An Introduction to Quantitative Genetics: Using the Single Locus Model to Understand the Foundations of Genome-Wide Association and Genomic Selection

Nicholas Santantonio

February 25th, 2020

Introductory plant breeding course

Section on quantitative genetics

- Basics of quantitve traits and selection
- Approx. 1/3 of course

Expectations from students at this point

- Understand basics of Mendelian inheritance
- Familar terms
  - gene
  - allele
  - dominance

Introduction

Quatitative genetics is statistical language

Introduction

Quatitative genetics is statistical language

- Population parameters are *estimated* from a sample of the population
    - Allele and genotypic frequencies
    - Gene “effects”
    - Means and variances

Introduction

Quatitative genetics is statistical language

- Population parameters are *estimated* from a sample of the population
    - Allele and genotypic frequencies
    - Gene “effects”
    - Means and variances

What does this language describe?

- Inheritance of traits, continuous and discrete
- Changes of a population through time

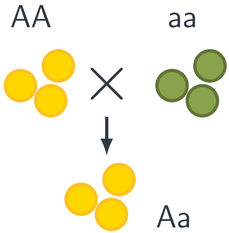
Gregor Mendel

Experiments in Plant Hybridization (1866)

- Single gene inheritance
- Qualitative traits
- “Complete” Dominance
- Independent assortment

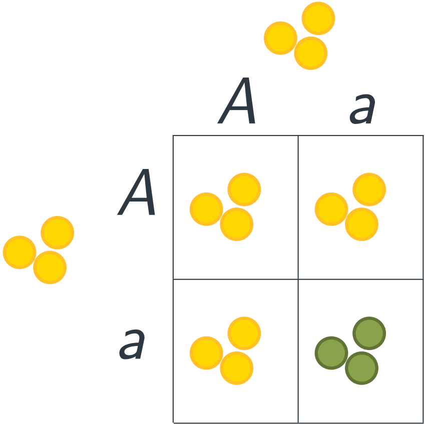
Lost for 34 years, rediscovered in 1900

“Complete” Dominance

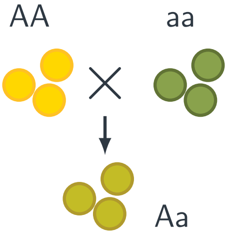


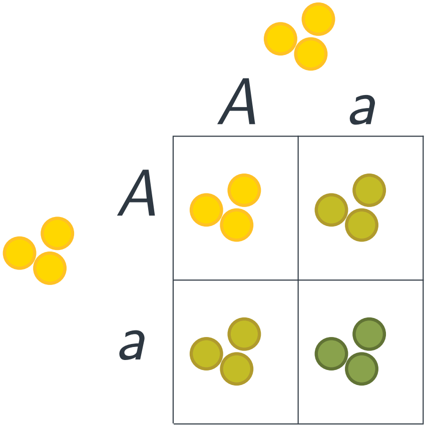
Mendel observed some traits were “hidden”

- reappeared when recombined
- inheritance of factors, i.e. “genes”



“Incomplete” Dominance





Not true for all traits

- some seemed to blend

Two waring sides

Mendelians

William Bateson

Biometricians

Karl Pearson

Two waring sides

Mendelians

William Bateson

Quantitative Genetics is born

Ronald Fisher

Biometricians

Karl Pearson

Two waring sides

Mendelians

William Bateson

Quantitative Genetics is born

Ronald Fisher

Biometricians

Karl Pearson

- 1918. The Correlation between Relatives on the Supposition of Mendelian Inheritance
- Used Mendelian genetics to explain continuous variation

Qualitative vs Quantitative traits

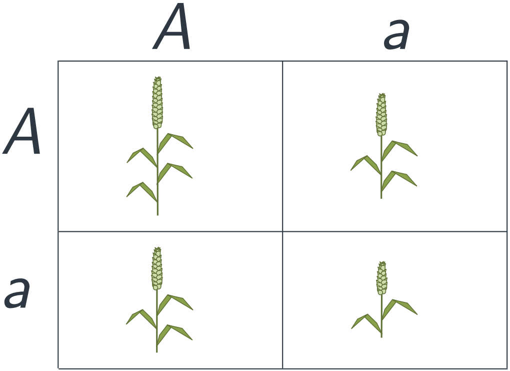
- Qualitative traits
  - yellow / green
  - tall / short
  - early / late
  - high / low yielding

- Quantitative traits
  - chlorophyll (g)
  - plant height (cm)
  - days to flowering
  - bushels acre^{-1}

We will treat the genetics of continuous traits as a (linear) mathematical problem!

