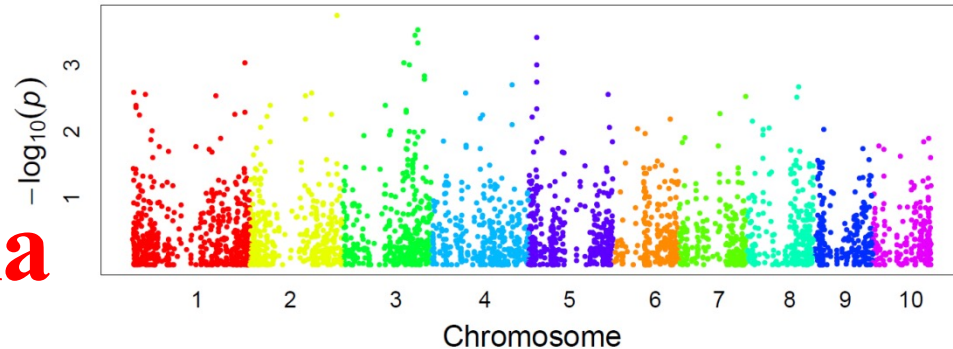


Polymorphisms and Association Tests

Alexander E. Lipka



Associate Professor of Biometry

Department of Crop Sciences

UIUC



This is what we will learn

- **How association tests fit into data analysis**
- **Basics: Simple linear regression**
- **Basics: Fixed and random effects**
- **Genome-wide association studies:**
 - **Introduction**
 - **Best practices**
- **Genomic selection:**
 - **Introduction**
 - **Best practices**
- **Examples of GWAS**
- **Examples of GS**

Role of the statistical genetics in polymorphism and association tests

G-to-P analyses

Genotypic data

Phenotypic data

Accurate G-to-P models help ensure that investments in high-throughput technologies lead to meaningful results in the field

<http://boort.com.au/gallery/drone-checking-corn-crop>
<https://www.pioneer.com/us/products/soybeans.html>

What can we do with statistical genetics?

- **Associate genotypes to phenotypes**
 - Basic statistical model: Y-variables are one or more traits; X-variables are one or more genomic markers
 - Models that accurately model the intricate relationship between genotype and phenotype
 - More accurate genomic selection models
- **Ramifications of this research**
 - Dissection of genetic sources of agronomically important traits in crops
 - Flexible statistical models that can analyze wider range of traits
 - Reduction in length of breeding cycles

G-to-P models are based on simple linear regression (SLR)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Diagram illustrating the Simple Linear Regression (SLR) model equation:

Y_i is the Y-value of i^{th} observation.

β_0 is the Intercept Parameter.

β_1 is the Slope Parameter.

x_i is the X-value of i^{th} observation.

ε_i is the Random Error Term.

- Models linear relationship between quantitative X and Y variables
- Parameters β_0 and β_1 are unknown constants
- Data sets of n (X, Y) observations used to estimate parameters

Assumptions of the error terms

- $\varepsilon_i \sim NID(0, \sigma_e^2)$
 - Normal
 - Independent

This framework can be used for X-variables that are categorical

- What can be done if assumptions are violated?
 - Transform the trait (e.g., Box-Cox procedure)
 - Implement a bootstrapping (or similar) procedure

Factorial Experiment

Factor A - Density

1

2

3

1

2

3

Factor B- Fertilizer

**Quantify two-way interaction
effect of Factors A and B**

2-way ANOVA model with fixed effects

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- Y_{ijk} : Y -value of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor A
- μ : Grand mean

Inferences of fixed effects apply only to the factor levels used in your experiment

- $(\alpha\beta)_{ij}$: Two-way interaction effect between receiving j^{th} level of Factor B and i^{th} level of Factor A
- ε_{ijk} : Error term of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor $A \sim NID(0, \sigma_e^2)$

2-way ANOVA model with random effects

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- Y_{ijk} : Y -value of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor A

Inferences of random effects apply to an entire population of factor levels

- β_j : Random main effect of j^{th} level of Factor B
 $\sim NID(0, \sigma_B^2)$
- $(\alpha\beta)_{ij}$: Random two-way interaction effect between receiving j^{th} level of Factor B and i^{th} level of Factor A
 $\sim NID(0, \sigma_{AB}^2)$
- ε_{ijk} : Error term of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor A $\sim NID(0, \sigma_e^2)$

Mixed model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

- Y_{ijk} : Y -value of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor A

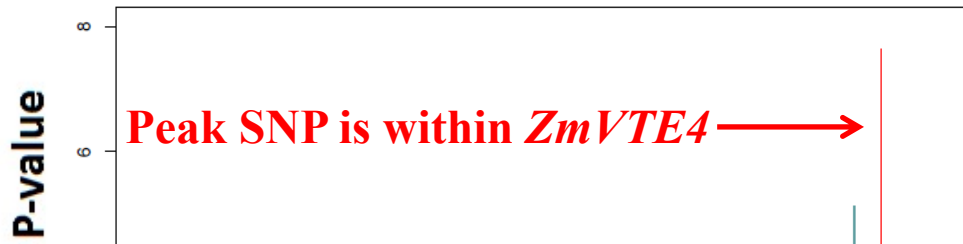
Mixed models are flexible and can be adapted for many different quantitative genetics analyses

receiving j^{th} level of Factor B and i^{th} level of Factor A
 $\sim NID(0, \sigma_{AB}^2)$

