Fundamentals of Genomic Prediction and Data-Driven Crop Breeding (August 4-8, 2025)

Basics: Mean, Variances and Linkage Disequilibrium

Module 2
August 5, 2025

Waseem Hussain and Mahender Anumalla
Rice Breeding Innovations Platform
IRRI

Quantitative Genetics

- Deals with complex traits determined by genes and environment
- Vary continuously and highly uninformative
- Effected by alleles at multiple loci
- Inherited as Mendelian Traits

- *Genes influencing Quantitative traits called as **QTLs***
- *Same principles of inheritance, however, more genes take part in the determination of quantitative characteristics.*

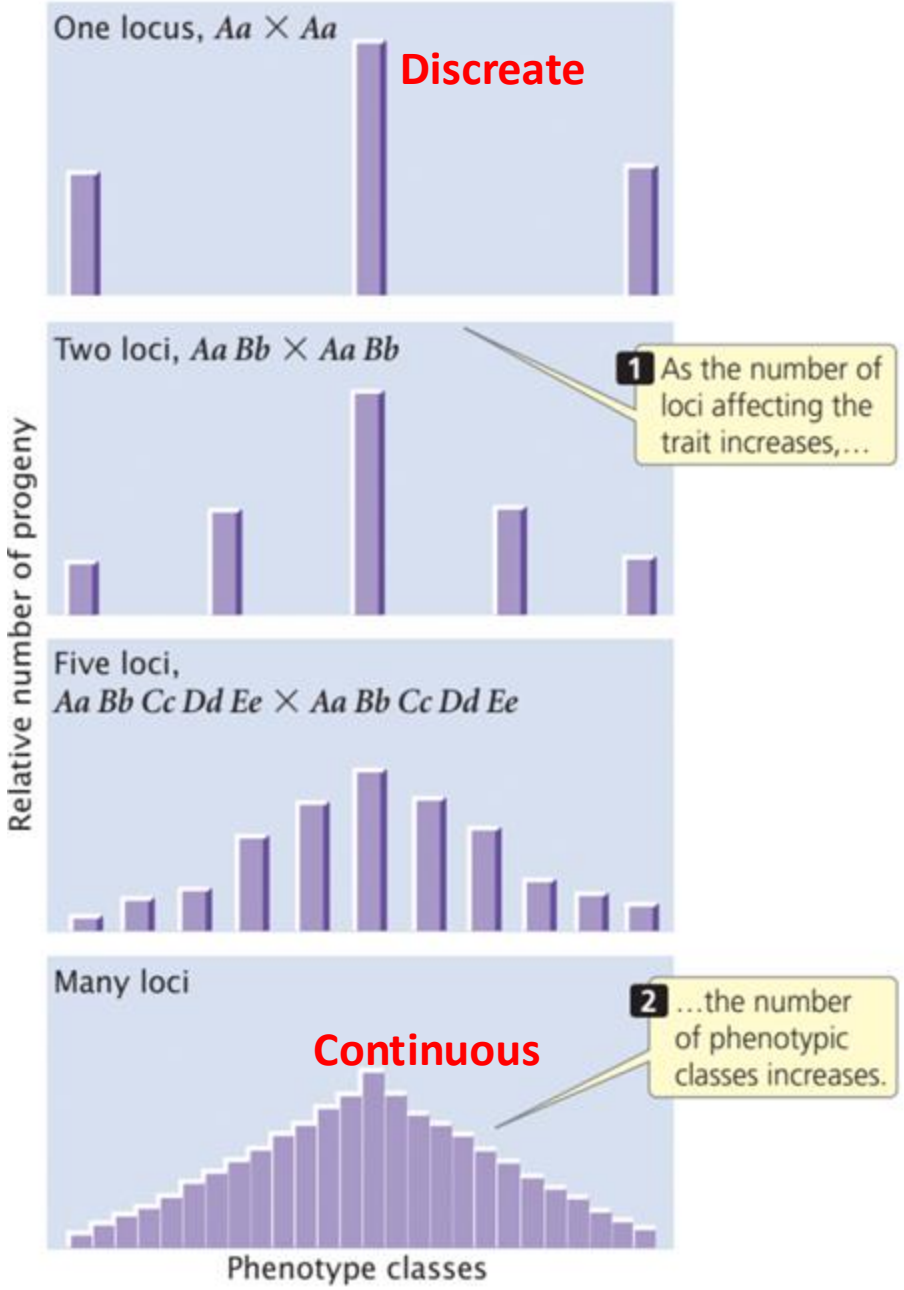


Figure adopted from Book Conceptual approaches to Genetics by Benjamin A. Pierce

Measurement of Quantitative Traits

Quantitative Traits are Measured at the Population Level using Statistical Tools

MEAN

Center of distribution of data

VARIANCE

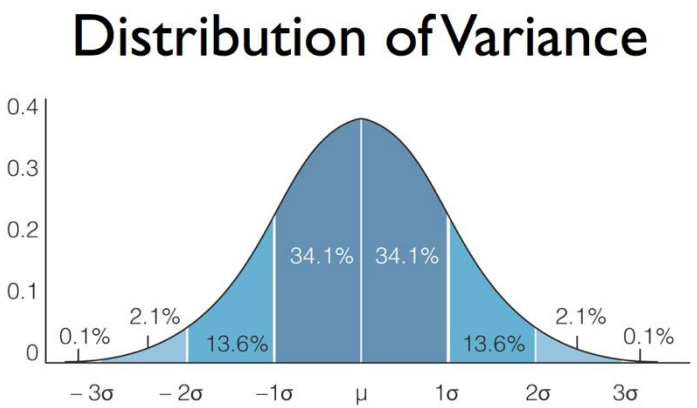
Spread of the data

CO-VARIANCES

Co-vary together

CORRELATIONS

Relationship between variables



REGRESSION

Relationship between variables

$$Mean = \frac{1}{n}\sum_{i}^{n} x_i$$

$$Variance = \frac{1}{n}\sum_{i}^{n} (x_i - \mu)^2$$