Additive Single Locus

Additive effects increase linearly with the total number of alleles



The Single Locus Model

Phenotype = Genotype + Environment

The Single Locus Model

Phenotype = Genotype + Environment

$$y_{ij} = G_i + e_{ij}$$

- Genotype $= G_i =$ genetic effect of the i^{th} individual
- residual $= e_{ij} =$ some deviation from the genetic effect

The Single Locus Model

Phenotype = Genotype + Environment

$$y_{ij} = G_i + e_{ij}$$

- Genotype $= G_i =$ genetic effect of the i^{th} individual
- residual $= e_{ij} =$ some deviation from the genetic effect

We will begin with the assumption that only one locus effects our phenotype

The Single Locus Model II - Matrix notation

$= AA$

$= Aa$

$= aa$

The Single Locus Model II - Matrix notation

$= AA$ $\qquad y_1 = x_1\beta_a + e_1$

$= Aa$ $\qquad y_2 = x_2\beta_a + e_2$

$= aa$ $\qquad y_3 = x_3\beta_a + e_3$

The Single Locus Model II - Matrix notation

$= AA$ $\qquad y_1 = x_1\beta_a + e_1$

$= Aa$ $\qquad y_2 = x_2\beta_a + e_2$

$= aa$ $\qquad y_3 = x_3\beta_a + e_3$

$$\mathbf{y} = \mathbf{x}_a\beta_a + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} aa \\ Aa \\ Aa \\ \vdots \\ AA \end{bmatrix} \begin{bmatrix} \beta_a \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 2 \end{bmatrix} \begin{bmatrix} \beta_a \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

The Single Locus Model III - Dominance

$$\mathbf{y} = \mathbf{x}_a \beta_a + \mathbf{x}_d \beta_d + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} aa & \text{hom} \\ Aa & \text{het} \\ Aa & \text{het} \\ \vdots & \vdots \\ AA & \text{hom} \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

The Single Locus Model III - Dominance

$$\mathbf{y} = \mathbf{x}_a\beta_a + \mathbf{x}_d\beta_d + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} aa & \text{hom} \\ Aa & \text{het} \\ Aa & \text{het} \\ \vdots & \vdots \\ AA & \text{hom} \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

