

**SYLLABUS**  
**PRINCIPLES OF QUANTITATIVE GENETICS**  
**SISG, Seattle, 17 - 19 July 2017**

**INSTRUCTORS:**

Bruce Walsh, Department of Ecology & Evolutionary Biology, University of Arizona  
[jbwalsh@u.arizona.edu](mailto:jbwalsh@u.arizona.edu)

Guilherme Rosa, Department of Animal Sciences, University of Wisconsin, Madison  
[grosa@wisc.edu](mailto:grosa@wisc.edu)

**LECTURE SCHEDULE**

**Monday, 17 July**

8:30 10:00 am	1. Population Genetics Framework (Walsh)
10:00 10:20 am	Break
10:20 12:00	2. Fisher's Variance Decomposition (Walsh)
12:00 2:00 pm	Lunch
2:00 3:20 pm	3. Resemblance Between Relatives, Heritability (Walsh)
3:20 3:50 pm	Break
3:50 5:00 pm	4. Artificial Selection (Walsh)

**Tuesday, 18 July**

8:30 10:00 am	5. Inbreeding and Crossbreeding (Walsh)
10:00 10:20 am	Break
10:20 12:00	6. Correlated Characters (Walsh)
12:00 2:00 pm	Lunch
2:00 3:20 pm	7. Estimation of basic genetic parameters (Rosa)
3:20 3:50 pm	Break
3:50 5:00 pm	8. Mixed Models, BLUP Breeding Values (Rosa)

**Wednesday, 19 July**

8:30 10:00 am	9. QTL/Association Mapping (Rosa)
10:00 10:20 am	Break
10:20 12:00	10. Binary and count traits, repeated measurements, maternal effects (Rosa)

## ADDITIONAL BOOKS ON QUANTITATIVE GENETICS

### General

- Falconer, D. S. and T. F. C. Mackay. *Introduction to Quantitative Genetics*, 4<sup>th</sup> Edition  
Lynch, M. and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer.  
Roff, D. A. 1997. *Evolutionary Quantitative Genetics*. Chapman and Hall.  
Mather, K., and J. L. Jinks. 1982. *Biometrical Genetics*. (3<sup>rd</sup> Ed.) Chapman & Hall.

### Animal Breeding

- Cameron, N. D. 1997. *Selection Indices and Prediction of Genetic Merit in Animal Breeding*. CAB International.  
Mrode, R. A. 1996. *Linear Models for the Prediction of Animal Breeding Values*. CAB International.  
Simm, G. 1998. *Genetic Improvement of Cattle and Sheep*. Farming Press.  
Turner, H. N., and S. S. Y. Young. 1969. *Quantitative Genetics in Sheep Breeding*. Cornell University Press.  
Weller, J. I. 2001. *Quantitative Trait Loci Analysis in Animals*. CABI Publishing.

### Plant Breeding

- Acquaah, G. 2007. *Principles of Plant Genetics and Breeding*. Blackwell.  
Bernardo, R. 2002. *Breeding for Quantitative Traits in Plants*. Stemma Press.  
Hallauer, A. R., and J. B. Miranda. 1986. *Quantitative Genetics in Maize Breeding*. Iowa State Press.  
Mayo, O. 1987. *The Theory of Plant Breeding*. Oxford.  
Sleper, D. A., and J. M. Poehlman. 2006. *Breeding Field Crops*. 5<sup>th</sup> Edition. Blackwell  
Wricke, G., and W. E. Weber. 1986. *Quantitative Genetics and Selection in Plant Breeding*. De Gruyter.

### Humans

- Khoury, M. J., T. H. Beaty, and B. H. Cohen. 1993. *Fundamentals of Genetic Epidemiology*. Oxford.  
Plomin, R., J. C. DeFries, G. E. McLearn, and P. McGuffin. 2002. *Behavioral Genetics* (4<sup>th</sup> Ed) Worth Publishers.  
Sham, P. 1998. *Statistics in Human Genetics*. Arnold.  
Thomas, D. C. 2004. *Statistical Methods in Genetic Epidemiology*. Oxford.  
Weiss, K. M. 1993. *Genetic Variation and Human Disease*. Cambridge.  
Ziegler, A., and I. R. König. 2006. *A Statistical Approach to Genetic Epidemiology*. Wiley.