- ε_{ijk} : Error term of k^{th} replicate receiving j^{th} level of Factor B and i^{th} level of Factor $A \sim NID(0, \sigma_e^2)$

Genome-wide association study (GWAS)

Association with Vitamin E Levels in Maize Grain

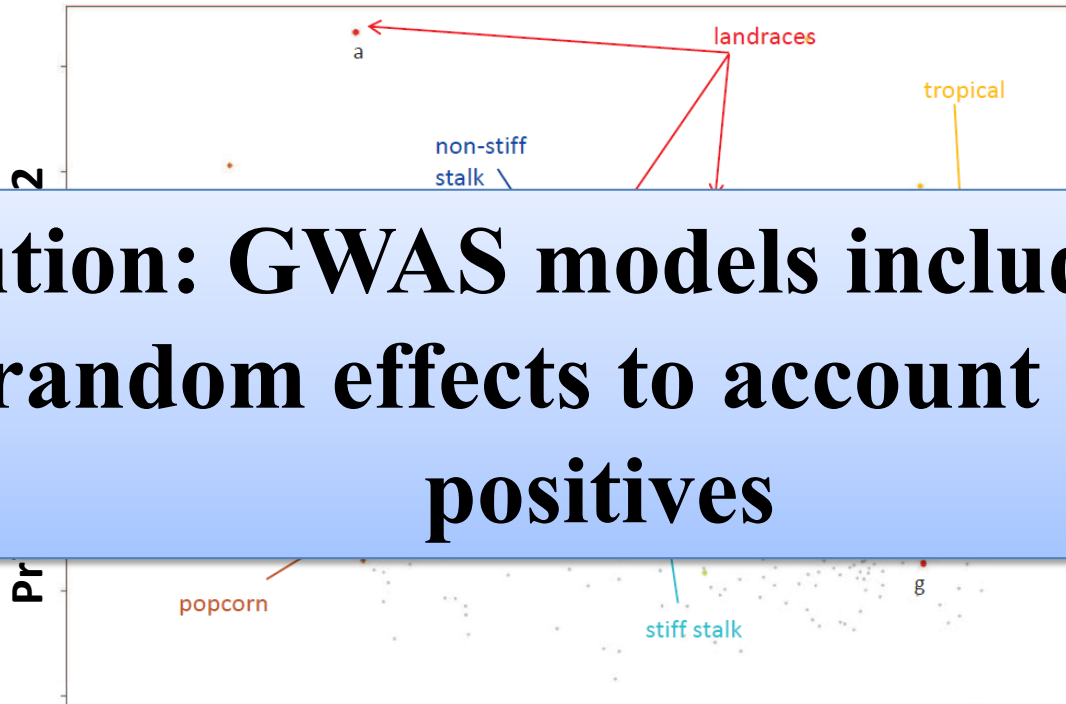


Markers exhibiting peak associations with traits are potential targets for marker-assisted selection (MAS)

- Identify genomic regions associated with a phenotype
- Fit a statistical model at each SNP in genome
- Use fitted models to test H_0 : No association with SNP and phenotype

Genetic diversity can lead to false positives in a GWAS

Genetic Diversity of 2,815 Maize Inbreds



Romay et al. (2013)

- **Solution: GWAS models include fixed and random effects to account for false positives**

- Two sources for false positives:
 - Population Structure
 - Familial Relatedness

Unified mixed linear model (MLM)

Grand Mean

Marker effect

Random effects:
account for familial
relatedness

$$Y_i = \mu + \left(\sum_{j=1}^p \beta_j PC_{ij} \right) + \alpha x_i + \left(Line_i \right) + \varepsilon_i$$

- Variance component estimation is computationally intensive
- Computational approaches are available to reduce this computational burden

- $(Line_1, \dots, Line_n) \sim \text{MVN}(\mathbf{0}, 2K\sigma_G^2)$

- $K =$ kinship matrix

Measures relatedness between individuals

- $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma_E^2)$

Approach 1: Compressed mixed linear model

$$Y_i = \mu + \sum_{j=1}^p \beta_j PC_{ji} + \alpha x_i + \text{Gin} \alpha_i \rho_i \varepsilon_i$$

Perform hierarchical

- Reduces computational time because it works with a smaller kinship matrix

- $(\text{Gin} \alpha_i \rho_i, \dots, \text{Li} \text{Gin} \alpha_i \rho_i) \sim \text{MVN}(\mathbf{0}, \mathbf{N} \times \mathbf{N}, 2K_c \sigma_G^2)$
- K_c = kinship ("matrix compressed") kinship matrix
- $\varepsilon_i \sim \text{i.i.d. N}(0, \sigma_E^2)$

Approach 2: Population parameters previously determined (P3D)

- Reduces computational time because intensive variance component estimation is conducted only once
- Approximation: tends to underestimate most significant associations

• Λ_C – group (compressed) kinship matrix

• $\varepsilon_i \sim \text{i.i.d. } N(0, \hat{\sigma}_E^2)$

Approach 3: GEMMA

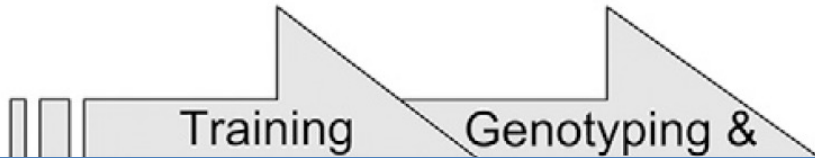
$$Y_i = \mu + \sum_{j=1}^p \beta_j P C_{ji} + \alpha x_i + \text{Line}_i + \varepsilon_i$$

