

Genomic Selection for Public Cotton Breeding

Amanda M Hulse-Kemp



Agricultural Research Service
U.S. DEPARTMENT OF AGRICULTURE



"Cotton Produced by Genomic Selection"
Generative AI

Meet the Hulse-Kemp Lab



Dr. Emily Delorean
Postdoc
Chili pepper Genomics



Jonathan Zirkel
PhD Student
Functional Genetics



Laide Rasaki
PhD Student
Advancement of Orphan
Crops



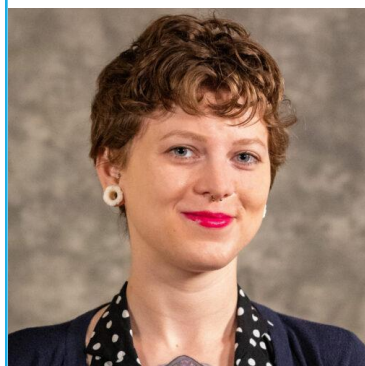
Matthew Willman
PhD Student
Wheat Genomics



Grant Billings
PhD Student
Cotton Genomics



Dr. Keo Corak
Associated Scientist
Breeding Informatics



Dr. Ash Yow
Postdoc
Hybrid Genomics



Jordan James
PhD Student
UT-Arlington



Ameaza Rodrigues
Intern



Romil Shah
Intern



Chandler Wilson
Intern



Dr. Heather Manching
Research Associate
Breeding Informatics

Software Development Team



Agricultural Research Service

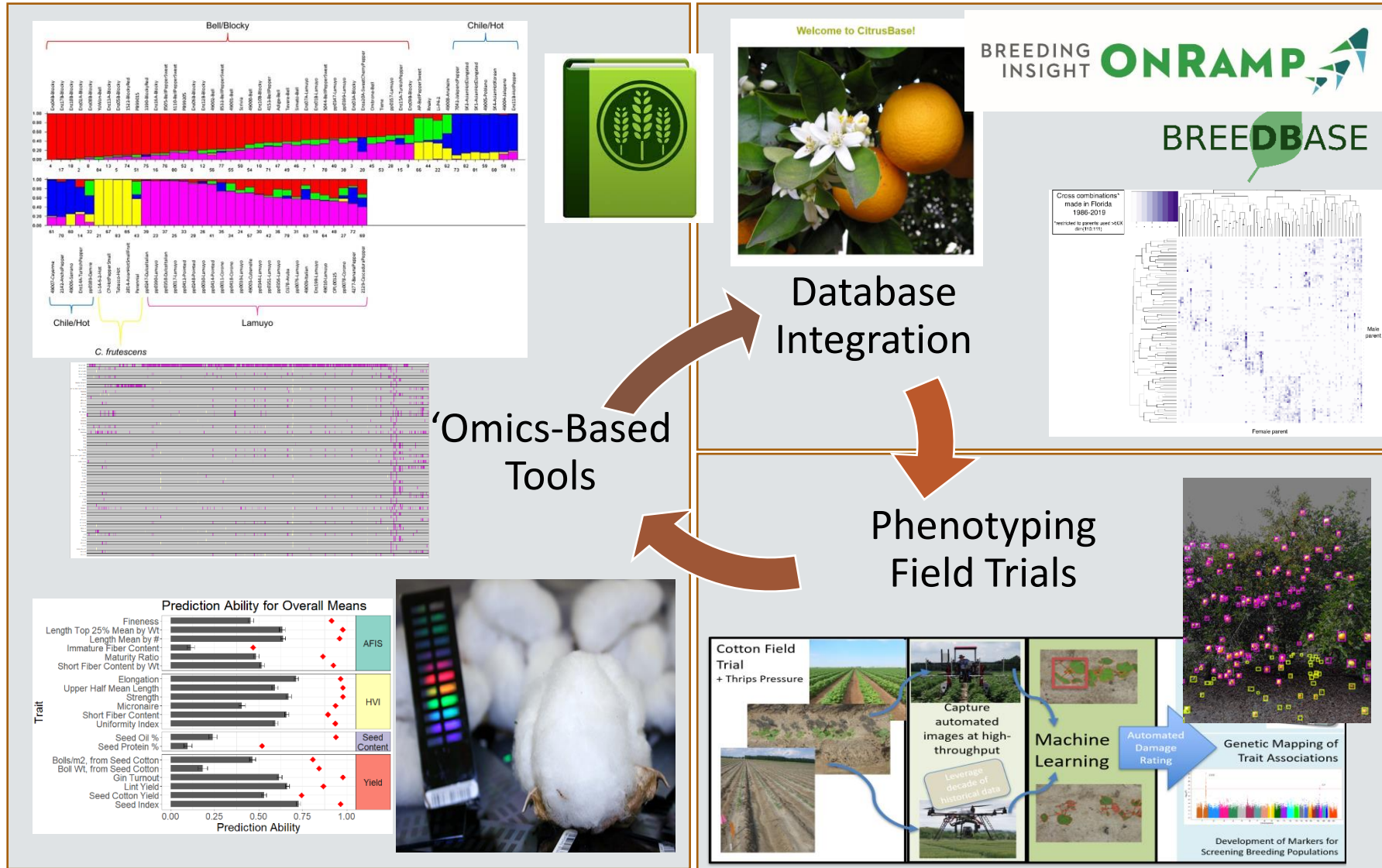
U.S. DEPARTMENT OF AGRICULTURE

NC STATE
UNIVERSITY

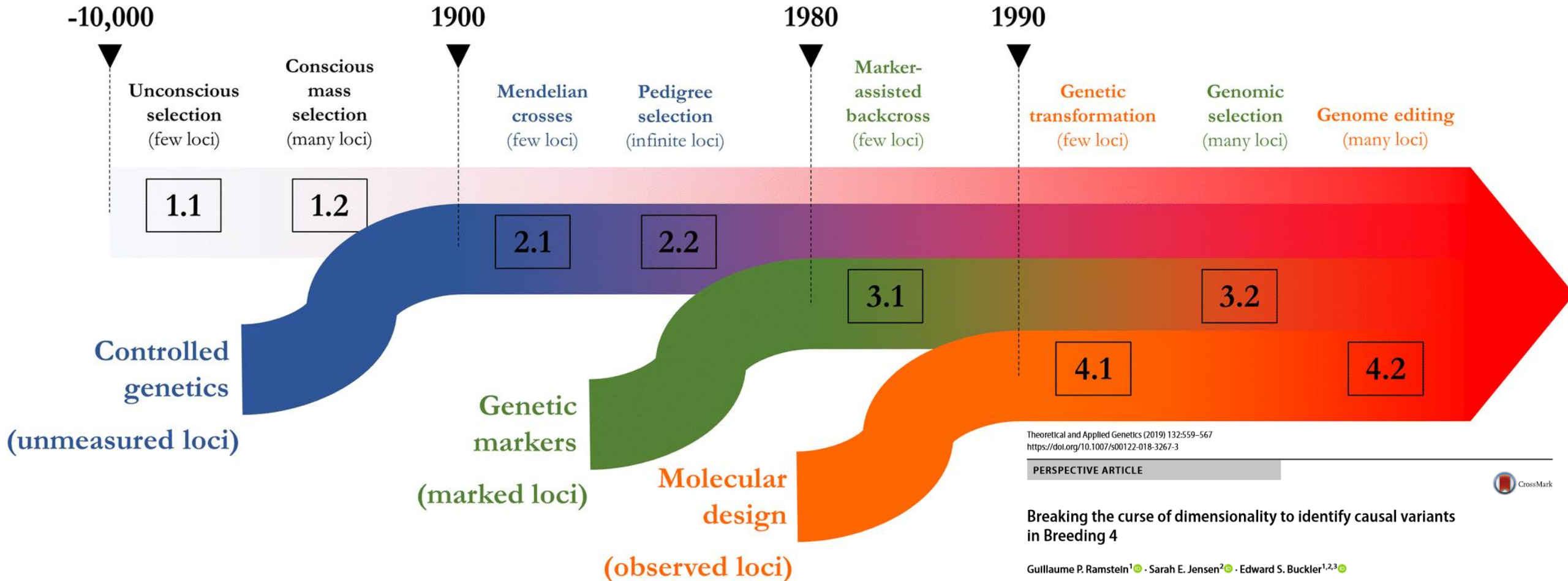
Cotton Team

As of July 2024

Developing Tools Across Crops for Breeders



TECHNOLOGICAL PHASES OF PLANT BREEDING



Theoretical and Applied Genetics (2019) 132:559–567
<https://doi.org/10.1007/s00122-018-3267-3>

PERSPECTIVE ARTICLE



Breaking the curse of dimensionality to identify causal variants in Breeding 4

Gillaume P. Ramstein¹ · Sarah E. Jensen² · Edward S. Buckler^{1,2,3}

Received: 15 November 2018 / Accepted: 7 December 2018 / Published online: 13 December 2018
 © The Author(s) 2018

MATH

Genetic Characteristics



Number of
Genes

Relationships between
Genes/Markers/Alleles

Genotyping Precision

Phenotyping Accuracy

Heritability



Relatedness

Population
Size

Population Characteristics

**Prediction
Model**



GENOMIC SELECTION IN BREEDING PROGRAMS

Genome Wide Association Model:

$$Y = \mu + X_S \beta_S + Zu + \varepsilon$$

Interest is in SNP effect $\hat{\beta}_S$

Genomic Prediction Model:

$$Y = \mu + Zu + \varepsilon$$

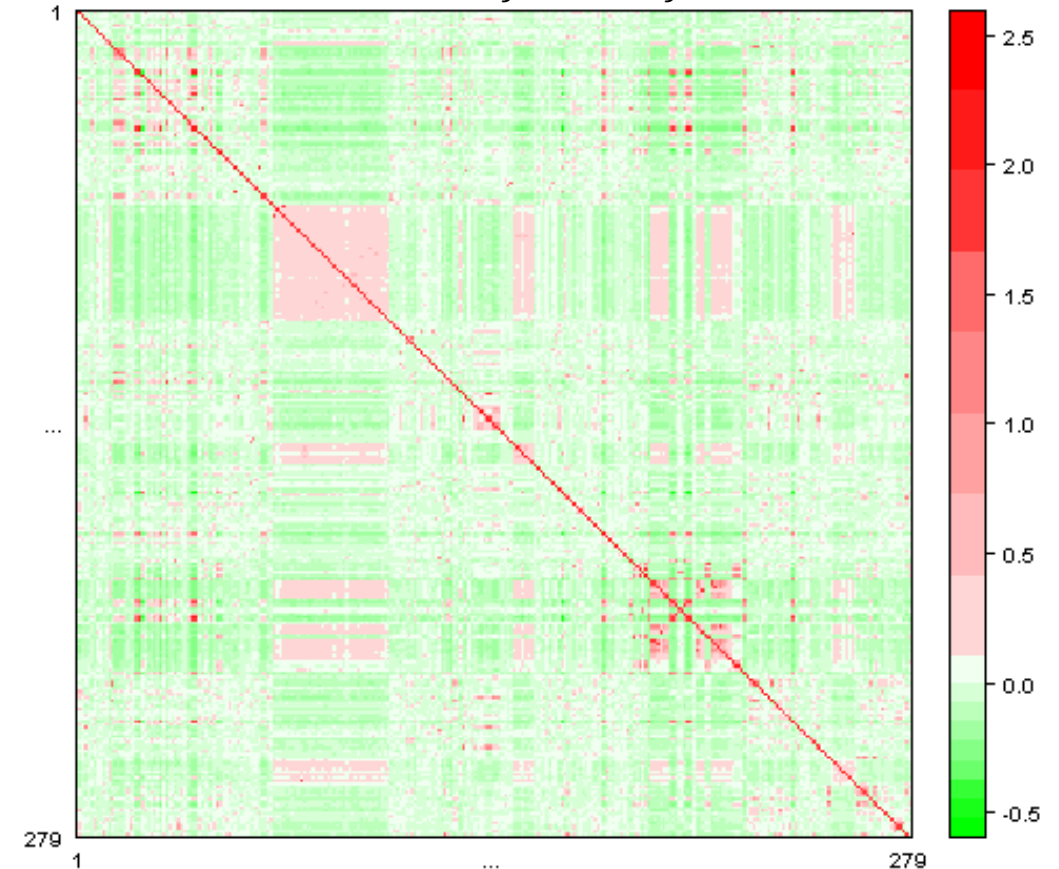
Interest is in prediction of u : $\hat{u}$