### Statistical and Technical Issues

- Bulmer, M. 1980. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press.

Kempthorne, O. 1969. *An Introduction to Genetic Statistics*. Iowa State University Press.

Saxton, A. M. (Ed). 2004. *Genetic Analysis of Complex Traits Using SAS*. SAS Press.

Sorensen, D., and D. Gianola. 2002. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer.

# Lecture 1

## Hardy-Weinberg equilibrium and key forces affecting gene frequency

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

1

## Outline

- Genetics of complex traits
- Stability of distributions over time
- Hardy-Weinberg
- Multilocus Hardy-Weinberg
- Population Structure
- Selection

2

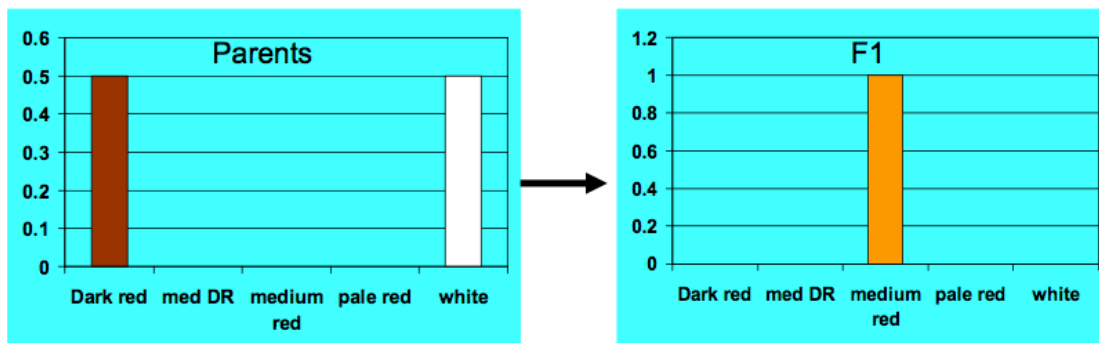


# Mendelian basis of complex traits

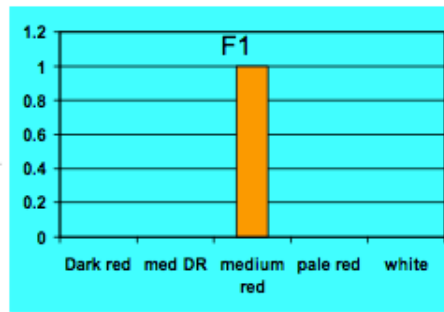
- Classic experiment of Nilsson-Ehle (1908) on wheat color
- “Simple” traits (green vs. yellow peas, etc.) had a single-gene basis
- Do complex traits have a different genetic basis?
  - Notion of **blending inheritance** (offspring = blended average of parents)

3

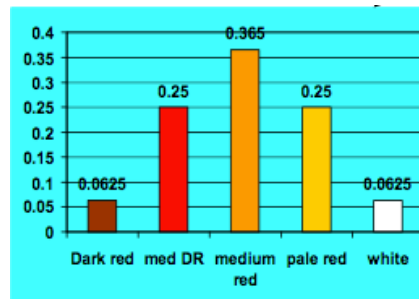
$F_1$  in a cross of dark red pure line x white pure line seems to support blending



