UF | UNIVERSITY of FLORIDA

(HOS 6932)– Survey of Breeding Tools and Methods

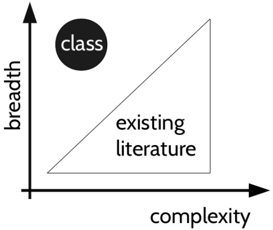
An introduction to Quantitative Genetics

Felipe Ferrão
Research Assistant Scientist
lferrao@ufl.edu
lfelipeferrao

January, 2023

Introduction

Main objectives

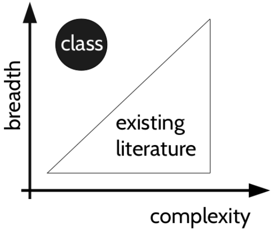


- This module address the practical implementation of genomic selection .
- What you can expect:
  - ▷ Understand general principles;
  - ▷ Keep mathematics to a minimum and focus instead on the intuition;
  - ▷ Fit a RR-BLUP predictive model

- What I expect from you:
  - ▷ Basic knowledge of statistics, genetics and breeding;
  - ▷ Some familiarity with R;

Introduction

Main objectives



- This module address the practical implementation of genomic selection .
- What you can expect:
  - ▷ Understand general principles;
  - ▷ Keep mathematics to a minimum and focus instead on the intuition;
  - ▷ Fit a RR-BLUP predictive model

- What I expect from you:
  - ▷ Basic knowledge of statistics, genetics and breeding;
  - ▷ Some familiarity with R;

Introduction

General Structure

- Genomic Selection can be presented under different perspectives
- Multiple courses at UF
- Quantitative Genetics: population genetics, resemble between relatives, pedigree analysis (BLUP), GBLUP
- Statistical Learning: normal distribution, regression (linear model), regularization

Introduction

Let's grab some coffee and discuss methods and techniques used in plant breeding!!

HOS 6932

University *of* Florida
Survey of Breeding Tools and Methods
(Genomic Selection)

SCAN ME

https://lfelipe-ferrao.github.io/teaching/

Quantitative Variation

Definition

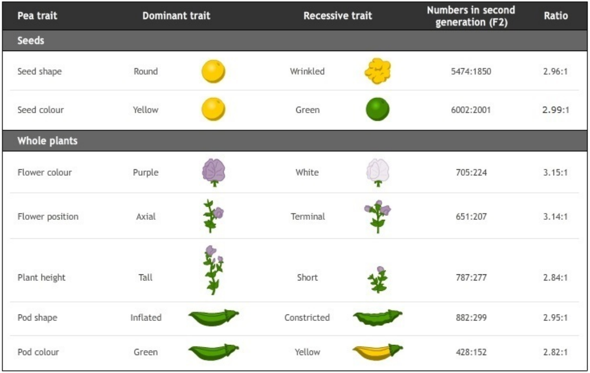
Quantitative genetics provides means for estimating the genetic architecture and predicting the evolutionary potential of complex traits.

- What is genetic architecture?
- Why we have complex and simple traits?
- How to study a complex trait?

Quantitative Variation

Background

- All traits measured by Mendel are very simplistic.
- Phenotypes were assumed to be completely determined by the genotype
- Discrete classes, with variation corresponding to single locus with two alleles



Pea trait	Dominant trait	Recessive trait	Numbers in second generation (F2)	Ratio
Seeds				
Seed shape	Round	Wrinkled	5474:1850	2.96:1
Seed colour	Yellow	Green	6002:2001	2.99:1
Whole plants				
Flower colour	Purple	White	705:224	3.15:1
Flower position	Axial	Terminal	651:207	3.14:1
Plant height	Tall	Short	787:277	2.84:1
Pod shape	Inflated	Constricted	882:299	2.95:1
Pod colour	Green	Yellow	428:152	2.82:1

© 2005-2011 The University of Waikato | www.biotechlearn.org.nz

Quantitative Variation

Question

- How many traits in your crop do you know that follow this discrete pattern?
- Is this type of genetic architecture a general rule or an exception?