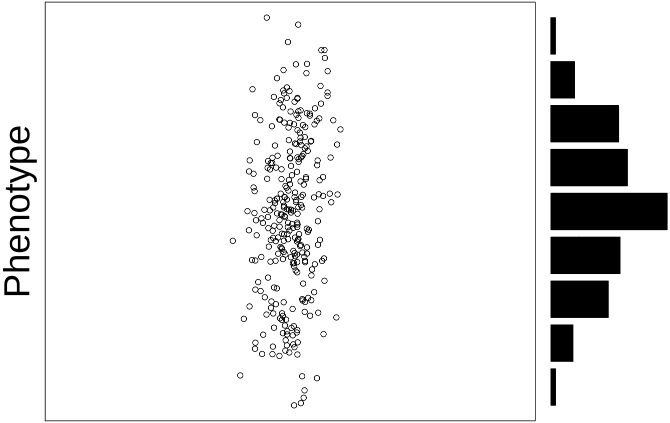
The Single Locus Model III - Dominance

$$\mathbf{y} = \mathbf{x}_a\beta_a + \mathbf{x}_d\beta_d + \mathbf{e}$$

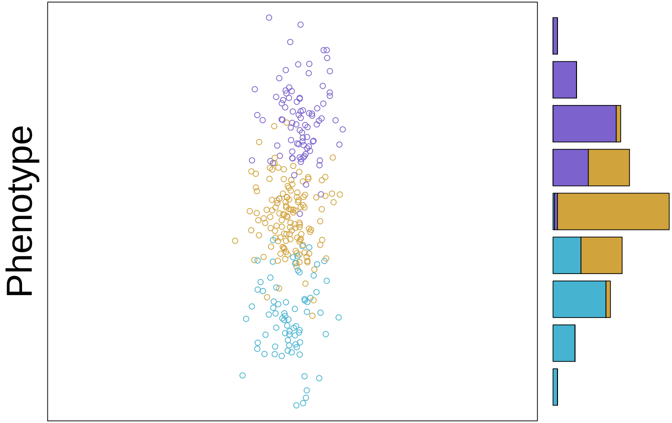
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} aa & \text{hom} \\ Aa & \text{het} \\ Aa & \text{het} \\ \vdots & \vdots \\ AA & \text{hom} \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

Lets see how this works...

