

Introduction
oooooooo

Quantitative Variation
oooooooooooo

Mean and Variance
oooooooooooo

Biometric Regression
oooooooooooooooooooo

Final Consideration
oooo



(HOS 6932)– Survey of Breeding Tools and Methods

An introduction to Quantitative Genetics

Felipe Ferrão
Research Assistant Scientist
lferrao@ufl.edu
 [@felipeferrao](https://twitter.com/felipeferrao)

January, 2022

Last Class

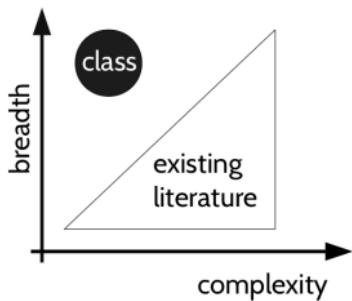
- Plant Breeding 101
 - Connection with the topics presented in the course

This class

- Introduction to Quantitative Genetics
 - First genomic model

Introduction

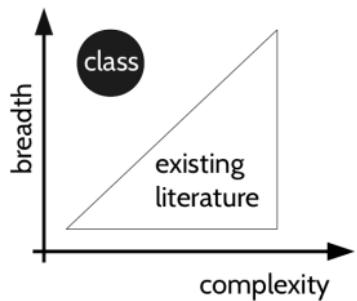
Main objectives



- 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

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

Plant Breeding

FONTE: Bernardo and Yu (2007). Prospects for Genomewide Selection for Quantitative Traits in Maize.

Introduction

Statistics

FONTE: de los Campos et al.(2013). Whole-Genome and Prediction Methods Applied to Plant and Animal Breeding.

Introduction

General Structure

- Genomic Selection can be presented under different perspectives
- Multiple courses at UF
- Important: it is a multidisciplinary approach

Quantitative Genetics (PCB 6555)

Molecular Marker Assisted Plant Breeding (HOS 6236)



Introduction

General Structure

- Genomic Selection can be presented under different perspectives
- Multiple courses at UF
- Important: it is a multidisciplinary approach

1) An overview of Quantitative Genetics

Introduction

Population Genetics

2) Computing genotypic means and variances

Variance and Means

Resemble between relatives

Genetics Desings

Quantitative Genetics (PCB 6555)

Molecular Marker Assisted Plant Breeding (HOS 6236)

BLUP

Genomic Selection

Bayesian Alphabet

5) Application

Supervised Learning

Linear Models (Regression)

Regularization

Kernel and Tress

Unsupervised Learning

Deep Learning

Machine Learning
Statistical Learning

3) Biometric Models

4) When $p >>n$

Introduction

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



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

Quantitative Variation

Motivation

- Role of micro-evolution and selection affect ecological and breeding dynamics either in natural or artificial population
- Understanding evolutionary, morphological, physiological, and life-history involves (i) measuring how genetics underpins variation in fitness, and (ii)the inheritance of traits over generations

Definition

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

Quantitative Variation

Motivation

- Role of micro-evolution and selection affect ecological and breeding dynamics either in natural or artificial population
- Understanding evolutionary, morphological, physiological, and life-history involves (i) measuring how genetics underpins variation in fitness, and (ii)the inheritance of traits over generations

