Design and Diversity for Complex Trait Dissection and Crop Improvement

Jianming Yu
Pioneer Distinguished Chair in Maize Breeding
Plant Sciences Institute Faculty Scholar
Raymond F. Baker Center for Plant Breeding
Quantitative Genetics & Maize Breeding

“To address significant questions in plant breeding by combining cutting-edge genomic technologies and quantitative genetics theories”

We believe the power of quantitative genetics to integrate genomics, biotechnology, phenomics, statistics, and ... into plant breeding.
Guiding Questions

1. How should we design the current plant breeding methods to better use the high throughput genotyping capacity, genetic resources, and high throughput phenotyping technologies?

2. How can we leverage genetic design, experimental design, and optimization thinking to plan out our research?

3. How can we efficiently identify genes underlying traits with agronomic and evolutionary importance so that the resulting empirical findings can guide our research in plant breeding?
Research and Training

• Complex Trait Dissection
• Breeding Strategy

• Genes and Genetics
• Genomes and Chromosomes

Griffiths et al. 2002
What is Genomic Selection?

• 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?

Individuals

\[ y = X\beta + Ig + e \]

\[ V = \text{Var}(y) = A\sigma_g^2 + l\sigma_e^2 \]

\[
\begin{bmatrix}
\hat{\beta}
\end{bmatrix} = \left[ X'X + \frac{1}{\sigma_g^2} A^{-1} \right]^{-1} \left[ X'y \right]
\]

\[
\sigma_g^2 = \frac{\sigma_u^2}{m}, \quad A = WW'/m
\]

\[
\hat{\gamma} = A\frac{\sigma_g^2}{\sigma_e^2} \hat{V}^{-1} (y - X\hat{\beta})
\]

\[
\hat{u} = W'A^{-1}\hat{\gamma}/m = W'(WW')^{-1} \hat{\gamma}
\]

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

\[ \hat{g}_{new} = W_{new}\hat{u} \]

\[ G_{EBV} = \sum_{k=1}^{m} (\alpha_i^k + \alpha_j^k) \]

Markers

\[ y = X\beta + Wu + e \]

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

\[
W_{ij} = \frac{(M_{ij} - 2p_i)}{\sqrt{2p_i(1-p_i)}}
\]

\[
\hat{\beta} = \left[ X'X \quad X'W \quad W'X \quad W'W + \frac{\sigma_e^2}{\sigma_u^2} I \right]^{-1} \left[ X'y \quad W'y \right]
\]

\[ \hat{\lambda} = \frac{\sigma_e^2}{\sigma_u^2} \]
IBM’s Design-Centered Strategy to Set Free the Squares
THE EVOLUTION OF DESIGN THINKING

IT'S NO LONGER JUST FOR PRODUCTS. EXECUTIVES ARE USING THIS APPROACH TO DEVISE STRATEGY AND MANAGE CHANGE.

PAGE 55
Case I: How to utilize the germplasm resources stored in gene banks?

Case II: How to design a training set for hybrid performance prediction?

Case III: How to dissect phenotypic plasticity and establish a unified framework “‘to explain’ and ‘to predict’”?
• Introgression, conversion, MAS
• Advanced backcross QTL
• Introgression library
• Association mapping, allele mining, GWAS
Genebank

race classification, site of collection, height uniformity, photoperiod sensitivity, and seed availability
Genomic prediction contributing to a promising global strategy to turbocharge gene banks

Xiaoqing Yu¹, Xianran Li¹, Tingting Guo¹, Chengsong Zhu¹, Yuye Wu², Sharon E. Mitchell³, Kraig L. Roozeboom², Donghai Wang², Ming Li Wang⁴, Gary A. Pederson⁴, Tesfaye T. Tesso², Patrick S. Schnable¹, Rex Bernardo⁵ and Jianming Yu¹* 