Quantitative Variation

Qualitative Traits

- Mendelian trait
- Fall into discrete categories
- One or few genes
- Example: Mendel's garden peas, color, insect and disease resistance

Quantitative Genetics

- Continues phenotype
- Greatly influenced by environment
- Join action of multiple genes (or QTL) – Infinitesimal Model
- Example: yield, height and weight

Nature of quantitative traits has two important aspects

- Phenotype is a function of genotype and the environment
- Continuous traits usually follows a normal distribution, that can be fully described with only two parameters: mean and variance.

Quantitative Variation

Qualitative Traits

- Mendelian trait
- Fall into discrete categories
- One or few genes
- Example: Mendel's garden peas, color, insect and disease resistance

Quantitative Genetics

- Continues phenotype
- Greatly influenced by environment
- Join action of multiple genes (or QTL) – Infinitesimal Model
- Example: yield, height and weight

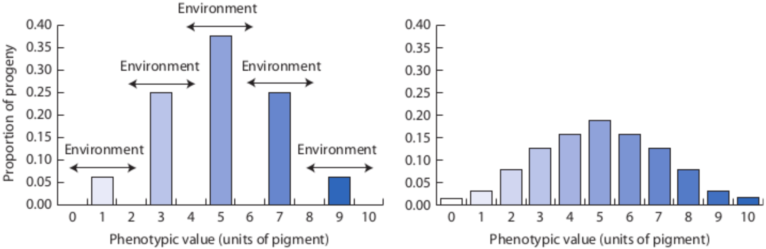
Nature of quantitative traits has two important aspects

- Phenotype is a function of genotype and the environment
- Continuous traits usually follows a normal distribution, that can be fully described with only two parameters: mean and variance.

Quantitative Variation

The environmental factor

- Ex: phenotypic distribution determined by two independent Mendelian loci.
- Environment can "increase" or "decrease" the phenotypic expression
- Even if there is only a single genotype, the phenotype expressed will change depending on the environmental conditions

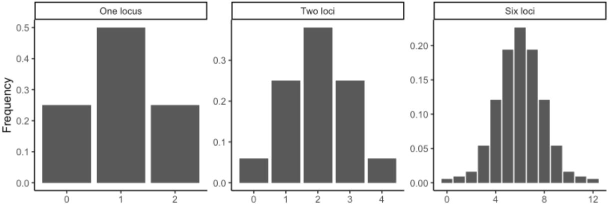


Quantitative Variation

Normal Distribution

Infinitesimal Model

Fisher in 1918 showed that a large number of Mendelian factors (genes) influencing a trait would cause a nearly continuous distribution of trait values. Therefore, mendelian genetics can lead to an approximately normal distribution



Quantitative Variation

- For a trait controlled by many genes and influenced by the environment, the measured character for any trait on an individual is called **phenotypic value**
- Formally, we can divide the phenotypic value

$$P = G + E$$

Environment (E)

- Include all non-genetic effects (systematic and non-systematic)
- In plant breeding: G × E is also important

Genotype (G)

- The particular set of genes in a given individuals
- Can be decomposed in additional terms

$$V_p = V_G + V_E$$

$$V_p = V_A + V_D + V_I + V_E$$

Quantitative Variation

- For a trait controlled by many genes and influenced by the environment, the measured character for any trait on an individual is called **phenotypic value**
- Formally, we can divide the phenotypic value

$$P = G + E$$

Environment (E)

- Include all non-genetic effects (systematic and non-systematic)
- In plant breeding: G × E is also important

Genotype (G)