Definition

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

Quantitative Variation

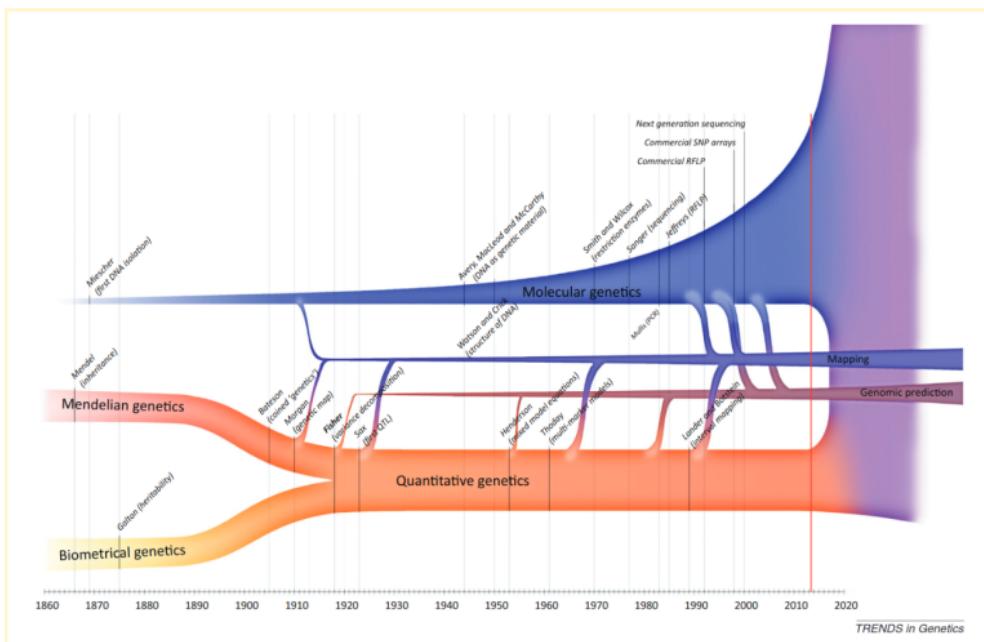
Historical Background

- Many brilliant scientists
 - Francis Galton (1869)
 - ▷ Founding the Biometrical School
 - ▷ Discrete vs. continuous traits
 - Fisher(1918)
 - ▷ Variance-component partition
 - ▷ Infinitesimal Model
 - Wright (1921)
 - ▷ Resamble between relatives
 - Henderson (1950)
 - ▷ BLUP theory



Quantitative Variation

Timeline



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
Seed colour	Yellow		Green		6002:2001
Whole plants					
Flower colour	Purple		White		705:224
Flower position	Axial		Terminal		651:207
Plant height	Tall		Short		787:277
Pod shape	Inflated		Constricted		882:299
Pod colour	Green		Yellow		428:152

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

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

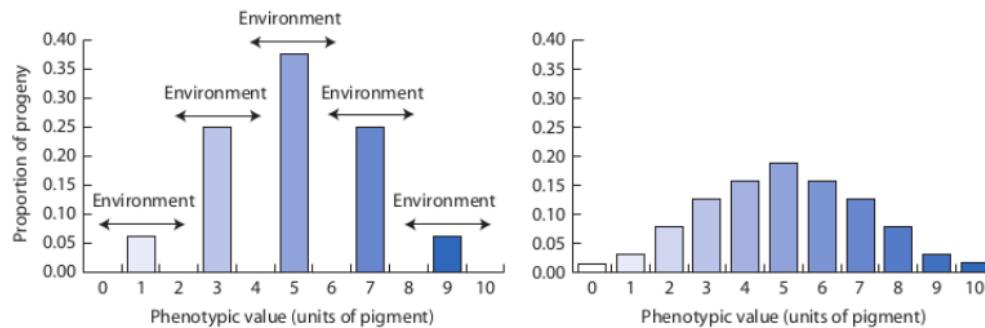
Nature of quantitative traits has two important aspects

- 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

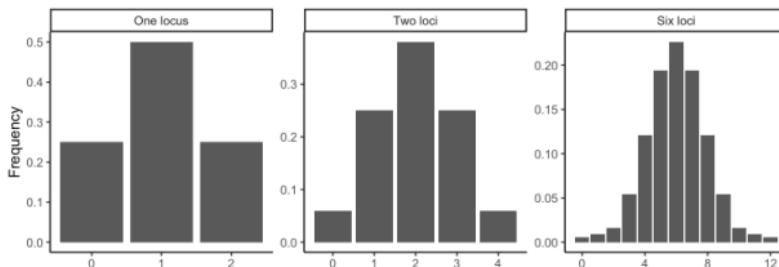


Quantitative Variation

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: $G \times E$ 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: $G \times E$ is also important

Genotype (G)

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

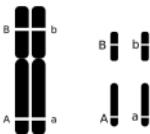
$$V_p = V_G + V_E$$

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

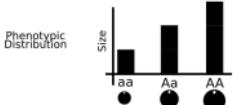
Quantitative Variation

Gene action

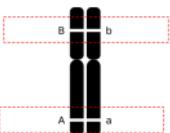
Additive



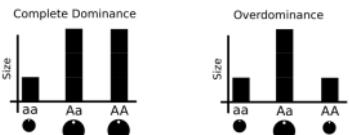
- * Cumulative phenotypic effects of alleles
 - * Phenotypic effect of each allele can be added



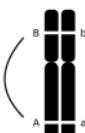
Dominance



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



Epistasis

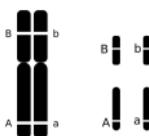


- * Combination of genotypes at two or more loci
 - * can be thought of as the "leftover" part of genotypic variance
 - * Different levels:
 - ** add-by-add
 - ** add-by-dom
 - ** dom-by-dom

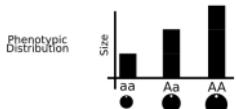
Quantitative Variation

Gene action

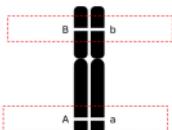
Additive



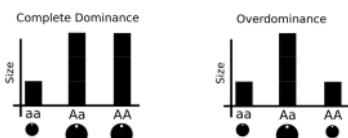
- * Cumulative phenotypic effects of alleles
 - * Phenotypic effect of each allele can be added