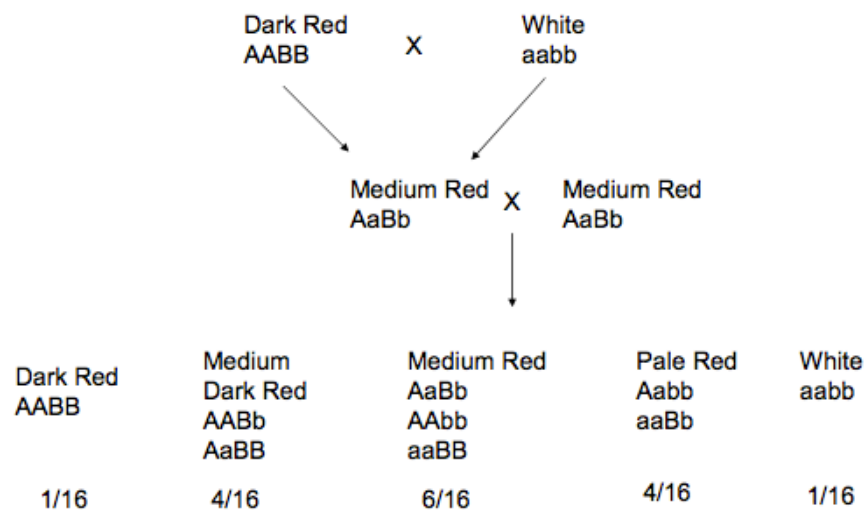
4



However, “**outbreak of variation**” in the  $F_2$  rules out blending

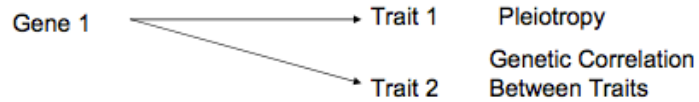
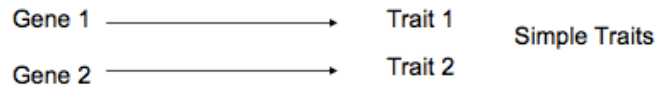


Hypothesis: 2 loci acting independently and cumulatively on one trait?



# Gene Effects

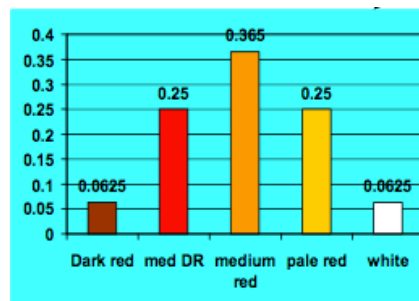
Usual Mendelian Concept



Stability of the phenotypic  
distribution over time

# Stability of the phenotype distribution

The parental lines, F1, and F2 all differ from each other. What happens to the distribution of F2 trait values in the F3, F4, Fx?



9

## Case 1: random mating

- Suppose the F2 are randomly mated. What are the genotype frequencies in the following generation?
- These are given by the Hardy-Weinberg theorem.
- If  $p = \text{freq}(A)$  and  $q = \text{freq}(a)$ , then
  - $\text{freq}(AA) = p^2$
  - $\text{freq}(Aa) = 2pq$
  - $\text{freq}(aa) = q^2$

10

- Here  $\text{freq}(A) = \text{freq}(a) = \frac{1}{2}$ , and  $\text{freq}(B) = \text{freq}(b) = \frac{1}{2}$ . Assuming the A and B loci are unlinked, then independent assortment gives
  - $\text{Freq}(\text{dark red}) = \text{Freq}(AABB) = \text{freq}(AA) * \text{freq}(BB) = (\frac{1}{4}) (\frac{1}{4}) = 0.0625$
  - $\text{Freq}(\text{white}) = \text{freq}(aabb) = \text{freq}(aa) * \text{freq}(bb) = 0.0625$
  - $\text{Freq}(\text{med red}) = \text{freq}(AAbb \text{ or } AaBb \text{ or } aaBB)$ 
    - $= (\frac{1}{4}) * (\frac{1}{4}) + (\frac{1}{2}) * (\frac{1}{2}) + (\frac{1}{4}) * (\frac{1}{4}) = 0.375$
- Hence, the distribution of phenotypes in the F3 is the same as the F2. What about in the F4? F5?