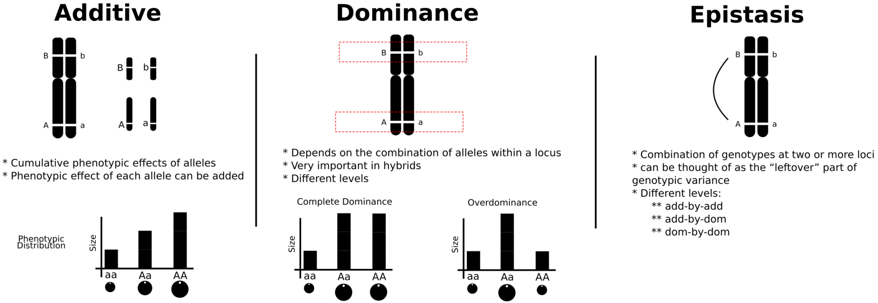
- The particular set of genes in a given individuals
- Can be decomposed in additional terms

$$V_p = V_G + V_E$$

$$V_p = V_A + V_D + V_I + V_E$$

Quantitative Variation

Gene action



Additive

* Cumulative phenotypic effects of alleles
* Phenotypic effect of each allele can be added

Dominance

* Depends on the combination of alleles within a locus
* Very important in hybrids
* Different levels

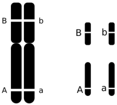
Epistasis

* Combination of genotypes at two or more loci
* can be thought of as the “leftover” part of genotypic variance
* Different levels:
** add-by-add
** add-by-dom
** dom-by-dom

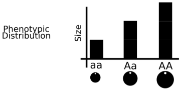
Quantitative Variation

Gene action

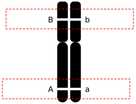
Additive



* Cumulative phenotypic effects of alleles
* Phenotypic effect of each allele can be added

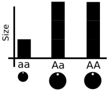


Dominance

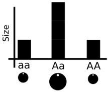


* Depends on the combination of alleles within a locus
* Very important in hybrids
* Different levels

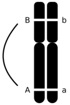
Complete Dominance



Overdominance

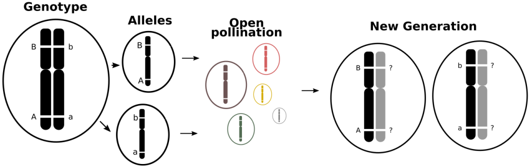


Epistasis



* Combination of genotypes at two or more loci
* can be thought of as the “leftover” part of genotypic variance
* Different levels:
  ** add-by-add
  ** add-by-dom
  ** dom-by-dom

Why are these concepts important?



Connecting the dots

- Differences between quantitative and qualitative traits
- Environment has an important impact on the phenotype expression
- Our unity of study is a population, with a normal distribution and mean and variance

What are we missing?

We need to define the means to study a complex trait

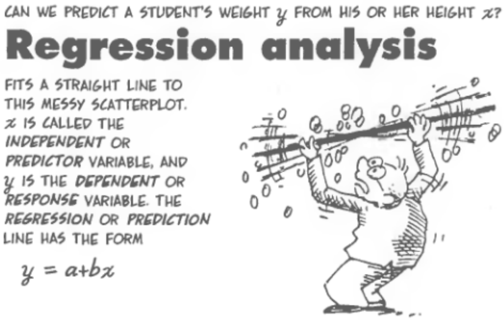
Connecting the dots

- Differences between quantitative and qualitative traits
- Environment has an important impact on the phenotype expression
- Our unity of study is a population, with a normal distribution and mean and variance

What are we missing?

We need to define the means to study a complex trait

Regression



CAN WE PREDICT A STUDENT'S WEIGHT y FROM HIS OR HER HEIGHT x?

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE ***INDEPENDENT*** OR ***PREDICTOR*** VARIABLE, AND y IS THE ***DEPENDENT*** OR ***RESPONSE*** VARIABLE. THE ***REGRESSION*** OR ***PREDICTION*** LINE HAS THE FORM

$$y = a+bx$$

Regression

Background

- Depending on the causal connections between two variables, their true relationship may be linear and can be described using a linear regression
- Examples
  - ▷ How gender (x) is associated with salary (y) income?
  - ▷ How fertilization (x) is associated to yield (y) in corn?
  - ▷ How the phenotypic value (x) is associated to the gene content (y)?

Regression

We can write such questions using a model

$$Y_i = \beta_0 + \beta_1 X_i + e$$

- The terms β_0 and β_1 are the intercept and slope of the model, respectively.
- Intercept is the point at which the line crosses the y axis at x = 0.
- The slope expresses the relationship between y and x.