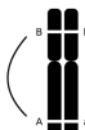
Dominance



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

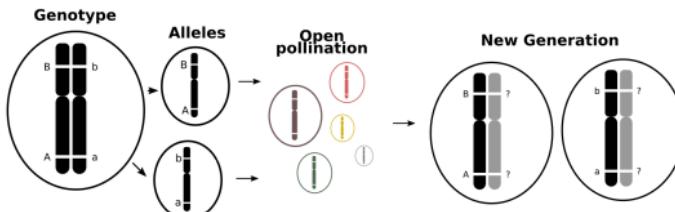


Epistasis



- * 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-hv-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 variation at the population level, in terms of means and variance

Connecting the dots



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

What are we missing?

We need to define the variation at the population level, in terms of means and variance

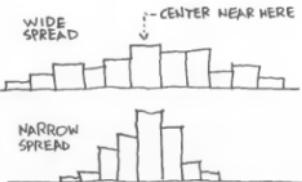
Mean and Variance

SUMMARY STATISTICS

NOW WE MOVE FROM PICTURES TO FORMULAS. OUR OBJECT IS TO GET SOME SIMPLE MEASUREMENTS OF THE CRUDEST CHARACTERISTICS OF A SET OF DATA.



ANY SET OF MEASUREMENTS HAS TWO IMPORTANT PROPERTIES: THE CENTRAL OR TYPICAL VALUE, AND THE SPREAD ABOUT THAT VALUE. YOU CAN SEE THE IDEA IN THESE HYPOTHETICAL HISTOGRAMS.



Mean and Variance

Population Mean

- How do gene frequencies affect the mean of a trait in a population?
- At the same way that we can compute mean values in our routine analyses, we can compute the population mean.

$$\bar{x} = \frac{\sum_{i=1}^N X_i}{N} \iff \bar{x} = \sum_{i=1}^N \frac{f_i \times X_i}{f_i}$$

Pop	1	2	3	4	5	Sum	Mean
1	10	20	15	30	10	85	17
2	10	10	10	100	10	140	28
3	17	17	17	16	18	85	17

Mean = freq x value

$$\bar{x}_1 = (10 \times 0.2) + (20 \times 0.2) + (15 \times 0.2) + (30 \times 0.2) + (10 \times 0.2) = 17$$

Mean and Variance

Population Mean

- How do gene frequencies affect the mean of a trait in a population?
 - At the same way that we can compute mean values in our routine analyses, we can compute the population mean.

$$\bar{x} = \frac{\sum_{i=1}^N X_i}{N} \iff \bar{x} = \sum_{i=1}^N \frac{f_i \times X_i}{f_i}$$

Pop	1	2	3	4	5	Sum	Mean
1	10	20	15	30	10	85	17
2	10	10	10	100	10	140	28
3	17	17	17	16	18	85	17

Mean = freq x value

$$\bar{x}_1 = (10 \times 0.2) + (20 \times 0.2) + (15 \times 0.2) + (30 \times 0.2) + (10 \times 0.2) = 17$$

Mean and Variance

In the genetic context

- We can also compute means
- This is not the phenotypic mean, but the genotypic mean

Genotype	Ind	Freq	Freq any pop	EHW	F2 pop	Value
AA	N11	N11/N	D	p^2	1/4	+a
Aa	N12	N12/N	H	$2pq$	2/4	d
aa	N22	N22/N	R	q^2	1/4	-a

Population in EHW:

$$\begin{aligned}\mu &= p^2 a + 2pqd - q^2 a \\ &= (p^2 - q^2)a + 2pqd \\ &= (p + q)(p - q)a + 2pqd \\ &= (p - q)a + 2pqd \\ \mu &= a(p - q) + 2pqd\end{aligned}$$

Mean and Variance

In the genetic context

- We can also compute means
- This is not the phenotypic mean, but the genotypic mean

Genotype	Ind	Freq	Freq any pop	EHW	F2 pop	Value
AA	N11	N11/N	D	p^2	1/4	+a
Aa	N12	N12/N	H	$2pq$	2/4	d
aa	N22	N22/N	R	q^2	1/4	-a

Population in EHW:

$$\begin{aligned}\mu &= p^2 a + 2pqd - q^2 a \\ &= (p^2 - q^2)a + 2pqd \\ &= (p + q)(p - q)a + 2pqd \\ &= (p - q)a + 2pqd \\ \mu &= a(p - q) + 2pqd\end{aligned}$$

Mean and Variance

At a single locus

$$\mu = a(p - q) + 2pqd$$

- More important than a final number, the equation shows that mean phenotypic values will vary to the level of dominance and the gene frequencies
- $a(p - q)$ represents the contribution of the homozygotes
- $2pqd$ represents the contribution of the heterozygotes
- For other populations:
 - ▷ For any population: $\mu = a(D - R) + Hd$
 - ▷ F2 population: $1/2d$

Mean and Variance

Why this is so important?