$$COV(X,Y) = \sum p_{xy}(x_i - \mu_x)\,(y_i - \mu_y)$$

Basic Quantitative Genetics Model

$$\boldsymbol{Phenotype}(\boldsymbol{P}) = \boldsymbol{Genotype}\ (\boldsymbol{G}) + \boldsymbol{Environment}(\boldsymbol{E}) + \boldsymbol{Residual}\ (\boldsymbol{e})$$

or

$$y_{ijk} = \mu_i + G_i + E_j + e_{ijk}$$

Trait value observed in Individual

Average phenotypic value over a range of environments $G = E(P)$

Environments Individuals evaluated

Residual: Uncontrolled

Dissecting Basic Quantitative Genetics Model

Genotype	A_2A_2	0	A_2A_1	A_1A_1
Genotype values	MP-a	MP	MP+d	MP+a
Code genotype	-a	0	d	a
Frequency	q^2	2pq		p^2

Mid parent (MP)= $\frac{A_1A_1+A_2A_2}{2}$ $(Half\ difference\ between\ homozygotes)$

$$Mean = q^2(MP - a) + 2pq(MP) + P^2(MP + a)$$

$$Mean = MP + a(p - q) + 2pqd$$

Heterozygote contribution

Homozygote contribution

a is additive effect
d is dominance
d=a, complete dominance
d=0, no dominance
0<d<a, partial dominance
d>a overdominance

For RILs, d=0, thus

$$\boldsymbol{Mean = MP + a(p - q)}$$

Multiple loci

$$Mean = \sum_{i=1}^{k} a_i(p_i - q_i) + 2\sum_{i=1}^{k} p_i q_i d_i$$

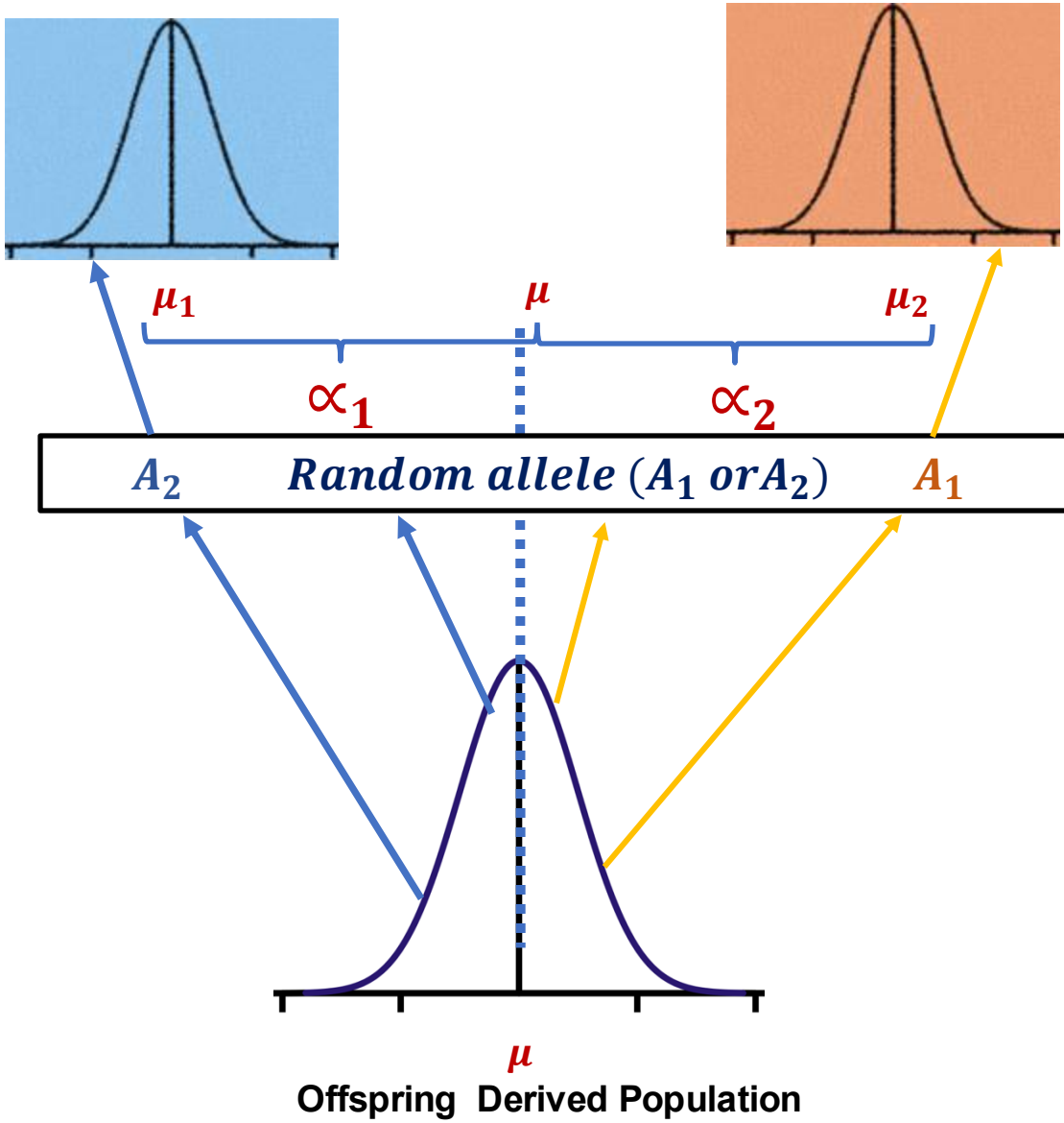
where mean is summation over all the loci and *k* is number of loci