Similar models, different
objectives

Best approach for breeding
depends on genetic architecture

$$u \sim N(0, G\sigma_A^2)$$

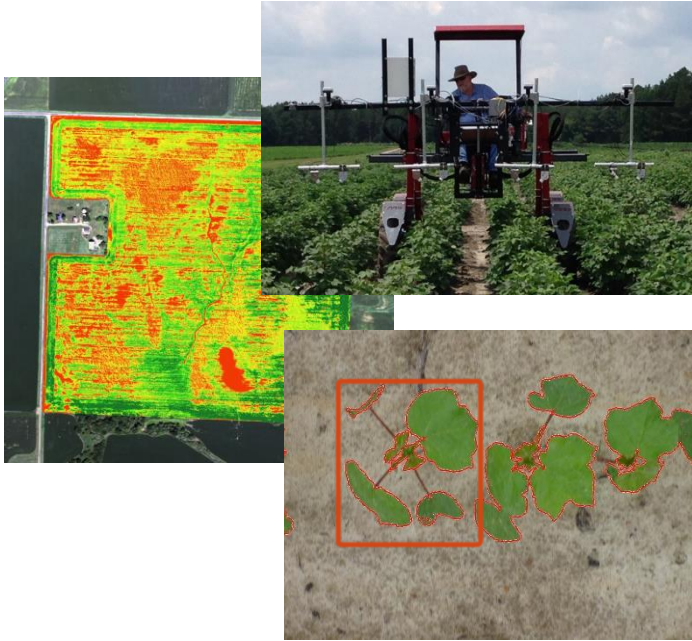
$$G_{ij} = 2\theta_{ij}$$



Pairwise realized genomic
relationship matrix for 279 maize
inbreds



INPUTS INTO GENOMIC SELECTION



Bioinformatics
for High-
throughput
Phenotyping

R Statistical
Software

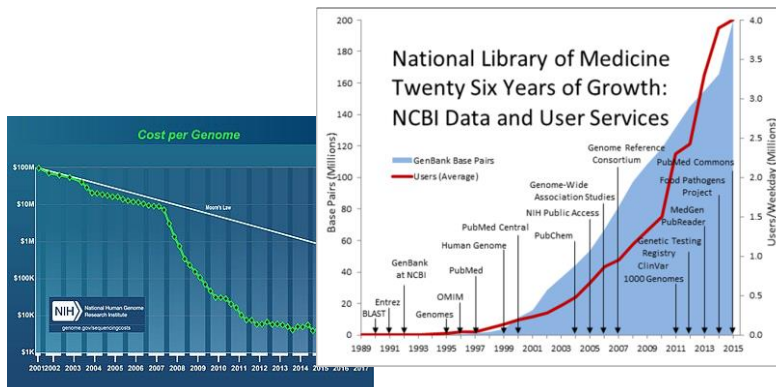


Genomic Estimated
Breeding Values
(GEBV) for
Selection

Bioinformatics
to Obtain
Inputs

Tassel Software

Many custom
pipelines



Predictive Breeding



Predictive Breeding

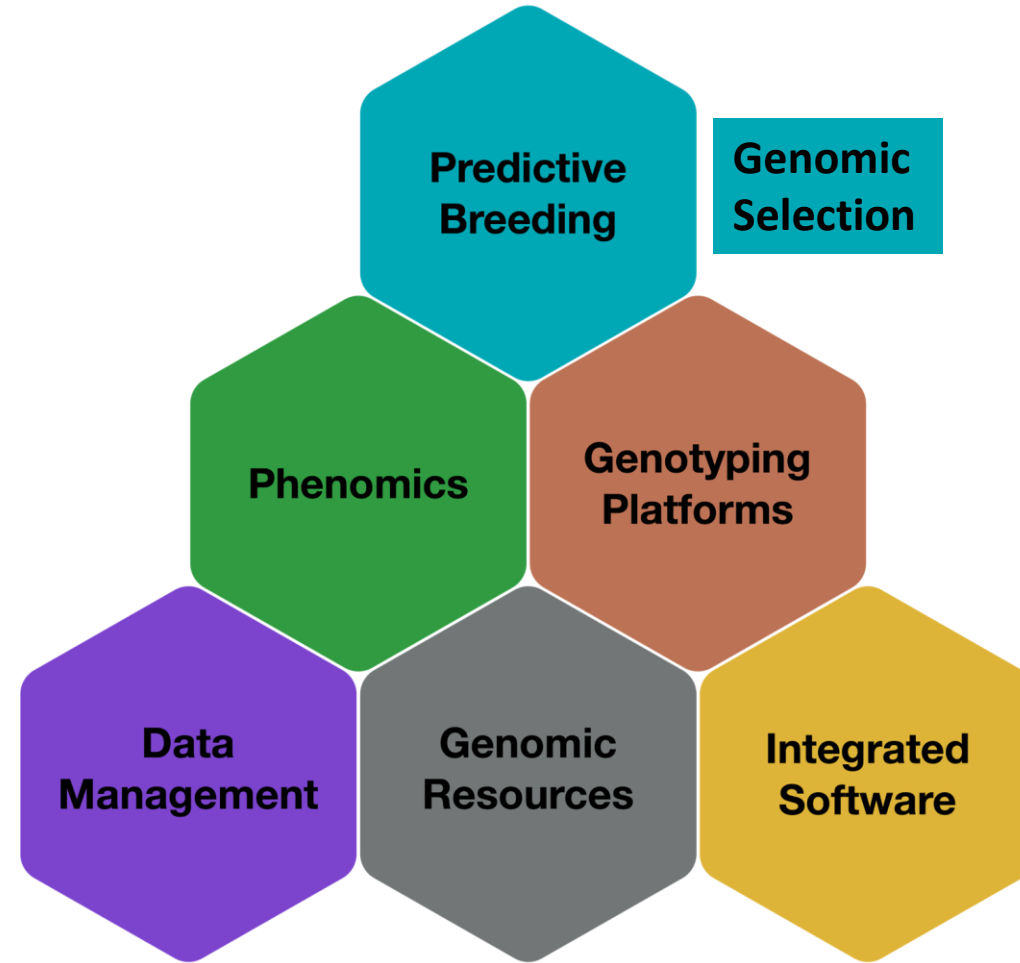
Single experiment - Estimating ability to estimate traits **WITHIN** the same experiment (same set of materials) = **GENOMIC PREDICTION**

- Usually gives an idea of theoretical maximums of many situations you may face - simplest path forward

Estimating traits across experiments - ie. in a breeding program (related, but **NOT** the same set of materials); THEN using those estimates to select what individuals to retain = **GENOMIC SELECTION**

- May be empirically similar or potentially very different in practice than Genomic Prediction
- Difficult to measure success until put in practice

Plant Breeding - Advanced Technologies



Advanced Breeding Technologies = Methods to Manipulate Breeders Equation



- Reduce generation intervals
- Improve genetic gain
- Standardized genotyping
- Advanced standardized phenotyping methods
- Data management systems to integrate multiple field trials

$$\text{Response } \mathbf{R_t} = h^2 \mathbf{S} = \frac{i \, r \, \sigma_A}{L}$$

h^2 = Narrow sense heritability

S = Difference between selected parents and population

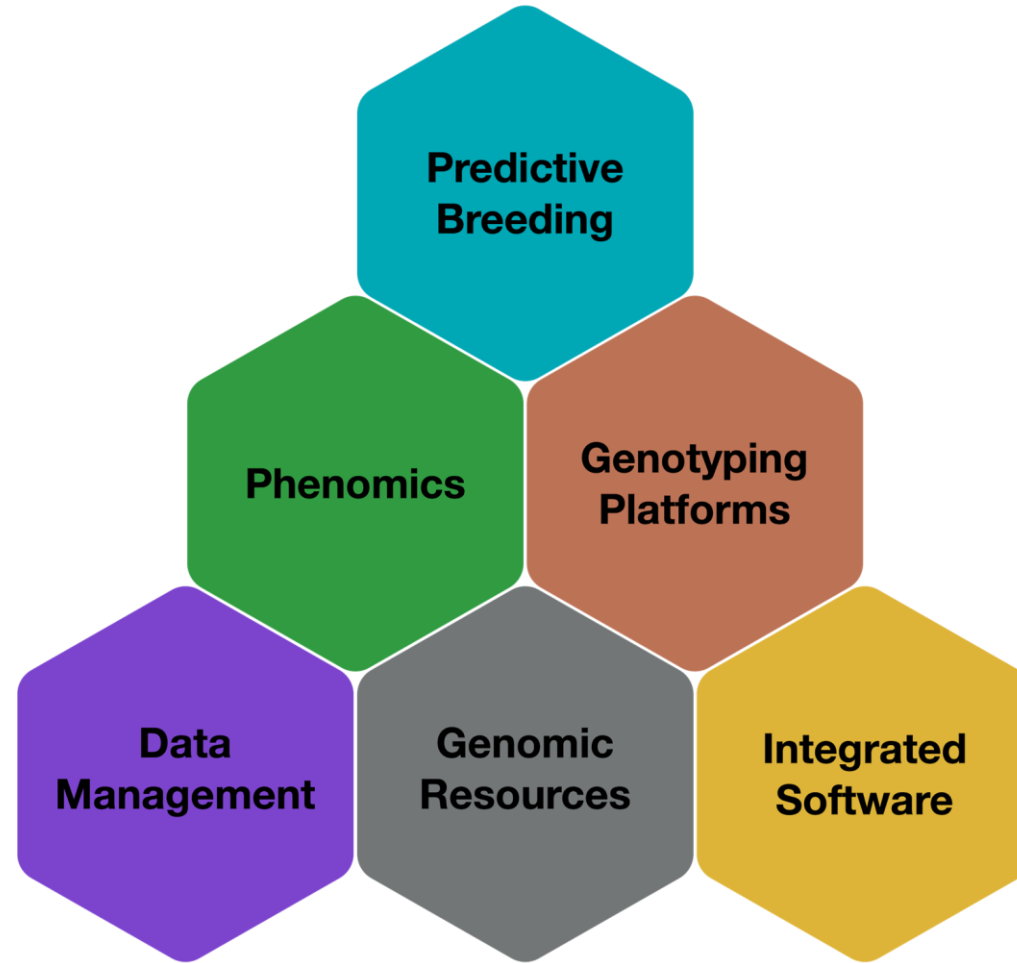
i = Selection intensity

r = Selection accuracy

σ_A = Genetic Variance

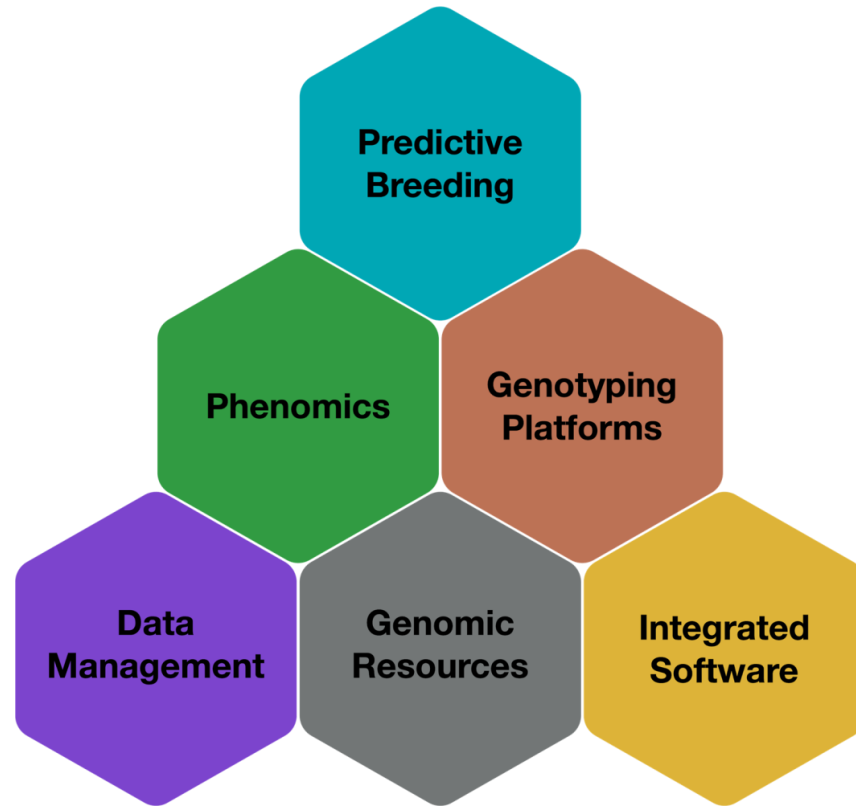
L = Generational interval

Plant Breeding - Advanced Technologies



All of these aspects can play into the various components of the breeders equation!