11

## Case 2: Inbred lines

- Suppose instead that each F2 is used to form an inbred line, and continually selfed over many generations. What happens to the distribution after complete selfing?
- Now each locus is a homozygote, with  $\text{Freq}(AA) = \text{freq}(aa) = \text{freq}(BB) = \text{freq}(bb) = \frac{1}{2}$ 
  - $AABB = \text{dark red (25\%)}$
  - $AAbb, aaBB = \text{medium red (50\%)}$
  - $aabb = \text{white (25\%)}$

12

# During selfing

- During selfing, an AA or aa line only produces AA /aa.  
However, an Aa line has probability  $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$  of producing AA : Aa : aa
- Hence, after one generation of selfing
  - $\text{Freq}(\text{AA}) = \text{Freq}(\text{AA} \mid \text{parent AA}) + \text{Freq}(\text{AA} \mid \text{parent Aa}) = 1 \cdot (\frac{1}{4}) + (\frac{1}{4}) \cdot (\frac{1}{2}) = \frac{3}{8}$
  - $\text{Freq}(\text{aa}) = \frac{3}{8}$ ,  $\text{freq}(\text{Aa}) = \frac{1}{4}$
  - Same for the B locus
- Resulting phenotypic (seed color) frequencies are
  - $\text{Freq}(\text{dark red}) = \text{Freq}(\text{AABB}) = \text{freq}(\text{AA}) \cdot \text{freq}(\text{BB}) = (\frac{3}{8}) \cdot (\frac{3}{8}) = 0.1406$
  - $\text{Freq}(\text{white}) = \text{freq}(\text{aabb}) = \text{freq}(\text{aa}) \cdot \text{freq}(\text{bb}) = 0.1406$
  - $\text{Freq}(\text{med red}) = \text{freq}(\text{AAbb or AaBb or aaBB})$ 
    - $= (\frac{3}{8}) \cdot (\frac{3}{8}) + (\frac{2}{8}) \cdot (\frac{2}{8}) + (\frac{3}{8}) \cdot (\frac{3}{8}) = 0.344$

13

# Hardy-Weinberg

# Importance of HW

- HW states that the distribution of genotypes in a population are stable under random mating, provided no
  - **Drift** (i.e., pop size is large)
  - **Migration** (i.e., no input of individuals from other populations/breeding programs)
  - **Selection** (no forces to systemically change allele frequencies)

15

## Derivation of the Hardy-Weinberg result

- Consider any population, where
  - $\text{Freq}(AA) = X$
  - $\text{Freq}(Aa) = Y$
  - $\text{Freq}(aa) = Z$
  - $\text{freq}(A) = p = \text{freq}(AA) + (1/2) \text{freq}(Aa) = X + \frac{1}{2} Y$
- What happens in the next generation from random mating?

16

## Frequency of matings

female genotype frequency		male genotype		
		AA (X)	Aa (Y)	aa (Z)
AA	(X)	$X^2$	XY	XZ
Aa	(Y)	XY	$Y^2$	YZ
aa	(Z)	XZ	YZ	$Z^2$

Random Mating=independence

## Genotype frequencies in next generation

Possible Matings	Frequency of Mating	Expected Frequency of Offspring		
		AA	Aa	aa
AA x AA	$X^2$	1	0	0
AA x Aa	2XY	1/2	1/2	0
AA x aa	2XZ	0	1	0
Aa x Aa	$Y^2$	1/4	1/2	1/4
Aa x aa	2YZ	0	1/2	1/2
aa x aa	$Z^2$	0	0	1

Conditional Probabilities given genotypes of parents

$$\text{Freq}(\text{AA}) = 1 * X^2 + \frac{1}{2} * 2XY + \left(\frac{1}{4}\right) Y^2 = (X + \frac{1}{2} Y)^2 = p^2.$$

$$\text{Freq}(\text{aa}) = 1 * Z^2 + \frac{1}{2} * 2YZ + \left(\frac{1}{4}\right) Y^2 = (Z + \frac{1}{2} Y)^2 = q^2.$$