Estimation

- β_0 and β_1 are two parameters estimated from the data
- Ordinary Least Squares (OLS) to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$

Regression

We can write such questions using a model

$$Y_i = \beta_0 + \beta_1 X_i + e$$

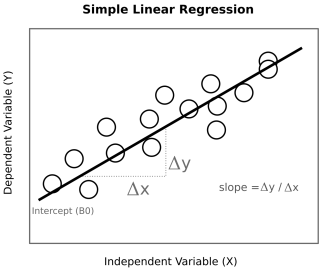
- The terms β_0 and β_1 are the intercept and slope of the model, respectively.
- Intercept is the point at which the line crosses the y axis at x = 0.
- The slope expresses the relationship between y and x.

Estimation

- β_0 and β_1 are two parameters estimated from the data
- Ordinary Least Squares (OLS) to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$

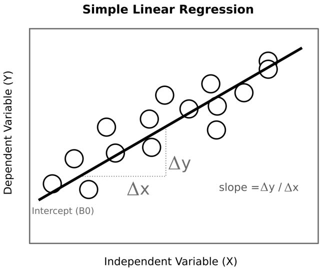
Regression

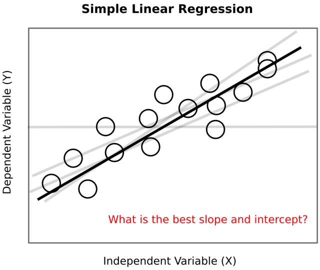
Geometric Representation



Regression

Geometric Representation

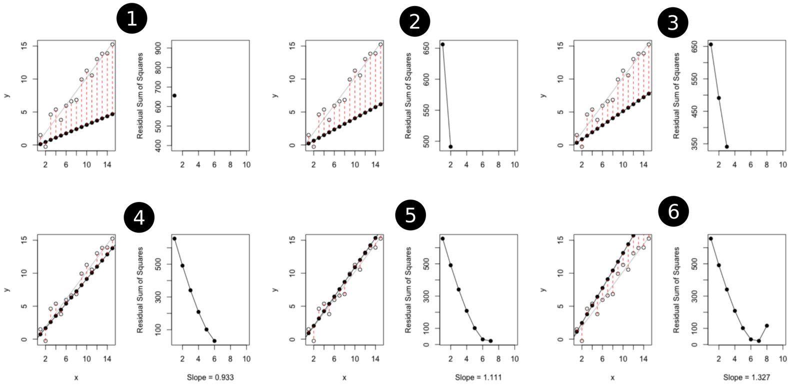




Regression

OLS in action

- Straight lines can be drawn in multiple ways – different slopes and intercepts.
- What is the best fit? Minimizes the error between the original and predicted values



Regression

Least-Squares Linear Regression

- We seek estimators for the intercept and slope that minimize the residual
- First, we need to define the errors
- Differentiate with respect to β_0 and β_1 and set the results equal to zero

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + e \\
\hat{e}'\hat{e} &= y_i - \hat{y} \\
\hat{e} = \sum \hat{e}_i{}^2 = \sum (y_i - \hat{y}_i)^2 &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
\frac{\partial \hat{e}'\hat{e}}{\partial \hat{\beta}_0} &= -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
\frac{\partial \hat{e}'\hat{e}}{\partial \hat{\beta}_1} &= -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0
\end{aligned}
$$

Simple Linear Regression Estimators

- Slope: $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regression

Least-Squares Linear Regression

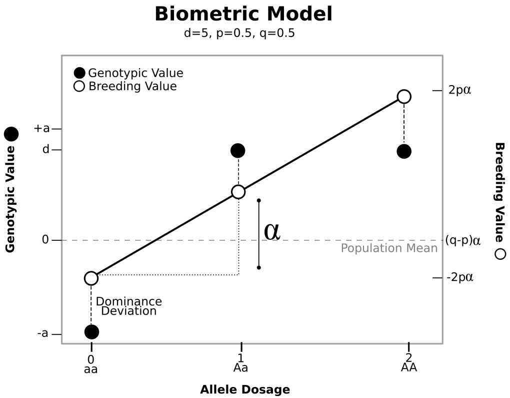
- We seek estimators for the intercept and slope that minimize the residual
- First, we need to define the errors
- Differentiate with respect to β_0 and β_1 and set the results equal to zero

