Genome Wide Association Study for Binomially Distributed Traits: A Case Study for Stalk Lodging in Maize

Esperanza Shenstone and Alexander E. Lipka
Department of Crop Sciences
University of Illinois at Urbana-Champaign

Unified Mixed Linear Model (MLM) in GWAS

Random effects: account for familial relatedness

Grand Mean

Marker effect

$$Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + Line_i + \varepsilon_i$$

Phenotype of ith individual

Fixed effects: account for population structure

Observed SNP alleles of ith individual

Random error term

- $(Line_1, \ldots, Line_n) \sim \mathrm{MVN}(\mathbf{0},\ 2K\sigma_G^2)$
  - K = kinship matrix
- $Residuals \sim NID\ (0, \sigma_e^2)$

Measures relatedness between individuals

Yu et al. (2006)

Adapted from A. Lipka

Assumptions of the Unified MLM

- ***Residuals*~ *NID* $(0,\sigma_e^2)$**
  - **Normal**
  - **Independent**
  - **Equal Variance**

What do we do if these assumptions cannot be met?

(Example: Binomially distributed data)

Yu et al. (2006)

Adapted from A. Lipka

Binomial Distribution: # Successes in n Independent Success/Failure Trials

Mixed Logistic Regression does not require normality or equal variances

Conduct

Problem: Fitting this model is extremely computationally intensive!!!

$\alpha = fixed\ additive\ effect\ of\ the\ tested\ marker$

$x_i = observed\ genotype\ of\ tested\ marker\ for\ plant\ with\ i^{th}\ genotype$

$\begin{cases} 0, if\ aa \\ 1, if\ Aa\ or\ aA \\ 2, if\ AA \end{cases}$

$Random\ effect\ of\ the\ i^{th}\ genotype\ where$

$(Line_1, \ldots, Line_n) \sim MVN(0, 2K\sigma_G^2)$

K = kinship matrix

$\beta_k = fixed\ effect\ of\ the\ k^{th}\ principal\ component\ (PC)$

$PC_{ik} = value\ of\ the\ k^{th}\ PC\ for\ plant\ with\ i^{th}\ genotype$

$Block_j = fixed\ effect\ of\ the\ j^{th}\ block$

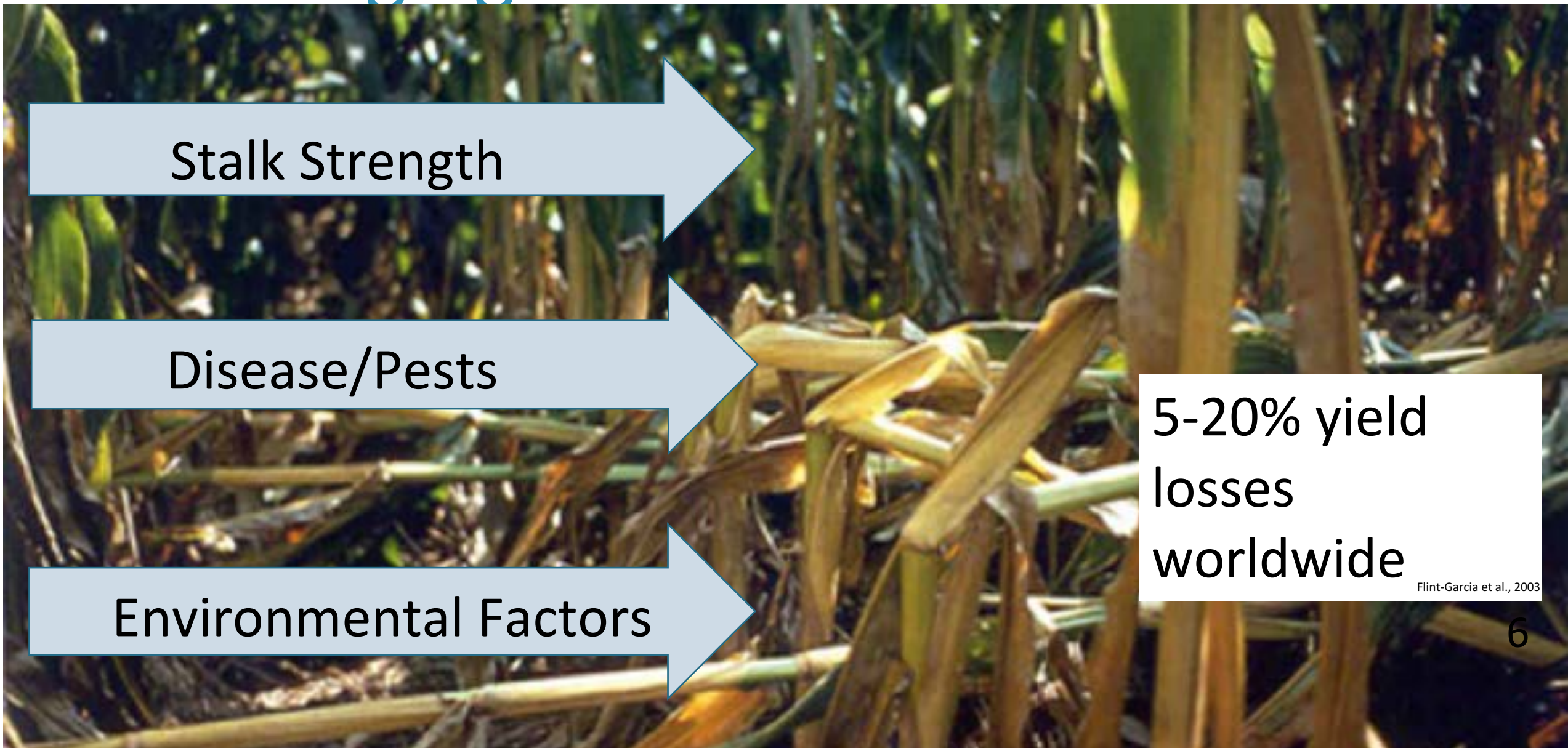
Logit Link function: the natural log-odds of a success

μ = the grand mean

Purpose

Develop a multi-model GWAS approach that will allow mixed model GWAS to be conducted on binomially distributed traits

Stalk Lodging In Maize

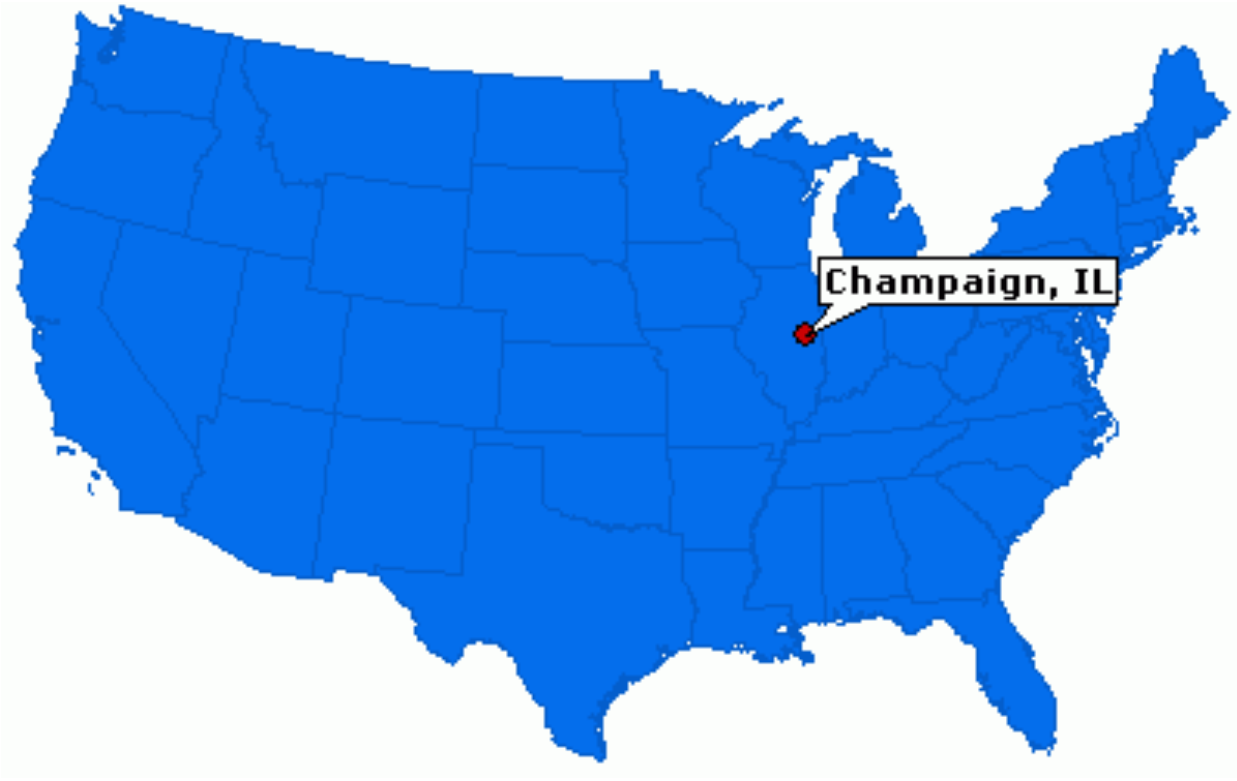


Data Collection- 2016

Two Reps of the Goodman-Buckler diversity panel were planted using incomplete block design

The entire experiment was inoculated with Goss's wilt

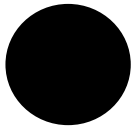
In this experiment there was no correlation between disease and lodging



The Jamann Lab at UIUC

Lodging Phenotyping

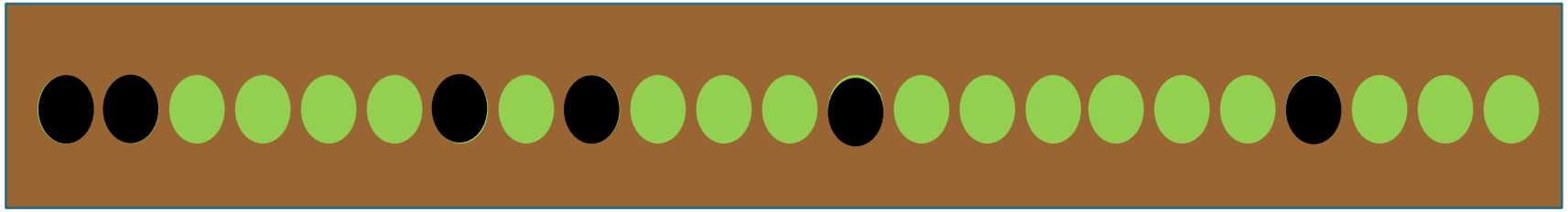
= 

= 

Standcount	Number of Plants Lodged	Number of plants Not lodged	Lodging Score (Percent Lodged)
23	6	17	26%

End of growing season

Beginning of Growing Season



Above: Diagram depicting one plot (rep) of one taxa in the field

Treat Lodging Data as a Binomial

Setup of Binomial	Why we think binomial is an appropriate approximation for lodging
The experiment consists of n repeated trials	Within each plot, each plant is a trial
Each trial has two outcomes: success or failure	**Success:** plant has lodged **Failure:** Plant has not lodged
The probability of success, π, is the same on every trial	The probability of a plant lodging, π, is the same within a plot
The trials are independent	One plant lodging will not change the likelihood of another plant lodging