The 7.4 million plant accessions in gene banks are largely underutilized due to various resource constraints, but current genomic and analytic technologies are enabling us to mine this natural heritage. Here we report a proof-of-concept study to integrate genomic prediction into a broad germplasm evaluation process. First, a set of 962 biomass sorghum accessions were chosen as a reference set by germplasm curators. With high throughput genotyping-by-sequencing (GBS), we genetically characterized this reference set with 340,496 single nucleotide polymorphisms (SNPs). A set of 299 accessions was selected as the training set to represent the overall diversity of the reference set, and we phenotypically characterized the training set for biomass yield and other related traits. Cross-validation with multiple analytical methods using the data of this training set indicated high prediction accuracy for biomass yield. Empirical experiments with a 200-accession validation set chosen from the reference set confirmed high prediction accuracy. The potential to apply the prediction model to broader genetic contexts was also examined with an independent population. Detailed analyses on prediction reliability provided new insights into strategy optimization. The success of this project illustrates that a global, cost-effective strategy may be designed to assess the vast amount of valuable germplasm archived in 1,750 gene banks.
Turbocharging gene banks through genomic prediction

7.4 million plant accessions are stored in about 1750 genebanks
Reliability & Upper Bound

• Reliability

\[- K_{ut} \left( K_{tt} + I \frac{\sigma_e^2}{\sigma_g^2} \right)^{-1} K'_{ut} \]

– where $K_{ut}$ is the genomic relationship among untested and tested individuals, $K_{tt}$ is the genomic relationship among tested individuals, $\sigma_e^2$ is the residual variance, and $\sigma_g^2$ is the genetic variance.

• Upper Bound for Reliability (U)

\[- U = \hat{v}' \hat{v} / v'v \]

– where $\hat{v} = M' (MM')^{-1} M v$; $M$ is a matrix of training set genotypes; and $v$ is a vector of untested (validation) genotype

Karaman et al., 2016 PLoS One 11, e0161054
Upper bound for reliability (U) and the predicted Yield

Optimal decision by considering the predictive value and the reliability of the prediction?

Would you bet on these?
Case II: Hybrid performance prediction

Design Thinking and Data Mining for Genomic Prediction in Hybrids
Case III: G x E & phenotypic plasticity

 Genome

 Type A
 Type B

 Environment

 Type I
 Type II

 Noisy development

 Type α
 Type β
 Type γ
 Type δ
 Type ε
 Type ζ
 Type η
 Type θ
 Type ι
 Type κ
 Type λ

 Phenotype
The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis

Marcos Malosetti¹ *, Jean-Marcel Ribaut² and Fred A. van Eeuwijk¹

doi:10.1093/jxb/erq095  Advance Access publication 16 April, 2010

REVIEW PAPER

Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops

Graeme L. Hammer¹ *, Erik van Oosterom¹, Greg McLean², Scott C. Chapman³, Ian Broad², Peter Harland⁴ and Russell C. Muchow⁴

¹ The University of Queensland
² Agri-Science Queensland, Lavarack
³ CSIRO Plant Industry, Queensland
⁴ Formerly CSIRO Sustainable Yields
* To whom correspondence should be addressed

CSIRO PUBLISHING
http://dx.doi.org/10.1071/CP14007

Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction

Mark Cooper⁴, Carlos D. Messina⁵, Dean Podlich⁶, L. Radu Totir⁵, Andrew Baumgarten⁶, Neil J. Hausmann⁶, Deanne Wright⁶, and Geoffrey Graham⁷

⁴DuPont-Pioneer, 7250 NW 62nd Avenue, PO Box 552, Johnston, IA 50131, USA.
⁵DuPont-Pioneer, 8305 NW 62nd Avenue, PO Box 7060, Johnston, IA 50131, USA.
⁶DuPont-Pioneer, 7300 NW 62nd Avenue, PO Box 1004, Johnston, IA 50131, USA.
⁷Corresponding author. Email: mark.cooper@pioneer.com
Gene Regulatory Network of Flowering Time