Average Effect and Why

- Parents **pass a single allele** to offspring at a given locus, and hence **PART of genotypic value** is passed, which is determined as the **Average Effect.**

- It is a deviation of individuals from the population mean who received a particular allele from one parent, and the other allele comes as random

$$\propto_x = \mu - \mu_x$$
$$\propto_1 = q(a + d(q - p))$$
$$\propto_2 = -p(a + d(q - p))$$
$$\propto = -p(a + d(q - p))$$

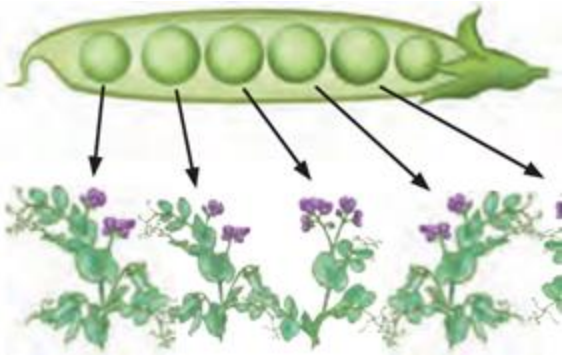


Average Effect and Why

- Alleles, not genotypes, are inherited
- Link the genotype values from one generation to another
- Average effect is population, depends upon allele frequency
- Average effect of random allele is 0?

Mendelian Trait

an allele causes certain phenotype



Quantitative Trait

an allele has a large average effect

Note

We don't measure individual loci average effects; we use Regression to get the effects.

Breeding value (Additive Genetic Value)

Deviations expected from the offspring of a particular genotype when it is mated with another individual of the same genetic worth

$$\boldsymbol{Genotype = mean + additive + dominance + epistais + error}$$

$$G_{ij} = \propto_i + \propto_j$$

A part of G (additive) is only transmitted

$$A = \sum_{k=1}^{n} \left(\propto_i^{(k)} \propto_j^{(k)} \right) = \sum_{k=1}^{n} a_{ik}$$

$\propto_i^{(k)}$ *and* $\propto_j^{(k)}$ = *is the effect of allele i and j and locus k*

A is called Breeding Value or Additive Genetic Value and

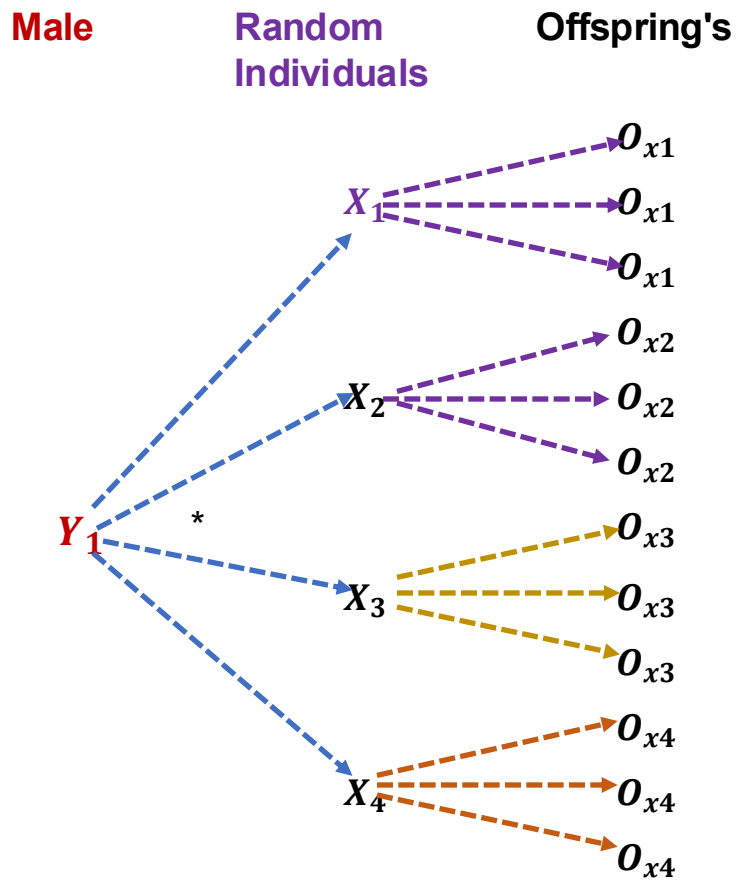
a_{ik} is additive genetic effect for k loci

Genetic value (Additive)= Sum of average effect of alleles

$$\boldsymbol{Mean\ BV\ of\ all\ genotypes = 2pq(p\alpha + q\alpha - p\alpha - q\alpha) = 0}$$

$$y \sim N(\mu, \sigma_a^2 + \sigma_\epsilon^2)$$
$$a \sim N(0, \sigma_a^2)$$
$$e \sim N(0, \sigma_\epsilon^2)$$