- **Same reduction in computational time as P3D**
 - **Exact: enables statistically optimal estimation of marker-trait associations**

- K = kinship matrix

- $\varepsilon_i \sim \text{i.i.d. } N(0, \hat{\sigma}_E^2)$

Genomic selection (GS)



- **Various frequentist and Bayesian models are commonly used**
- **Most produce approximately the same prediction accuracies**

Heffner et al. (2009)

- Predict phenotypic values using markers distributed throughout the genome
- Enables selection without phenotyping individuals
- Developed to speed up breeding cycles

Basic GS statistical model

- Trait is the response variable (Y_i)
- All markers are the explanatory variables (x_{1i}, \dots, x_{pi})

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

- Number of markers (p) typically exceeds sample size (n): $n \ll p$

Issues with $n \ll p$

- Problem:
 - Unique estimates of marker effects do not exist
- Solution 1 (Non-Bayesian):
 - Add a penalty that restricts values of the marker effects (e.g., ridge regression, LASSO)
- Solution 2 (Bayesian):
 - Assign a “prior distribution” on the marker effects (e.g., Bayes A, Bayes B, ...)

Ridge regression best linear unbiased prediction (RR-BLUP) for genomic selection

Observed SNP alleles
of k^{th} marker at i^{th}
individual

Random marker effect
 $\sim N(0, \sigma_G^2)$

All predicted marker values are equally penalized so that they are shrunk to zero

Best linear unbiased predictors (BLUPs) of the β'_k s are subjected to the ridge regression penalty:

$$J(\beta) = \sum_{k=1}^p \beta_k^2$$

Cross-validation



- **Repeat so that each fold gets a chance to be the test set**
- **Prediction accuracy: average correlation between observed and predicted traits across folds**

Differences between GS and GWAS

- The overall objectives differ:
 - Main objective of GWAS is to find genomic regions associated with a trait
 - Main objective of GS is to determine how well marker sets predict trait breeding values
- The statistical models differ:
 - Typical GWAS models in plants test one marker at a time
 - Typical GS models include all markers in the model at once

Differences between GS and marker-assisted selection (MAS)

- GS:
 - Uses genome-wide marker sets for predictions
 - Can account for both major- and minor-effect QTL
 - Ideal for predicting complex traits
- MAS:
 - Focuses on marker(s) linked to genes of major effect
 - Accounts for only major-effect QTL
 - Adequate for predicting simple and oligogenic traits

How to choose the best model for GWAS

- It is critical to account for population structure and familial relatedness in a typical GWAS:


Suggested strategy:

- Use unified mixed linear model
- **If/when quantifying interesting GWAS result needs further refinement, use more sophisticated models**
 - Accounting for variance heterogeneity
 - Multi-locus, multi-trait models

Examples of GWAS in crops

- Rincker et al. (2016): Targets for brown stem rot resistance in soybean
- Owens/Lipka et al. (2014): Targets for boosting provitamin A and other carotenoid levels in maize grain
- Fernandes and Lipka (2020) and Fernandes et al (2021): Simulations to test performance of multi-trait GWAS models

Example: Rincker et al. (2016)

- Brown stem rot (BSR) and 
 - **Three genes associated with BSR resistance, *Rbs1-3*, have been identified in previous studies**
 - **Critical need to obtain a more precise location of these loci**
 - **Result in more efficient MAS for BSR resistance**

Source: cornandsoybeandigest.com/

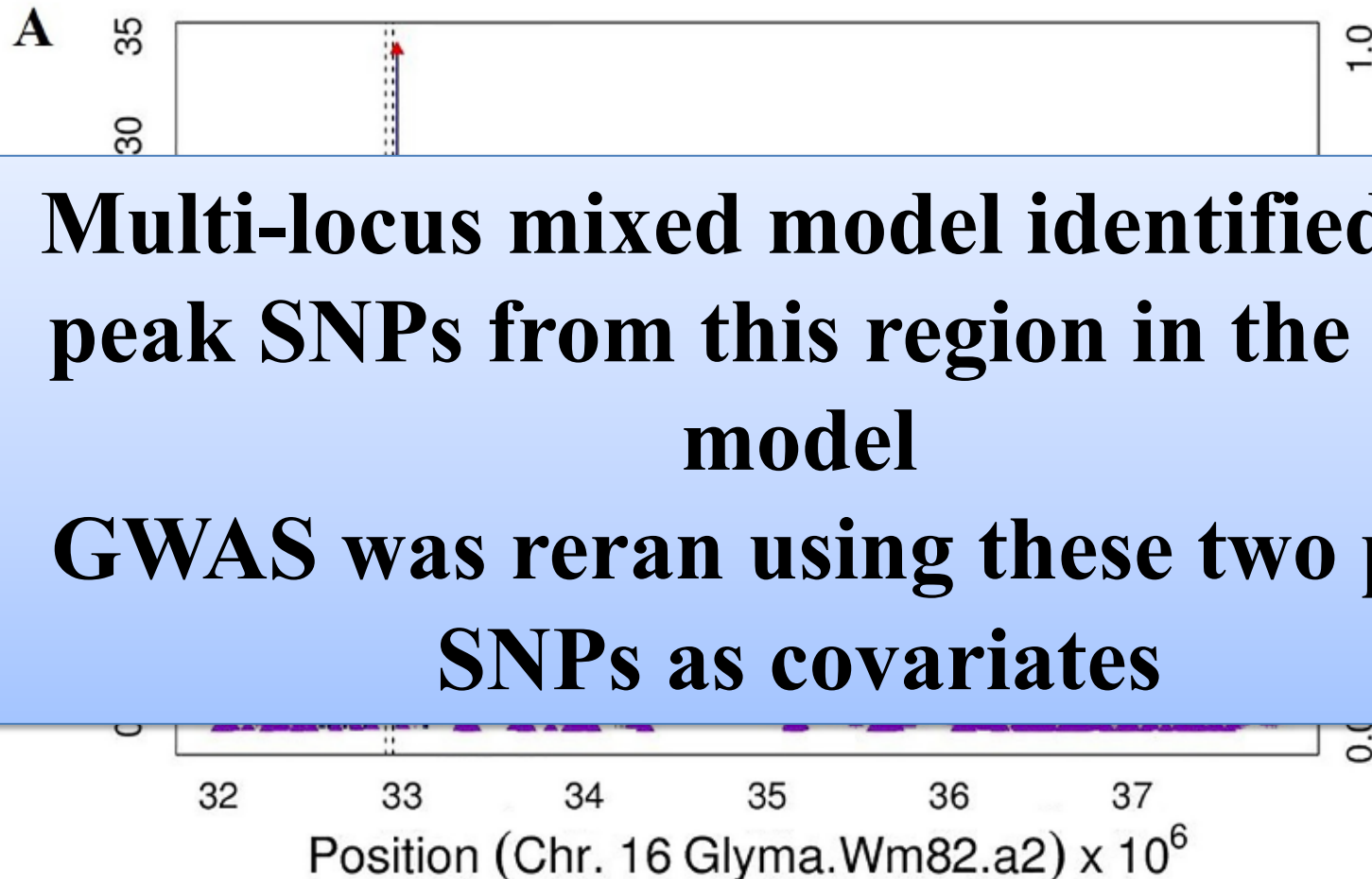
Separate GWAS performed on four association panels