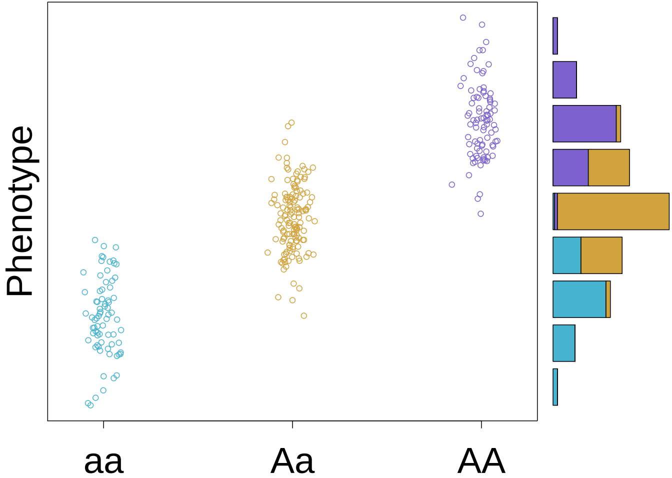
Additive only



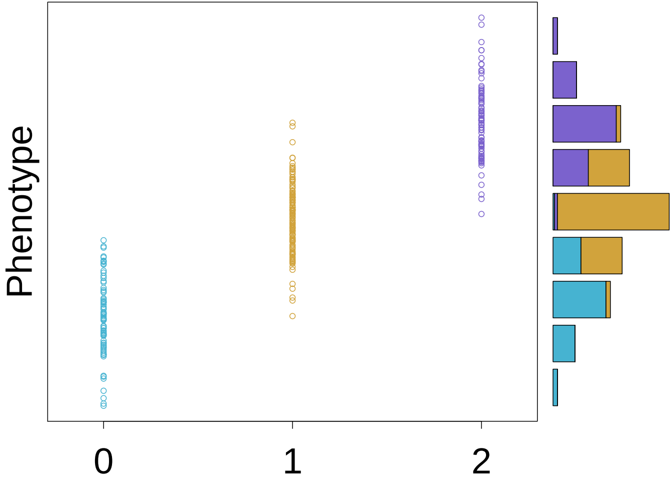
Additive only



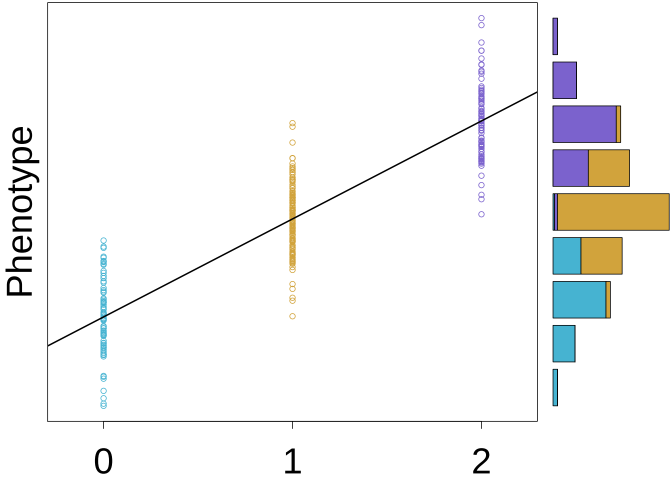
Additive only



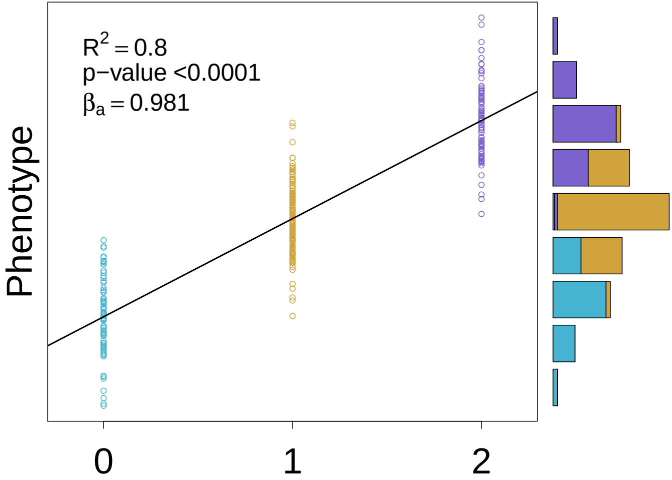
Additive only



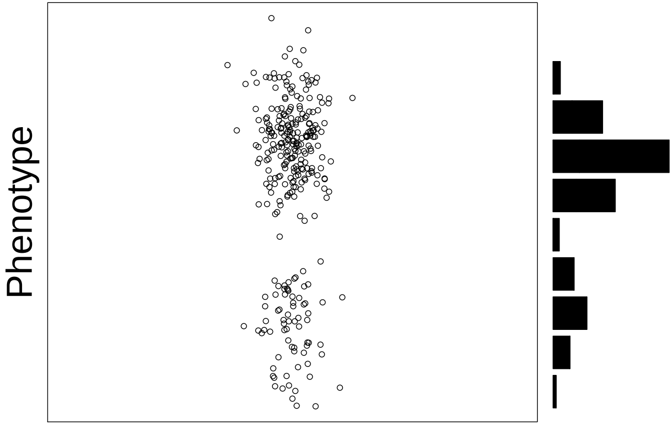
Additive only



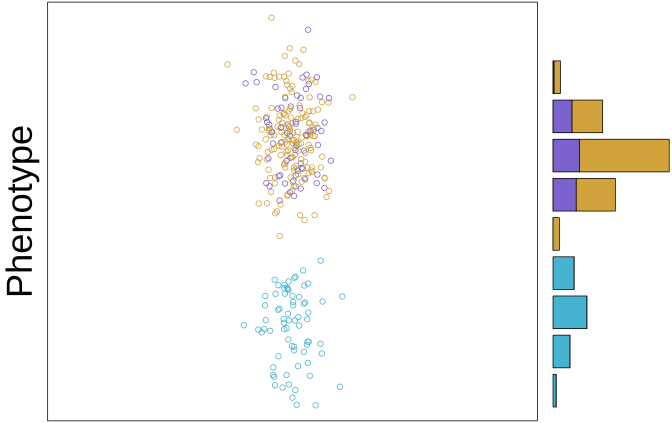
Additive only



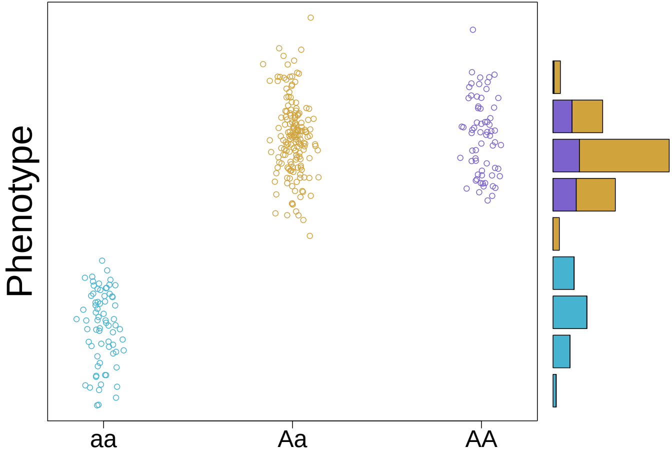
Additive with Dominance



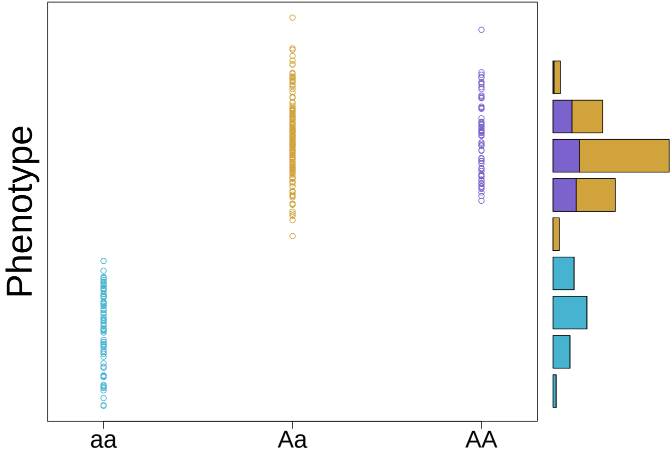
Additive with Dominance



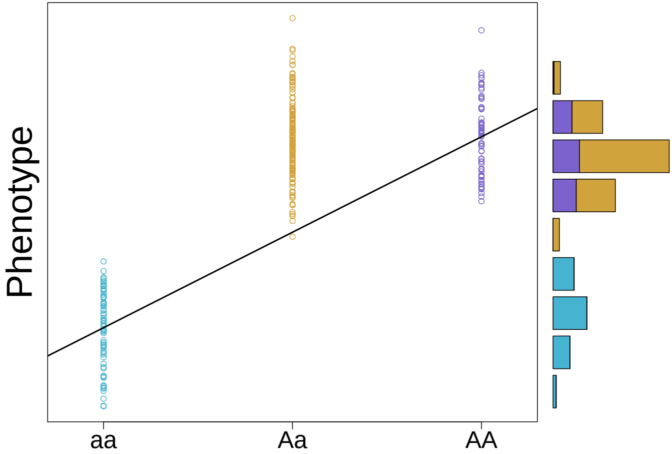
Additive with Dominance



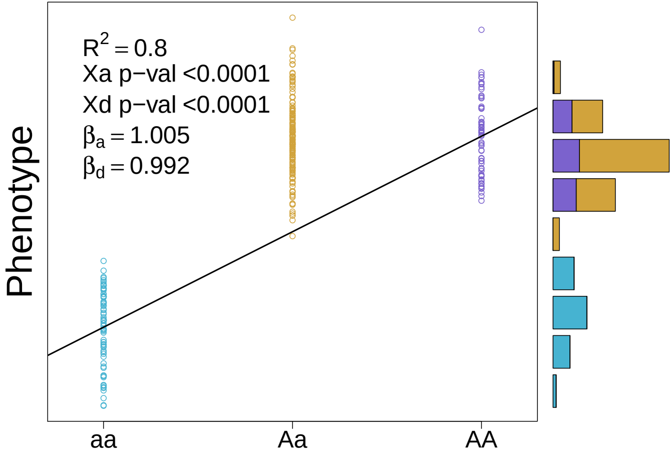
Additive with Dominance



Additive with Dominance



Additive with Dominance



Lets simulate it