Plant Breeding - Advanced Technologies



Response $R_t = h^2 S = \frac{i r \sigma_A}{L}$

h^2 = Narrow sense heritability

S = Difference between selected parents and population

i = Selection intensity

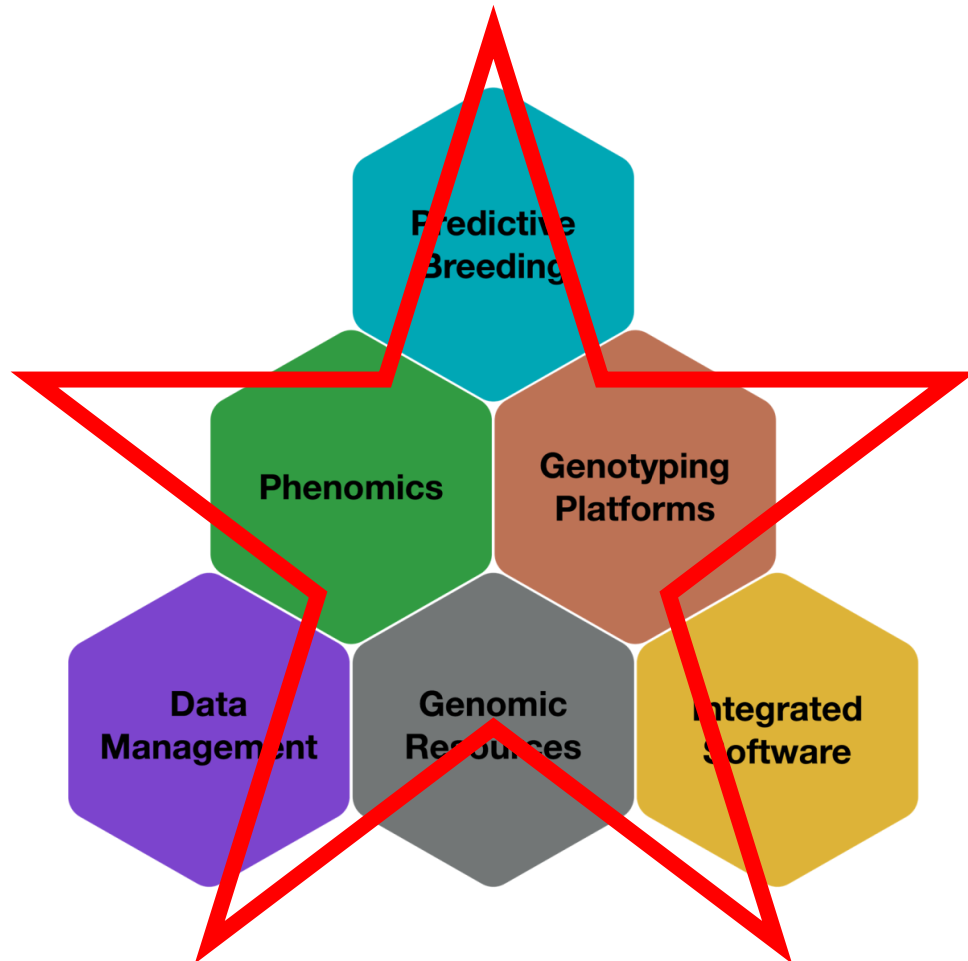
r = Selection accuracy

σ_A = Genetic Variance

L = Generational interval

Biological aspect can measure - changes depending on how you measure traits of interest

Plant Breeding - Advanced Technologies



$$\text{Response } R_t = h^2 S = \frac{i r \sigma_A}{L}$$

h^2 = Narrow sense heritability

S = Difference between selected parents and population

i = Selection intensity

r = Selection accuracy

σ_A = Genetic Variance

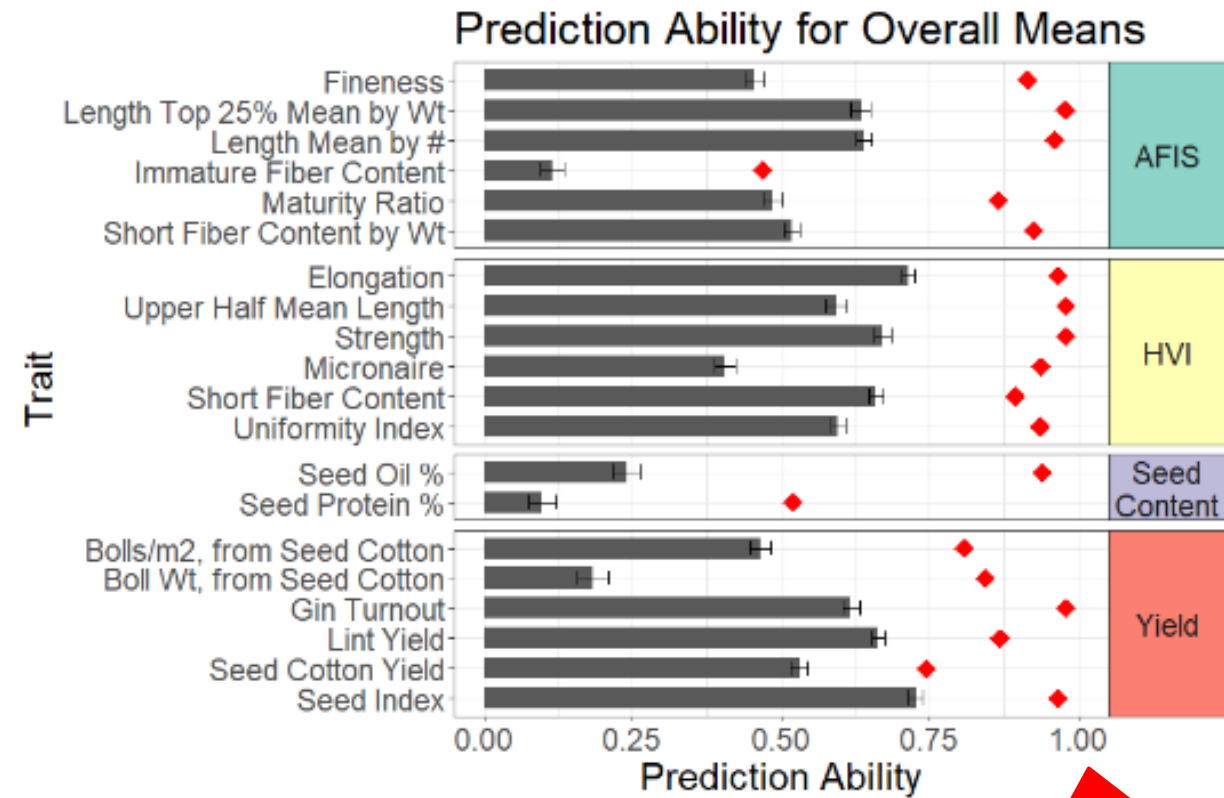
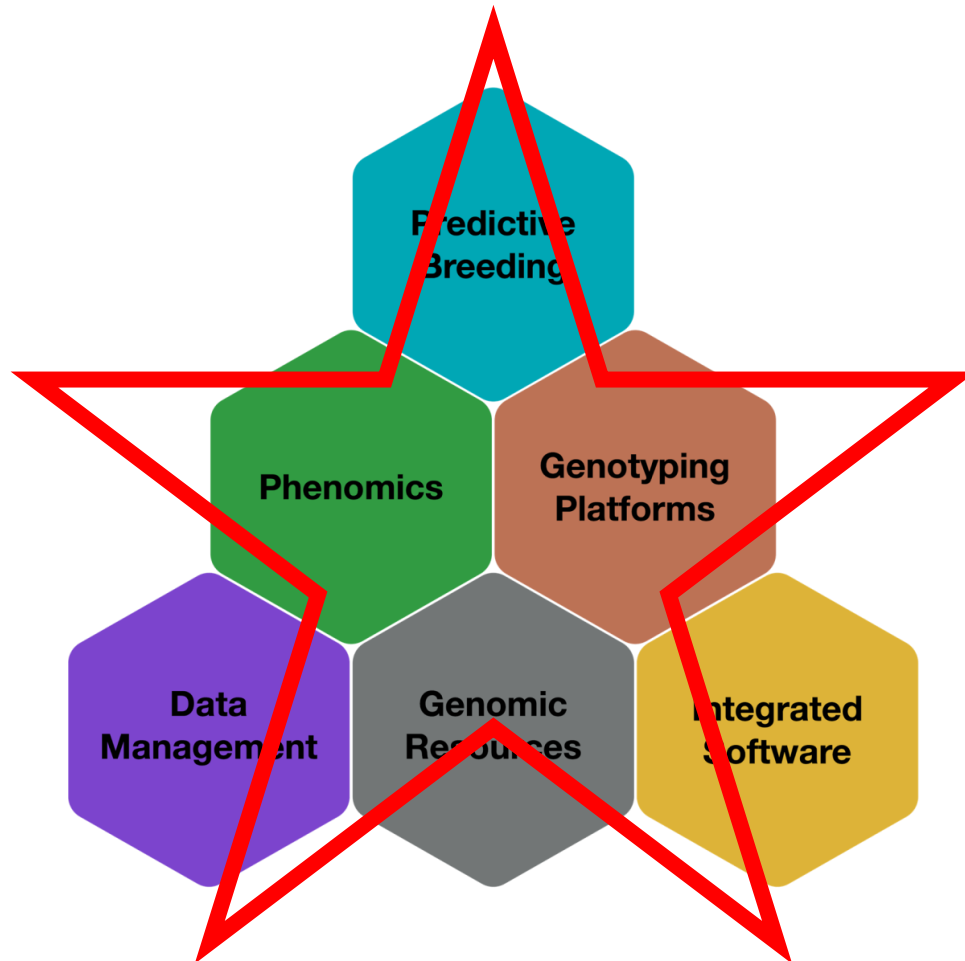
L = Generational interval

ACCURACY!!!!

Important in ALL aspects of the system

Affects our ability to reach theoretical maximums

Prediction Accuracies



$$\text{Response } R_t = h^2 S = \frac{i r \sigma_A}{L}$$

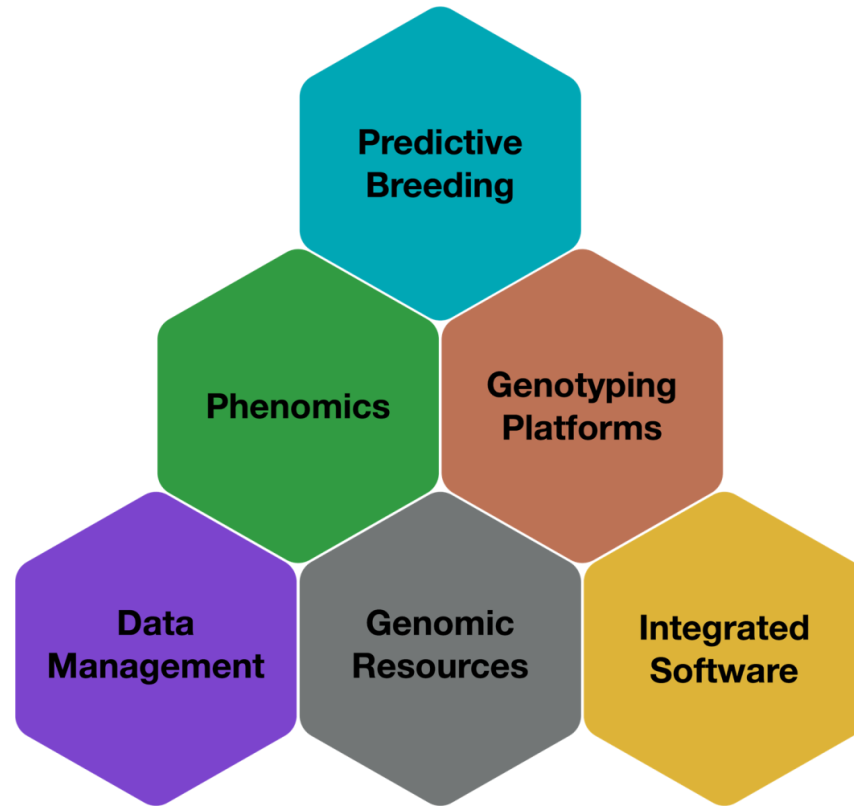
h^2 = Narrow sense heritability

Theoretical maximum on prediction = h^2

Changes per trait!

Difference can be many things but impacted by models developed and accuracy on measurements
- genotype & phenotype

Plant Breeding - Advanced Technologies



Response $R_t = h^2 S = \frac{i r \sigma_A}{L}$

h^2 = Narrow sense heritability

S = Difference between selected parents and population

i = Selection intensity

r = Selection accuracy

σ_A = Genetic Variance

L = Generational interval

Quickest way to increase gain

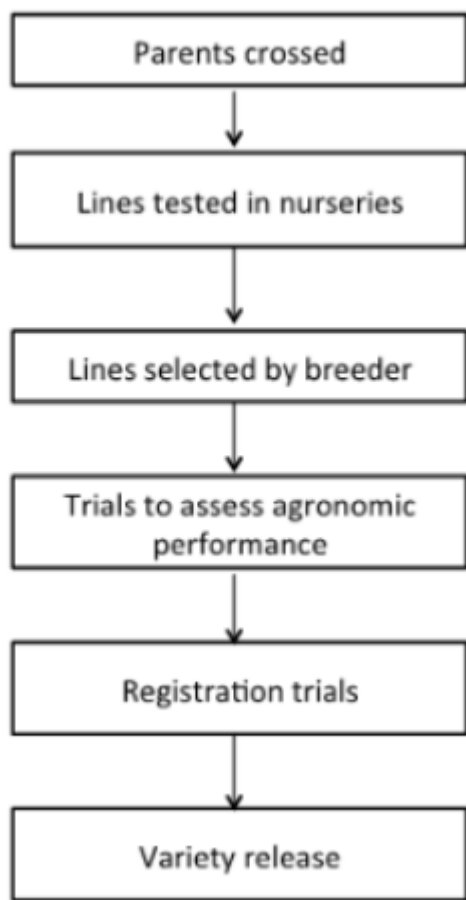
Biggest effect on whole equation -> spend a lot of time here