Multi-Model Approach

Model 1

Fit Logistic Regression Model

Controls for population structure only

Identify peak SNPs

Model 3

Fit Mixed Logistic Regression Model

Using Peak SNPs from Model 1 and Model 2

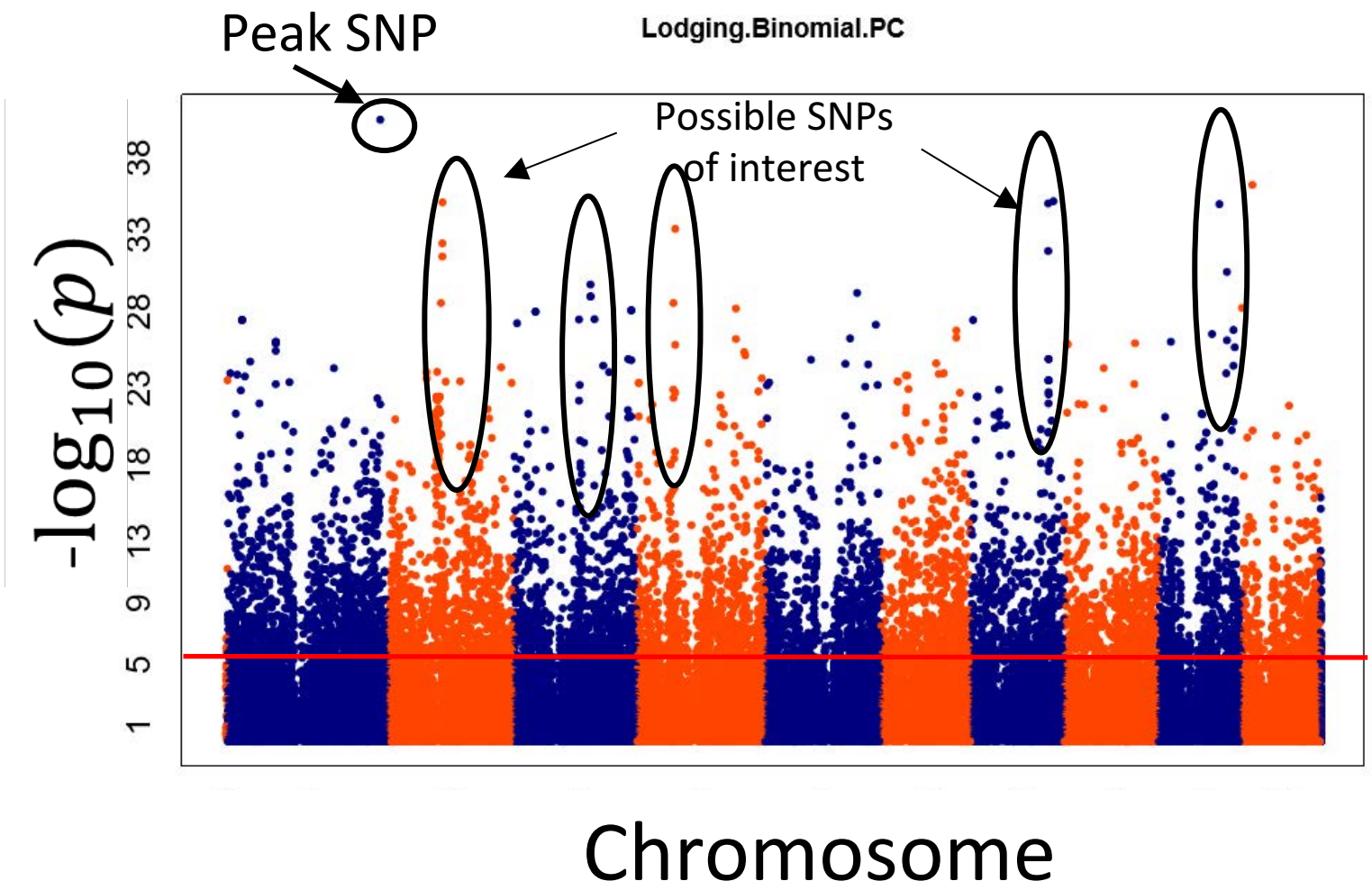
Model 2

Fit a Mixed Linear Model

Controls for population structure and relatedness

Identify peak SNPs

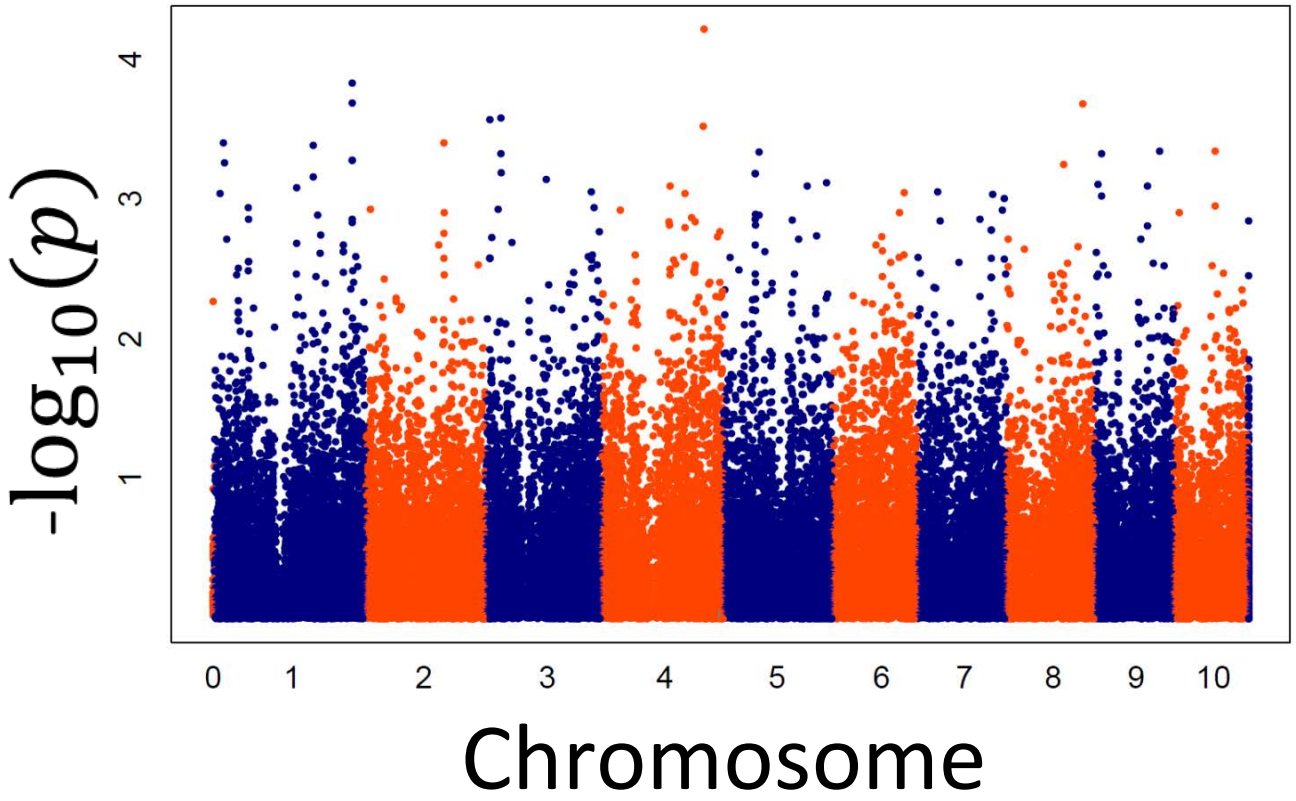
Logistic Regression Identified ~50% of Markers to be Significant



The top 2,796 SNPs from this model were subset

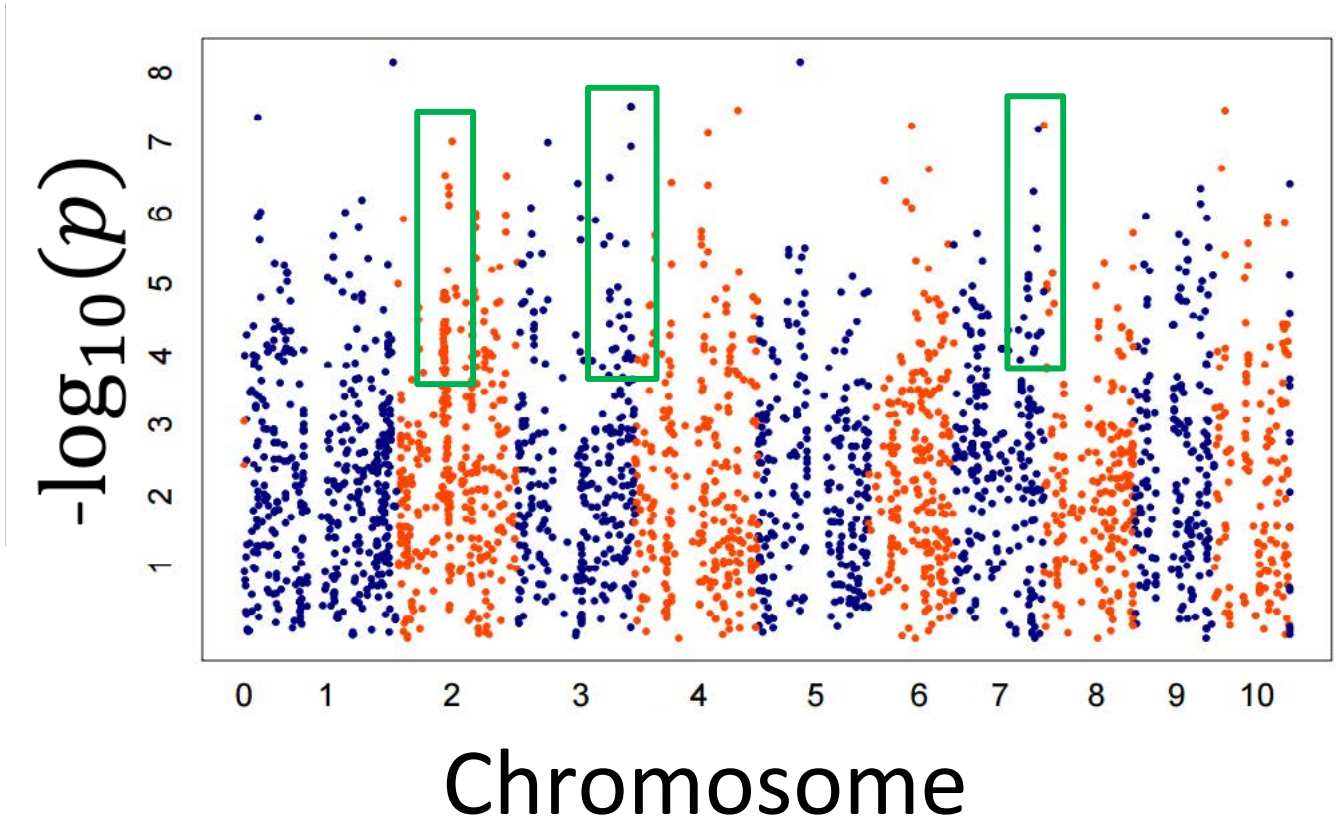
Motivation: mixed logistic regression model can fit 2,796 models in < 1 day

Unified MLM Identified No Significant Signals



GAPIT
Lipka et al., 2012

Mixed Logistic Regression Identifies 68% of SNPs Identified in Logistic Regression to Be Significant



Accounting for familial relatedness helped refine location of putative genomic regions

Signals coincide with those previously identified for traits related to lodging

SAS 9.4 PROC GLIMMIX

Simulation Study in Goodman-Buckler Diversity panel: Determine which parameters of the binomial distribution contribute the most to identification of genomic signals

Proposed Methodology
for Simulation Study

Assign SNP from 4K Set to be QTN

↓

Simulate binomial distributed trait

QTN Effect size | Stand count per plot | Grand mean

↓

For each trait in each setting:
Assessed genomic positions of "top 100" markers with strongest associations

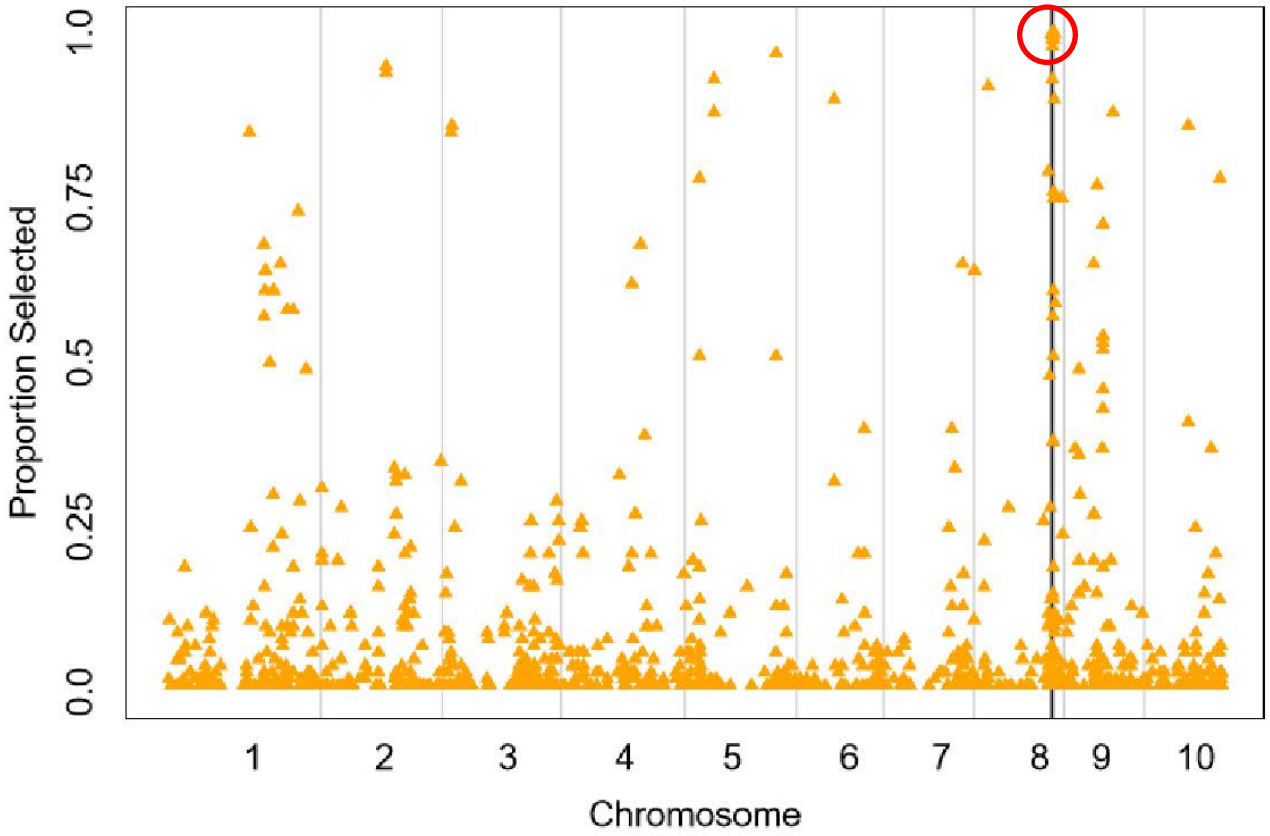
↓

Fit logistic regression model at each of 55K SNPs

How does the total number of plants in a plot affect QTN detection?