# What about the next generation?

Possible Matings	Frequency of Mating	Expected Frequency of Offspring		
		AA	Aa	aa
AA x AA	$p^4$	1	0	0
AA x Aa	$4p^3q$	1/2	1/2	0
AA x aa	$2p^2q^2$	0	1	0
Aa x Aa	$4p^2q^2$	1/4	1/2	1/4
Aa x aa	$4pq^3$	0	1/2	1/2
aa x aa	$q^4$	0	0	1

$$\text{Freq(AA)} = 1 * p^4 + \frac{1}{2} * 4p^3q + \left(\frac{1}{4}\right) 4p^2q^2 = p^2(p+q)^2 = p^2.$$

Genotype frequencies unchanged

19

## Hardy-Weinberg

genotype	gen 0	gen 1	gen 2
P( AA )	X	$p^2$	$p^2$
P( Aa )	Y	$2pq$	$2pq$
P( aa )	Z	$q^2$	$q^2$

After one generation of random mating, genotype frequencies remain unchanged and are given by HW proportions

Assuming random mating, no migration, drift, or selection, then allele frequencies remain unchanged

More generally, for any number of alleles,  $\text{freq}(A_i A_i) = p_i^2$ ,  
 $\text{freq}(A_i A_j) = 2p_i p_j$ .

# Hybridization

- Hardy-Weinberg assumes allele frequencies are the same in both sexes. If not, then after one generation of random mating, the frequencies of autosomal alleles is the same in both sexes, and HW is obtained on the second generation
- Suppose  $\text{Freq}(A \text{ in males}) = p_m$ ,  $\text{Freq}(A \text{ in females}) = p_f$ . Average allele frequency  $p = (p_m + p_f)/2$ .
- In generation one,
  - $\text{Freq}(AA) = p_m * p_f$  which is different from  $p^2$  if  $p_m$  &  $p_f$  differ
  - $\text{Freq}(Aa) = p_m (1 - p_f) + (1 - p_m) p_f$

21

## Example

- Cross females from a pop where  $p_f = 0.4$  with males from a pop where  $p_m = 0.6$ . Average frequency = 0.5.
  - Under random-mating,  $\text{freq}(Aa) = 0.5$
  - Here,  $\text{Freq}(Aa) = p_m (1 - p_f) + (1 - p_m) p_f = 0.4 * 0.4 + 0.6 * 0.6 = 0.52$
  - Hence, with crosses between populations where allele frequencies differ, we see **an excess of heterozygotes**.
  - Excess in F1, Hardy-Weinberg values in F2.
  - Implications for persistence of heterosis.

22

## Crosses vs. synthetics

- In a **cross**, males and females are always from different populations.  
Example of **nonrandom mating**!
- In a **synthetic**, all individuals are randomly-mated, therefore F2 is in HW
- Example: equal mix of P1 X P2
  - In a synthetic, 25% of crosses are P1 X P1, 50% P1 x P2, 25% P2 x P2.

23

## Multi-locus Hardy-Weinberg

24

# Multi-locus HW

- When following multiple loci, we need to consider gametes, rather than alleles
  - For example, an AaBb parent gives four distinct gametes AB, Ab, aB, ab
  - While allele frequencies do not change under random mating, gamete frequencies can.
  - Concept of linkage disequilibrium

25

## Genotypic frequencies under HW

- Under multi-locus HW,
  - $\text{Freq(AABB)} = \text{Freq(AA)} * \text{Freq(BB)}$
  - i.e., can use single-locus HW on each locus, and then multiply the results
- When D is non-zero (LD is present), cannot use this approach
  - Rather, must follow gametes

26

# Linkage Disequilibrium

- Under linkage equilibrium, the frequency of gametes is the product of allele frequencies,
  - e.g.  $\text{Freq}(AB) = \text{Freq}(A) * \text{Freq}(B)$
  - A and B are **independent** of each other
- If the linkage phase of parents in some set or population departs from random (alleles not independent) , linkage disequilibrium (LD) is said to occur
- The amount  $D_{AB}$  of disequilibrium for the AB gamete is given by
  - $D_{AB} = \text{Freq}(AB) \text{ gamete} - \text{Freq}(A) * \text{Freq}(B)$
  - $D > 0$  implies AB gamete more frequent than expected
  - $D < 0$  implies AB less frequent than expected