Let's start with the single locus:
nsantantonio.shinyapps.io/singlelocus/

Genetic Variance

Let n = number of individuals
Let P = frequency of AA's
Let 2Q = frequency of Aa's
Let R = frequency of aa's

such that $P + 2Q + R = 1$

$$\mathrm{E}[\mathbf{x}] = \frac{1}{n}\sum_{i=1} x_i$$

$$\mu = Pa + 2Qd - Ra$$

$$\begin{aligned}\mathrm{Var}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}(x_i - \mu)^2 \\ &= P(a-\mu)^2 + 2Q(d-\mu)^2 + R(-a-\mu)^2\end{aligned}$$

Genetic Variance

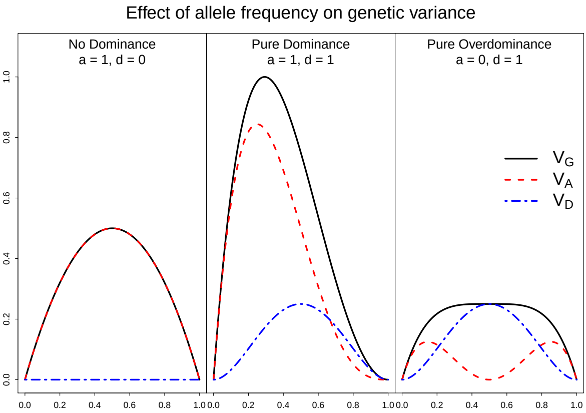
Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$

Genetic Variance

Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$



Finding causal genes

How do we find causal genes / variants?