Stand Count: 10

Top 100 SNPs from each trait used to create this figure



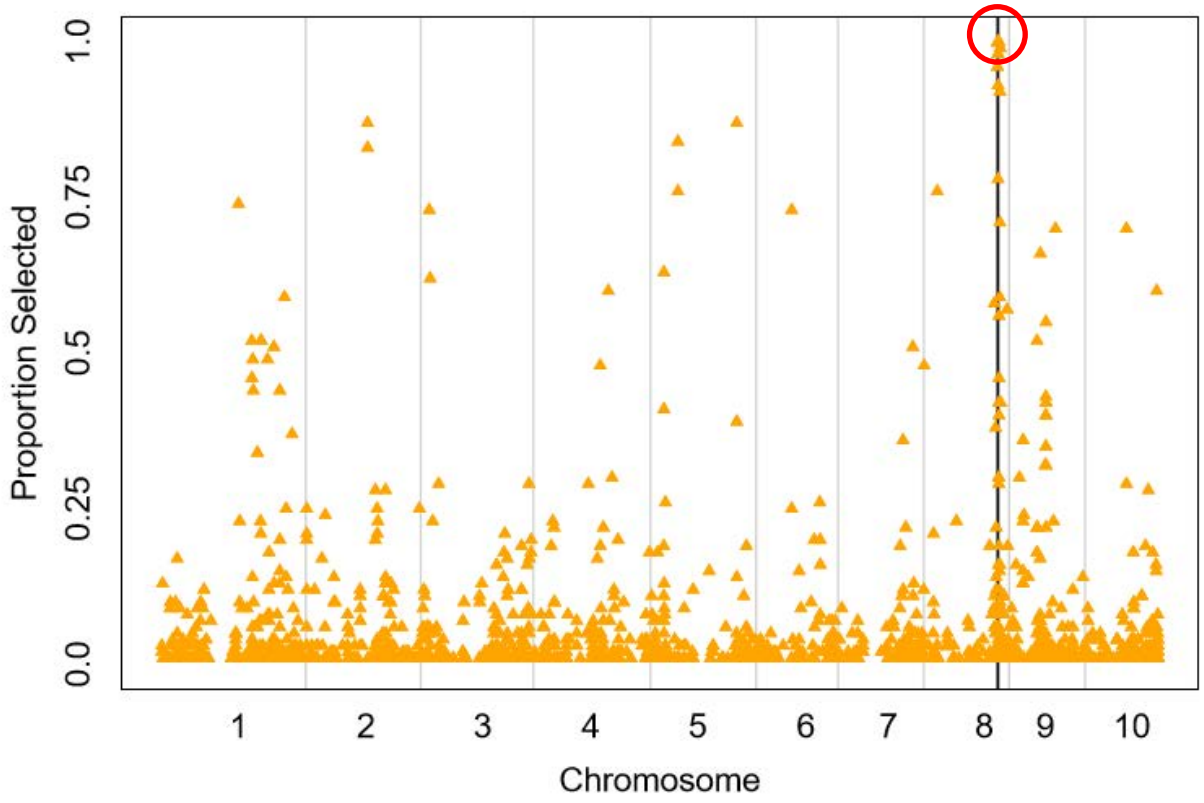
Proportion of times detected: 1.0

Model 1

How does the total number of plants in a plot affect QTN detection?

Stand Count: 15

Top 100 SNPs from each trait used to create this figure



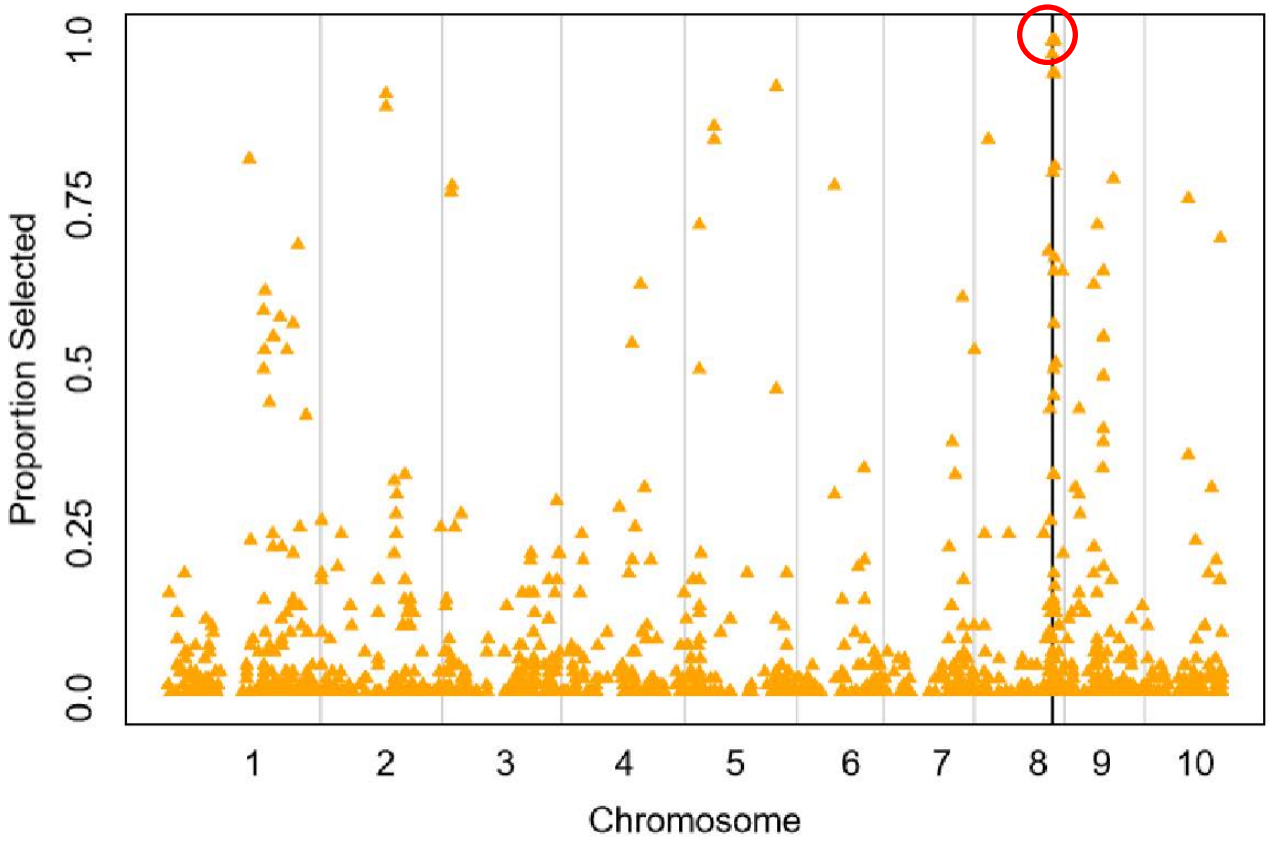
Proportion of times detected: 1.0

Model 1

How does the total number of plants in a plot affect QTN detection?

Stand Count: 20

Top 100 SNPs from each trait used to create this figure



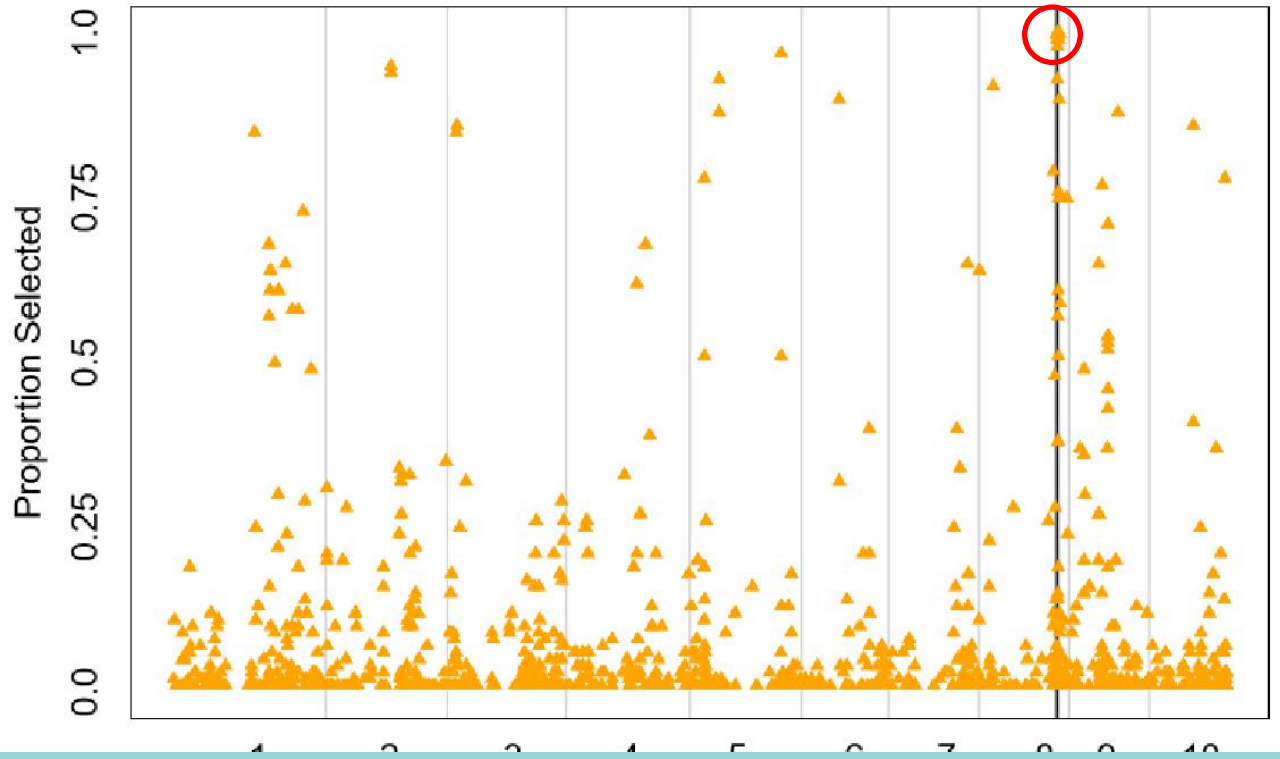
Proportion of times detected: 1.0

Model 1

How does the total number of plants in a plot affect QTN detection?