Table 1. Characteristics of association panels analyzed with genome-wide association study and stepwise procedures.

Panel	Data type	Symptoms measured	Accessions	SNP† markers	Box-Cox lambda	BSR Score‡		
						Mean	SD§	$h^2¶$
N-1989	Binary	Foliar and stem	2773	33,240	na	na	na	na#
B-1997	Proportion 0–1	Foliar	540	33,486	log	0.09	0.15	0.49
B-1997	Proportion 0–1	Stem	540	33,486	1	0.38	0.20	0.61
B-2000	Proportion 0–1	Foliar	825	32,150	0.25	0.33	0.29	0.93
P-2003	Proportion 0–1	Stem	606	29,815	0.75	0.39	0.25	0.68

- N-1989 panel:
 - Binary phenotype: logistic regression + stepwise model selection
- Other panels:
 - Quantitative phenotype: Unified MLM + multi-locus mixed model

Unified MLM GWAS identifies signals near *Rbs1-Rbs3*



Multi locus mixed model (MLMM)

quantifies associations of multiple markers

Grand Mean

Additive effect of k^{th}

Random effects:
account for familial

- “Final” model selected by MLMM consisted of exactly 2 SNPs, both in *Rbs1-Rbs3* region
- We will revisit usage of the MLMM in a future example

$K =$ kinship matrix

Measures relatedness between individuals

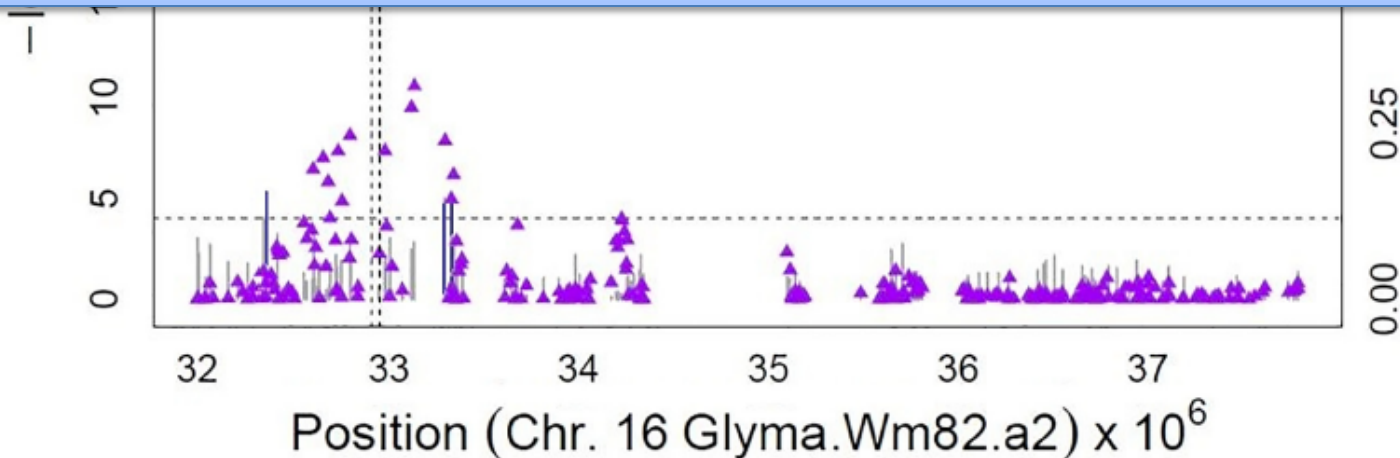
$I =$ subset of markers selected from stepwise regression