- As a breeder, we can project some practical results based on the mean values

Population Type	Population Mean	Notes
	$M = a(p-q) + 2pqd$	Allelic frequency is constant for population in HW equilibrium
	$M = (M_1 + M_2)/2 + d(p-p')^2$	M_1 = mean parent 1 M_2 = mean parent 2 $d(p-p')^2$ = hybrid vigor
	$M = M_1 - 2pqF$	F = inbreeding coefficient Depression in most of the cases

Mean and Variance

The standard measure of spread is the

STANDARD DEVIATION

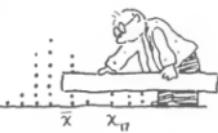
UNLIKE THE IQR, WHICH IS BASED ON MEDIANs, THE STANDARD DEVIATION MEASURES THE SPREAD FROM THE MEAN. YOU CAN THINK OF IT, ROUGHLY SPEAKING, AS THE AVERAGE DISTANCE OF THE DATA FROM THE MEAN \bar{x} .



EXCEPT THAT WE USE THE SQUARES OF THE DISTANCES INSTEAD. THAT IS, IF THE SQUARED DISTANCE OF POINT x_i TO \bar{x} IS $(x_i - \bar{x})^2$, THEN

$$\text{AVERAGE SQUARED DISTANCE} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

FOR TECHNICAL REASONS, WE USE $n-1$ IN THE DENOMINATOR RATHER THAN n , AND DEFINE THE SAMPLE VARIANCE s^2 AS



$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Mean and Variance

Variances

$$\begin{aligned}\sigma^2 &= \frac{\sum(X - \mu)^2}{N} = \frac{\sum(X^2 - 2\mu X + \mu^2)}{N} \\&= \frac{\sum X^2}{N} - 2\mu \frac{\sum X^2}{N} + \frac{\sum \mu^2}{N} \\&= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2 \\&= \frac{1}{N} \sum X^2 - \mu^2\end{aligned}$$

- $freq \times value^2 - mean^2$

Pop	1	2	3	4	5	Σ	\bar{x}	$\Sigma(X - \bar{x})^2$	σ^2
1	10	20	15	30	10	85	17	280	56
2	10	10	10	100	10	140	28	6480	1296
3	17	17	17	16	18	85	17	2	0.4

Mean and Variance

In the genetic context

- We can also compute the variance for different populations

Genotype	Ind	Freq	Freq any pop	EHW	F2 pop	Value
AA	N11	N11/N	D	p^2	1/4	+a
Aa	N12	N12/N	H	$2pq$	2/4	d
aa	N22	N22/N	R	q^2	1/4	-a

- Any population: $V_G = a^2[(D + R) - (D - R)^2] + d^2H(1 - H) - 2adH(D - R)$
- F2 population: $V_G = 1/2a^2 + 1/4d^2$
- Population EHW

$$V_G = 2pq\alpha + (2pqd)^2$$

$$\alpha = a + d(p - q)$$

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
- Dominance: $(2pqd)^2$
- Why breeders need to estimate genetic variances?

▷ Additive variance is the cause of resemblance between parents

$$\text{▷ Heritability: } h^2 = \frac{V_G}{V_P}$$

▷ Genetic gain is function of the h^2

▷ Genetic correlations

▷ Direct and indirect selection

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
- Dominance: $(2pqd)^2$
- Why breeders need to estimate genetic variances?
 - ▷ Additive variance is the cause of resemblance between parents
 - ▷ Heritability: $h^2 = \frac{V_G}{V_P}$
 - ▷ Genetic gain is function of the h^2
 - ▷ Genetic correlations
 - ▷ Direct and indirect selection

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
- Dominance: $(2pqd)^2$
- Why breeders need to estimate genetic variances?
 - ▷ Additive variance is the cause of resemblance between parents
 - ▷ Heritability: $h^2 = \frac{V_G}{V_P}$
 - ▷ Genetic gain is function of the h^2
 - ▷ Genetic correlations
 - ▷ Direct and indirect selection

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
- Dominance: $(2pqd)^2$
- Why breeders need to estimate genetic variances?
 - ▷ Additive variance is the cause of resemblance between parents
 - ▷ Heritability: $h^2 = \frac{V_G}{V_P}$
 - ▷ Genetic gain is function of the h^2
 - ▷ Genetic correlations
 - ▷ Direct and indirect selection