Implementation of Genomic Selection

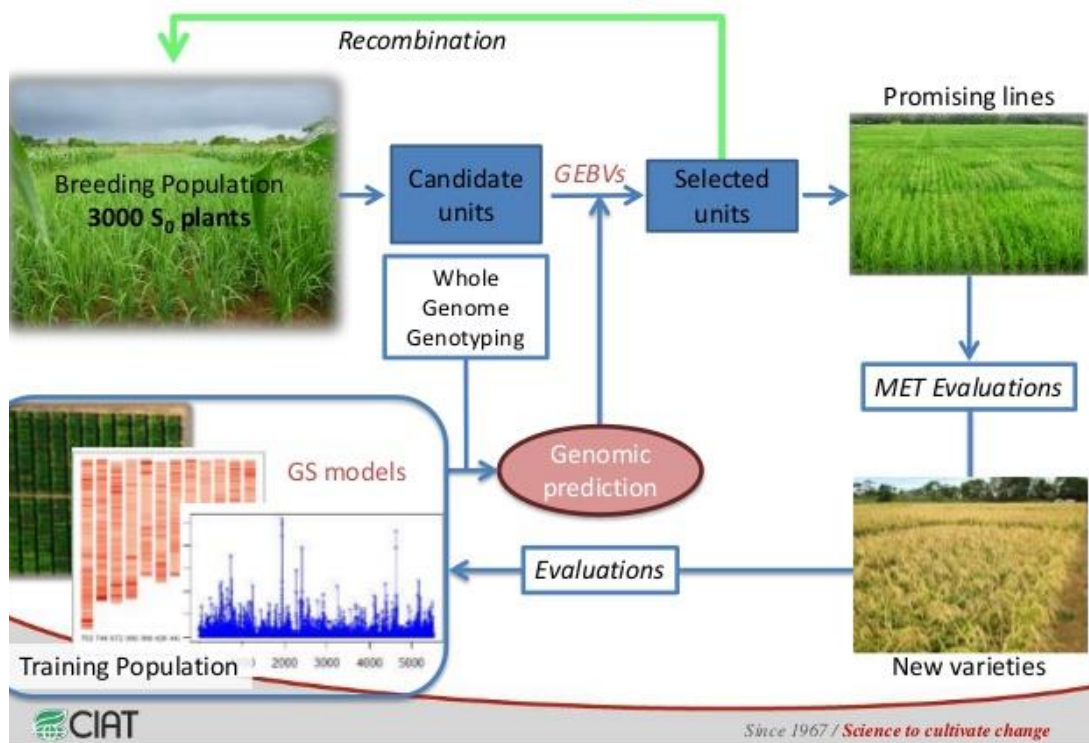


Outline of breeding programs

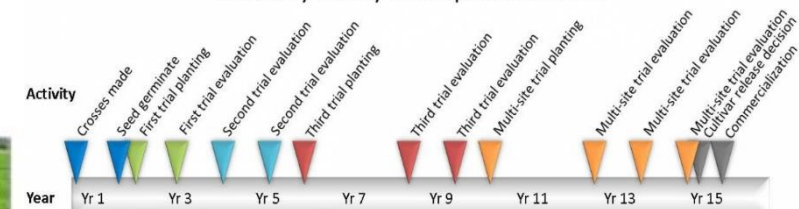
Traditional Breeding Program



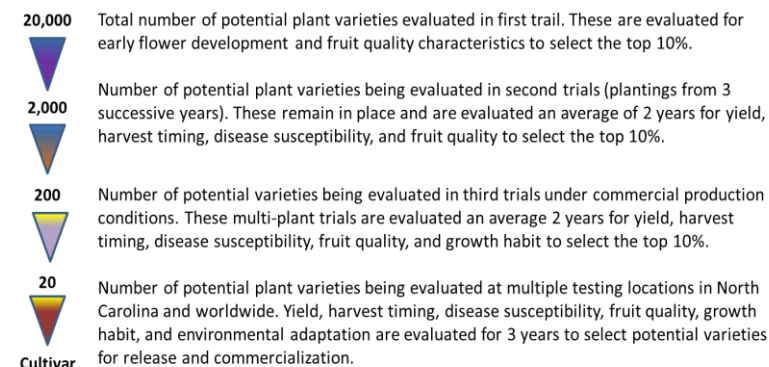
The RGS breeding scheme



Blueberry Variety Development Timeline



Evaluations/Selections Made Each Year



Predictive Breeding



Estimating traits across experiments - ie. in a breeding program (related, but NOT the same set of materials); THEN using those estimates to select what individuals to retain = **GENOMIC SELECTION**

- May be empirically similar or potentially very different in practice than Genomic Prediction
- Difficult to measure success until put in practice

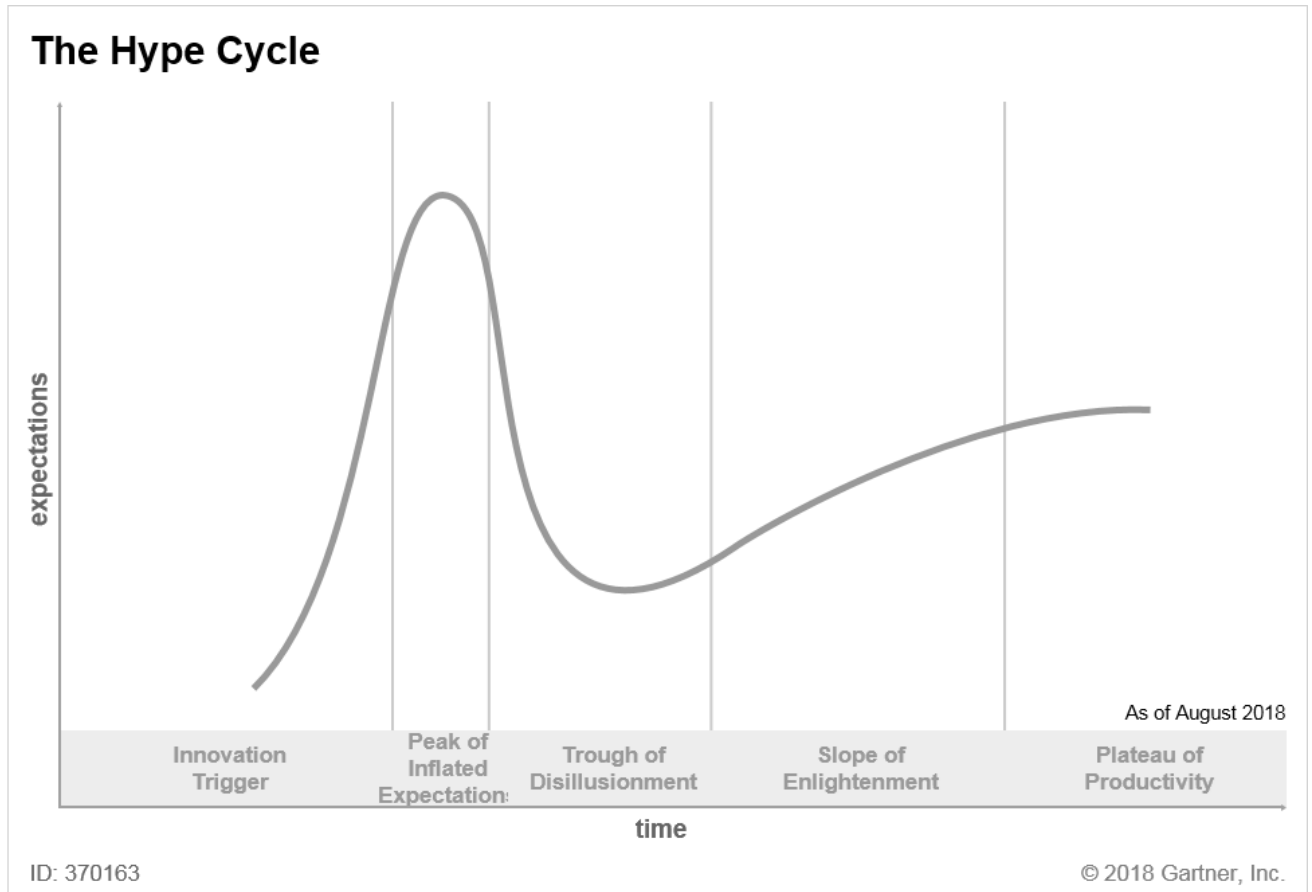
Population Improvement
Select: Parent Value



Product Development
Select: Variety Value

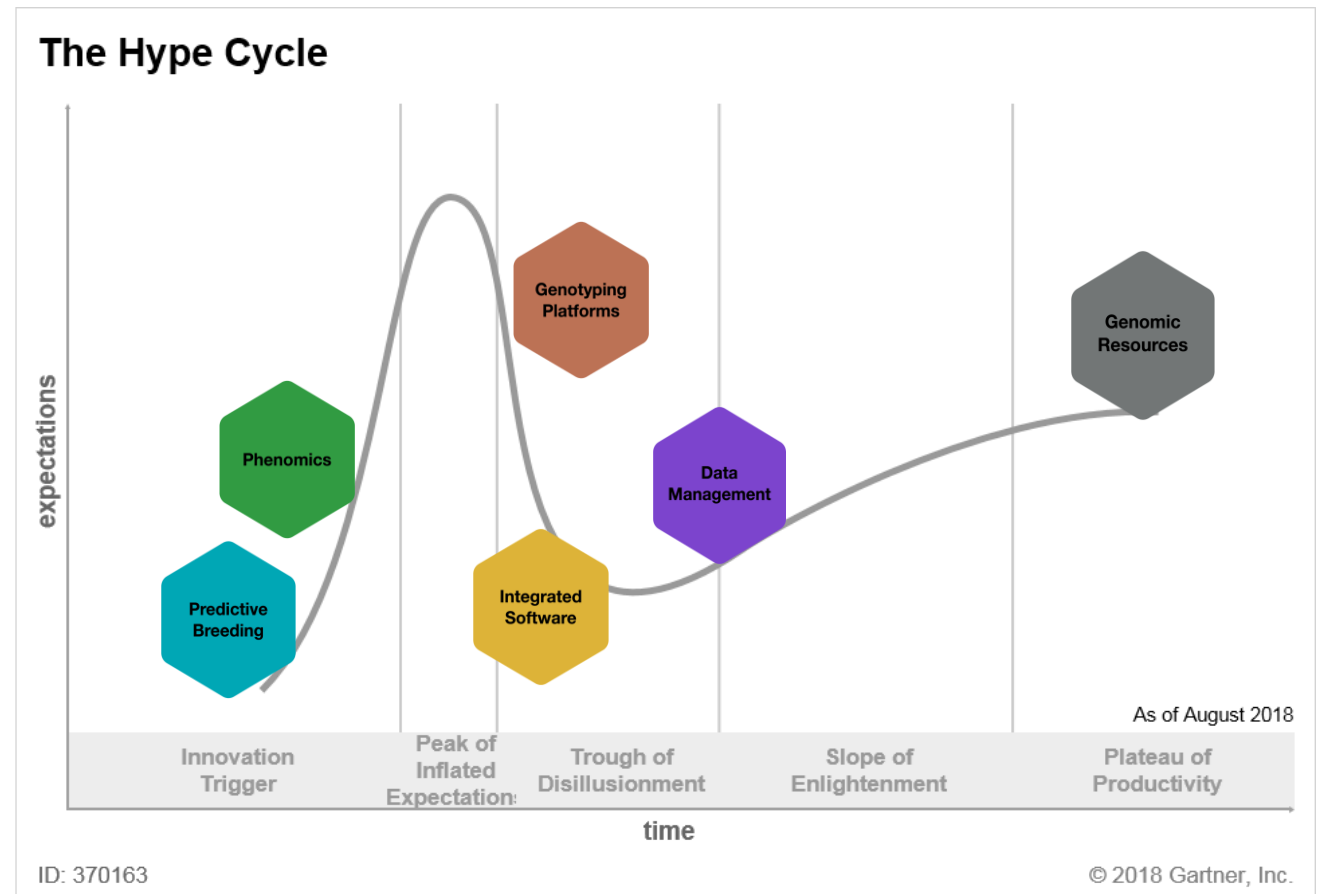
Technology Researchers - Generalists Evaluating and Developing Any Type of Technology