Murphy et al. PNAS 2011;108:16469-16474

Yang et al. PLoS ONE 2014; 9:e105352
Patterns Underlying the Apparently Complex G x E

Can we replace this x-axis?
General Context of Predictive Phenomics
Pattern Detection & Model Fitting -> Relationship Building-> Predicting & Forecasting

<table>
<thead>
<tr>
<th>Tested Genotypes</th>
<th>1</th>
<th>Untested Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tested Environments</td>
<td>2</td>
<td>Untested Environments</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Environmental Conditions

A. Day length

B. Temperature

Day length (hours)

Month

Cumulative growing degree days (GDD)

Days after planting (DAP)
Clustering Environments

Temperature Performance of RILs
Photothermal Time (PTT) = GDD x DL
Environmental Index Defined by Photothermal Time

A

B

C

D

End of window

Population mean

Flowering time

Photothermal time

Beginning of window

Photothermal time
CERIS
Critical Environmental Regressor through Informed Search

1 - 4 days after planting

Population mean

Environmental index (day 1 - 4)

$\rho = 0.64$

End of window

Beginning of window
CERIS
Critical Environmental Regressor through Informed Search

1 - 4 days after planting

100 200 300 500
Environmental index (day 1 - 4)

1600 2400
Population mean

$ r = 0.64 $

End of window

Beginning of window
Joint Genomic Regression Analysis (JGRA) through Reaction-Norm Parameter Estimation
Joint Genomic Regression Analysis (JGRA) through Genome-Wide Marker Effect Continuum
Empirical Validation of Performance Prediction

(A) Predicted flowering time vs. Photothermal time. The red line represents P898012, and the blue line represents Tx430.

(B) Scatter plot of IA15 observed flowering time vs. predicted flowering time, with a correlation coefficient of $r = 0.80$.

(C) Scatter plot of IA16 observed flowering time vs. predicted flowering time, with a correlation coefficient of $r = 0.88$. 
Genetic Determinants underlying G x E

prr37

ghd7-1

ghd7-2
Genetic Mapping of Reaction Norm Parameters
Genotype-by-Environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environments

David L. Des Marais,1,* Kyle M. Hernandez,1,* and Thomas E. Juenger1,2

1Department of Integrative Biology, University of Texas at Austin, Austin, Texas 78712; email: tjuenger@austin.utexas.edu
2Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712

AP: Antagonistic Pleiotropy
CN: Conditional Neutrality
DS: Differential Sensitivity
• Differential sensitivity, conditional neutrality, antagonistic pleiotropy, and no genetic effect x environment interaction are emergent properties of the plant perceptron interacting with diverse environmental and developmental cues
R. A. Fisher, Lancelot Hogben, and the Origin(s) of Genotype–Environment Interaction

Lancelot Hogben and the "Interdependence of Nature and Nurture"

Figure 2. R. A. Fisher. Fisher papers, Barr Smith Library, University of Adelaide Library, MSS 0013/Series 25. Reproduced with the permission of the University of Adelaide Library.

Figure 3. Lancelot Thomas Hogben, Hogben papers, special collections, University of Birmingham Library. Reproduced with the permission of the University of Birmingham Library.
Genomic and environmental determinants and their interplay underlying phenotypic plasticity

Xin Li^a, Tingting Guo^a, Qi Mu^a, Xianran Li^a, and Jianming Yu^a

Department of Agronomy, Iowa State University, Ames, IA 50011

PNAS 2018, 115:6679-6684

Observed phenotypic variation in living organisms is shaped by genomes, environment, and their interactions. Flowering time under natural conditions can showcase the diverse outcome of the gene-environment interplay. However, identifying hidden patterns and specific factors underlying phenotypic plasticity under natural field conditions remains challenging. With a genetic population showing dynamic changes in flowering time, here we show that the integrated analyses of genomic responses to diverse environments is powerful to reveal the underlying genetic architecture. Specifically, the effect continuum of individual genes (Ma, MaC, FT, and ELF3) was found to vary in size and in direction along an environmental gradient that was quantified by photothermal time, a combination of two environmental factors (photoperiod and temperature). Gene-gene interaction...
• Improved understanding of gene × environment interaction could lead to better-informed decisions in personalized medicine and optimized breeding.