$\varepsilon_i \sim$ i.i.d. $N(0, \sigma_E^2)$

Peak SNPs from MLMM reduces explains most of *Rbs1-Rbs3* signal



- Similar findings were obtained in the other association panels



Breeding Ramifications



Source: blogs.ext.vt.edu

- Previous *Rbs1-Rbs3* signals been refined to a 0.3 Mb region on Chromosome 16
- Should facilitate both MAS-based approaches and gene cloning efforts
- Demonstrates the utility of GWAS in soybean

Biofortification

- Identify target genes with nutrients

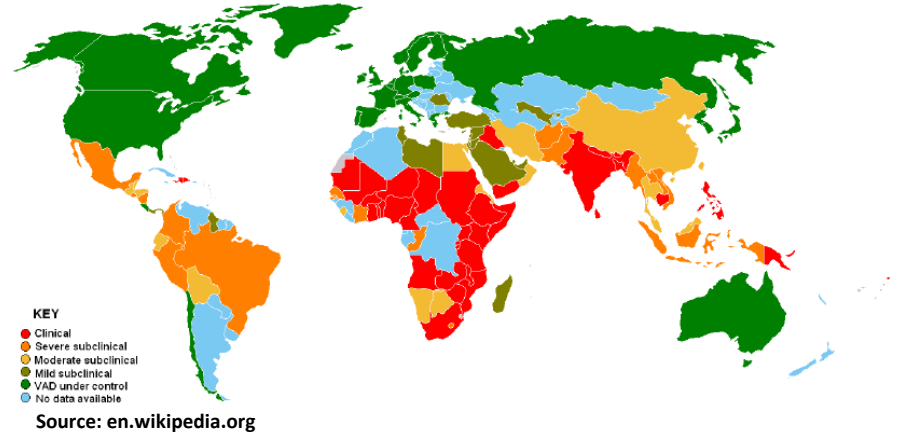


Source: www.aboutharvest.com

- Increase nutritional value of local crop varieties by selecting on these target genes
- Results in increased availability of essential nutrients

Targeting vitamin A deficiency through biofortification

- Vitamin A deficiency (VAD):
 - Affects 17-30% of children under 5
 - 250-500,000 children become blind every year
 - Infant morbidity and mortality



- Maize is a primary food source in many vitamin A deficient regions
- Biofortification: breed locally-adapted maize lines for increased provitamin A levels in grain

Work in maize provitamin A biofortification prior to Owens/Lipka et al. (2014)

- Candidate gene studies identified loci in maize (Harjes et al., 2008; Vallabheneni et al., 2010; Yan et al. 2010)

Owens/Lipka et al (2014):

- 1.) Conduct a GWAS to identify new candidate genes
- 2.) Determine a minimal marker set to accurately predict carotenoid levels

1. Heterotopy identified among metabolite
QTL (Kandianis et al., 2013)



Source: Chandler/Lipka et al., 2013

Data analyzed in Owens/Lipka et al. (2014)



- **Maize lines with white kernels do not produce measureable carotenoids**
- **We only analyzed a subset of 201 lines that range from light yellow to dark orange kernel color**

- Here's what we did:
- Compound levels quantified in grain:
 - Carotenoids for 252 lines

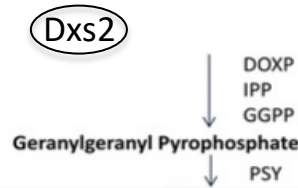
GWAS found significant marker-trait associations near carotenoid pathway genes




- Adjusting for multiple testing at the genome-wide level was conservative
- We also conducted a pathway-level analysis, where only markers near 58 *a priori* genes were considered

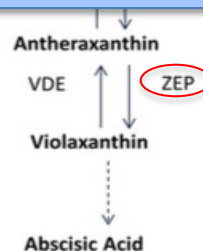


GWAS found significant marker-trait associations near carotenoid pathway genes



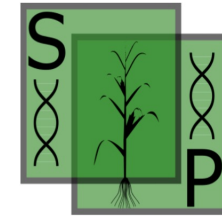
 = Significant at the genome-wide level

- This work identified potential targets for marker-assisted selection (MAS)
 - Are selecting for these target loci sufficient for improving provitamin A content in maize grain?





Prof. Samuel Fernandes



- **Multivariate quantitative genetics approaches have great potential**

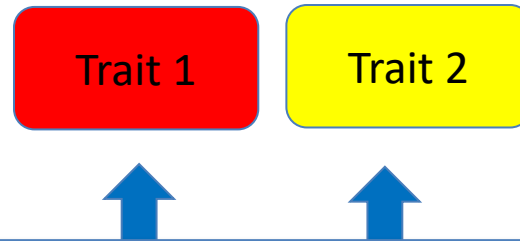
**CRAN/R package: simplePHENOTYPES:
simulate univariate and multivariate traits
based on user-inputted marker data**

utility

- We know genetic architectures of simulated traits
- We directly assess true and false positive identification rates

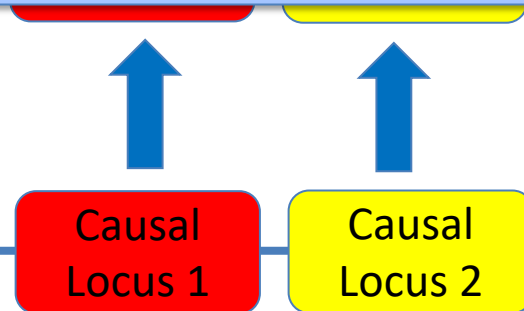
How well can GWAS models differentiate between pleiotropy and linkage?