The breeding value is population specific (depends upon allele frequency)



$$BV = 2\ (mean\ progeny_{\ A1A1} - Mean\ (Population)$$

$$E(BV) = \frac{p_1 + p_2}{2} + \delta$$

δ **is Mendelian Sampling** (variation between offspring between same parent)

Breeding value (Additive Genetic Value)

Expresses the value transmitted from parent to offspring

- **We don't observe breeding values, we estimate it**
- **Breeding values (EBVs) are estimated from phenotypic (True values) and relative information using a mixed model (Regression); BLUP**
- **Conditional on Phenotypic value** $(E(a|y)$

How Selection Accuracy and Reliability are related to EBVs

$$E(a|y) = \underbrace{cov(a,y)[var(y)]^{-1}}_{\text{Regression coefficient (b)}} \underbrace{(y-\mu)}_{\text{Adjusted mean}}$$

Breeding value is adjusted mean weighted by the coefficient that is heritability

Under a simple genetic model

$$y_i = \mu + a_1 + \varepsilon_i$$

$$b_{(a,y)} = \frac{Cov[y_i, a_i]}{Var[_{y_i}]} = \frac{\sigma_a^2}{\sigma_y^2} = h^2$$

$$\hat{a}_i = h^2(y_i - \mu)$$

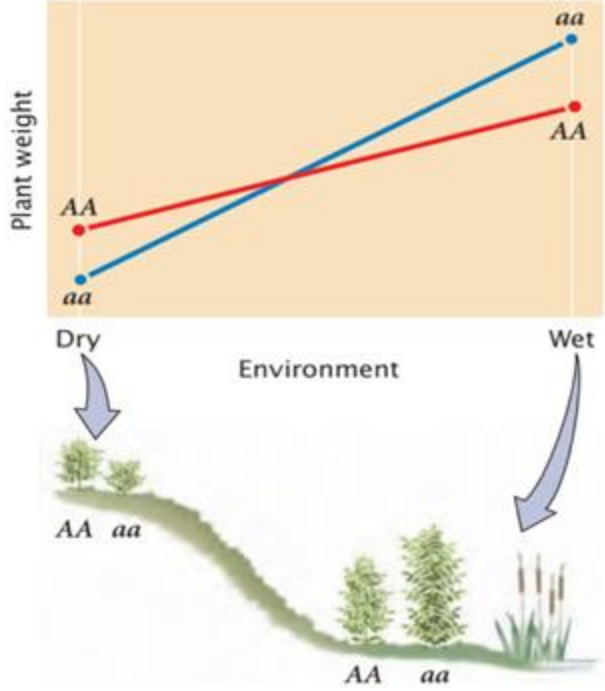
EBV is adjusted mean multiplied by the heritability

Variances and Co-variances

$$\boldsymbol{Phenotype(P)} = \boldsymbol{Genotype\ (G)} + \boldsymbol{Environment(E)} + \boldsymbol{Residual\ (e)}$$

$$V_P = V_g + V_e + V_{ge}$$

$$\boxed{V_a + V_d + V_i}$$



Heritability is used to measure the proportion of variance

The proportion of variation that is due to genetics is heritability

Broad Sense heritability(H^2) $= \frac{V_g}{V_p}$

Broad Sense heritability(h^2) $= \frac{V_a}{V_g}$

$$Cov(a_x, a_y) = A_{xy}\sigma_a^2$$

where, a_x *and* a_y are the breeding BV's and σ_a^2 is the additive variance

And A is the additive genetic relationship related to IBD (*identity –by-descent*)

Dominance Deviation

$$V_P = V_g + V_e + V_{gxe}$$

V_P: *total varaince*
V_g: *Genotype vara*
V_e: *environmental*
V_{gxe}: variance due to interactions

$$V_g = V_a + V_d + V_i$$

Dominance deviations are obtained by taking the difference between the genotypic value and the breeding value for each genotype.

Points to Note:

- When d=0, a genotype has a breeding value that is identical to its genotypic value
- Without dominance, genotypic values in progeny as measured by the breeding value can be predicted perfectly from the combination of average effects.

Covariance Between Individuals

$$Cov(X,Y) = 2f_{xy}\sigma_a^2 (assuming\ Dominance = 0)$$

Where f_{xy} is the coefficient of co-ancestary between two individuals and σ_a^2 is the additive genetic variance

Example of Variance and Co-varaince

Genotypes	1	2	3	4
1	Var (1)	cov(1,2)	cov(1,3)	cov(1,4)
2		Var (2)	cov(2,3)	cov(2,4)
3			Var (1)	cov(3,4)
4				Var (4)

$$\mathrm{var}(u) = A * \sigma_a^2$$

Elements of **the additive relationship matrix (A)** has the elements given by the $\mathrm{u(x,y)} = 2f_{xy}$