Stand Count: 25

Top 100 SNPs from each trait used to create this figure



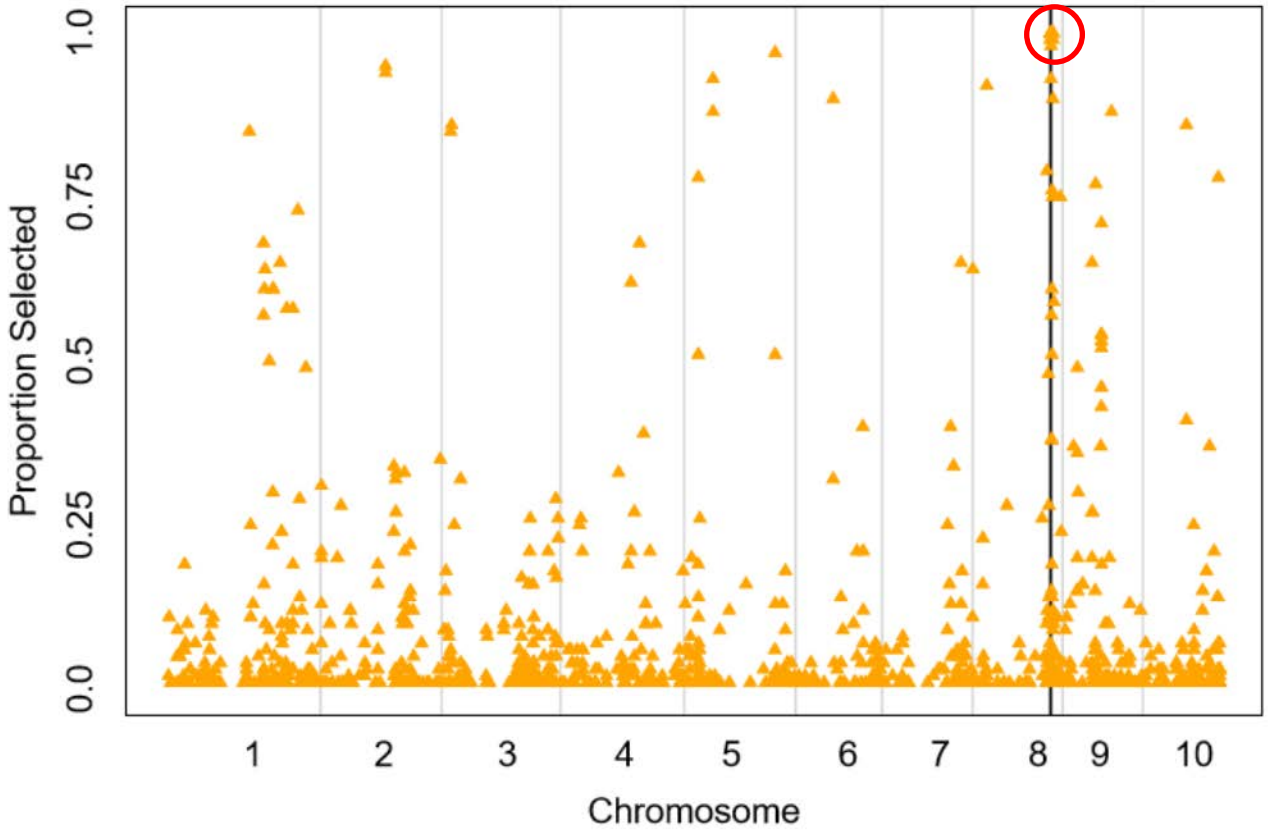
Proportion of times detected: 1.0

Model 1

Stand count does not appear to affect our ability to detect QTN

How does grand mean affect QTN detection?
Grand Mean = 0 ; P{Success} = 0.5

Top 100 SNPs from each trait used to create this figure



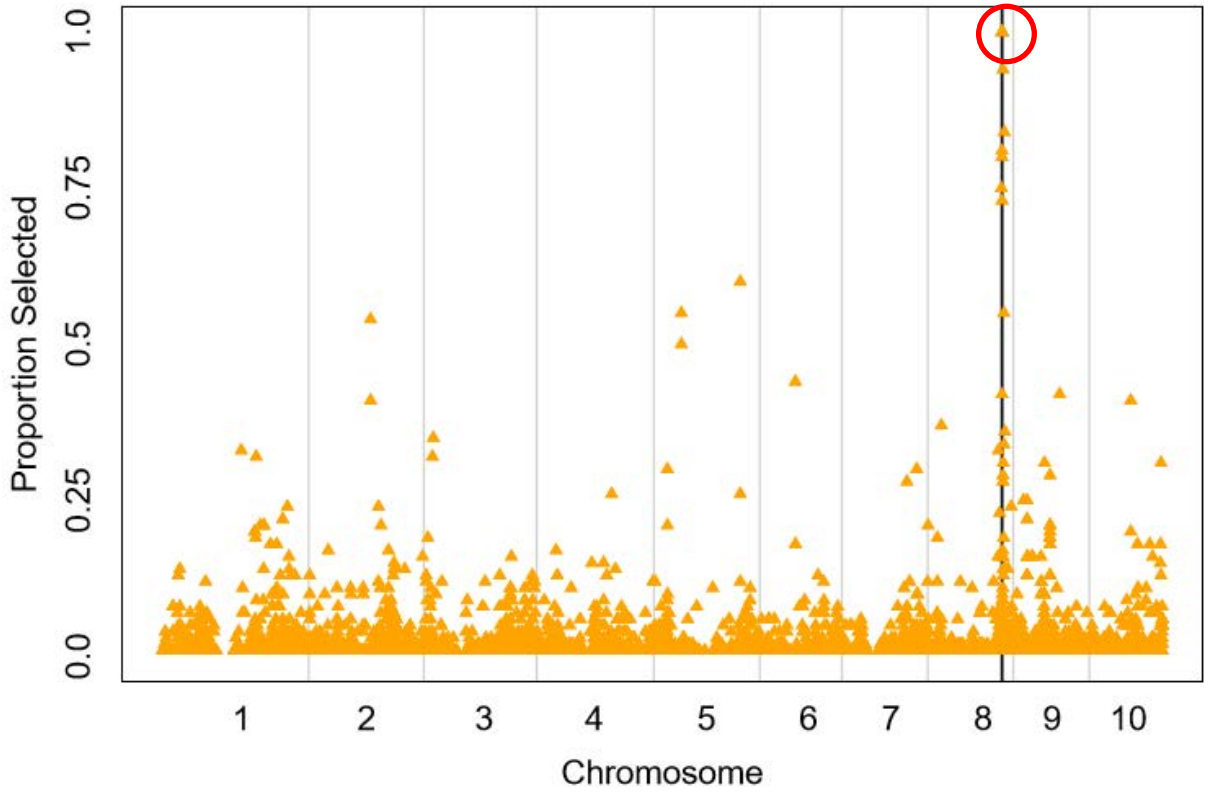
Proportion of times detected: 1.0

Model 1

How does grand mean affect QTN detection?

Grand Mean = 1; P[Success] = 0.73

Top 100 SNPs from each trait used to create this figure

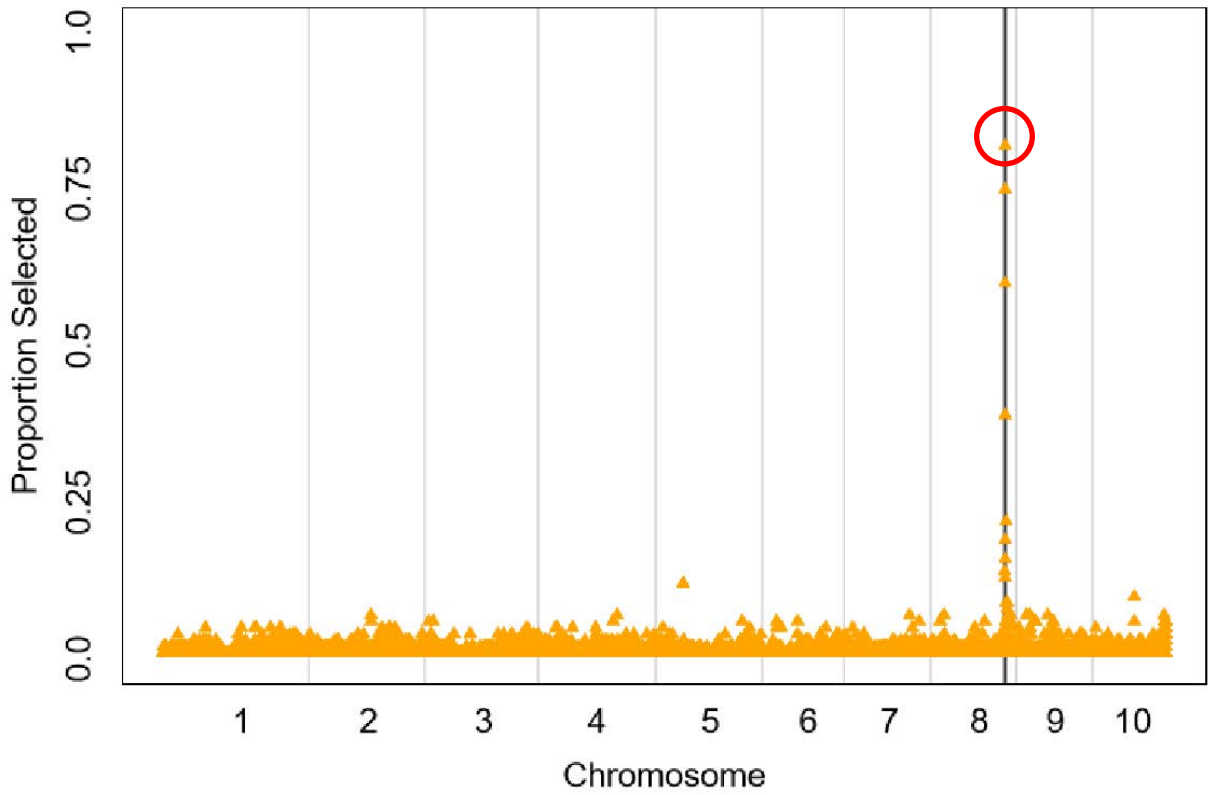


Proportion of times detected: 1.0

Model 1

How does grand mean affect QTN detection?
Grand Mean = 3; P{Success} = 0.95

Top 100 SNPs from each trait used to create this figure

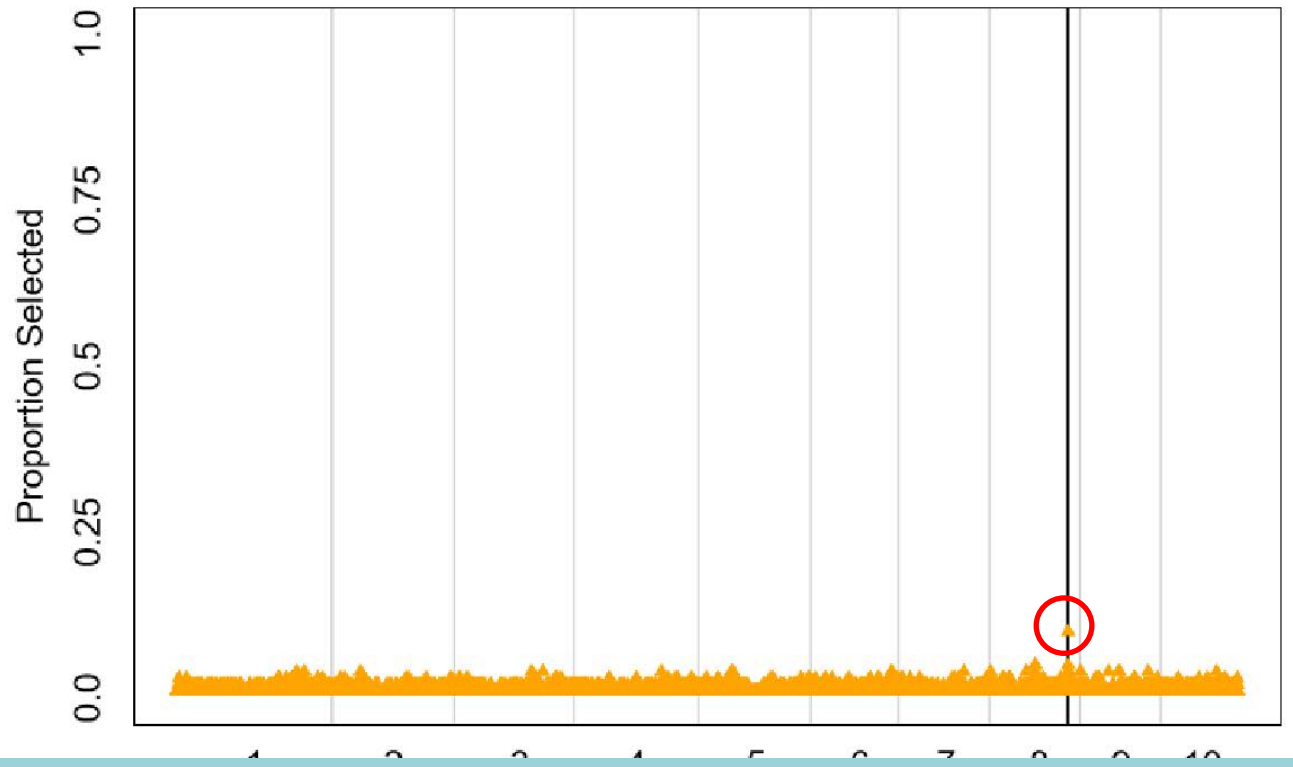


Proportion of times detected: 0.82

Model 1

How does grand mean affect QTN detection?
Grand Mean = 5; P{Success} = 0.99

Top 100 SNPs from each trait used to create this figure



Proportion of times detected: 0.10

Model 1

Grand mean values affects our ability to detect QTN

Future Directions

Any phenotype that measures # successes in a plot of *n* plants could theoretically use these approaches

- *Try to design experiments that result in a baseline probability of success of 0.5*

How can we fit mixed linear models in a computationally efficient manner on a Windows/Mac computer?