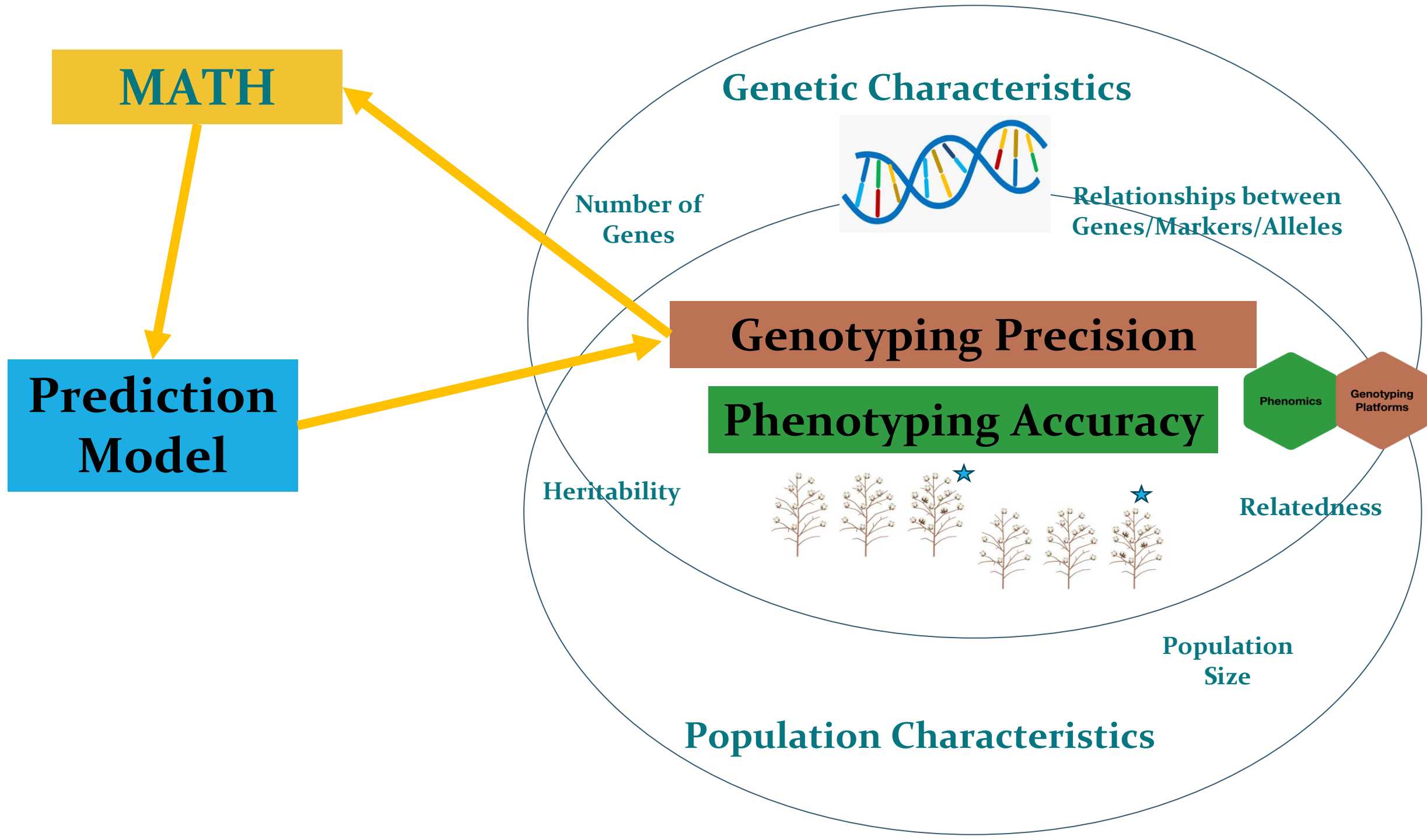
- Evaluate emerging technology
- Research to investigate integration of tools into breeding programs
- Develop new tools to fit holes where nothing is available



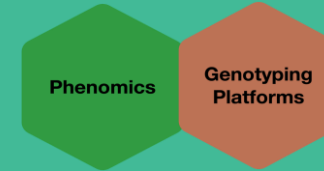
Technology Researchers - Generalists Evaluating and Developing Any Type of Technology

- Evaluate emerging technology
- Research to investigate integration of tools into breeding programs
- Develop new tools to fit holes where nothing is available



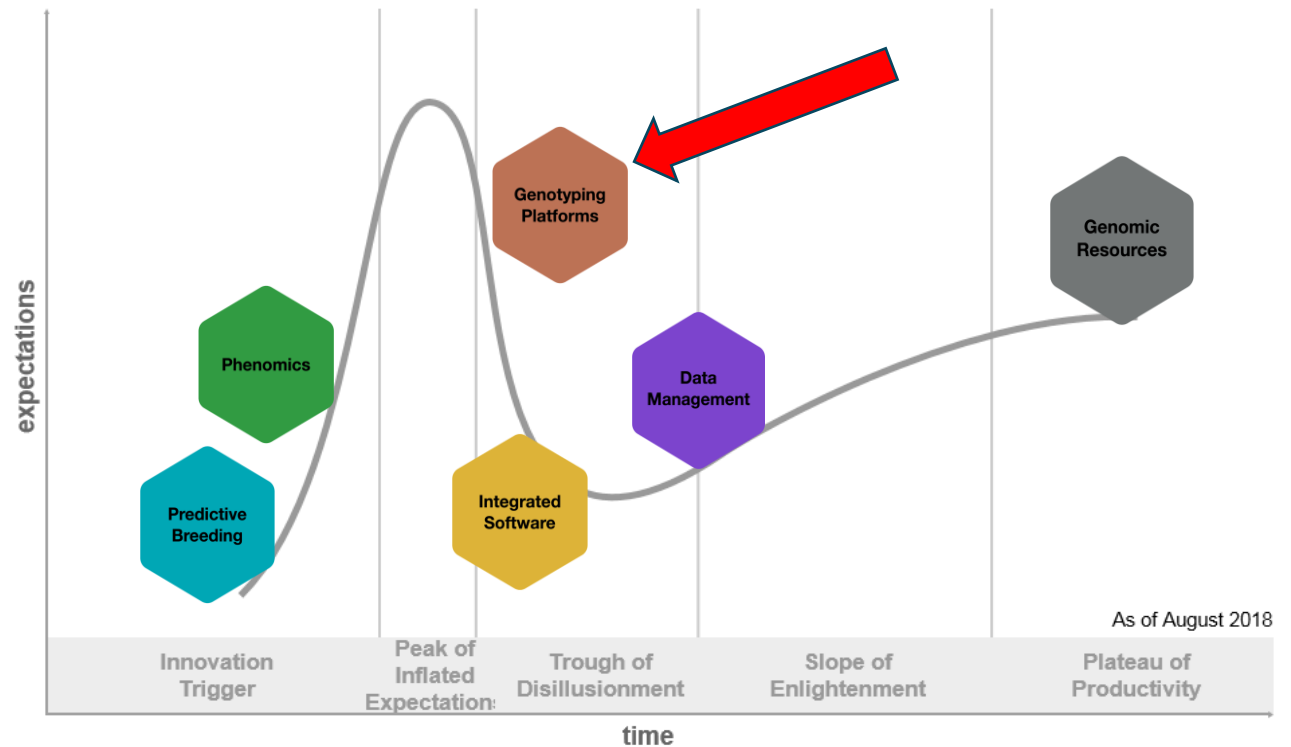


Predictive Breeding =



- Evaluate emerging technology
- Research to investigate integration of tools into breeding programs
- Develop new tools to fit holes where nothing is available

The Hype Cycle

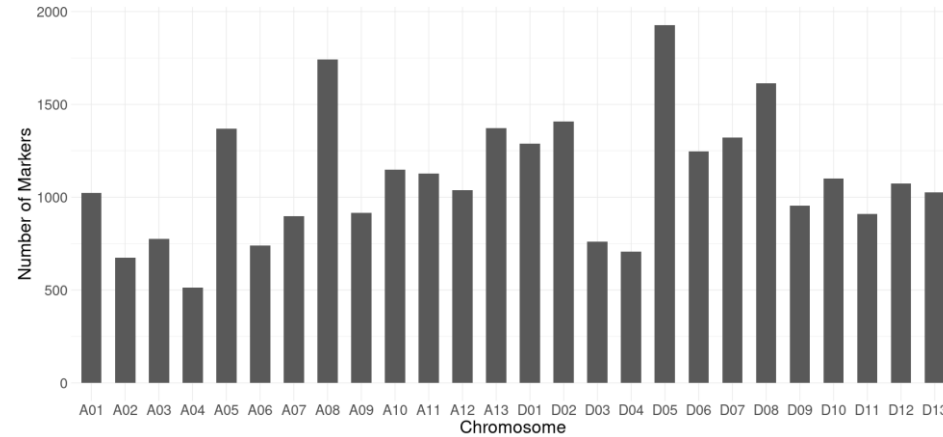


Various Options

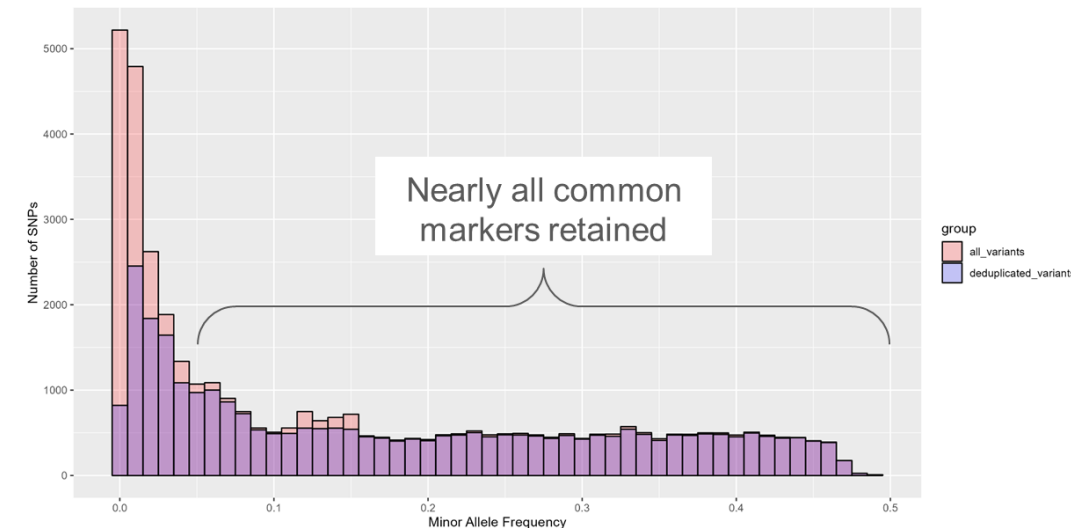
- ~~CottonSNP63K (Available 2015-2023)~~
- CottonSNP30K
 - Public – should be available by end of 2024
- 18K XT Infinium Cotton Array
 - Available now through SGS

CottonSNP30K

- \$32/sample + processing
- Includes all markers on SGS 18K
- Like the old array, add-on content is available for private use
- Currently producing cluster file
- ~100 markers for agronomic traits provided by breeders and geneticists



CottonSNP30K



Trait-Specific Markers - Pending Validation



CottonSNP30K

Pulled from Publications

- A10 Cotton blue disease (Fang et al, 2010a - 1 [A/T] type)
- D01 Okra leaf locus (Andres et al, 2016)
- D02 Bacterial blight (Fang et al, 2010b - 4 [A/T])
- D03 FOV4 resistance gene (Liu et al, 2021)
- D06 markers for photoperiodicity (Gowda et al, 2023)

Contributed Confidentially Pre-Publication by Various Collaborators

- 95 markers for various agronomic/fiber quality traits (Billings/Hulse-Kemp et al, unpub. - [A/T] and [C/G] type)
- 13 other markers retained from 63K due to suggestive mapping results (Chee et al, unpub.; Kuraparthi et al, unpub.)
- Other markers (Thyssen/Fang et al.) for:
 - Reniform Nematode
 - STR/UI/SFI
 - Root Knot Nematode
 - Immature Fiber Content, Elongation, Micronaire, Upper Half Mean Length, Fiber Strength

Public Resources along with CottonSNP30K



CottonSNP30K

- Genetic map positions for each marker (from publications on old array)
- Physical TETRAPLOID genome positions on Coker 312
- Cluster file to enable automated processing like the old array
- “Immortal set” of 24 lines - data public
 - Enabling future development of new technologies and technology integration
- R/Shiny App for streamlined analysis & deposition of data to CottonGen
 - <https://gbru-ars.shinyapps.io/iCottonQTL/>



Article

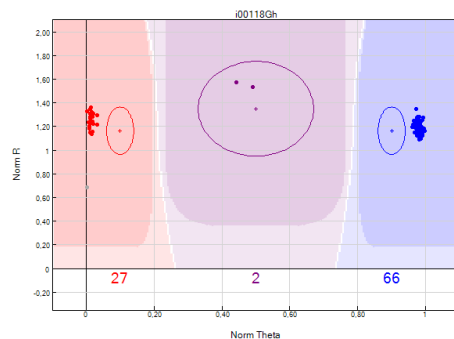
Detecting Cotton Leaf Curl Virus Resistance Quantitative Trait Loci in *Gossypium hirsutum* and iCottonQTL a New R/Shiny App to Streamline Genetic Mapping

Ashley N. Schoonmaker ^{1,2}, Amanda M. Hulse-Kemp ^{1,2,3,*}, Ramey C. Youngblood ⁴, Zainab Rahmat ^{5,6}, Muhammad Atif Iqbal ⁶, Mehboob-ur Rahman ⁶, Kelli J. Kochan ⁷, Brian E. Scheffler ⁸ and Jodi A. Scheffler ^{9,*}

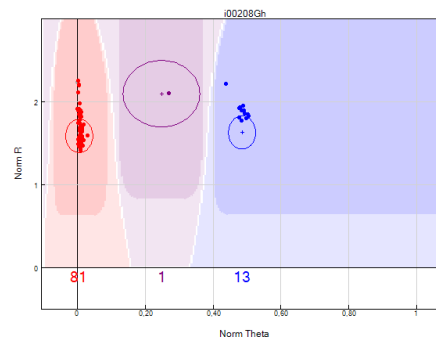
Design Illumina 18K XT Infinium Cotton Array

- **18,438** SNP markers were selected from the 63K HD Infinium Array based on Indian and North-American material
- Selection criteria: marker quality (cluster resolution), genome distribution and redundancy