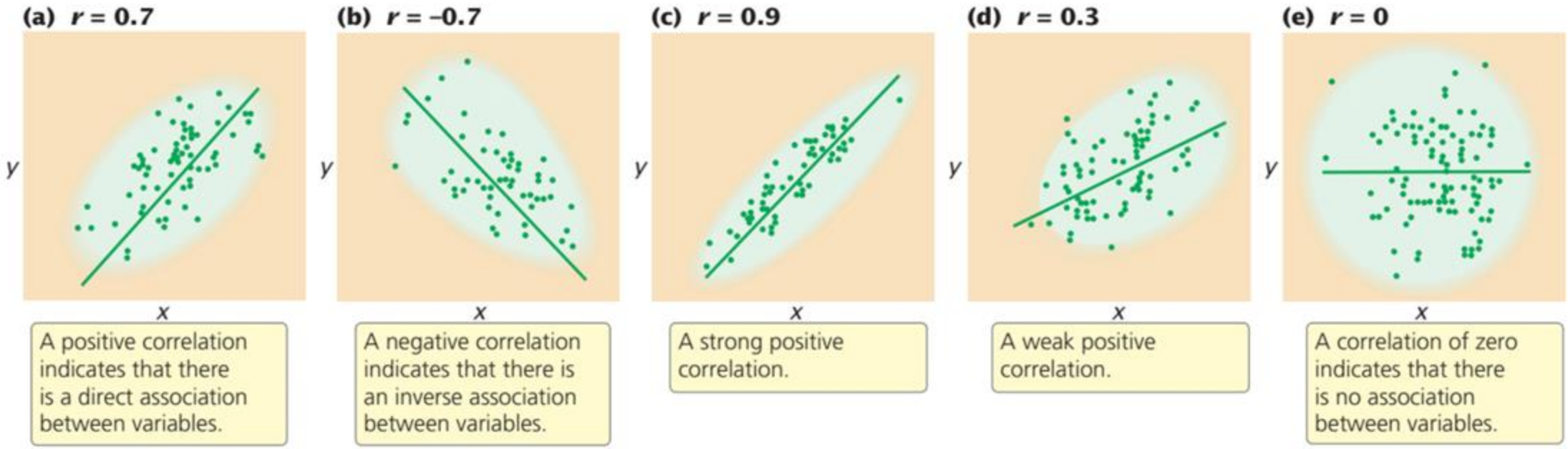
Note:

- **The diagonal represents variance and off-diagonal covariance**
- **If genotypes are independent off-diagonal elements=0**
- **Non-zero elements of off-diagonal reflect use of relative information for BLUP estimation**

Correlations

(*with more than one character*)

Relationship between two variables



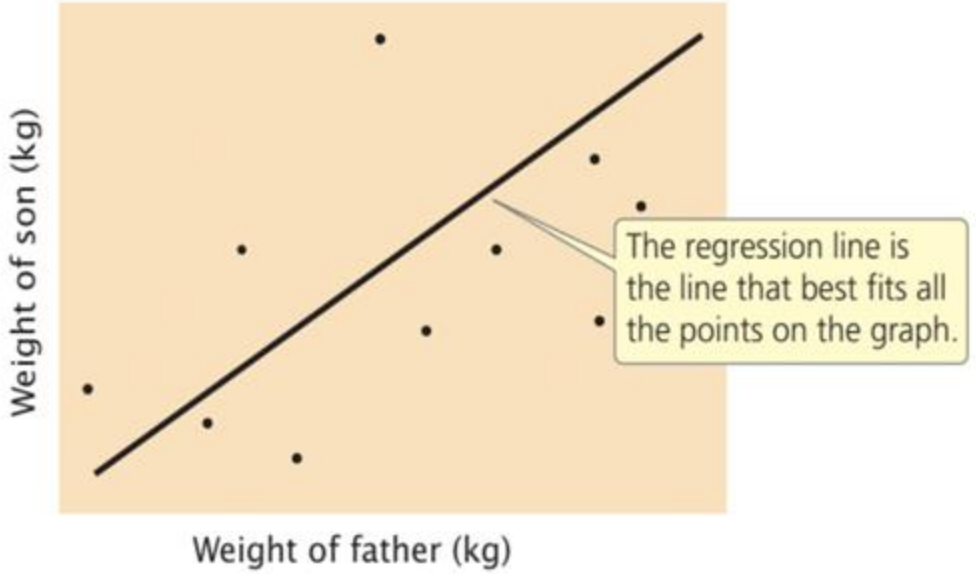
24.11 The correlation coefficient describes the relation between two or more variables.

Linkage disequilibrium is simply correlation between the SNPs

Figure adopted from Book Conceptual approaches to Genetics by Benjamin A. Pierce

Regression

- Correlation provides information only about the strength and direction of association between variables.
- Whether two variables are associated.
- Basically, to predict the value of one variable, given a value of the other.



24.13 A regression line defines the relation between two variables. Illustrated here is a regression of the weights of fathers against the weights of sons. Each father–son pair is represented by a point on the graph: the *x* value of a point is the father's weight and the *y* value of the point is the son's weight.

Dissect the Regression line

$$y = a + bx$$

- x and y represent the x and y variables
- a is intercept
- B is slope or regression coefficient.