One pleiotropic locus



We are now at a stage where we should deploy multivariate GWAS models more readily

Two loci in linkage



Genomic position

Unified mixed linear model can become multivariate

$$Y_i = \underbrace{\mu}_{\text{Grand Mean}} + \underbrace{\sum_{j=1}^p \beta_j PC_{ij}}_{\text{Marker effect}} + \alpha x_i + \underbrace{Line_i}_{\text{Random effects: account for familial relatedness}} + \varepsilon_i$$

- **Univariate GWAS: Y-variable consists of one trait**
- **Multivariate GWAS: Y-variable consists of two or more traits**

- $(Line_1, \dots, Line_n) \sim \text{MVN}(\mathbf{0}, 2K\sigma_G^2)$

- $K =$ kinship matrix

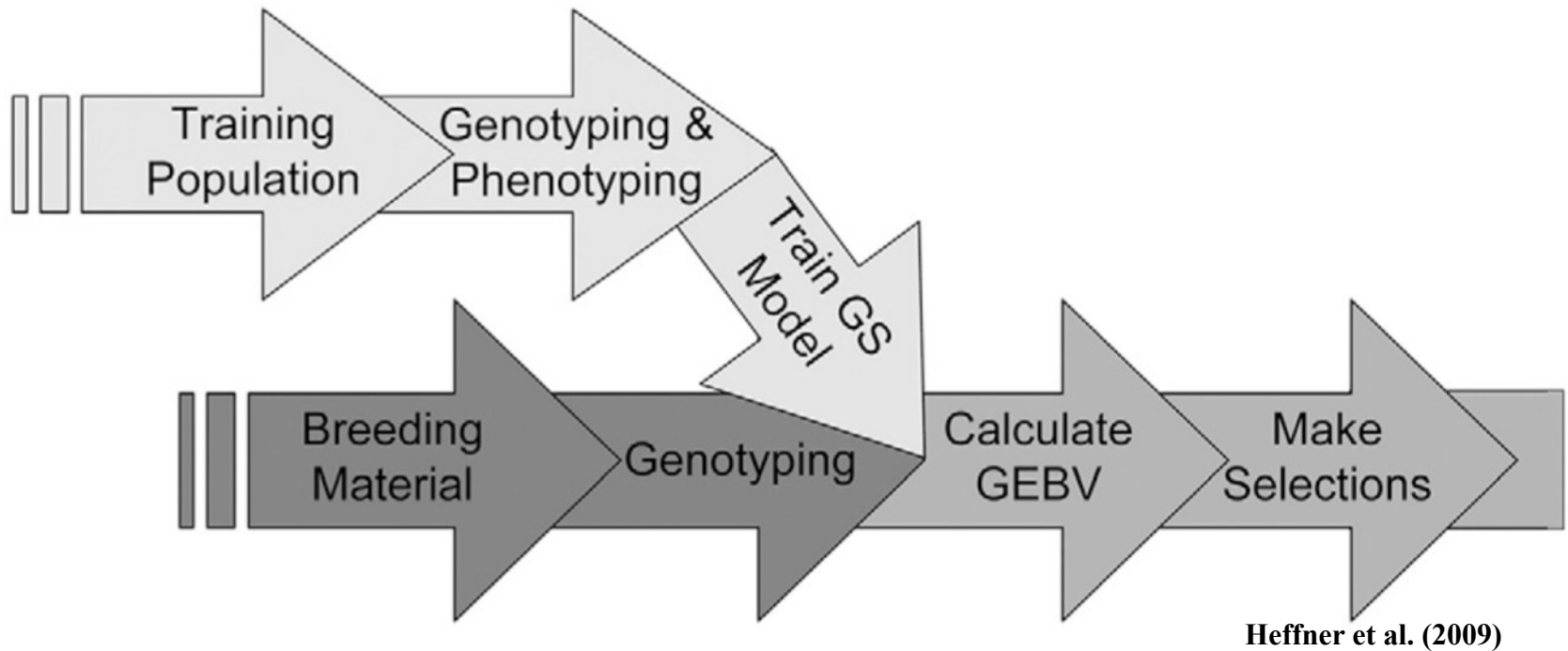
Measures relatedness between individuals

- $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma_E^2)$

Examples of GS in crops

- Lipka et al. (2014): Basic GS example in switchgrass
- Olatoye et al. (2020): More advanced GS example in *Miscanthus*

Genomic selection (GS) could speed up switchgrass breeding cycle



- GS on simple-to-measure traits approximating biomass yield could revolutionize switchgrass breeding efforts
- We evaluated the potential of GS using the latest genotypic and phenotypic resources

Switchgrass Association Panel



Photo taken 17 August 2010; Caldwell Field
Cornell University, Ithaca NY

- 515 members
- Grown in Ithaca, NY in 2009-2011
- Tetraploids and octoploids included
- Predominantly northern-adapted upland germplasm

Genotypic and Phenotypic Data

- 7 morphological traits
- 13 biomass quality traits (Vogel et al., *Bioenergy Resources*, 2011)
- 16,669 SNPs using genotyping-by-sequencing (GBS) techniques
- SNPs were anchored to the *Panicum virgatum* v1.1 reference genome
 - Used to impute missing SNP values