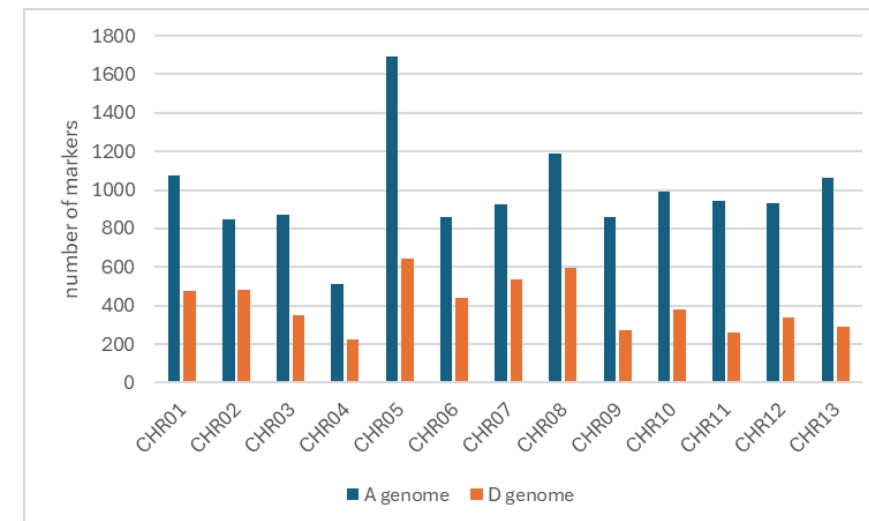
	A genome	D genome
CHR01	1074	477
CHR02	847	480
CHR03	874	352
CHR04	509	223
CHR05	1695	641
CHR06	858	437
CHR07	924	538
CHR08	1187	594
CHR09	859	270
CHR10	991	379
CHR11	946	259
CHR12	933	336
CHR13	1066	290
sum	12763	5276
unmapped	399	



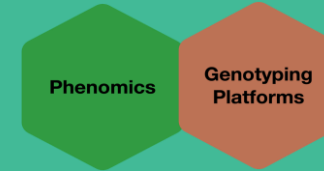
genome specific marker



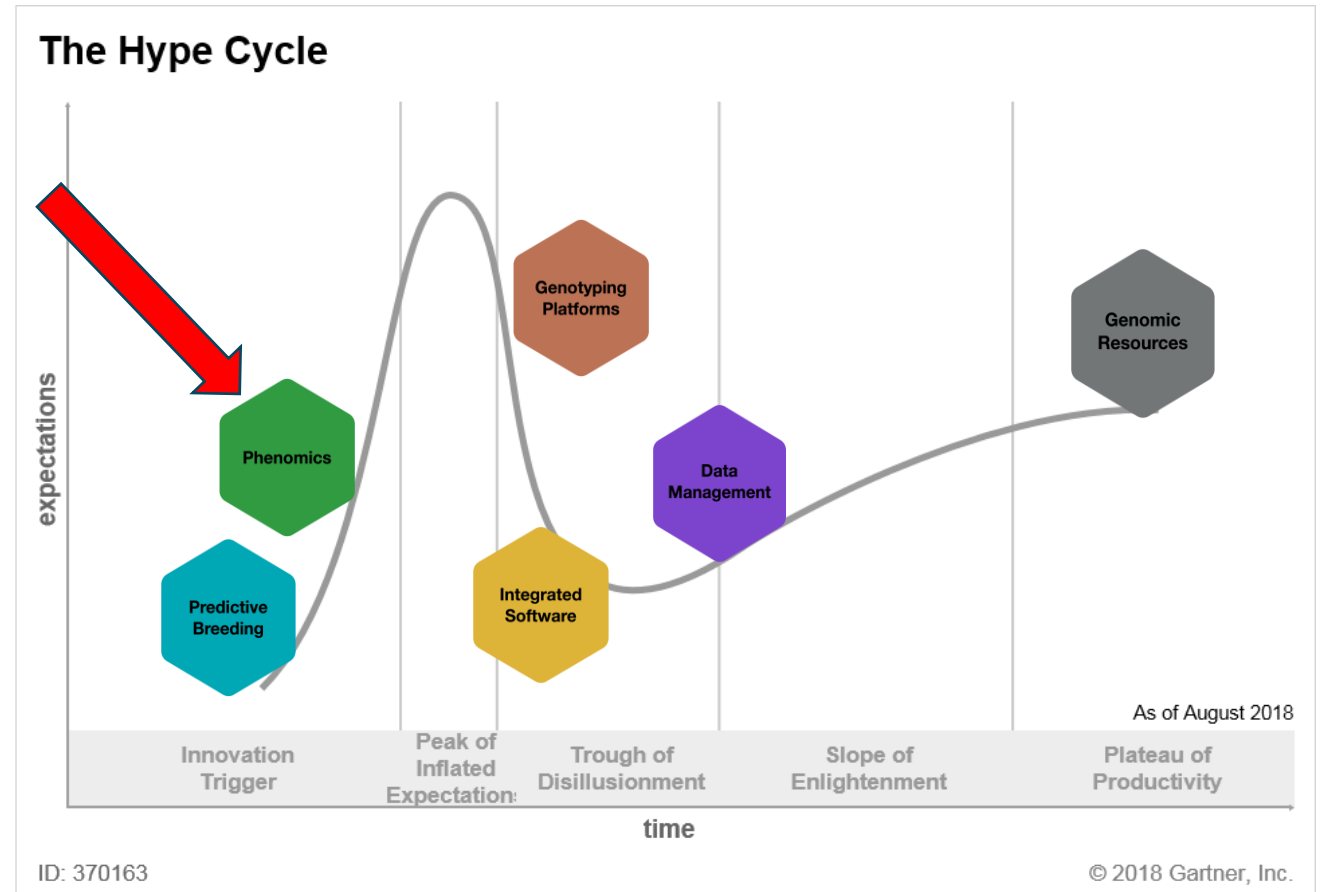
marker detecting both genomes



Predictive Breeding =



- Evaluate emerging technology
- Research to investigate integration of tools into breeding programs
- Develop new tools to fit holes where nothing is available



What would be helpful

- Capturing additional trait-specific markers
 - QTL/Fine-mapping
- Validating function of identified loci
 - Transformation, VIGS

VALIDATED FUNCTION $\rightarrow$ FIXED EFFECTS

Genome Wide Association Model:

$$Y = \mu + X_S\beta_S + Zu + \varepsilon$$

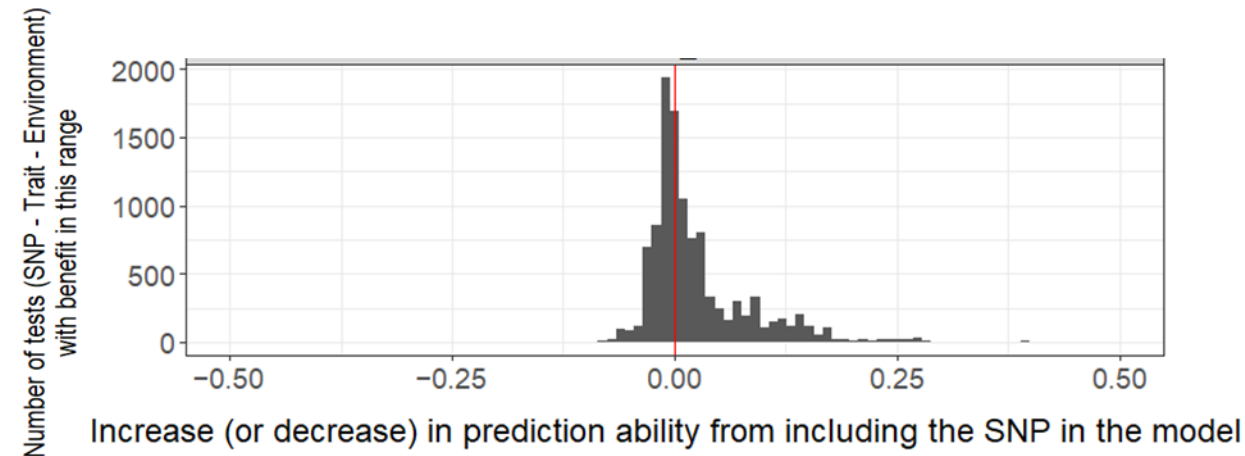
Interest is in SNP effect $\hat{\beta}_S$

Genomic Prediction Model:

$$Y = \mu + Zu + \varepsilon$$

Interest is in prediction of u : $\hat{u}$

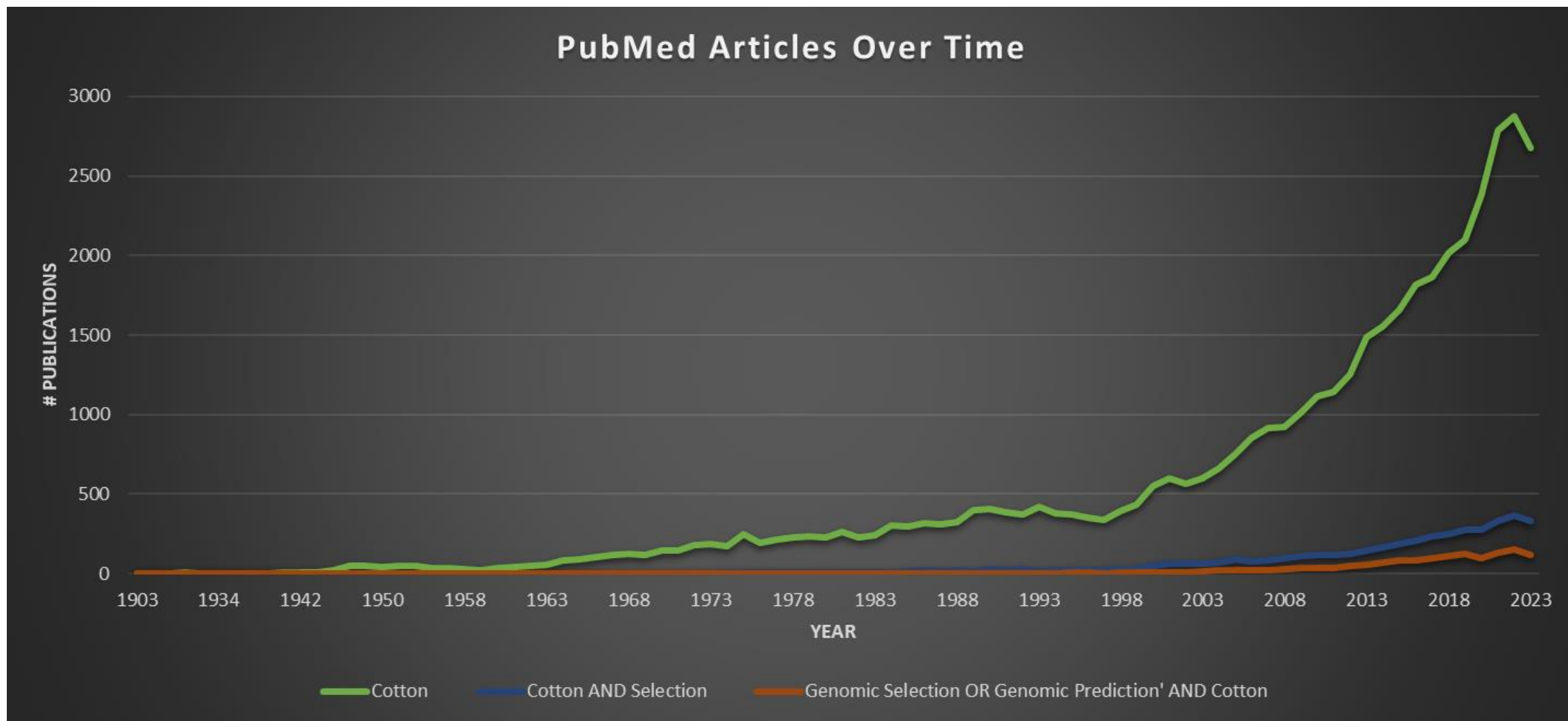
Can add fixed effects to the model for loci with known function



What would be helpful

- Capturing additional trait-specific markers
 - QTL/Fine-mapping
- Validating function of identified loci
 - Transformation, VIGS
- Standardized formatting of publication results

WHY?



FORMATTING ENABLES BIOINFORMATICS



aka...
Helping
ourselves!!!

FORMATTING ENABLES BIOINFORMATICS



aka...
Helping
ourselves!!!

- Full summary statistics for any analysis (GWAS or differential expression)
- Supplemental tables listing each differentially expressed gene
- Clearly indicate genome version utilized in analysis
- Include marker names in the main text
 - Only main text is indexed for automated processing
- Raw field data included not averages

FORMATTING ENABLES BIOINFORMATICS

aka...
Helping
ourselves!!!

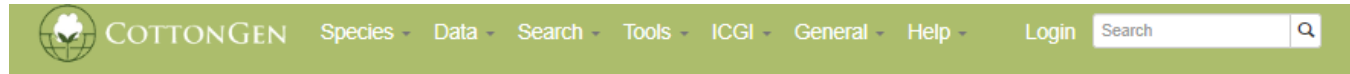
- Full summary statistics for any analysis (GWAS or differential expression)
- Supplemental tables listing each differentially expressed gene
- Clearly indicate genome version utilized in analysis
- Include marker names in the main text
 - Only main text is indexed for automated processing
- Raw field data included not averages

BONUS -> Work directly with CottonGen as a part of your publication process and ensure result is available through that avenue directly



FORMATTING ENABLES BIOINFORMATICS

aka...
Helping
ourselves!!!



