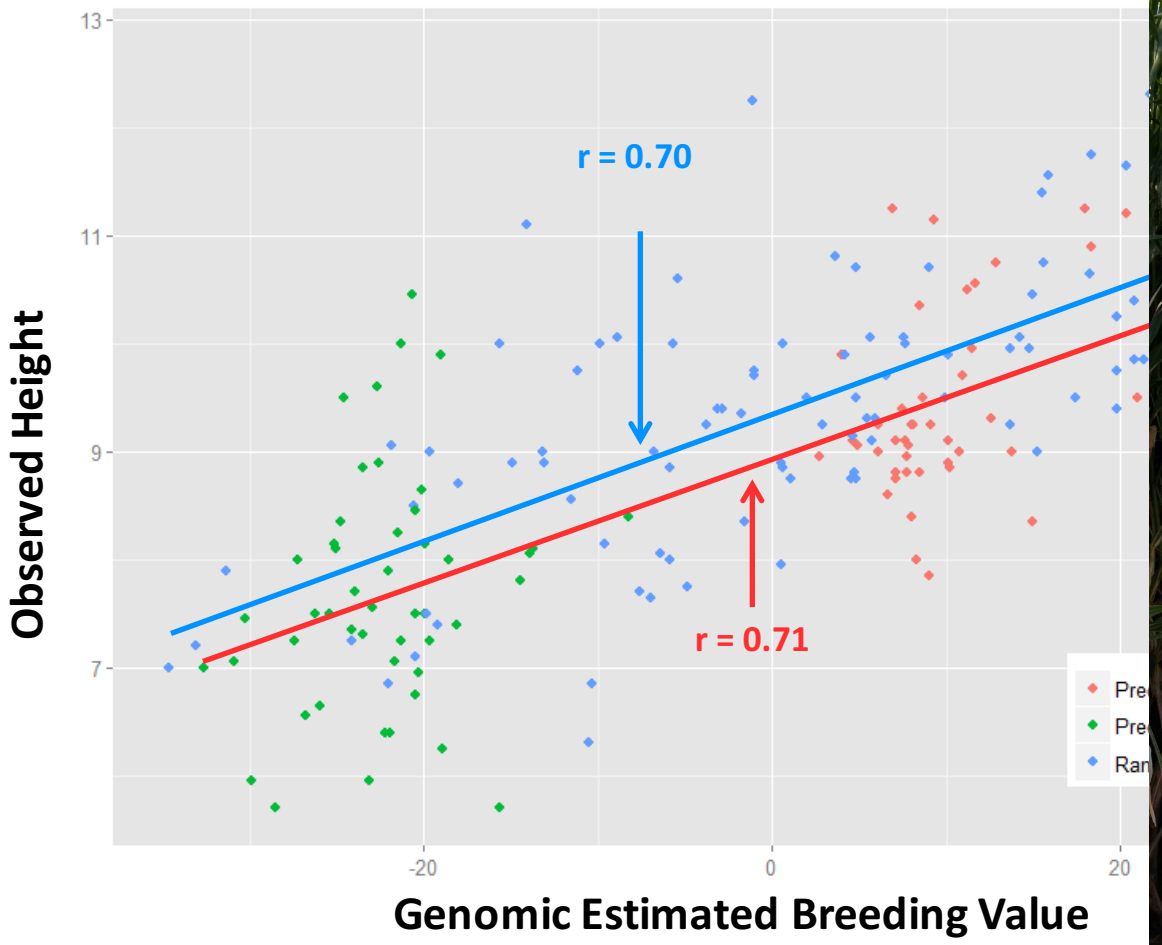
Empirical Findings

# Biomass



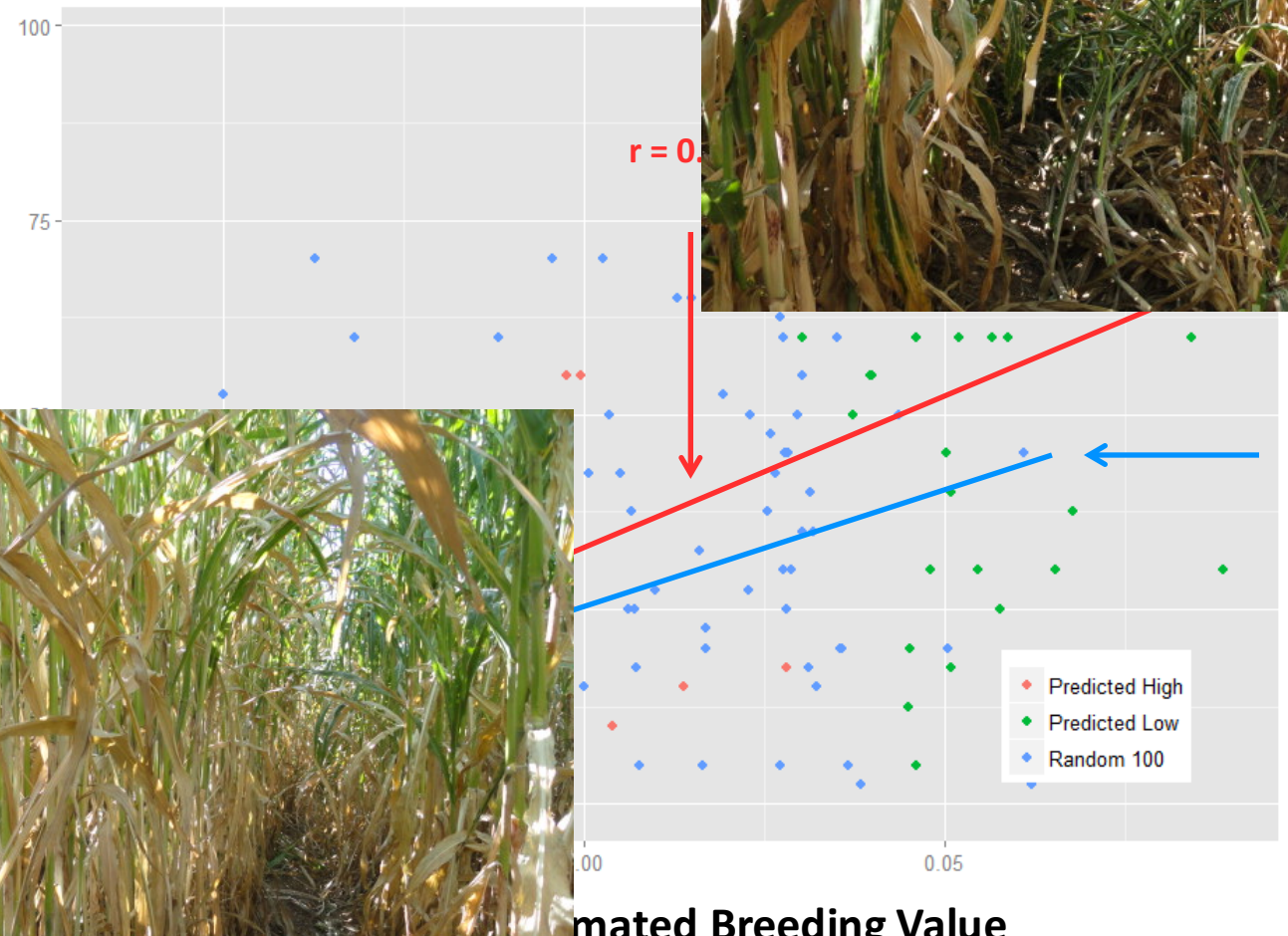
Empirical Findings

# Plant Height



Empirical Findings