27

## The Decay of Linkage Disequilibrium

The frequency of the AB gamete is given by

$$\text{freq}(AB) = \underbrace{\text{freq}(A) \text{freq}(B)}_{\text{LE value}} + \underbrace{D_{AB}}_{\text{Departure from LE}}$$

If recombination frequency between the A and B loci is  $c$ , the disequilibrium in generation  $t$  is

$$D(t) = \underbrace{D(0)}_{\text{Initial LD value}} (1 - c)^t$$

Note that  $D(t) \rightarrow \text{zero}$ , although the approach can be slow when  $c$  is very small

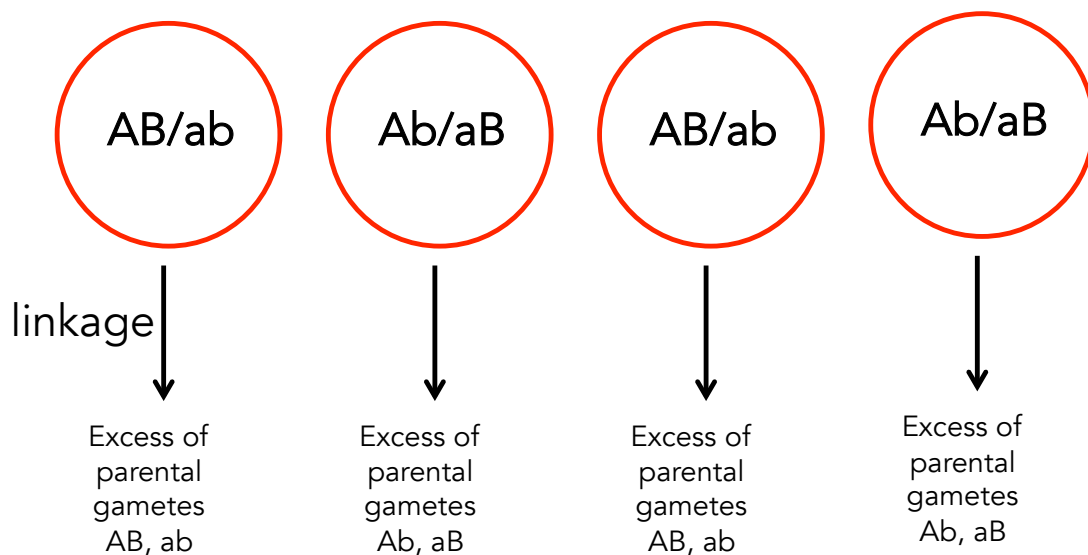
28

# Dynamics of D

- Under random mating in a large population, allele frequencies do not change. However, gamete frequencies do if there is any LD
- The amount of LD decays by  $(1-c)$  each generation
  - $D(t) = (1-c)^t D(0)$
- The expected frequency of a gamete (say AB) is
  - $\text{Freq}(AB) = \text{Freq}(A) \cdot \text{Freq}(B) + D$
  - $\text{Freq}(AB \text{ in gen } t) = \text{Freq}(A) \cdot \text{Freq}(B) + (1-c)^t D(0)$