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e \\ \hat{e}'\hat{e} &= y_i - \hat{y} \\ \hat{e} = \sum \hat{e}_i{}^2 = \sum (y_i - \hat{y}_i)^2 &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ \frac{\partial \hat{e}'\hat{e}}{\partial \hat{\beta}_0} &= -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial \hat{e}'\hat{e}}{\partial \hat{\beta}_1} &= -2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{aligned}$$

Simple Linear Regression Estimators

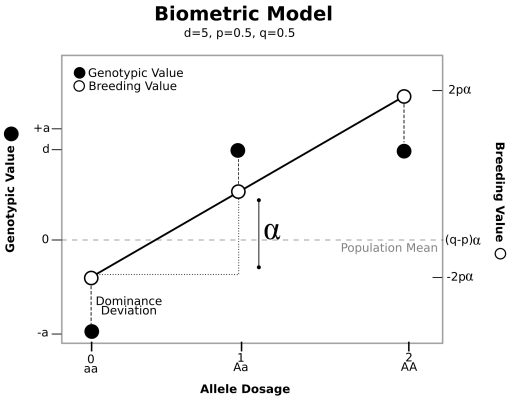
- Slope: $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regression



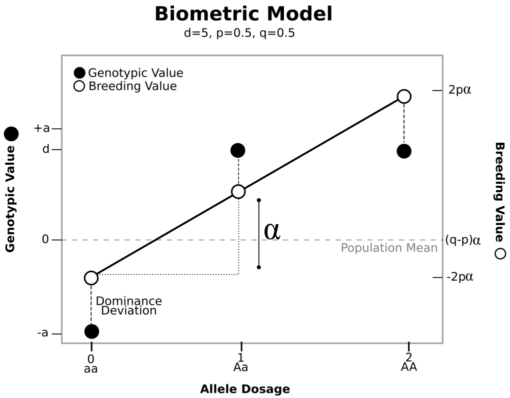
- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \alpha$ (average effect of allelic substitution)
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -2p\alpha$ (breeding value for aa)

Regression



- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \alpha$ (average effect of allelic substitution)
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -2p\alpha$ (breeding value for aa)

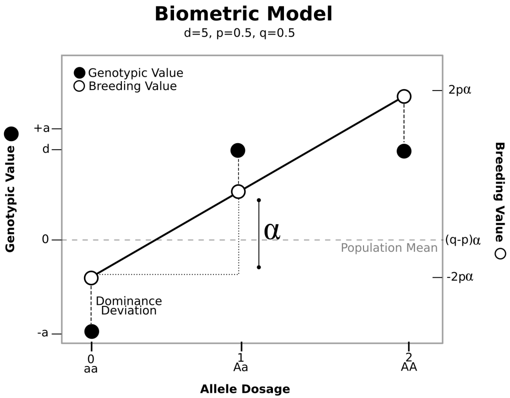
Regression



Partition the Genetic Variance

- SSTotal = SSRegression + SSdeviation
- $SSTotal = \sum f_i y_i^2 = p^2(y_1)^2 + 2pq(y_2)^2 + q^2(y_3) = \sigma_g^2$
- $SSReg = \hat{\beta}_1 \sum f_i x_i y_i = 2pq\alpha^2 = \sigma_a^2$
- $SSDe = \sum f_i \hat{e}_i^2 = (2pqd)^2 = \sigma_d^2$