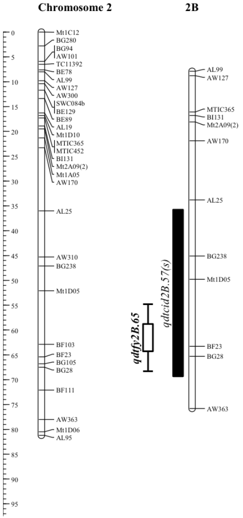
Finding causal genes

How do we find causal genes / variants?

DNA markers!

- genotype individuals with genome-wide markers
- statistical association between marker and trait
  - $H_0 : \beta_a = 0$ and $\beta_d = 0$
  - $H_0 : \beta_a \neq 0$ or $\beta_d \neq 0$

Finding causal genes



How do we find causal genes / variants?

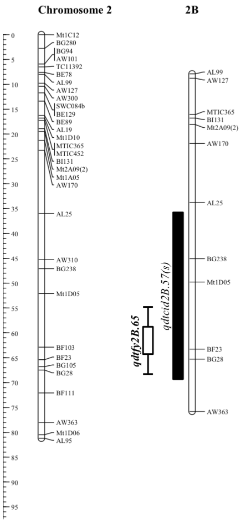
DNA markers!

- genotype individuals with genome-wide markers
- statistical association between marker and trait
  - $H_0 : \beta_a = 0$ and $\beta_d = 0$
  - $H_0 : \beta_a \neq 0$ or $\beta_d \neq 0$

Bi-parental mapping populations

- maximize allele frequencies
  - Statistical power

Finding causal genes



How do we find causal genes / variants?
DNA markers!

- genotype individuals with genome-wide markers
- statistical association between marker and trait
    - $H_0 : \beta_a = 0$ and $\beta_d = 0$
    - $H_0 : \beta_a \neq 0$ or $\beta_d \neq 0$

Bi-parental mapping populations

- maximize allele frequencies
    - Statistical power
- maximize linkage
    - poor precision...

Genome-Wide Association Studies

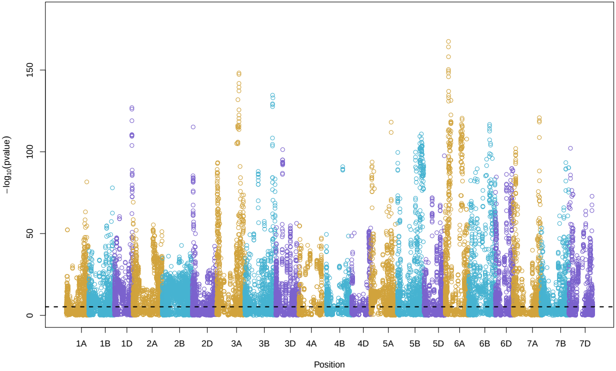
Association Mapping Population

- Take advantage of historical recombination events (low linkage)

Genome-Wide Association Studies

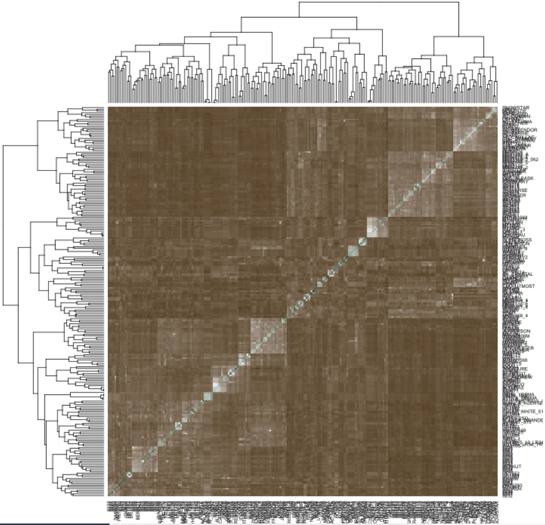
Association Mapping Population

- Take advantage of historical recombination events (low linkage)
- However, more closely related individuals will share functional alleles, as well as *many* other alleles