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
- Dominance: $(2pqd)^2$
- Why breeders need to estimate genetic variances?
 - ▷ Additive variance is the cause of resemblance between parents
 - ▷ Heritability: $h^2 = \frac{V_G}{V_P}$
 - ▷ Genetic gain is function of the h^2
 - ▷ Genetic correlations
 - ▷ Direct and indirect selection

Mean and Variance

$$V_G = 2pq[a + d(p - q)] + (2pqd)^2$$

- Additive: $2pq[a + d(p - q)]$
 - Dominance: $(2pqd)^2$
 - Why breeders need to estimate genetic variances?
 - ▷ Additive variance is the cause of resemblance between parents
 - ▷ Heritability: $h^2 = \frac{V_G}{V_P}$
 - ▷ Genetic gain is function of the h^2
 - ▷ Genetic correlations
 - ▷ Direct and indirect selection

Hands-on

Genetic Variance

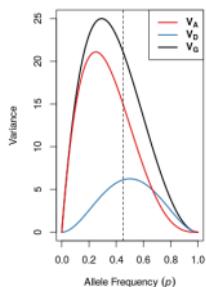
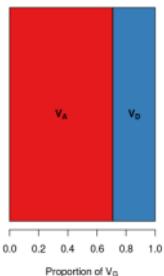
This simulation will assess the impact of allele frequency and coded genotypic values on the level of genetic variance at a single gene.

Starting Allele Frequency (p):


Additive Effect (a):


Dominance Effect (d):


Partition of Genetic Variance



Source of Variance	Value	Proportion of Genetic Variance
Additive	14.97	0.71
Dominance	6.13	0.29

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
 - ▷ Biological sex (x) associated with salary (y)
 - ▷ Product attributes (x) associated to number of sales (y)
 - ▷ Fertilization (x) associated to yield (y)
 - ▷ Genotypic value (x) associated to the gene content (y)

Regression

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

Model

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

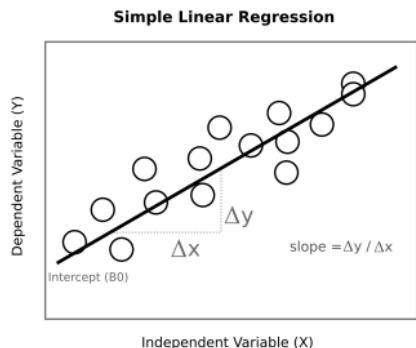
- The terms β_0 and β_1 are the intercept and slope of the model, respectively.
- Intercept is the point at which the line crosses the y axis at $x = 0$.
- The slope expresses the relationship between y and x.

Estimation

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

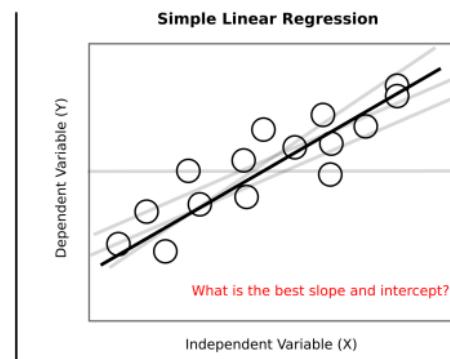
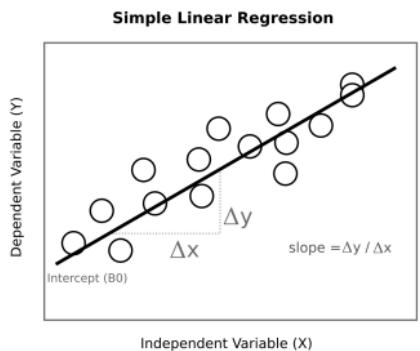
Regression