- *Temporary solution: multi-model approach is reasonable*
- *Try to strive for: write software that uses the score test*

Acknowledgements

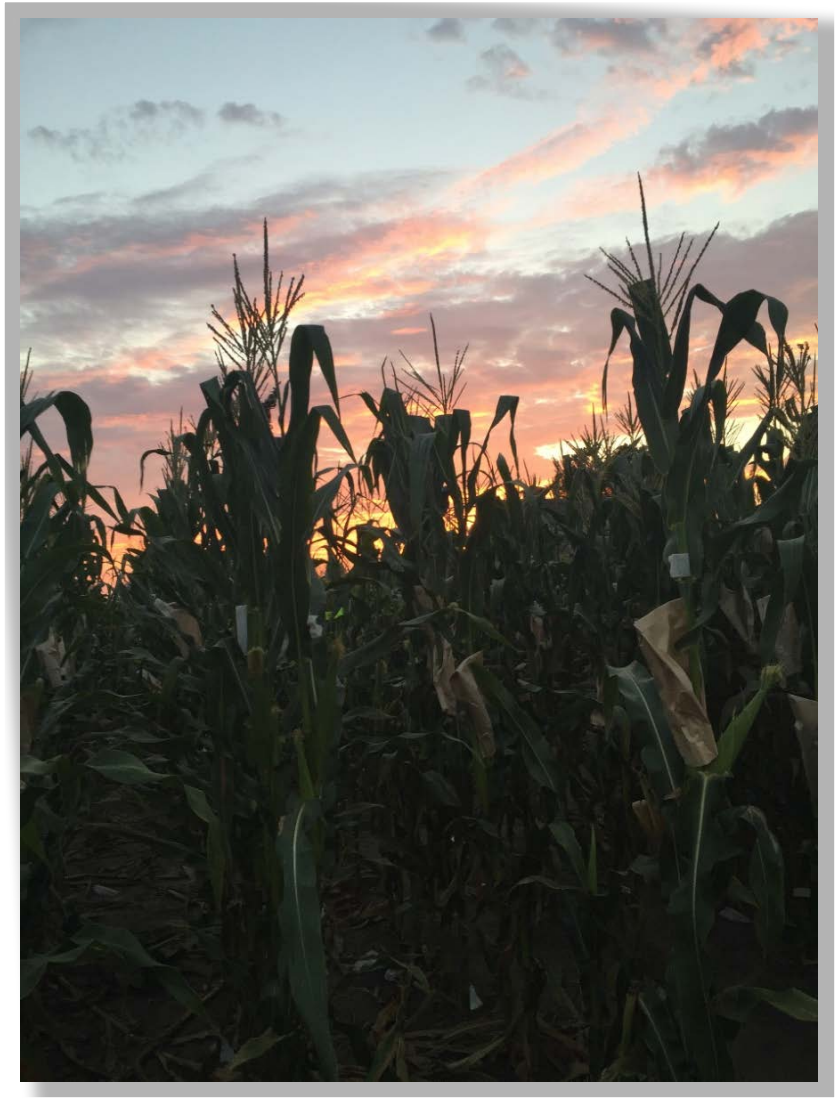
Committee Members
Dr. Alexander E. Lipka
Dr. Tiffany Jamann
Dr. Martin Bohn
Dr. Pat Brown

The Jamann Lab
Julian Cooper

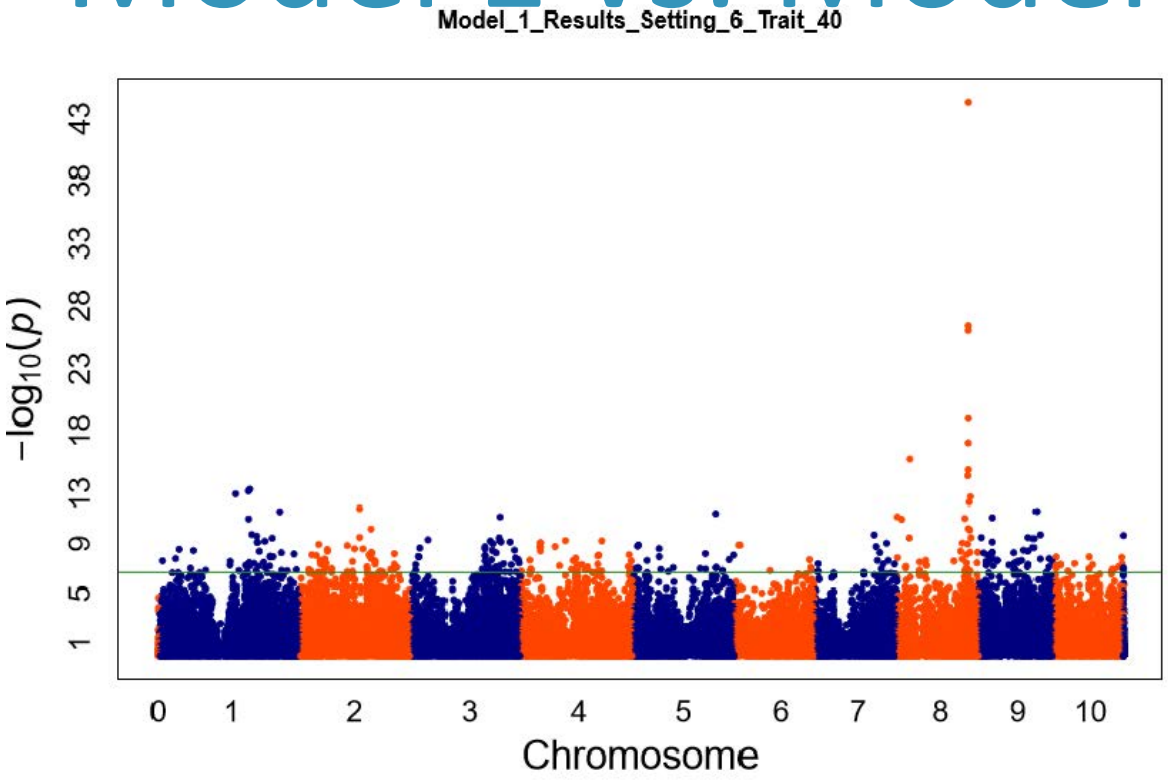
The Lipka Lab
Brian Rice
Angela Chen

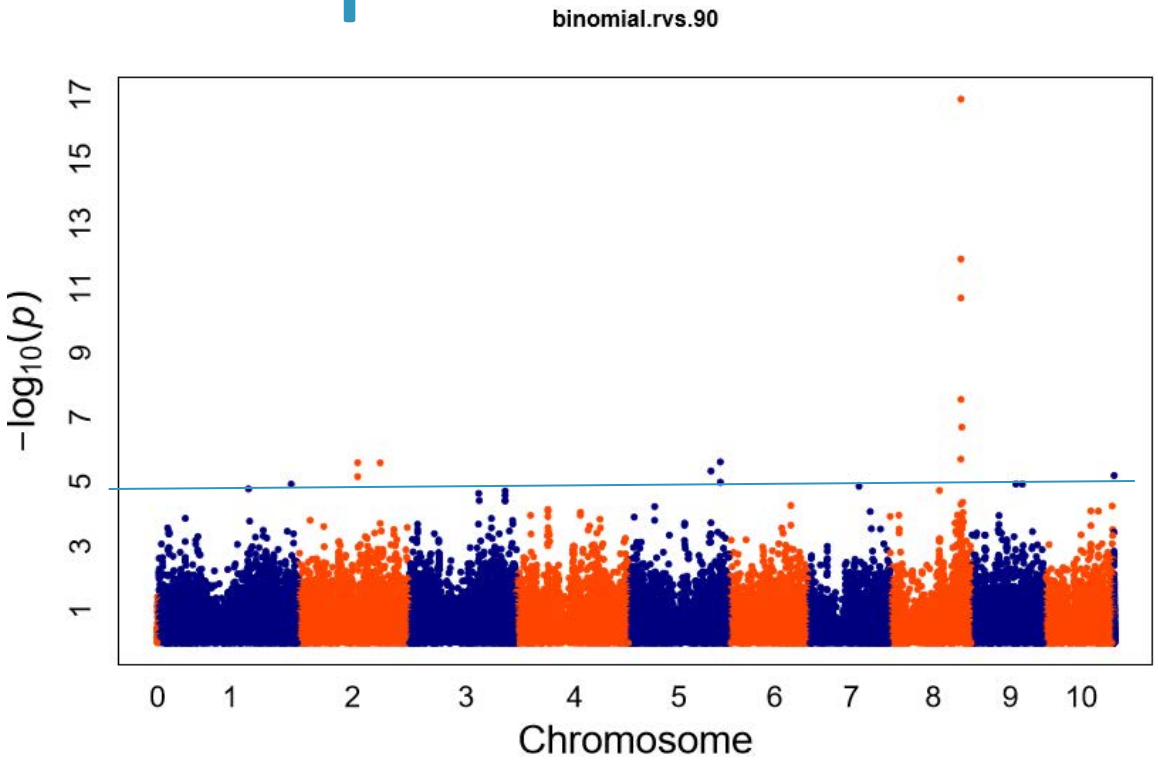
Graduate Students
Amanda Owings

Department of Crop Sciences at UIUC

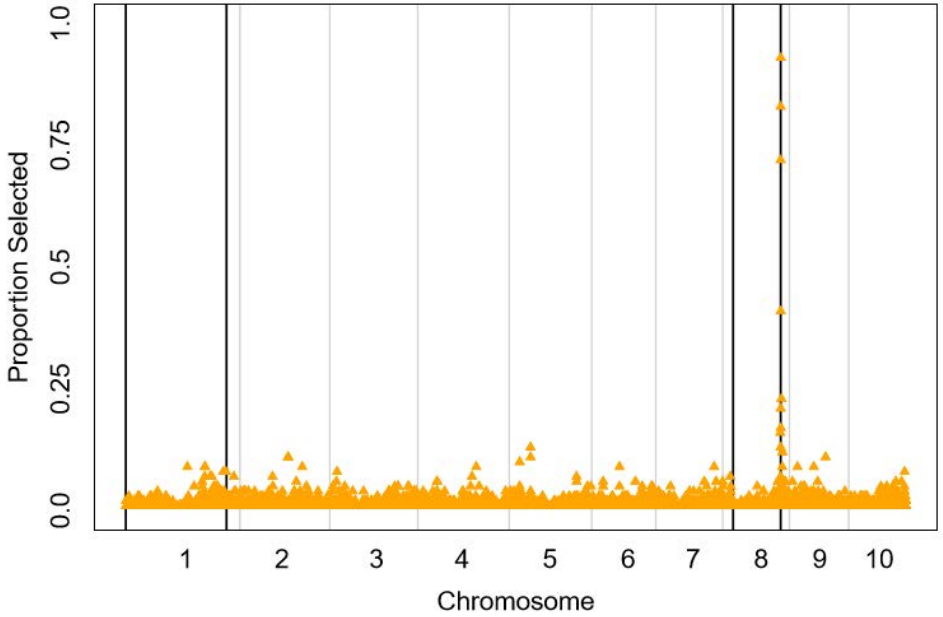


Model 1 vs. Model 2 Comparison



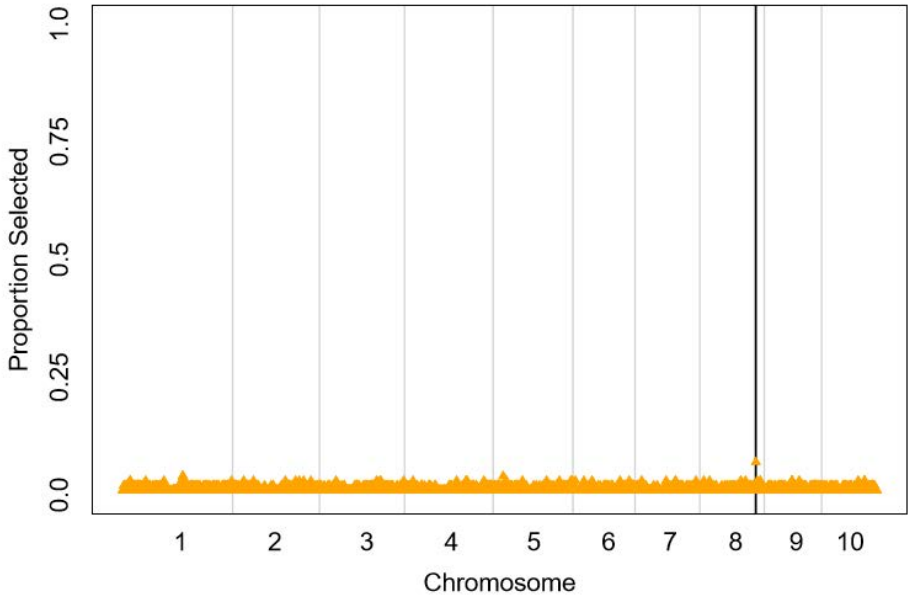


Varying Additive Effect Sizes (Same Assigned QTN)



Additive effect size 0.5 on chromosome 8

Proportion time detected: 0.93



Additive effect size 0.1 on chromosome 8

Proportion of times detected:0.07

Summary of Results

Able to identify two significant SNPs in the BP region of Maize Stalk Strength QTL