GS study

- Three popular GS models:
 - RR-BLUP
 - Elastic net
 - LASSO
- RR-BLUP should perform best for complex traits
- LASSO should perform best for simple traits
- 10-fold cross validation to evaluate performance

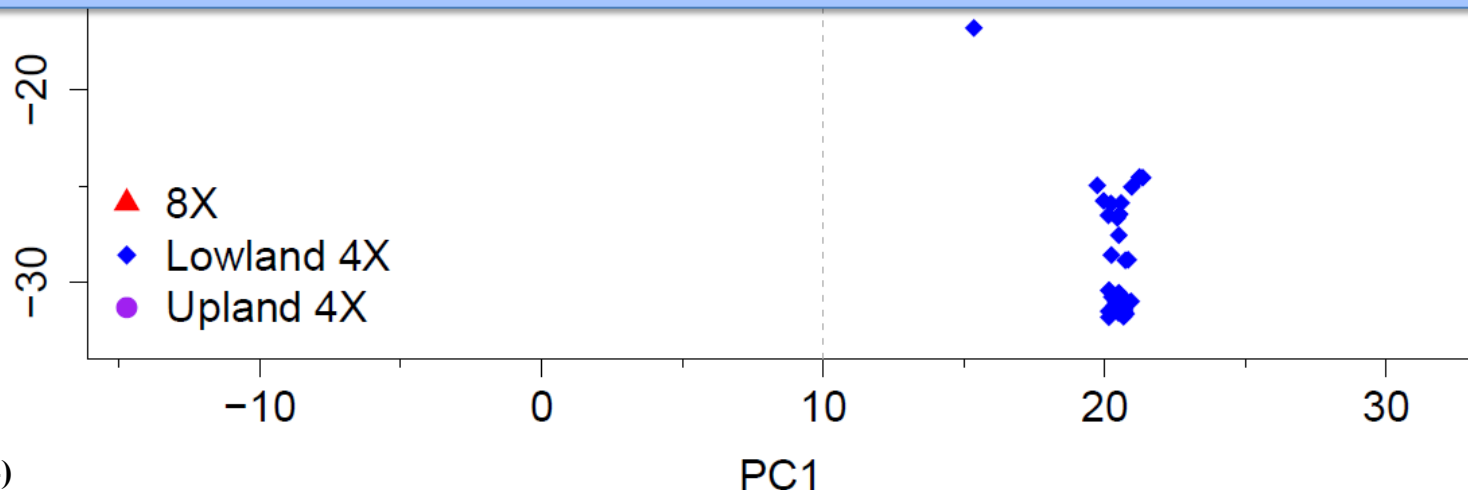
Main finding: GS appears to work

- Three GS models produced similar prediction accuracies
- High prediction accuracies obtained for most traits
 - Standability had the highest (0.52)
- Morphological traits generally had higher prediction accuracies than the biomass quality traits

Do we need to “account” for population structure in GS?

First two principal components (PCs) of 16,669 SNPs

- We used first two PCs to factor out SNP effects from population structure
- No longer agree that we need to factor out population structure from GS models

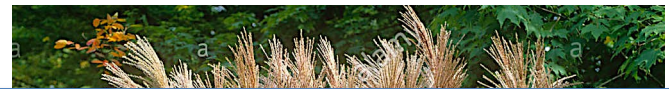


Miscanthus is a sustainable source of lignocellulosic ethanol biofuel production

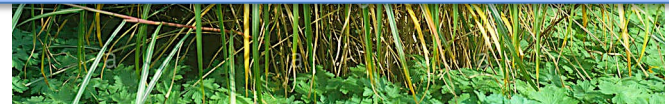
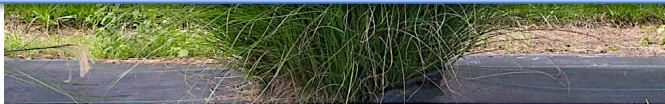
Species 1: *M. sinensis*



Species 2: *M. sacchariflorus*



**One clone of an interspecific cross of
Species 1 × Species 2 (M × g)
used for biofuel purposes in North America
and Europe**



Miscanthus sp.: Perennial grass from eastern Asia

Both species have substantial subpopulation structure

N China
W Korea
Eastern Yangtze
Qinling Mtns.
Hange Mtns.
Yangtze diploids
North China diploids

Purpose of GS is to quantify total genetic merit, whether it is:

- From genes underlying a trait
- Other genetic differences arising from subpopulation structure

Honshu
US nat.
S Honshu

diploids

N Japan tetraploids

Subpopulations in Species 1: Clark et al. 2014

Subpopulations in Species 2:

Contribution of pop structure to prediction accuracy?