Geometric Representation



Regression

Geometric Representation

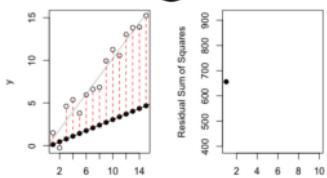


Regression

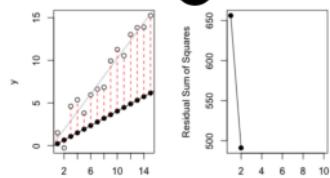
OLS in action

- 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

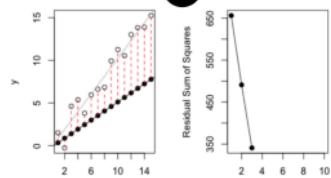
1



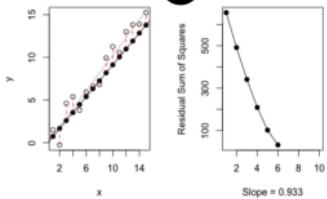
2



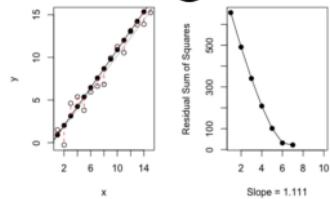
3



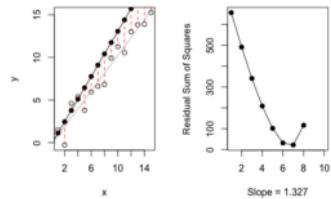
4



5



6



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

$$y_i = \beta_0 + \beta_1 x_i + e$$

$$\hat{e}' \hat{e} = y_i - \hat{y}$$

$$\hat{e} = \sum \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial \hat{e}' \hat{e}}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \hat{e}' \hat{e}}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

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

$$y_i = \beta_0 + \beta_1 x_i + e$$

$$\hat{e}' \hat{e} = y_i - \hat{y}$$

$$\hat{e} = \sum \hat{e}_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial \hat{e}' \hat{e}}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \hat{e}' \hat{e}}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

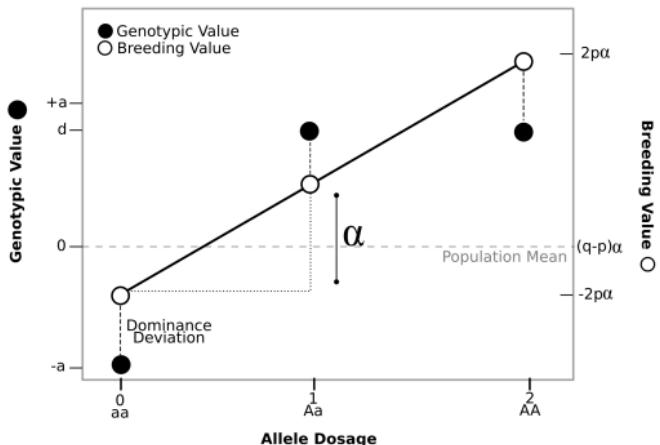
Simple Linear Regression Estimators

- 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

Biometric Model

$d=5, p=0.5, q=0.5$

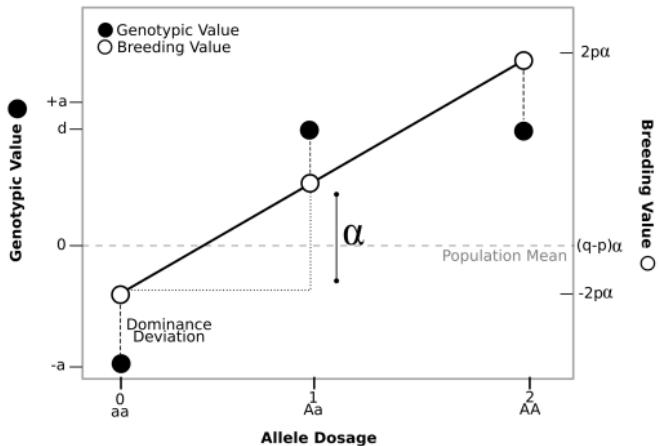


- $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \alpha$ (average effect of allelic substitution)
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -2p\alpha$ (breeding value for aa)