Li et al., 2014, Flint-Garcia et al., 2003, Hu et al., 2012

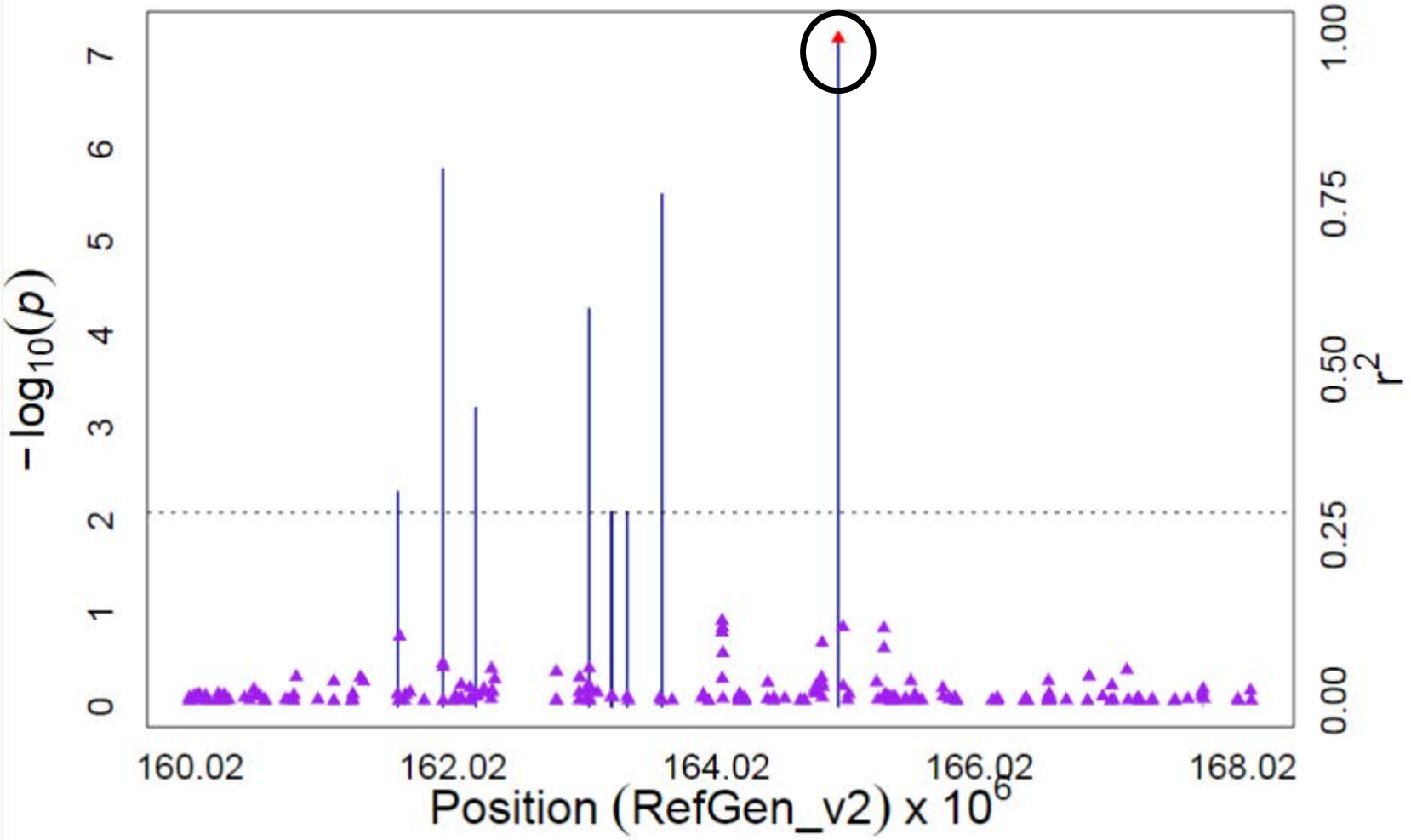
Peak SNPs on Chromosome 7 were in the same location as the most robust marker association with RPR

Pieffer et al., 2013

A significant SNP on Chromosome 1 was in the same region as a candidate gene for Mediterranean Corn Borer stalk destruction susceptibility

Samayoa et al., 2015

High LD Decay Observed AroundPeak SNP on Chromosome Seven



Limiting Factors of This Study

Stalk lodging is a putatively low heritability trait

- No repeatability across replications

Only one year of data included in this analysis

- Only one environment

Missing data

- Various factors contributed

Summary of Project

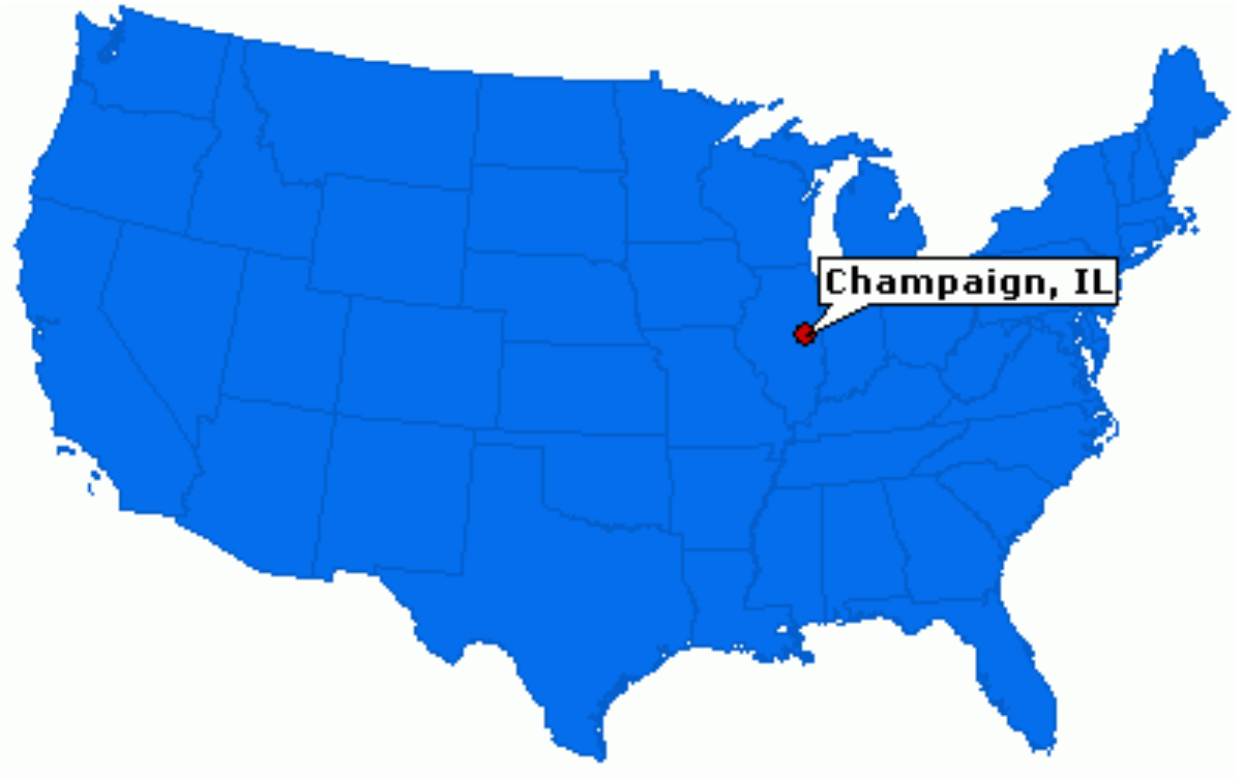
- Logistic Regression is computationally intensive
  - Approximately 30 seconds to run 1 SNP in SAS
  - ~17.36 days to run 50,000 SNPs
- Model 1 and Model 2 are used to identify which SNPs are fit using the complete logistic regression model (Model 3)
  - The number of SNPs to include is dependent on computational power available
- Stalk Lodging data was used to test this approach
  - Some Peak SNPs identified are in the same region as QTL associated with stalk strength, and a candidate gene for MCB Stalk Damage

Data Collection- 2016

2 Reps of the 282 diversity panel were planted using incomplete block design

The entire experiment was inoculated with Goss's wilt

In this experiment there was no correlation between disease and lodging



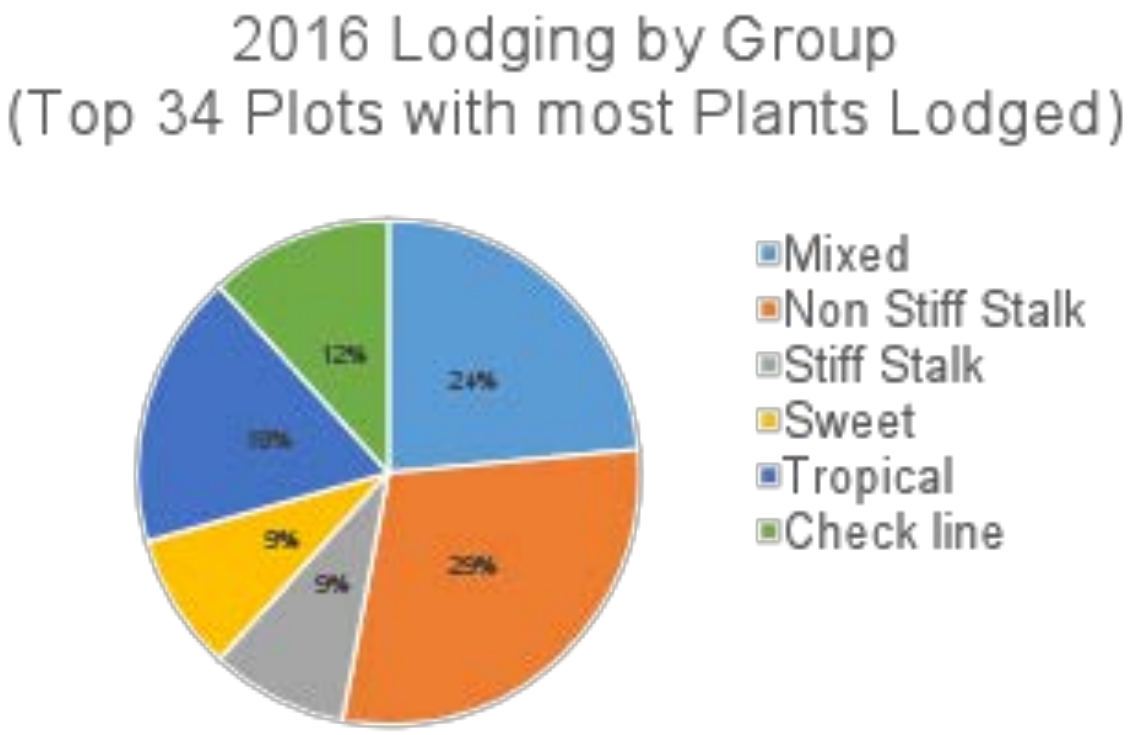
The Jamann Lab

Observed Lodging in the Field

Taxa classified as non stiff stalk were lodged more often

Taxa classified as stiff stalk were lodged the less often

All plots represented in this graph had at least 10 plants lodged

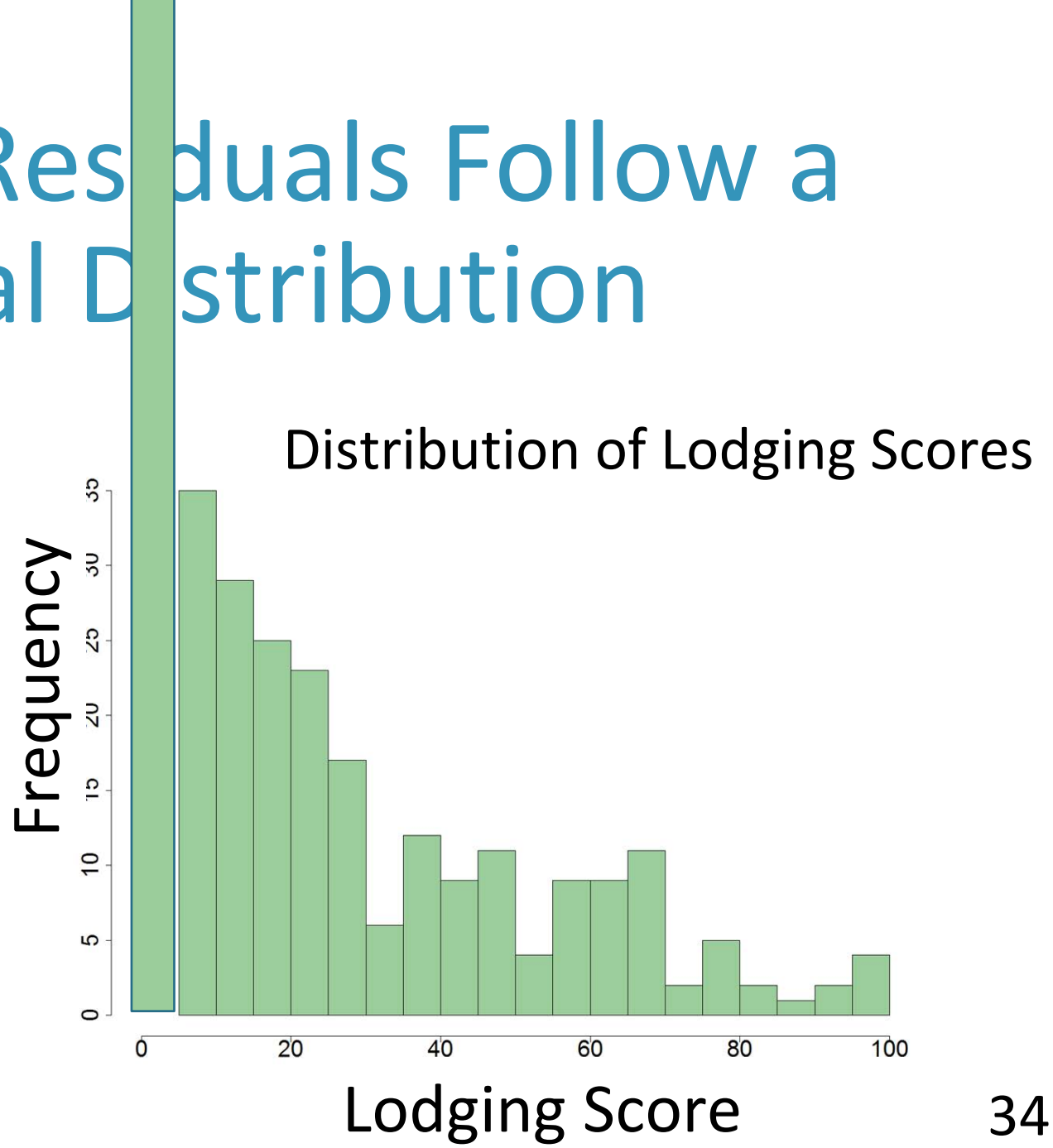


Lodging Score Residuals Follow a Non-Normal Distribution

The Box-Cox procedure was implemented, and λ=-0.6 was the suggested transformation