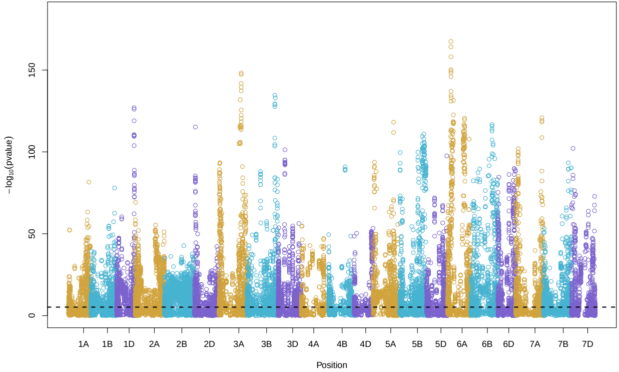
Population Structure Problem

Population structure “inflates” significance. Correct with kinship.



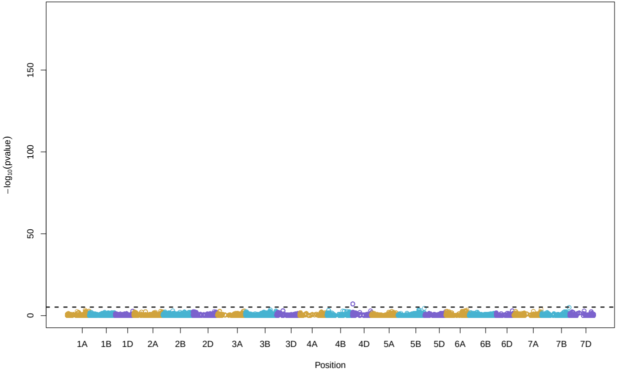
Finding causal genes

Population structure “inflates” significance.



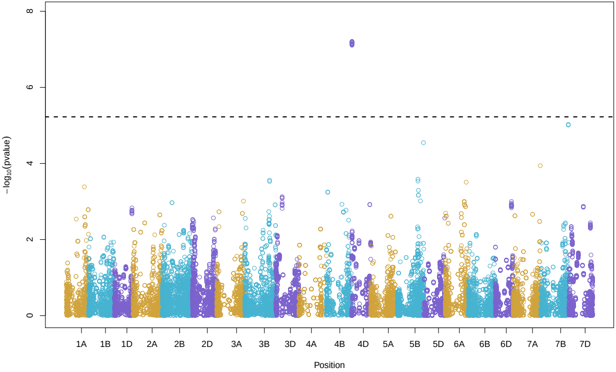
Finding causal genes

Population structure “inflates” significance.



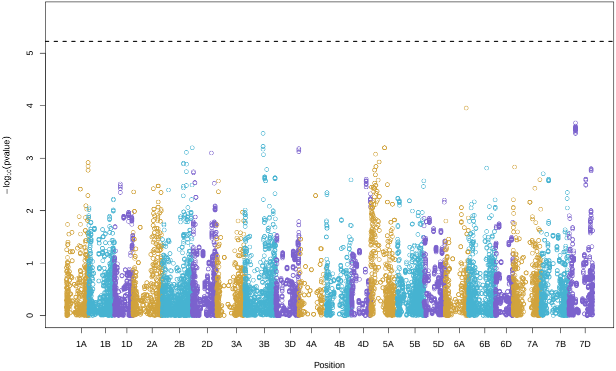
Finding causal genes

Population structure “inflates” significance.