Dr. Marcus Olatoye

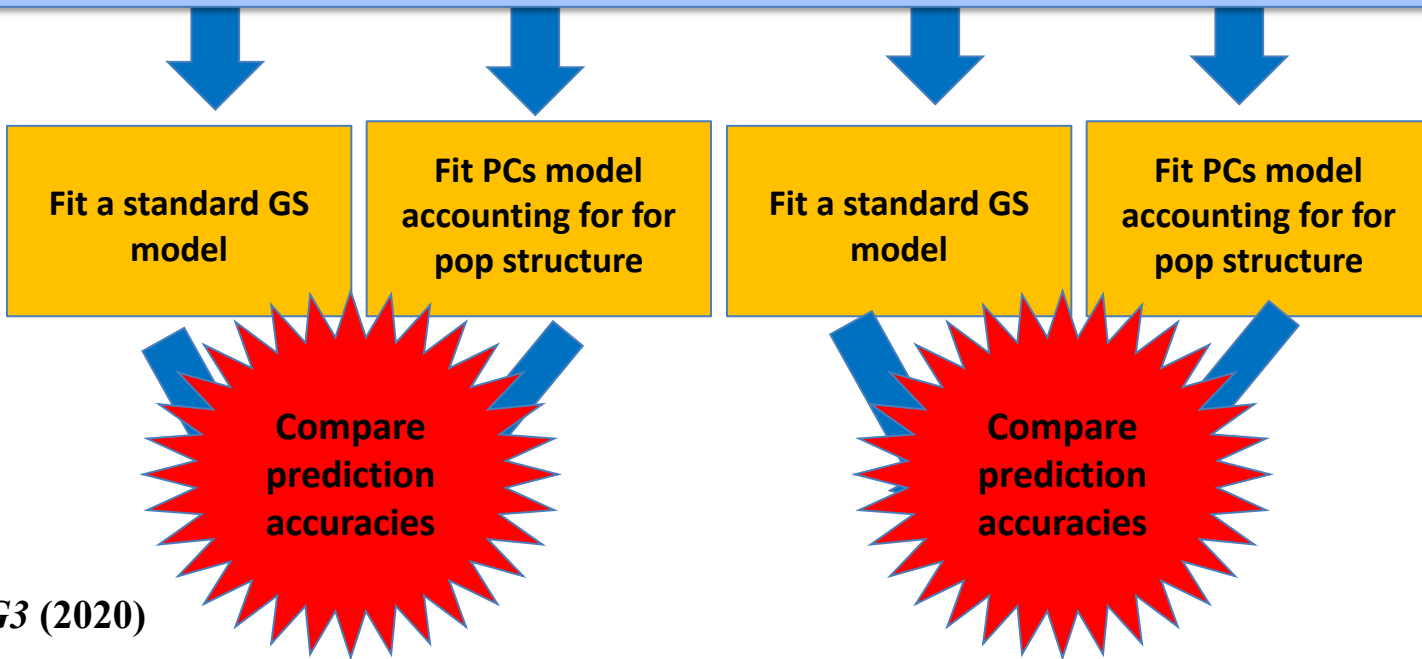
Panel 1: *Msi* panel



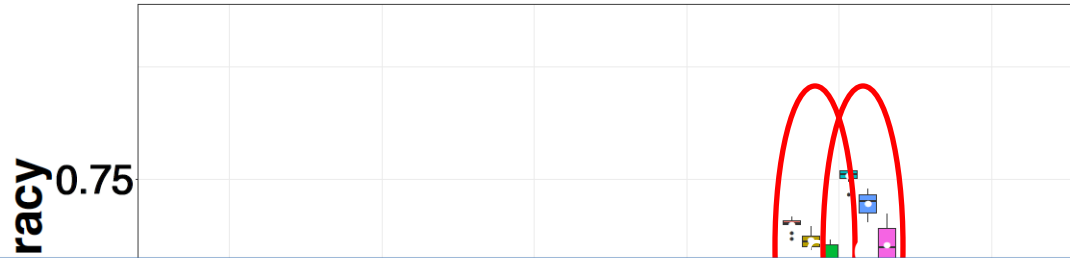
Panel 2: *Msa* panel



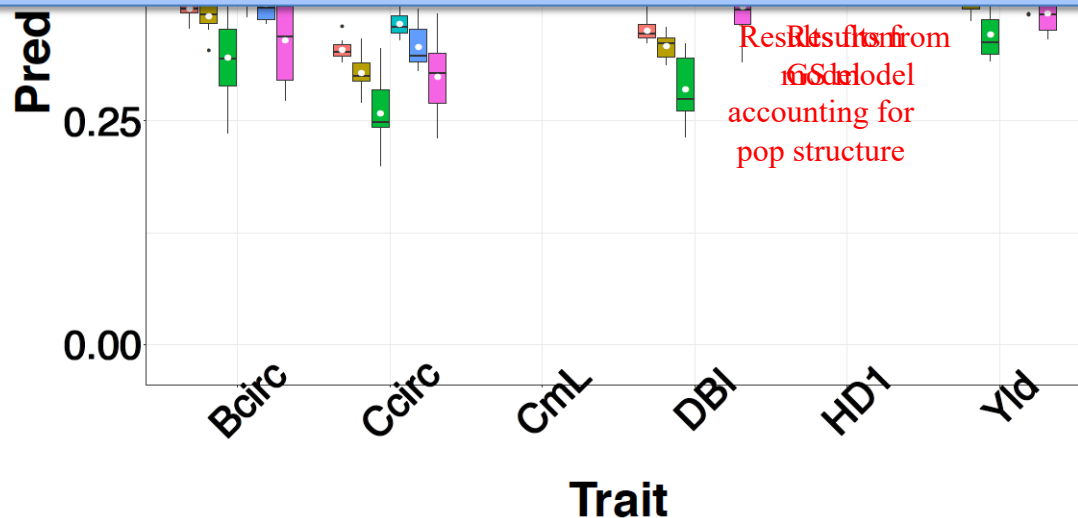
If pop structure is important, then prediction accuracy of PCs model should be close to GS model



Population structure accounts for substantial portion of GS prediction accuracy



My current opinion: do not “factor out” population structure in GS models



What did we just learn, and why is it important?

- **What we learned:**
 - **The basics of association tests**
 - **GWAS**
 - **GS**
- **GWAS and GS are the two most widely used applications association tests**