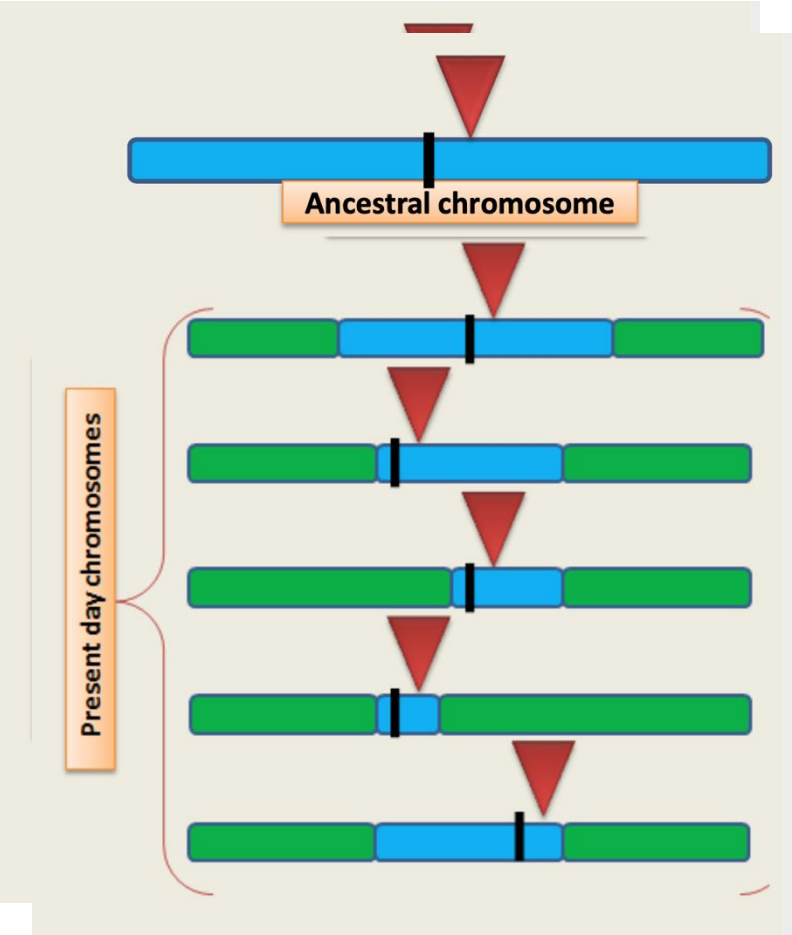
$$b = \frac{cov(x,y)}{varaince(x)}$$

Regression coefficient indicates how much increases, on average, per increase in x

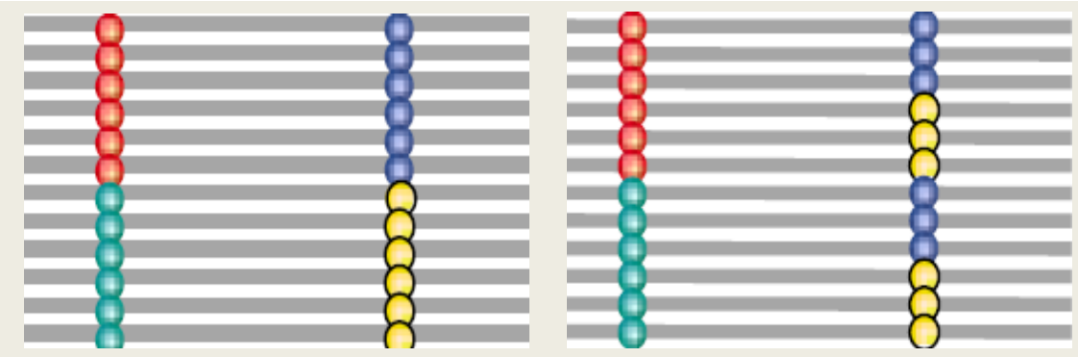
Figure adopted from Book Conceptual approaches to Genetics by Benjamin A. Pierce