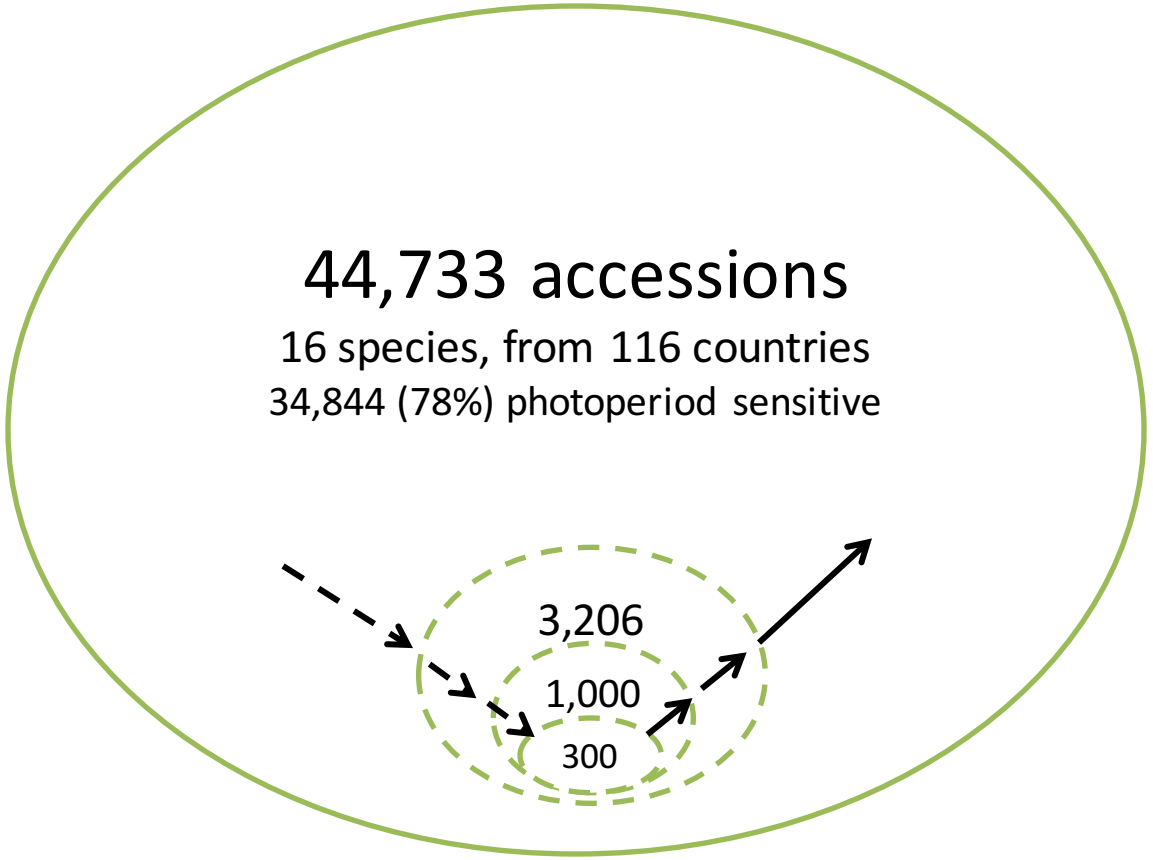
# Root Lo



Bioenergy crops
provide a great opportunity to
implement and test new strategies of
genomics-assisted breeding

Perspectives

# *Why did Things Work out Nicely?*




Germplasm Information
Genotyping-by-Sequencing
Selective Phenotyping
Genome-Wide Prediction
Genome-Wide Association Study