Transformation was unsuccessful

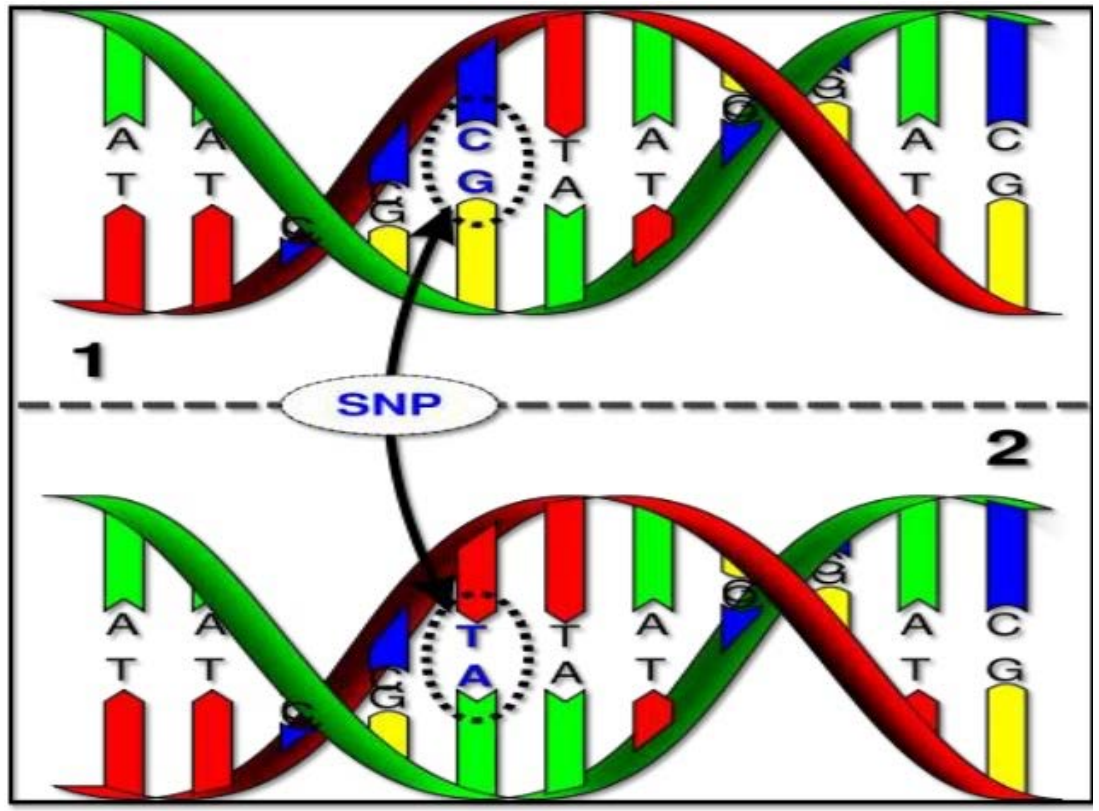
352 plots had no lodging



Genome-Wide Association Study (GWAS)

Search the genome for genetic markers significantly associated with your trait of interest

Allows for the identification of QTLs region of the genome associated with the trait

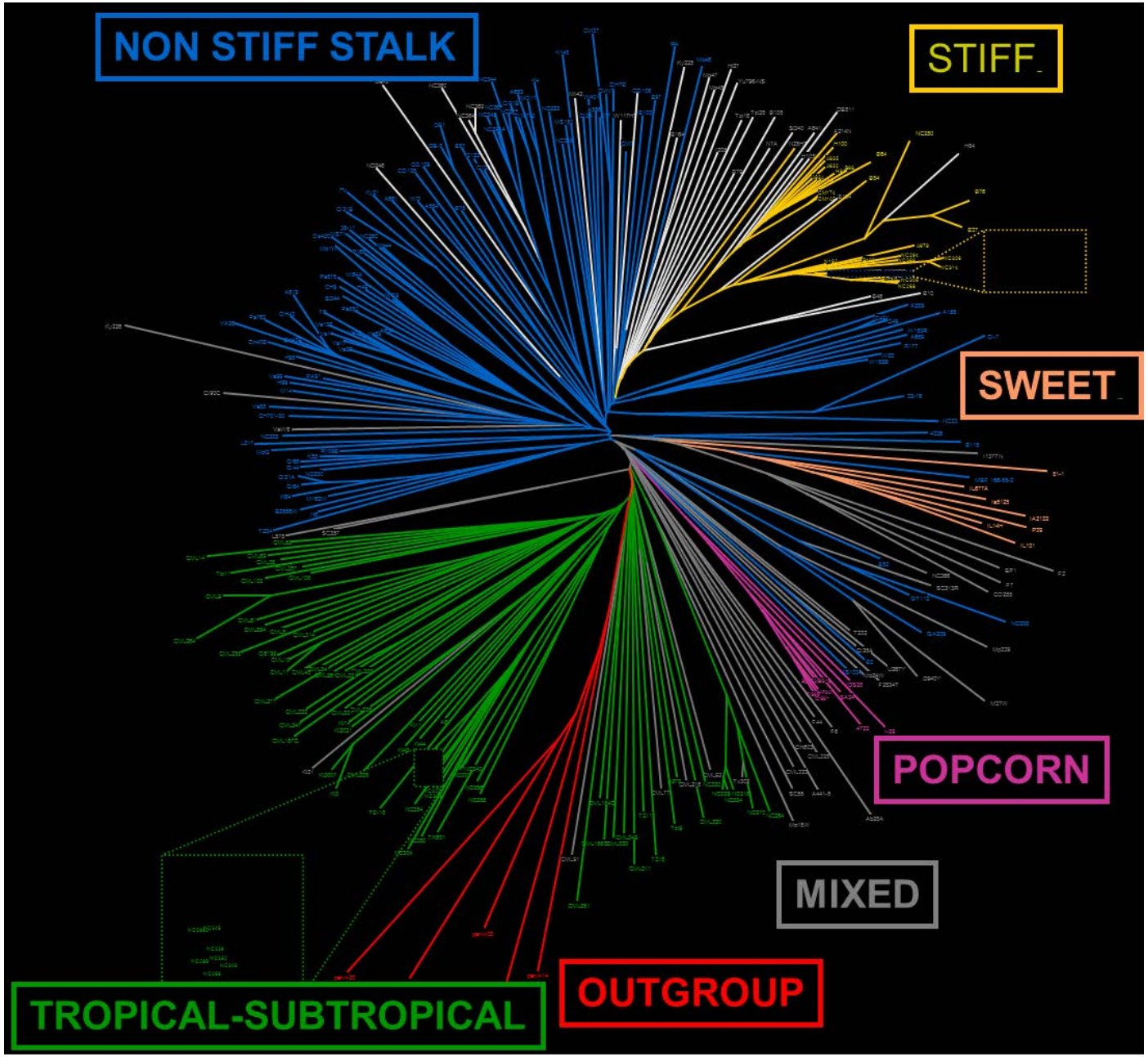


http://knowgenetics.org/snps/

Single Nucleotide Polymorphism (SNP): A type of genetic marker

282 Diversity Panel

~75% of all allelic diversity in Maize



Romay et al., 2013

Adapted from Flint-Garcia et al., 2005

Outline

Introduction

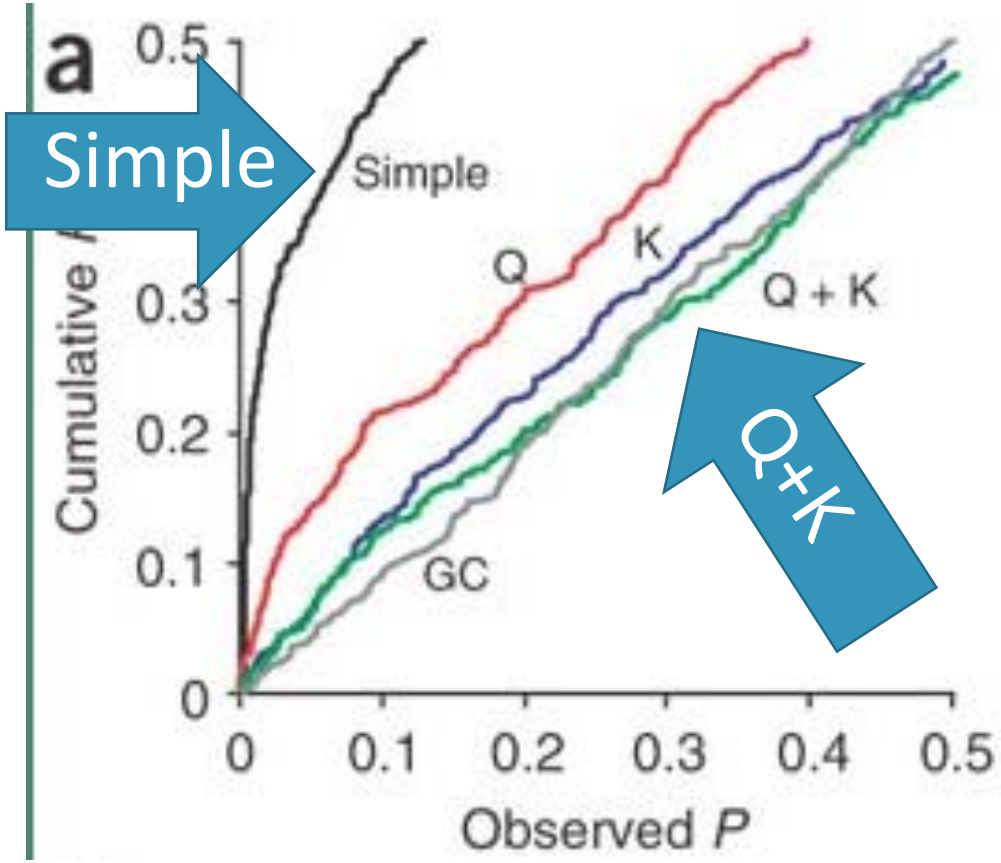
Genome-Wide Association on Stalk Lodging in Maize

Simulation Study

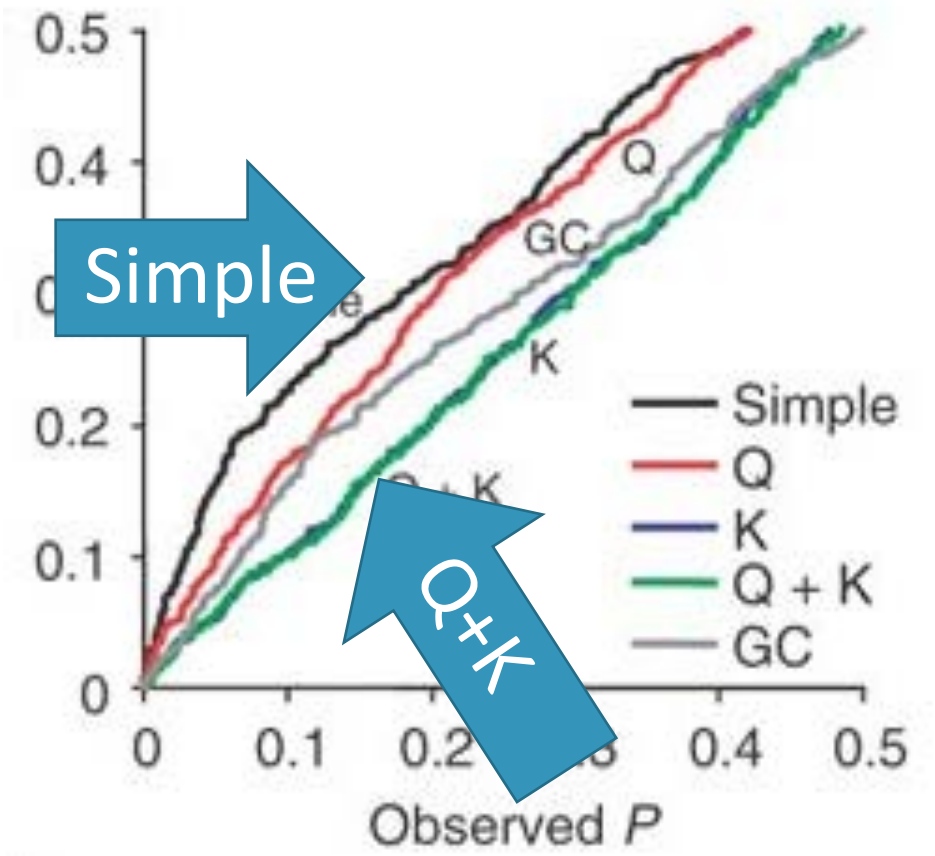
Conclusions

Unified Mixed Linear Model Controls for False Positives

Flowering time of Maize
(High population structure)



Ear Diameter of Maize
(Low Population Structure)



Yu et al., 2006

Stalk Lodging in Maize

Stanger and Lauer, 2006

Predicting lodging is challenging

Most methods are destructive and/or use other traits as proxies

Can phenotyping lodging still yield interesting results?