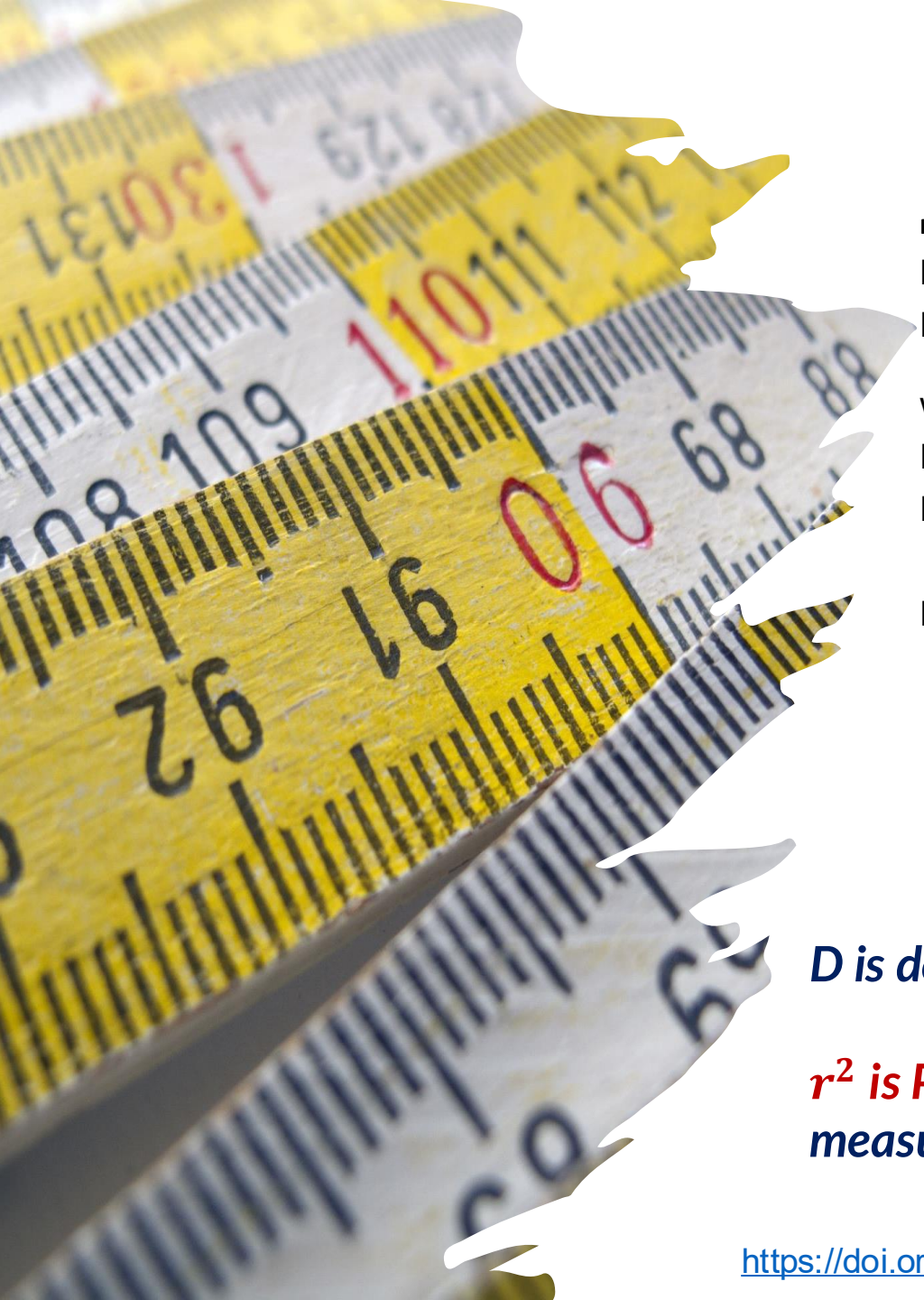
Linkage Disequilibrium

- Non-random association of alleles at adjacent locus
- Closer the markers higher is the LD
- The resolution with which the QTL can be mapped is function of LD.
- Very Important in Genomic Selection to determine the prediction accuracy and number of markers.



LD across the Historical Combinations





How to Measure LD

Mathematically,

Linkage equilibrium **PAB= PA x PB**

Linkage disequilibrium PAB ≠ PA x PB

where A and B are alleles at two different loci,

PAB is the frequency of haplotypes having both alleles at the two loci,

PA and PB are the frequency of haplotypes having only A allele and B allele, respectively.

D ranges from 0 -1 (At equilibrium, D= 0)

$$D` = \frac{D}{|D|}$$

$$r^2 = \frac{D^2}{p_A p_B (1-p_A)(1-p_B)}$$ **(Hill and Robertson, 1968)**

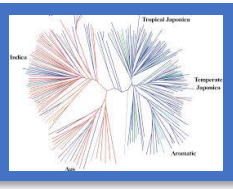
D is determined by the range of allele frequency

r^2 is Pearson's correlation coeffenicit, and is the most relevant LD measurement (0-1)

https://doi.org/10.1534/genetics.118.300642; https://doi.org/10.1105/tpc.150730

Factors Affecting the LD

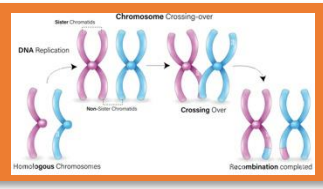
01



Germplasm

Longer in narrow base germplasm. 1 kb in cultivated and 2kb in landraces.

02

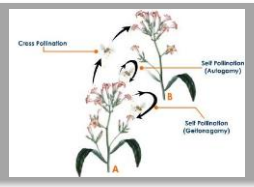


Recombination

Recombination decreases LD. LD decays high with each generation or mating

$$D_{t+1} = (1 - c)D_t$$

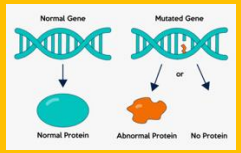
03



Mating System

Longer in self-pollinated crops.
Less recombination in Self pollinated
LD decays faster in Cross Pollinated crops