Gohir.A03G073700.1

Transcript Overview
Alignments
Analyses
Annotated Terms
Contact
Cross References
Homology
InterPro
Orthologs
Publications
Relationships
Sequences

Publications

Year	Publication
2023	Osborne AN, Osagiede A, Storm AR, Hulse-Kemp AM, Stoeckman AK. <i>Gossypium hirsutum</i> gene of unknown function Gohir.A03G0737001 encodes a potential Chaperone-like Protein of protochlorophyllide oxidoreductase (CPP1).. <i>microPublication biology</i> . 2023; 2023.

microPublication
BIOLOGY μ P

7/30/2023 - Open Access

***Gossypium hirsutum* gene of unknown function Gohir.A03G0737001 encodes a potential Chaperone-like Protein of protochlorophyllide oxidoreductase (CPP1)**

Alana N. Osborne¹, Andrew Osagiede¹, Amanda R. Storm^{2,8}, Amanda M. Hulse-Kemp^{3,4,8}, Angela K. Stoeckman^{5,8}

¹Chemistry, Bethel University, Saint Paul, Minnesota, United States

²Biology, Western Carolina University, Cullowhee, NC

³Genomics and Bioinformatics Research Unit, USDA-ARS, Raleigh, NC

⁴Department of Crop and Soil Sciences, North Carolina State University, Raleigh, North Carolina, United States

⁵Chemistry Department, Bethel University, Saint Paul, Minnesota, United States

⁸To whom correspondence should be addressed: arstorm@wcu.edu; amanda.hulse-kemp@usda.gov; a-stoeckman@bethel.edu

BONUS -> Work directly with CottonGen as a part of your publication process and ensure result is available through that avenue directly



What would be helpful

- Capturing additional trait-specific markers
 - QTL/Fine-mapping
- Validating function of identified loci
 - Transformation, VIGS
- Standardized formatting of publication results
- Movement to digital data and databases



Field Book



Breedbase allows fast data queries without advanced programming knowledge!

“I want to analyze Brix data for check varieties used in advanced yield trials conducted in 2011-2019 on muck soils”

Search Wizard

Don't see your data? [Refresh Lists](#) [Update Wizard](#)

Trial Types

Search

Select All 1/20 Clear

+ Clonal Evaluation

+ crossing_block_trial

+ crossing_trial

+ genetic_gain_trial

+ genotyping_trial

× Advanced Yield Trial

Match ANY ALL

Add to List... Add

Create New List... Create

Locations

Search

Select All 3/6 Clear

+ Okeelanta Corporation

+ Pahokee Produce Inc.

+ Townsite Farm

× Duda and Son's Inc.

× Hillard Brothers of Florida Ltd.

× Sugar Farms Cooperative North - Osceola Region

Match ANY ALL

Add to List... Add

Create New List... Create

Years

Search

Select All 11/11 Clear

× 2011

× 2012

× 2013

× 2014

× 2015

Match ANY ALL

Add to List... Add

Create New List... Create

Accessions

Search

Select All 5/1386 Clear

+ CP09-1390

+ CP09-1418

+ CP09-1430

+ CP09-1503

+ CP09-1512

× CP00-1101

× CP78-1628

× CP89-2143

× CP96-1252

× CPCL05-1201

Add to List... Add

Create New List... Create

Load/Create Datasets using Match Columns

Load Dataset Load Make Public Delete

S3 Create

Related Genotype Data

Related Trial Metadata

Related Trial Phenotypes

Too few trials

Fast (Improves speed but may miss recent) CSV Plots

☐ Include timestamps ☐ Suppress user defined phenotype outliers

Trait Name Contains Min Value Max Value

Brix -∞ ∞

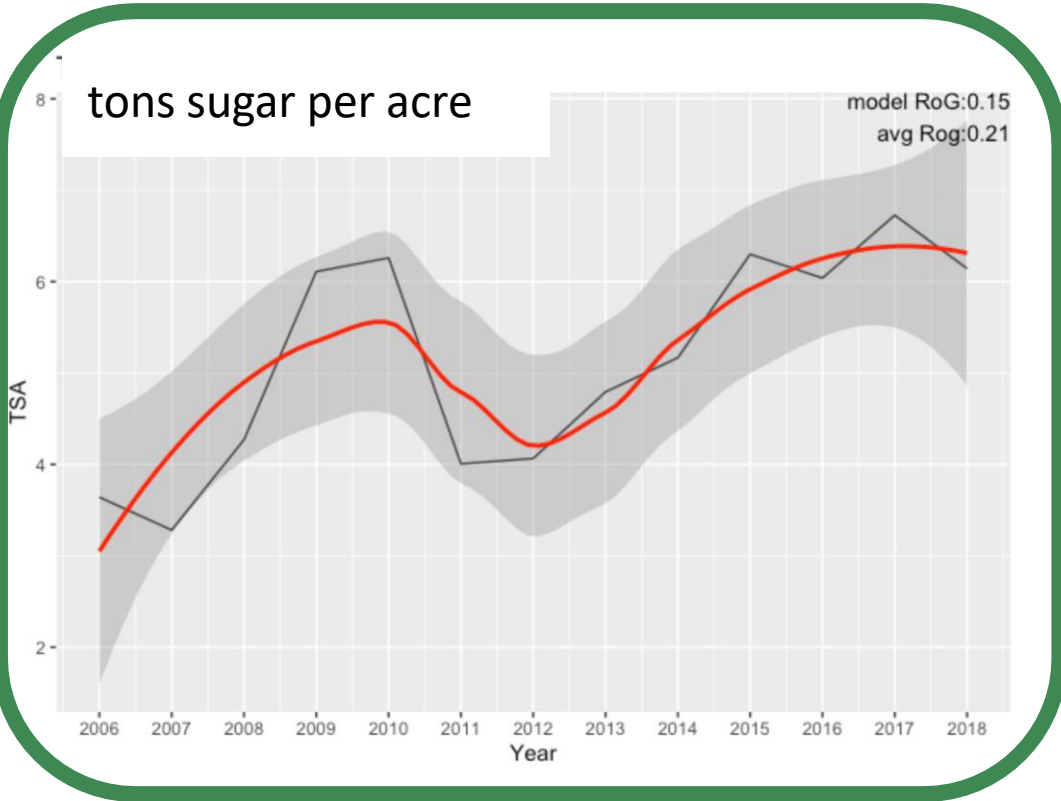
Download Phenotypes

What would be helpful

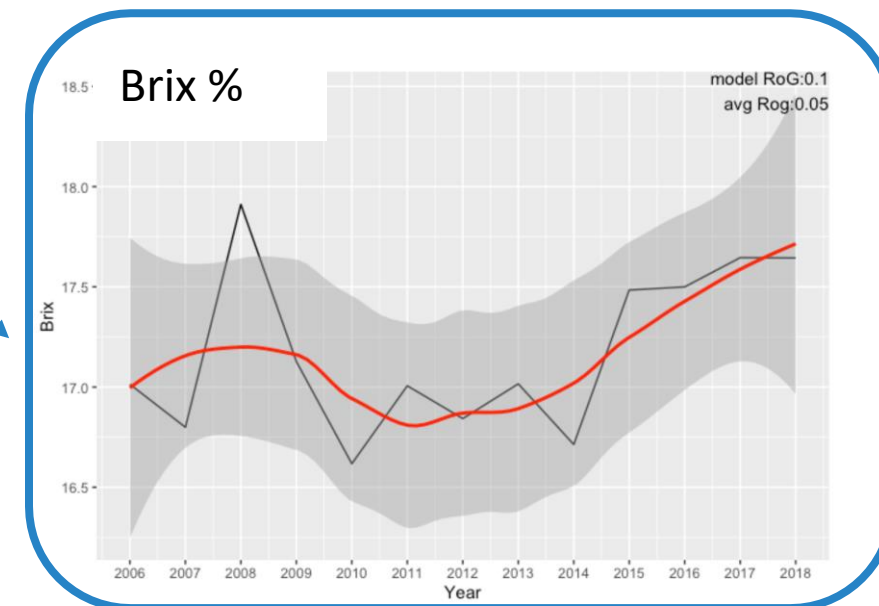
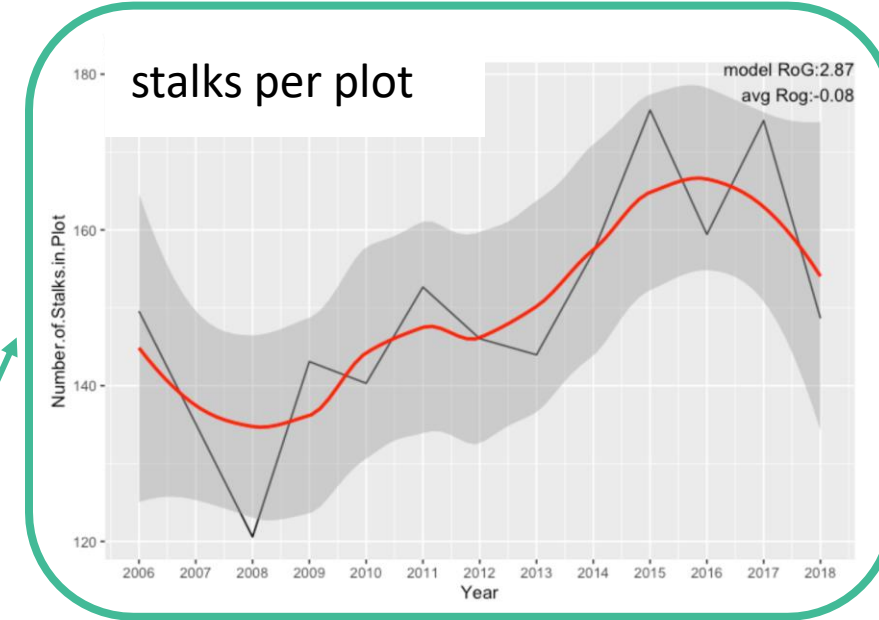
- Capturing additional trait-specific markers
 - QTL/Fine-mapping
- Validating function of identified loci
 - Transformation, VIGS
- Standardized formatting of publication results
- Movement to digital data and databases
- Increased shared use of checks
 - Enable cross location comparison of early generation materials
 - Use transformation lines as checks

MAKES EVALUATION OVER TIME POSSIBLE

USDA Sugarcane



Yield is up,
but that's
driven by
increases in
biomass,
not sugar

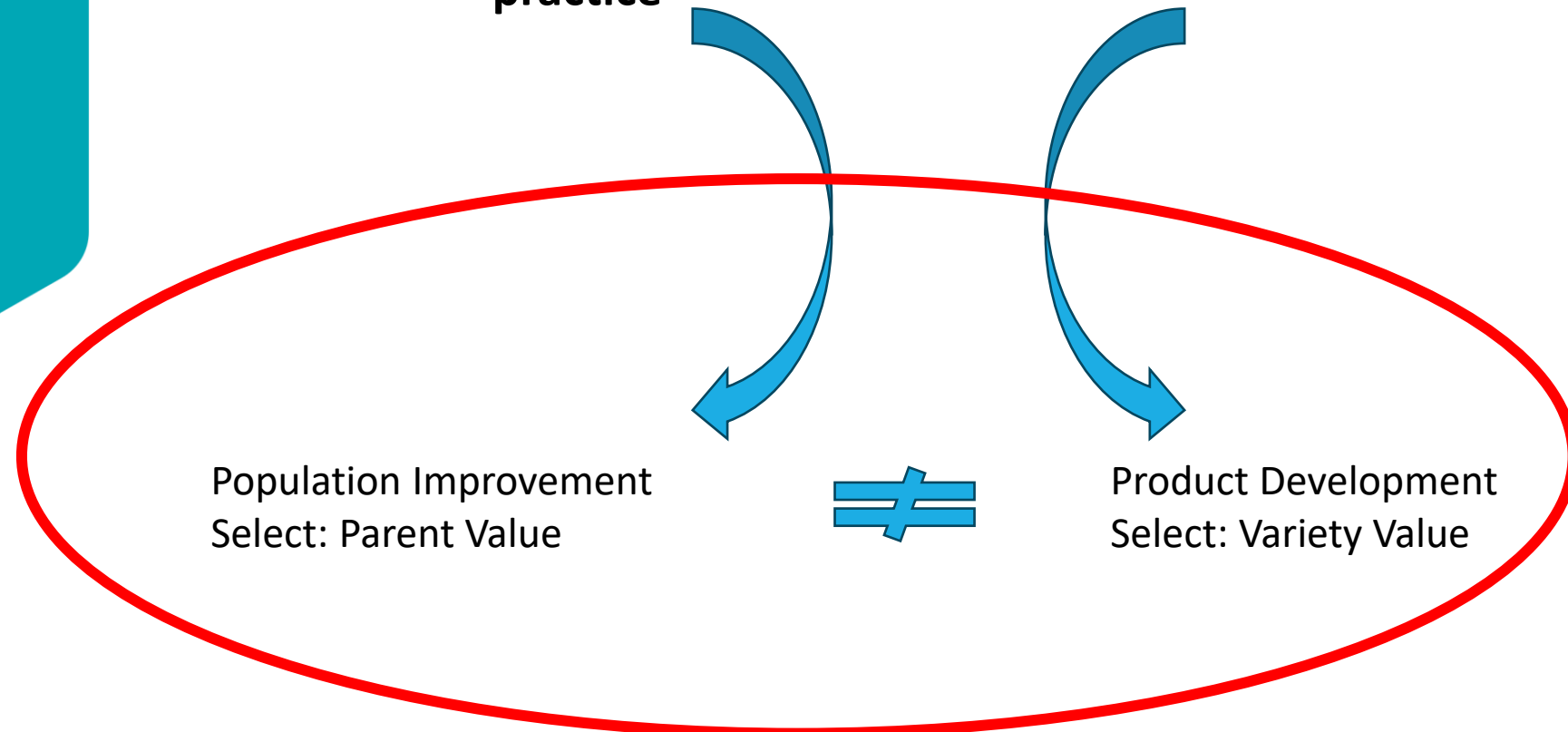


Predictive Breeding



Estimating traits across experiments - ie. in a breeding program (related, but NOT the same set of materials); THEN using those estimates to select what individuals to retain = **GENOMIC SELECTION**