Regression



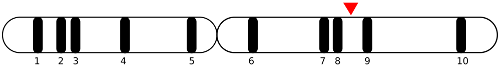
Partition the Genetic Variance

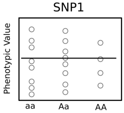
- SSTotal = SSRegression + SSdeviation
- $SSTotal = \sum f_i y_i^2 = p^2(y_1)^2 + 2pq(y_2)^2 + q^2(y_3) = \sigma_g^2$
- $SSReg = \hat{\beta}_1 \sum f_i x_i y_i = 2pq\alpha^2 = \sigma_a^2$
- $SSDe = \sum f_i \hat{e}_i^2 = (2pqd)^2 = \sigma_d^2$

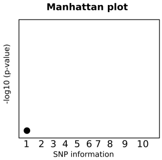
Regression

Including Molecular Markers

- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(marker)$ and testing $H_0 : \beta_1 = 0$



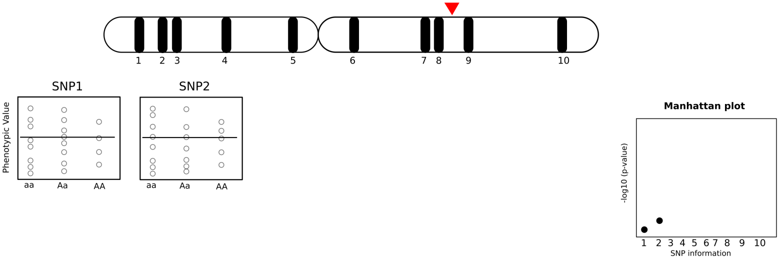




Regression

Including Molecular Markers

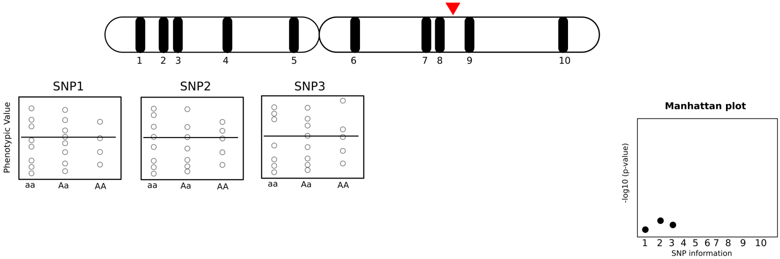
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(marker)$ and testing $H_0 : \beta_1 = 0$



Regression

Including Molecular Markers

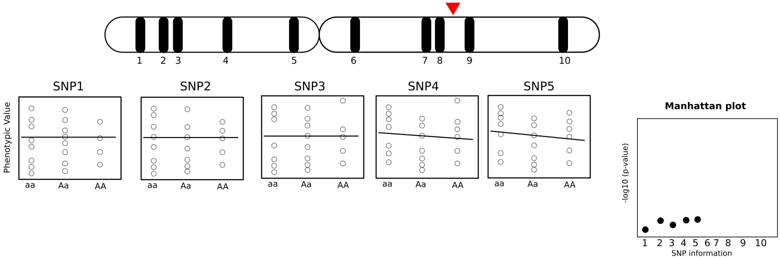
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(marker)$ and testing $H_0 : \beta_1 = 0$



Regression

Including Molecular Markers

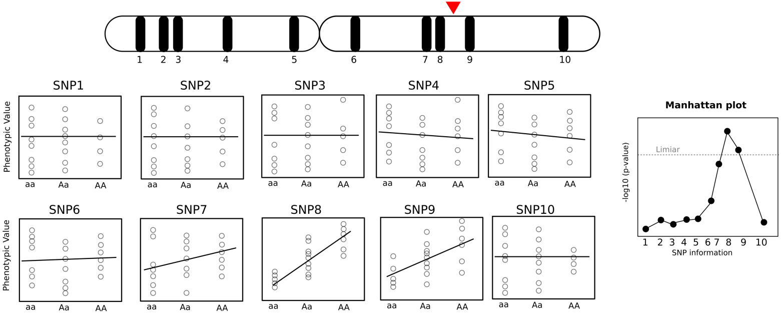
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(marker)$ and testing $H_0 : \beta_1 = 0$