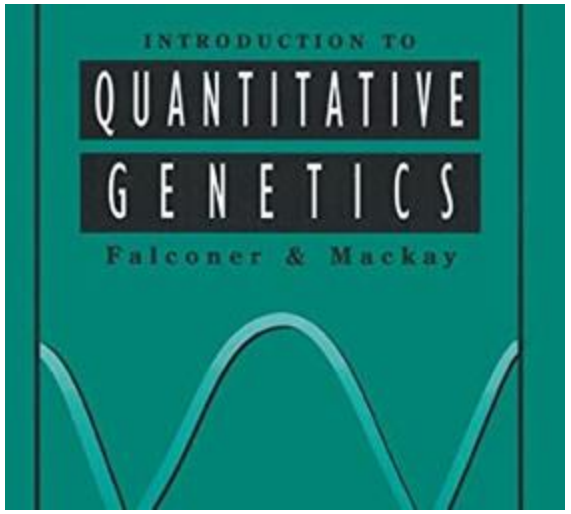
04

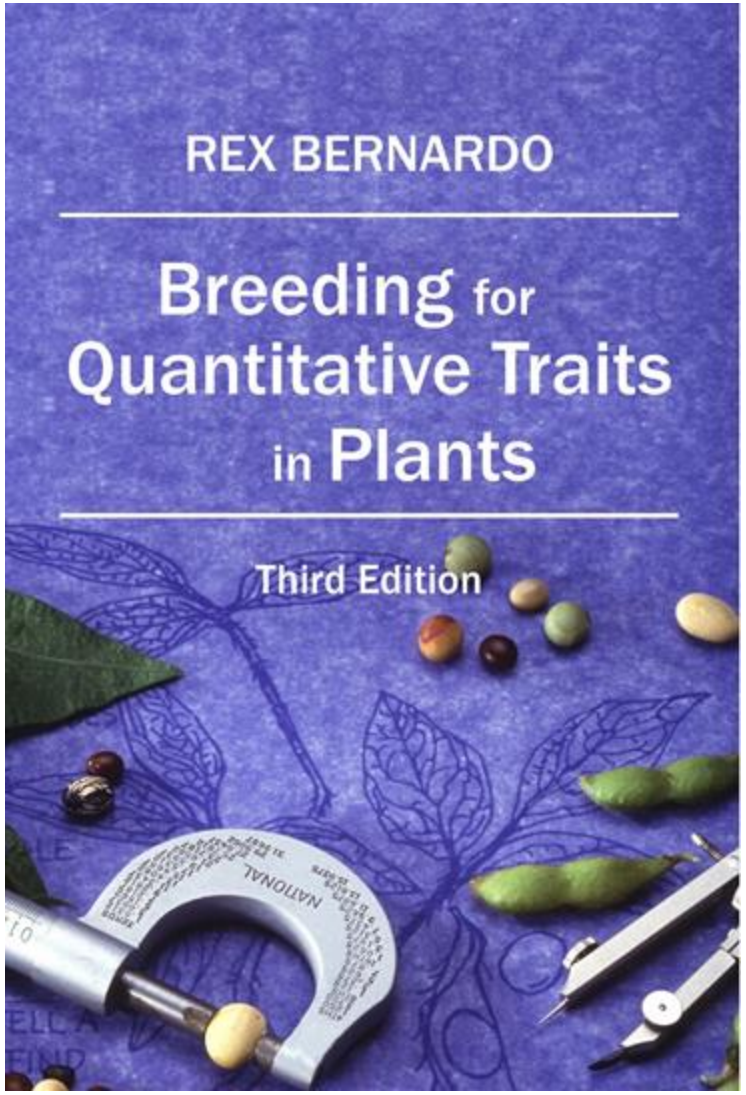


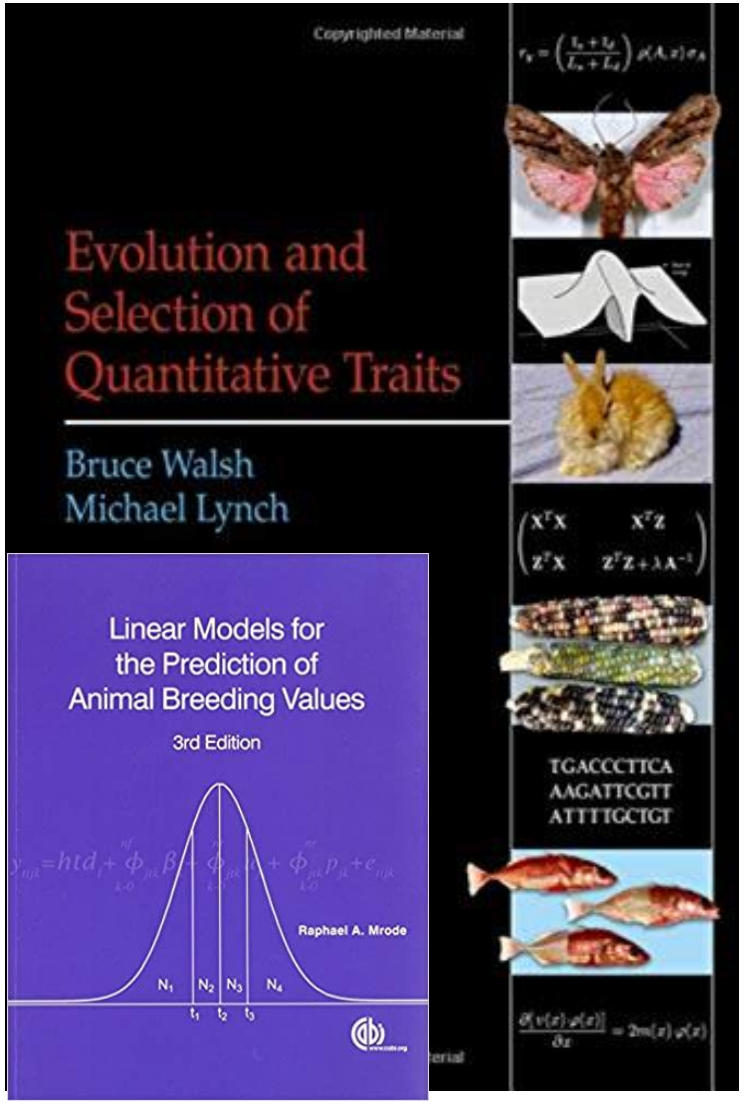
Mutations

Creation of LD and is eroded by Recombination.


Genetic Data Analysis for Plant and Animal Breeding


INTRODUCTION TO
QUANTITATIVE
GENETICS
Falconer & Mackay


REX BERNARDO
Breeding for Quantitative Traits in Plants
Third Edition


Copyrighted Material
Evolution and Selection of Quantitative Traits
Bruce Walsh
Michael Lynch
Linear Models for the Prediction of Animal Breeding Values
3rd Edition
Raphael A. Mrode
TGACCCTTCA
AAGATTCGTT
ATTTTGCTGT