Regression

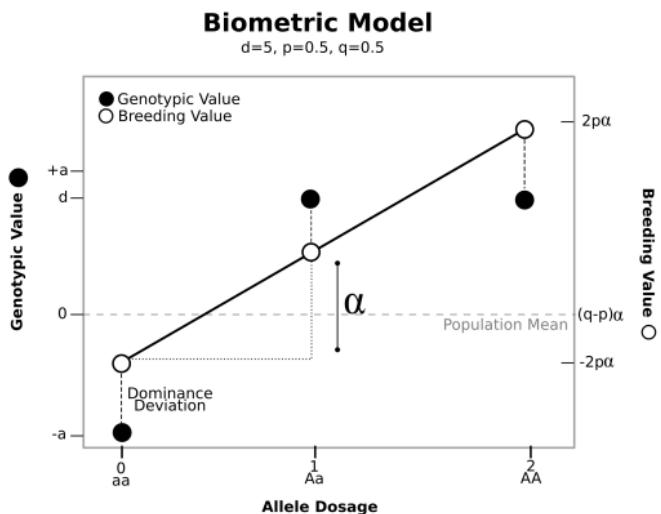
Biometric Model

$d=5, p=0.5, q=0.5$



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

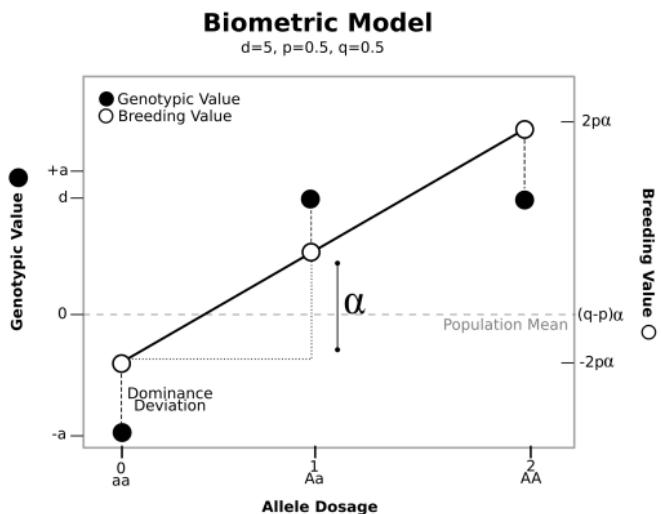
Regression



Partition the Genetic Variance

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

Regression



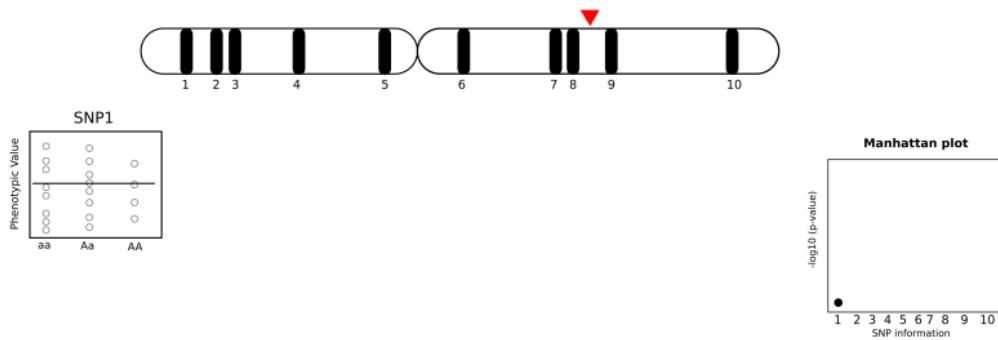
Partition the Genetic Variance

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

Regression

Including Molecular Markers

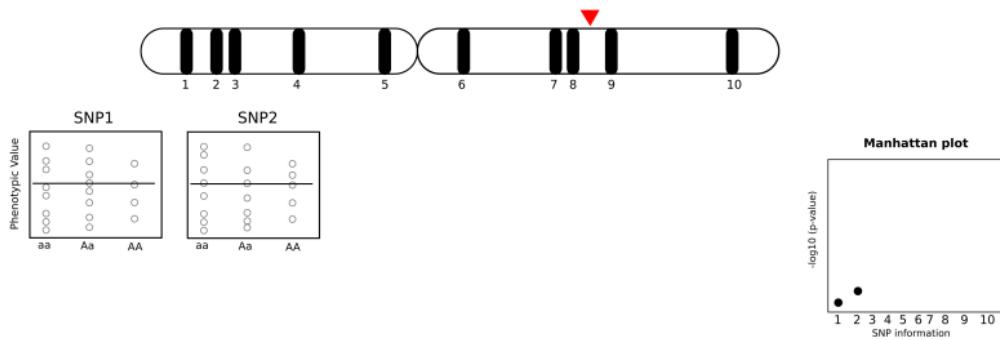
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(\text{marker})$ and testing $H_0 : \beta_1 = 0$



Regression

Including Molecular Markers

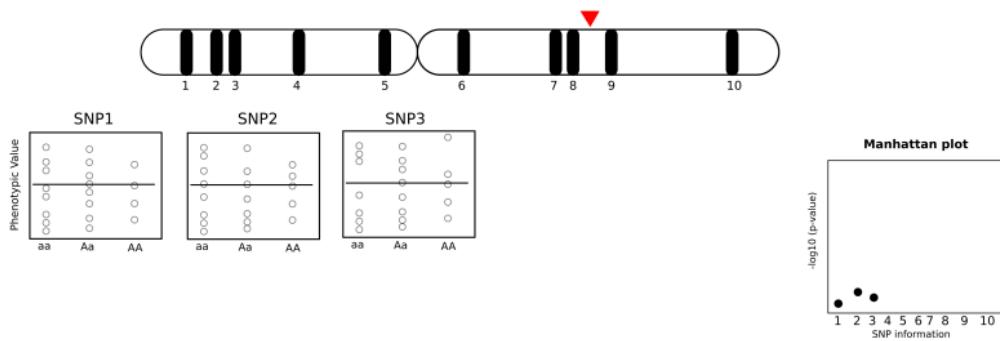
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(\text{marker})$ and testing $H_0 : \beta_1 = 0$