29

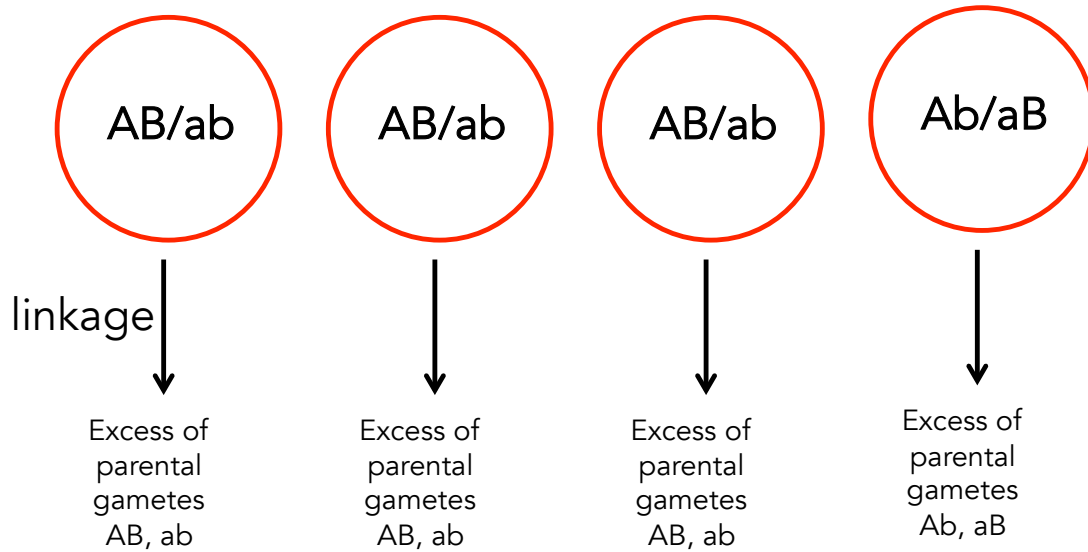
No LD: random distribution of linkage phases



Pool all gametes: AB, ab, Ab, aB equally frequent

30

With LD, nonrandom distribution of linkage phase



Pool all gametes: Excess of AB, ab due to an excess of AB/ab parents

31

## Example

- Suppose  $\text{Freq}(A) = 0.4$ ,  $\text{freq}(B) = 0.3$ ,  $D = 0.1$
- $\text{Freq}(AB)$  gamete is  $\text{freq}(A) \cdot \text{freq}(B) + D$ 
  - $\text{Freq}(AB) = 0.4 \cdot 0.3 + 0.1 = 0.22$
- $\text{Freq}(AABB) = \text{Freq}(AB) \cdot \text{Freq}(AB) = 0.22^2 = 0.0484$
- At multilocus HW,
  - $\text{Freq}(AABB) = \text{Freq}(AA) \cdot \text{freq}(BB) = 0.4^2 \cdot 0.3^2 = 0.0192$
- Suppose  $c = 0.2$ . In next generation,
  - $D(1) = (1 - 0.2) \cdot D(0) = 0.8 \cdot 0.1 = 0.08$ ,
  - $\text{Freq}(AB) = 0.20$ ;  $\text{freq}(AABB) = 0.04$

32

# Population structure

33

## Population Structure

Populations often show **structure**, with an apparently single random-mating population instead consisting of a collection of several random-mating **subpopulations**

Suppose there are  $n$  subpopulations, and let  $w_k$  be the probability that a random individual is from population  $k$

Let  $p_{ik}$  denote the frequency of allele  $A_i$  in subpopulation  $k$ .

The overall frequency of allele  $A_i$  is

$$p_i = \sum_{k=1}^n w_k * p_{ik}$$

34



The frequency of  $A_i A_i$  in the population is just

$$\text{freq}(A_i A_i) = \sum_{k=1}^n w_k p_{ik}^2$$

Expressed in terms of the population frequency of  $A_i$ ,

$$\begin{aligned} \text{freq}(A_i A_i) &= p_i^2 - \left( p_i^2 - \sum_{k=1}^n w_k p_{ik}^2 \right) \\ &= p_i^2 + \text{Var}(p_i) \end{aligned}$$

Thus, unless the allele has the same frequency in each population ( $\text{Var}(p_i) = 0$ ), **the frequency of homozygotes exceeds that predicted from HW**

35

Similar logic gives the frequency of heterozygotes as

$$\text{freq}(A_i A_j) = 2p_i p_j + \text{Cov}(p_i, p_j)$$

Hence, when the population shows structure, **homozygotes are more common than predicted from HW**, while heterozygotes can be more (or less) common than expected under HW, as the covariance could be zero, positive, or negative