Regression

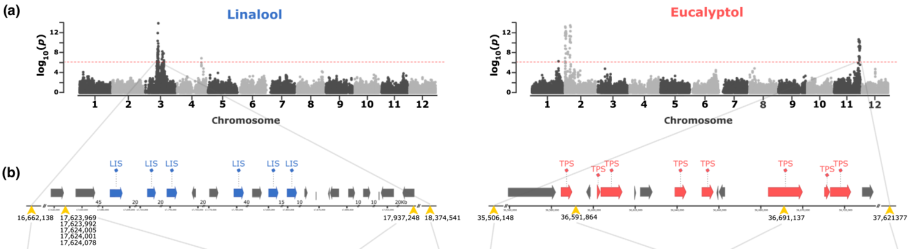
Including Molecular Markers

- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(marker)$ and testing $H_0 : \beta_1 = 0$



Regression

Genome-wide association of volatiles reveals candidate loci for blueberry flavor



Ferrão et al., 2020. New Phytologist doi: 10.1111/nph.16459

Regression

Including Molecular Markers

- Single Marker Regression
- Version of the Biometric Model: we can compute genetic parameters
- Theoretical basis for GWAS models and QTL mapping
- Precursor of genomic selection methods
- Problems:
  - ▷ Testing millions of markers, one at a time, inflate type I error
  - ▷ Lack of power: small effect can rarely be detected
  - ▷ Beavis (or winner's curse) effect: noises will occur in analysis with many markers, and this biases the estimates, making it look much larger than real

Regression

Including Molecular Markers

- Single Marker Regression
- Version of the Biometric Model: we can compute genetic parameters
- Theoretical basis for GWAS models and QTL mapping
- Precursor of genomic selection methods
- Problems:
  - ▷ Testing millions of markers, one at a time, inflate type I error
  - ▷ Lack of power: small effect can rarely be detected
  - ▷ Beavis (or winner's curse) effect: noises will occur in analysis with many markers, and this biases the estimates, making it look much larger than real

Final Considerations

Summary

- Differences between qualitative and quantitative traits
- How to compute means and variance using genetic information
- Key concepts: additive, dominance, epistasis, additive variance, dominance variance and average effect of allelic substitution
- First marker-assisted selection model !!

Next class

- Why use one marker at a time in the regression analyses?
- What happens if we use all markers simultaneously?

Final Considerations

Summary

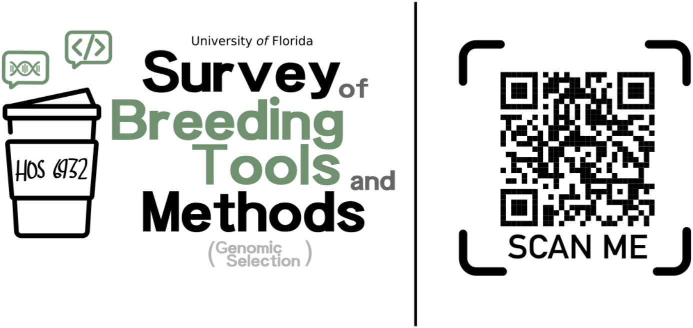
- Differences between qualitative and quantitative traits
- How to compute means and variance using genetic information
- Key concepts: additive, dominance, epistasis, additive variance, dominance variance and average effect of allelic substitution
- First marker-assisted selection model !!

Next class

- Why use one marker at a time in the regression analyses?
- What happens if we use all markers simultaneously?

Final Considerations

Hands-on 1



https://lfelipe-ferrao.github.io/teaching/

References

References

- Hamilton, 2009. Population Genetics.
  Book – *Basic knowledge*
- Bernardo, 2010. Breeding for Quantitative Traits in Plants
  Book – *Intermediate knowledge, more focuses in breeding*
- Lynch and Walsh, 1998. Genetics and Analysis of Quantitative Traits
  Book – *Complete book, modern approach, advanced knowledge*
- Falconer and Mackay, 1996. Introduction to Quantitative Genetics
  Book – *Main reference in the field, advanced knowledge*
- Cruz, 2005. Princípios de Genética Quantitativa.
  Book – *Portuguese version*