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

Perspectives

# 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?*

Perspectives

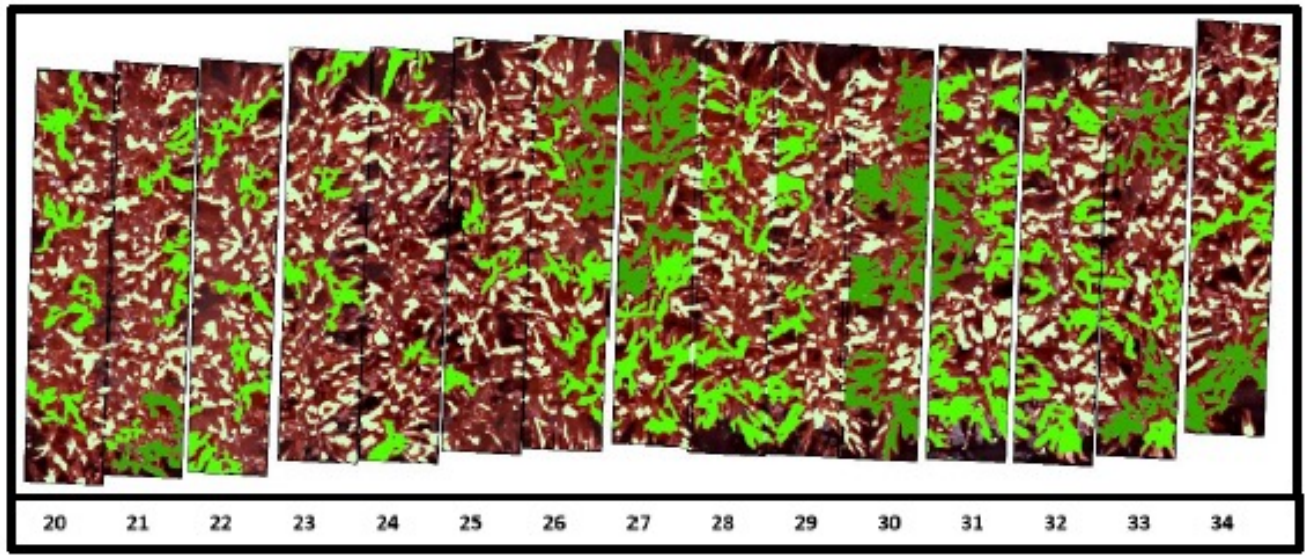
# Phenotyping

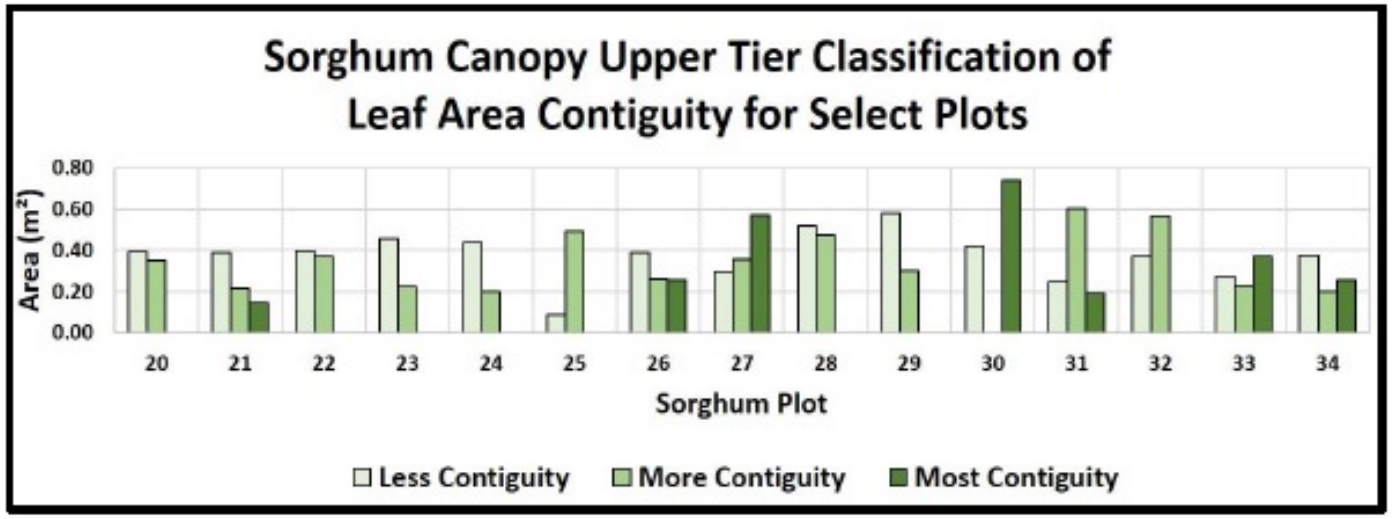
- Breeder's Tenet 3: **Uniform plots** strengthen the association between **phenotype** and **genotype**
- Breeder's Tenet 4. Use of **heritability** of a trait to achieve maximum **gain per year**
  - *Is it important to make sure our phenotypic data for training population reliable? Validation population?*
  - *Can we achieve higher accuracy with High Throughput Phenotying?*
  - *Will there be many startup companies that are focused on generating high accuracy field-level phenotypes?*