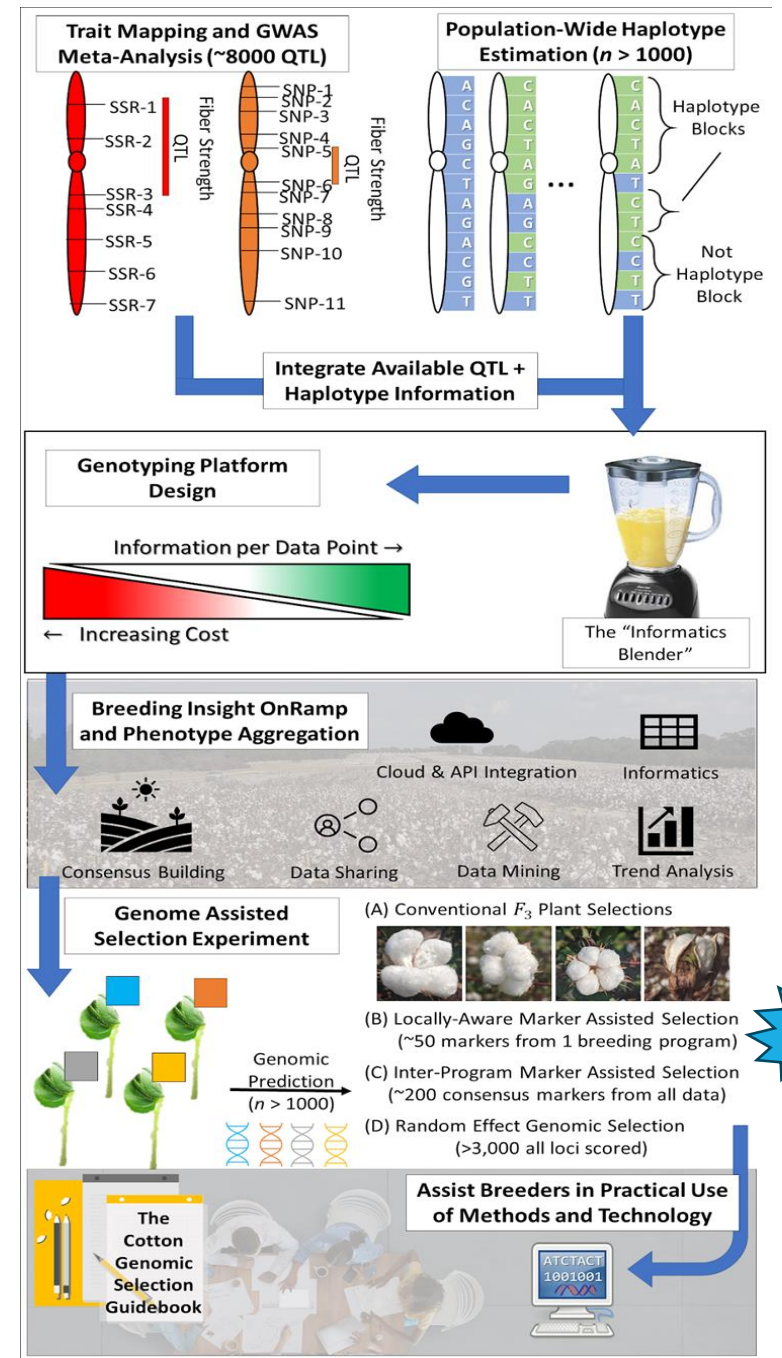
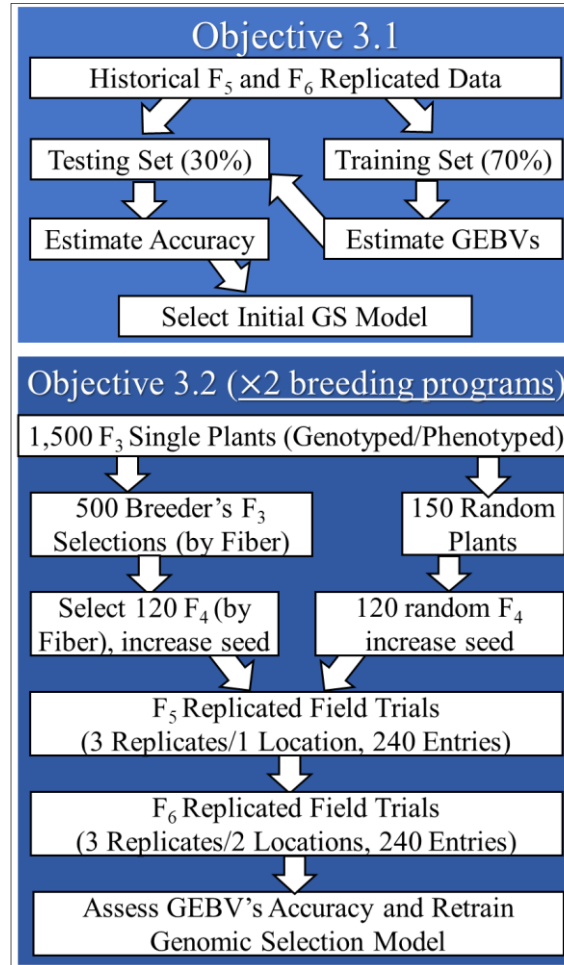
- May be empirically similar or potentially very different in practice than Genomic Prediction
- Difficult to measure success until put in practice



PRODUCT DEVELOPMENT

SELECT: VARIETY VALUE

- Empirically test impact of selection against traditional selection methods
- Initiate and compare genomic selection methods in two US public cotton breeding programs and develop recommendations.



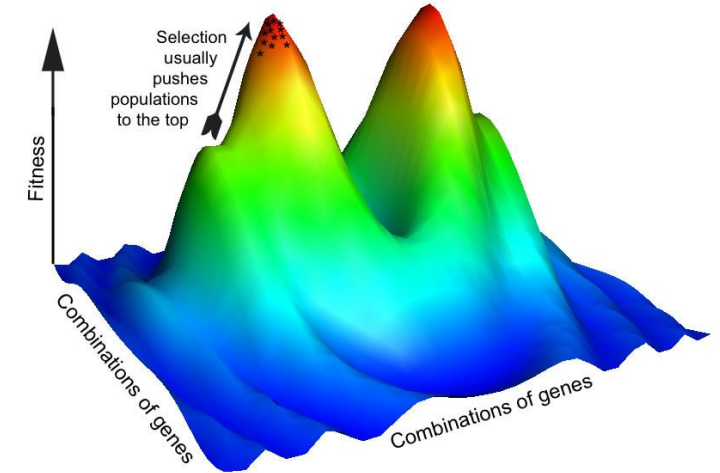
POPULATION IMPROVEMENT SELECT: PARENT VALUE

- Investigate benefit of earlier recycling of materials for parents
- Are mid-parent values fully predictive of progeny performance?
- Target recombination recalcitrant areas
- General combining ability vs. Specific combining ability

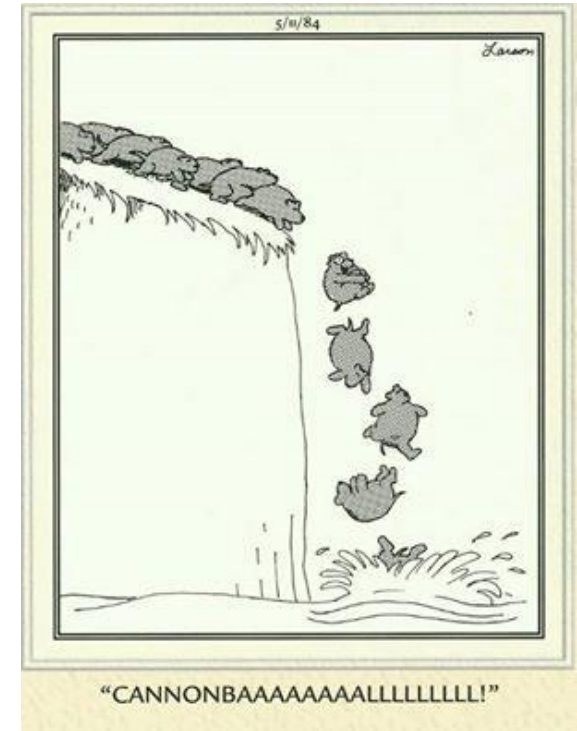
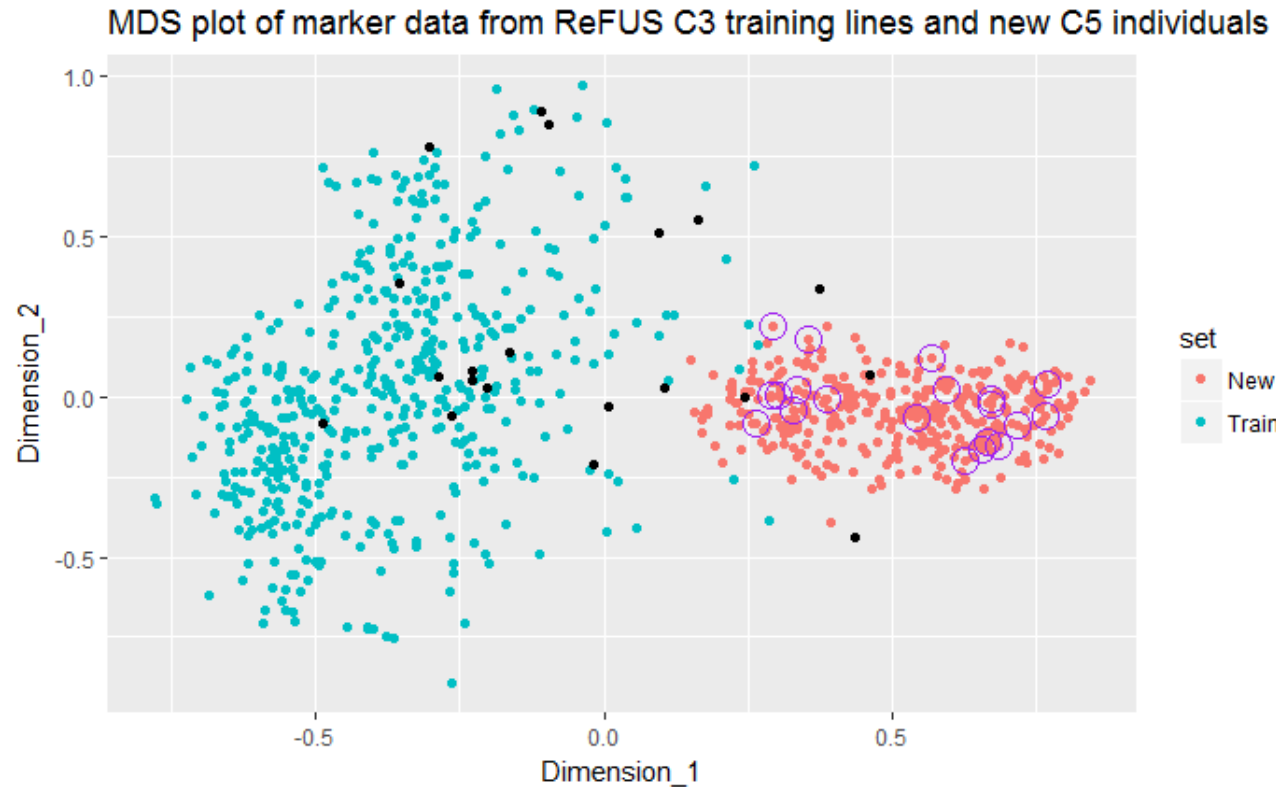


WHAT DOES THE FUTURE LOOK LIKE?

- Moving from single crop function to leveraging across crops
 - Informatics to Identify NEW adaptive peaks
- Advancing Genomic Selection
 - Leverage data across breeding programs
 - Continued access to performant high-quality genotyping platforms
- Better integrated Systems can use in the Field!
 - Sharing phenotyping technologies



GENOTYPIC COMPOSITION OF GS POPULATION IS CHANGING VERY FAST!



Gary Larson

Minimizing inbreeding with 'Optimal Contribution' method and genetic algorithms!

Courtesy of Jim Holland et al.

Accuracy of prediction model breaks down over generations



ACKNOWLEDGEMENTS

Grant Billings
Jonathan Zirkel



**Agricultural
Research
Service**

SGS (Formerly Trait Genetics) - Martin Ganai, Ed Bruggemann

David Stelly

Kelli Kochan

Gregory Thyssen

David Fang

Vasu Kuraparthi

Anjan Gowda

Peng Chee

Todd Campbell

Jodi Scheffler



**Cotton
Incorporated**

Nos.18-274, 19-872, 21-734



MISW-2021-11369
MISW-2021-07681