Binomial Data Allows for Logistic Regression

Y_i are independent binomial random variables with expected values

$$E\{Y_i\} = n * \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)$$

and variance of

$$\boxed{Var(Y_i) = n * \pi}(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged)\ (1 - \pi(plant\ with\ genotype\ i\ in\ block\ j\ has\ lodged))$$

Methods

One SNPs from 4K marker set was assigned to be QTN

Taxa from the 282 diversity panel were simulated to experience lodging

The 55K marker set was used to genotype the taxa used in the simulation

Objectives

Evaluate the efficacy of the three model approach to mixed logistic regression

Evaluate the use of the diversity panel for use in logistic regression GWAS

Examine how variables within the data set effect the ability to detect a QTN

Simulation Study Settings

Setting	Grand Mean	Stand Count	Additive effect size
1	0	10	0.9
2	1	10	0.9
3	3	10	0.9
4	5	10	0.9
5	0	15	0.9
6	0	20	0.9
7	0	25	0.9

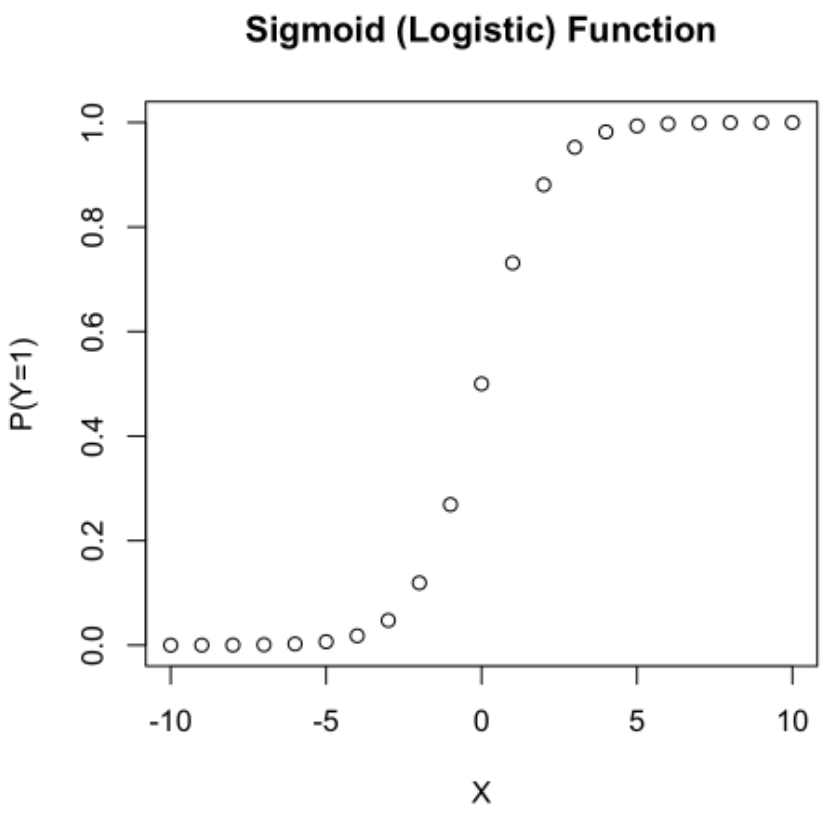
Model 1 identifies Peak SNPs While Accounting for Population Structure

$$\log\left(\frac{\pi(number\ lodged)}{\pi(not\ lodged)}\right) = \beta_o + PCs + \alpha x + Blocks$$

Intercept Affects the Baseline Trait Probability $\pi(x_i)$

Sigmoid Function:

$$\pi(x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$



What does changing the intercept do to our data?

$\beta_o = 5$

4722	10	10	10	10
33-16	10	10	10	10
38-11	10	10	10	10
A188	10	10	10	10
A239	10	10	10	10
A441-5	10	10	10	10
A554	10	10	10	10
A6	10	10	10	10
A619	10	10	9	10
A632	10	10	10	10
A634	10	10	10	10
A635	10	10	10	10
A641	10	10	10	10
A654	10	10	10	10
A659	10	10	10	10
A661	10	10	10	9

$\pi(x_i) = 0.99$

$\pi(x_i) = 0.73$

$\beta_o = 1$

4722	10	5	5	6	7
33-16	10	9	10	10	10
		10	9	10	10
		10	9	10	9
		9	9	9	7
		10	10	8	9
		8	5	5	8
		8	6	9	7
A654	10	10	10	9	10
A659	10	9	6	8	7

Model 3 Failed to Converge in SAS Proc GLIMMIX

Possible reasons for this failure:

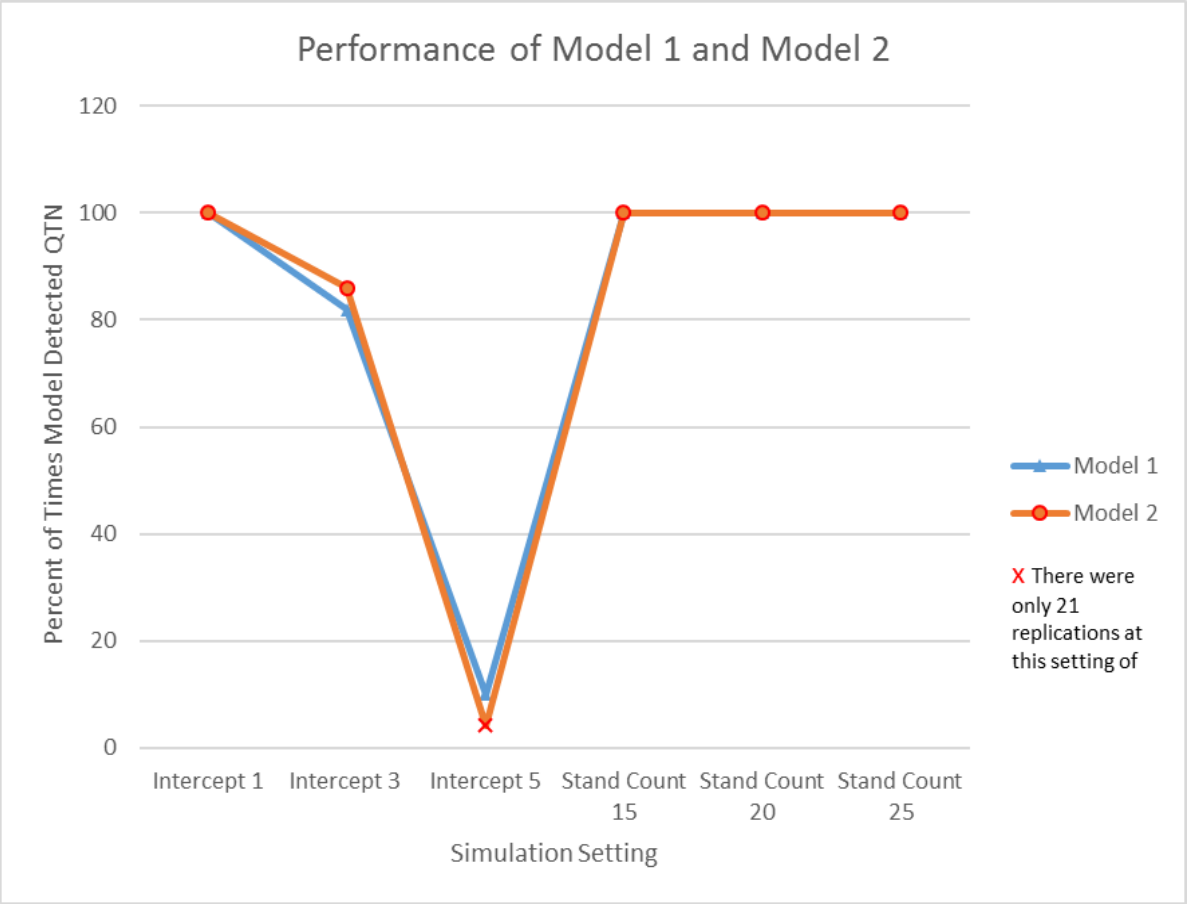
- "there was not enough variation in the response to attribute any variation to the random effect"
- Estimated G matrix is not positive definite: "procedure converged to a solutions where the variance of the random effect is 0"

Alternative Solution:

- Use the GMMAT package (Chen et al. 2015) (Only runs on UNIX OS)

Model 2

- Model 2 may have had enough power to successfully detect QTN despite model assumptions being violated
- Previous studies have shown that linear models can sometimes be approximated by logistic regression models



Conclusion

- Traditional GWAS requires normal data
- Logistic regression has the potential to analyze non-normally distributed traits
- The biggest limitation of using logistic regression is the computational power required
- Simulation Study show the need for increased variability of phenotypic data- this is especially hard to achieve in a binary trait

Model 1 identifies Peak SNPs While Accounting for Population Structure

$$\log\left(\frac{\pi(number\ lodged)}{\pi(not\ lodged)}\right) = \beta_o + PCs + \alpha x + Blocks$$

Binomial Data Allows for Logistic Regression

Logistic Regression does not require normality or equal variances

Conduct GWAS by fitting a logistic regression model at each SNP

$$\log\left(\frac{\pi(\text{plant with genotype } i \text{ in block } j \text{ has lodged})}{\pi(1-\pi(\text{plant with genotype } i \text{ in block } j \text{ has lodged})}\right) = \mu + \sum_{k=1}^{3} \beta_k PC_{ik} + \alpha x_i + Line_i + Block_j$$

$\alpha = \text{fixed additive effect of the tested marker}$

$x_i = \text{observed genotype of tested marker for plant with } i^{th} \text{ genotype}$

$= \begin{cases} 0, & \text{if } aa \\ 1, & \text{if } Aa \text{ or } aA \\ 2, & \text{if } AA \end{cases}$

$Block_j = \text{fixed effect of the } j^{th} \text{ block}$

$Line_i = \text{Random effect of the } i^{th} \text{ genotype where}$

$(Line_1, \ldots, Line_n) \sim MVN(0, 2K\sigma_G^2)$

K = kinship matrix

$\beta_k = \text{fixed effect of the } k^{th} \text{ principal component } (PC)$

$PC_{ik} = \text{value of the } k^{th} PC \text{ for plant with } i^{th} \text{ genotype}$

Logit Link function: The natural log-odds of a plant is lodged or not lodged

The grand mean

Model 2 Identifies Peak SNPs While Controlling for Population Structure and Relatedness

Grand Mean

Marker effect

Random effects: account for familial relatedness

$$Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha x_i + Line_i + \varepsilon_i$$

Phenotype of ith individual

Fixed effects: account for population structure

Observed SNP alleles of ith individual

Random error term

$$(Line_1, \ldots, Line_n) \sim \mathrm{MVN}(\mathbf{0},\ 2\mathrm{K}\sigma_G^2)$$

K = kinship matrix

Measures relatedness between individuals

$$Residuals \sim NID\ (0, \sigma_e^2)$$

Yu et al. (2006)

Adapted from A. Lipka

Model 3 is Fit Using Subset of Peak SNPs

$$log\left(\frac{\pi(number\ lodged)}{\pi(not\ lodged)}\right) = \beta_o + PCs + \alpha x + \text{Individuals+Blocks}$$

SAS 9.4
PROC
GLIMMIX

Model 3 is fit using top SNPs from Model 1

Recommendation: Number of SNPs that can be run in approximately 24 hours

Results of Simulation Study in Context of Stalk Lodging Data

- It is possible that our model's ability to accurately detect QTL was compromised because of an observed low rate of lodging
- Can we control
- If this baseline probability occurs, then the inability of our model to detect QTL may have been exacerbated by an intercept value that is far removed 0.

Peak SNPs that Coincide with Signals Associated with Related Traits				
Type of Region identified	Chr	Location in Literature	Location in Model 3	Notes
Marker	7	159.4 Mb	161.9 Mb 155.8 Mb 164.9 Mb	Three most significant SNPs on Chr 7
qRPR2 QTL	2	236.4-237.0 Mb	236.8 Mb	14th most significant SNP on Chr 2
qRPR3-1 QTL	3	181.1 Mb-184.7	181.7 Mb 182.0 Mb	92nd and 98th most significant SNP On Chr 3