Regression

Including Molecular Markers

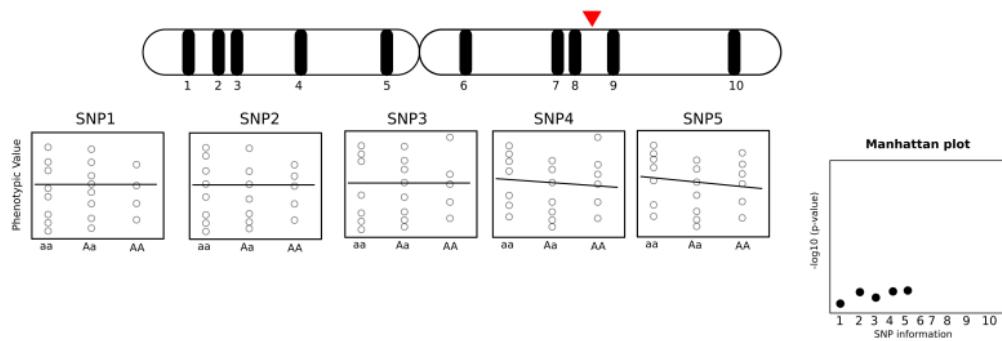
- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(\text{marker})$ and testing $H_0 : \beta_1 = 0$



Regression

Including Molecular Markers

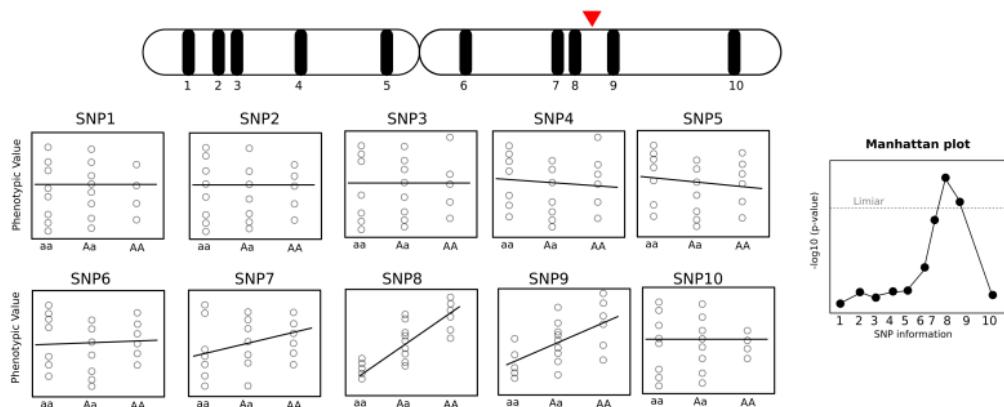
- 10 markers (allele dosage) and a single QTL (red)
 - 20 individuals measured for a given phenotypic trait (ex: yield)
 - Regression Model: $y \sim f(\text{marker})$ and testing $H_0 : \beta_1 = 0$



Regression

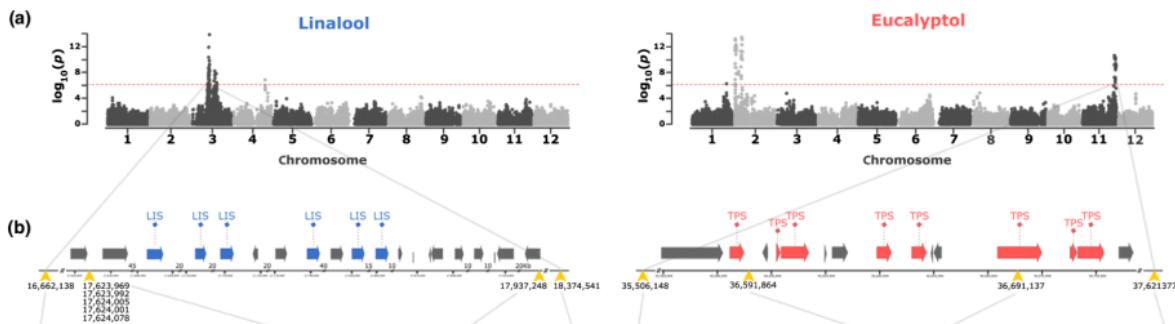
Including Molecular Markers

- 10 markers (allele dosage) and a single QTL (red)
- 20 individuals measured for a given phenotypic trait (ex: yield)
- Regression Model: $y \sim f(\text{marker})$ 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*