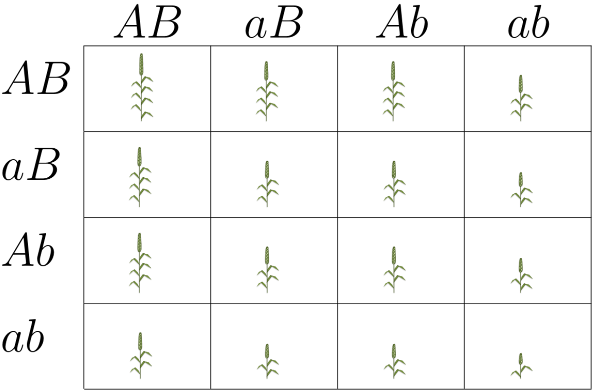


Finding causal genes

Population structure “inflates” significance.



Additive Two Loci



Lets see what happens when we have many loci

Let’s start with the single locus:
nsantantonio.shinyapps.io/quantitative/

Genomic Prediction

$$G_i = \sum_{i=1}^{m} \mathbf{x}_{a_i} \beta_{a_i}$$

- Genetic value of an individual is the sum of its allele effects
- Interestingly, same as modeling kinship between individuals!

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \varepsilon$$

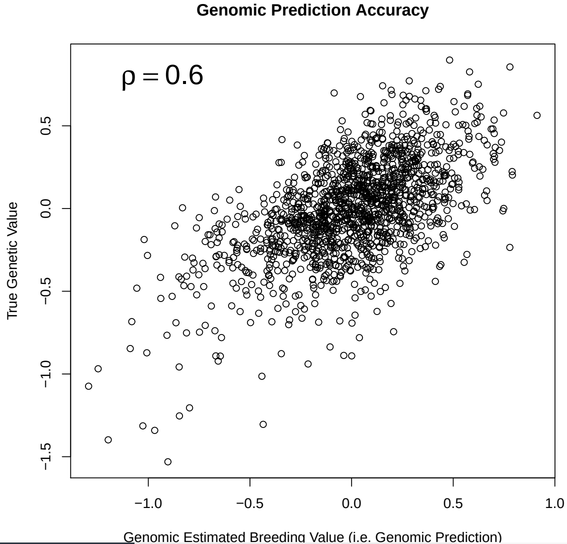
- $\mathbf{1}_n\mu$ is the global mean
- $\mathbf{X}$ is the design matrix
- β is the vector of fixed environmental effects.
- $\mathbf{Z}$ is the incidence matrix
- $\mathbf{g} \sim \mathcal{N}(0, \sigma_a^2\mathbf{K})$, random genetic effects
- $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{R})$, error

Mixed Model Equations

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \varepsilon$$

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\mathsf{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\mathsf{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\mathsf{T}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1}\left(\frac{\sigma_e^2}{\sigma_g^2}\right) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^\mathsf{T}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^\mathsf{T}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Genomic Prediction of Grain Yield



Genomic Selection

Genomic Selection

Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$

Genomic Selection

Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$

- Can select without observing phenotypes!

Genomic Selection

Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$

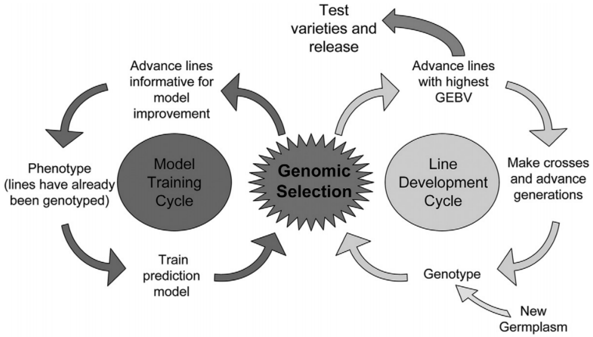
- Can select without observing phenotypes!
- make crosses in (winter) greenhouse to decrease cycle time (c)!

Genomic Selection

Breeder's Equation:

$$\Delta_R = \frac{ir\sigma_a}{c}$$

- Can select without observing phenotypes!
- make crosses in (winter) greenhouse to decrease cycle time (c)!


Test varieties and release
Advance lines informative for model improvement
Advance lines with highest GEBV
Phenotype (lines have already been genotyped)
Model Training Cycle
Genomic Selection
Line Development Cycle
Make crosses and advance generations
Train prediction model
Genotype
New Germplasm