• A comprehensive framework for gene × environment interaction and gene × gene interaction can be established (even) if different sets of individuals are assessed at different environments.

• For plants, with a robust G × E modeling framework, observations at winter nurseries may be leveraged to enhance selection gain per unit time.

• Imbalanced yield trials across environments can be projected onto a series of representative environmental panels for selection, and optimized testing strategies for genotypes and environments can be designed.

• Introducing the environmental dimension to GWAS and GS allows us to investigate the Infinitesimal Model or "Omnigenic Model".
Performance Forecasting: CERIS-JAGR SS-IS-OT-WG-SS
"Site Specific"

• Rather than a generic prediction for a mega-environment, specific prediction values can be made for genotypes at a location where the detailed site information and historical weather records are available. How to combine site-specific information for untested sites with performance trials data already obtained at tested sites is the 1st critical link.
"In Season"

- Beyond the historical weather records, once the growing season starts, forecasting should be updated for the specific year and planting date. Moreover, as soon as the weather data come in for the initial growth stages, performance prediction can be further updated, but it certainly should be ahead of the actual observation of the phenotype.
"On Target"

- Rather than just the relative ranks of different hybrids, the actual performance values should be predicted. The key to be "On-Target" is to incorporate the environmental information so that both the $E$ effect and the $G \times E$ effect are all included in the performance prediction.
"Whole Genome"

• Rather than relying on a few major effect genes, the best performance prediction is achieved by mining the genome-wide fingerprinting data. Genomics is the 2nd critical link to make performance prediction of untested individuals possible, and to be "Site-Specific", "In-Season, and "On-Target".
“Simple Straightforward"

• Integration of genomics and crop models is a long-term goal. However, it is highly desirable to design an initial open system that contains no black boxes with differential equations to exploit the essence of genomics and crop models.
Design Thinking and Data Mining in Predictive Phenomics
## Acknowledgements

- Xianran Li (ISU)
- Tingting Guo (ISU)
- Xin Li (ISU)
- Xiaoqing Yu (ISU)
- Matt Dzievit (ISU)
- Jinyu Wang (ISU)
- Qi Mu (ISU)
- James McNellie (ISU)
- Greg Schoenbaum(ISU)
- Cathy Zhang (ISU)
- Haozhe Zhang (ISU)
- Adam Vanous (ISU)
- Paul White (ISU)
- Jode Edwards (USDA-ARS)
- Sotiris Archontoulis (ISU)
- Fernando E. Miguez (ISU)
- Kevin Price (AgPixel)

- Patrick Schnable (ISU)
- Junping Chen (USDA-ARS)
- Rex Bernardo (U of Minn.)
- MingLi Wang (USDA-ARS)
- Gary Pederson (USDA-ARS)
- Tesfaye Tesso (KSU)
- Kraig Roozeboom (KSU)
- Donghai Wang (KSU)
- Guihua Bai (USDA-ARS)
- Kerry Mayfield (Chromatin)
- Larry Lambright (Chromatin)
- Scott Staggenborg (Chromatin)
- Sharon Mitchell (Cornell)

- Mike Scanlon (Cornell)
- Patrick Schnable (ISU)
- Gary Muehlbauer (U of Minn.)
- Marja Timmermans (CSHL)
- Dean DellaPenna (MSU)
- Mike Gore (Cornell)
- Robin Buell (MSU)
- John McKay (CSU)
- Candy Gardner (USDA-ARS)
- Dave Peters (USDA-ARS)
- Elliot Heffner (Pioneer)
- Tabare Abadie (Pioneer)