Perspectives

AgPixel

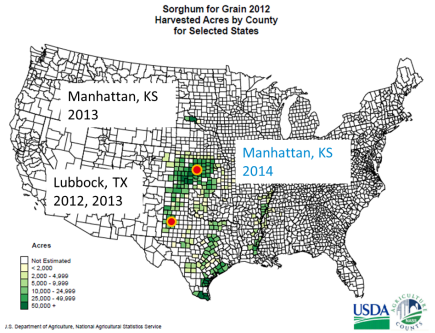
# High Throughput Phenotyping



Perspectives

# Environment

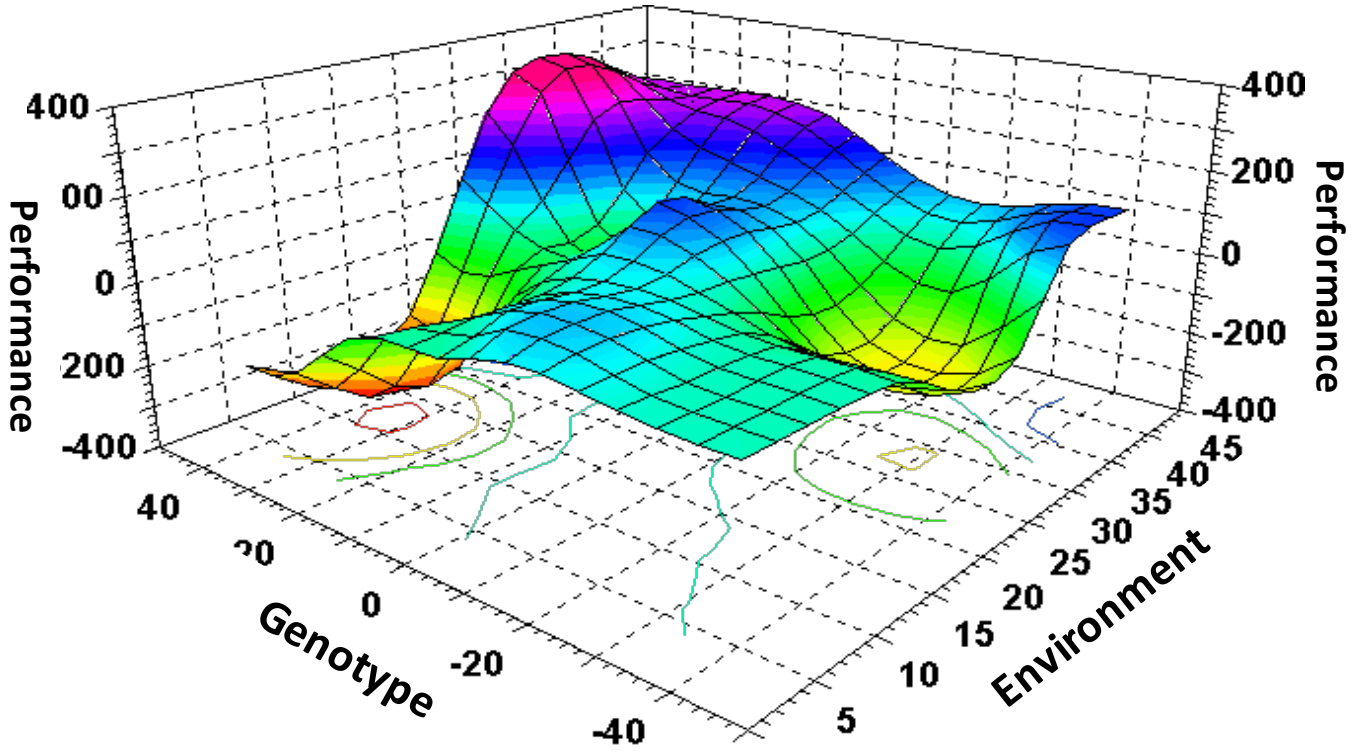



- Breeder's Tenet 8: Include **check genotypes** in breeding experiments and know how you will **analyze** an experiment before you plant it.
  - *Should we rely on environment variables to generate the kernel (variance-covariance) among different environments, particularly training and validation stage?*
  - *Should we include a small subset of training population to mark up the testing environments at validation stage?*

Perspectives

# Optimization

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



Perspectives

# 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?*

Perspectives

# Positive Perspectives

- *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 Phenotying?*
- *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?*

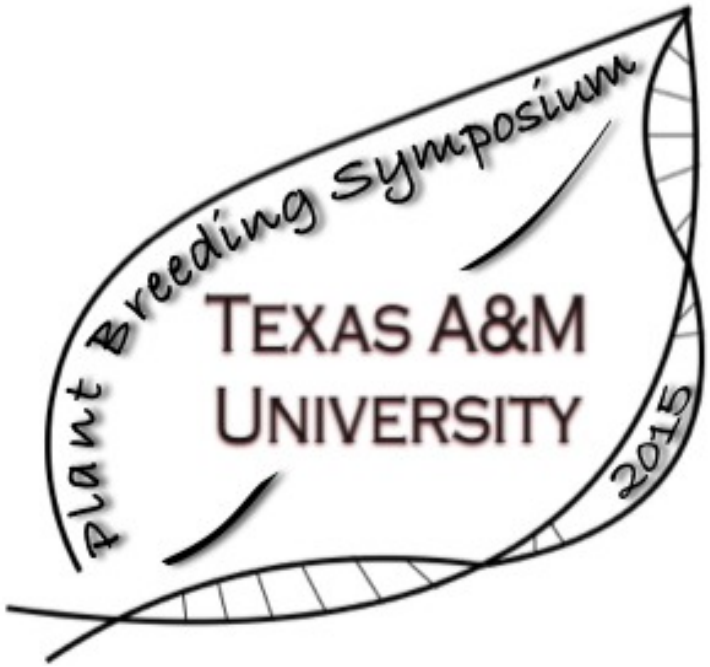
Yes to all these questions!

# 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:

DUPONT

PIONEER

AGRICULTURE & LIFE SCIENCES
TEXAS A&M UNIVERSITY

BIOLOGY
TEXAS A&M UNIVERSITY

Texas A&M Horticulture
The ultimate growth experience

Cotton Incorporated

SOIL & CROP SCIENCES
TEXAS A&M UNIVERSITY

GRADUATE AND PROFESSIONAL STUDIES
TEXAS A&M UNIVERSITY

Plant Breeding Symposium
Texas A&M University
2015