36

Population structure also generates disequilibrium

Again suppose there are  $k$  subpopulations, each in linkage equilibrium

The population frequency of  $A_i B_j$  gametes is

$$\text{Freq}(A_i B_j) = \sum_{k=1}^n w_k * p_{A_{ik}} * p_{B_{jk}}$$

The population-wide disequilibrium becomes

$$\begin{aligned} D_{ij} &= \text{Freq}(A_i B_j) - \text{Freq}(A_i) * \text{Freq}(B_j) \\ &= \sum_{k=1}^n w_k * p_{A_{ik}} * p_{B_{jk}} - \left( \sum_{k=1}^n w_k * p_{A_{ik}} \right) \left( \sum_{k=1}^n w_k * p_{B_{jk}} \right) \end{aligned}$$

37

Consider the simplest case of  $k = 2$  populations

Let  $p_i$  be the frequency of  $A_i$  in population 1,  
 $p_i + \delta_i$  in population 2.

Likewise, let  $q_j$  be the frequency of  $B_j$  in population 1,  
 $q_j + \delta_j$  in population 2.

The expected disequilibrium becomes

$$D_{ij} = \delta_i * \delta_j * [w_1(1 - w_1)]$$

Here,  $w_1$  is the frequency of population 1

38

## $F_{ST}$ , a measure of population structure

- One measure of population structure is given by **Wright's  $F_{ST}$  statistic** (also called the fixation index)
- Essentially, this is the fraction of genetic variation due to between-population differences in allele frequencies
- Changes in allele frequencies can be caused by evolutionary forces such as genetic drift, selection, and local adaptation
- Consider a biallelic locus (A, a). If p denotes overall population frequency of allele A,
  - then the overall population variance is  $p(1-p)$
  - $\text{Var}(p_i)$  = variance in p over subpopulations
  - **$F_{ST} = \text{Var}(p_i) / [p(1-p)]$**

39

## Example of $F_{ST}$ estimation

Population	Freq(A)
1	0.1
2	0.6
3	0.2
4	0.7

Assume all subpopulations contribute equally to the overall metapopulation

Overall freq(A) =  $p = (0.1 + 0.6 + 0.2 + 0.7)/4 = 0.4$

$$\text{Var}(p_i) = E(p_i^2) - [E(p_i)]^2 = E(p_i^2) - p^2$$

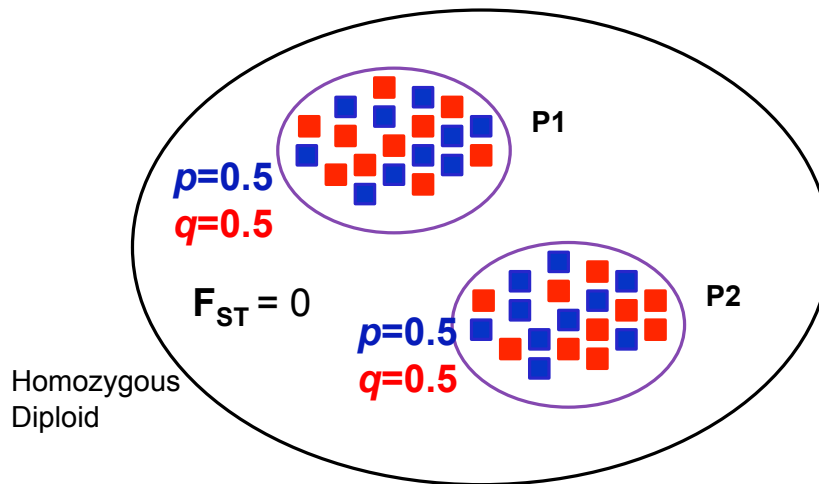
$$\text{Var}(p_i) = [(0.1^2 + 0.6^2 + 0.2^2 + 0.7^2)/4] - 0.4^2 = 0.065$$

$$\text{Total population variance} = p(1-p) = 0.4(1-0.4) = 0.24$$

$$\text{Hence, } F_{ST} = \text{Var}(p_i) / [p(1-p)] = 0.065/0.24 = 0.27$$

40

## Graphical example of $F_{ST}$

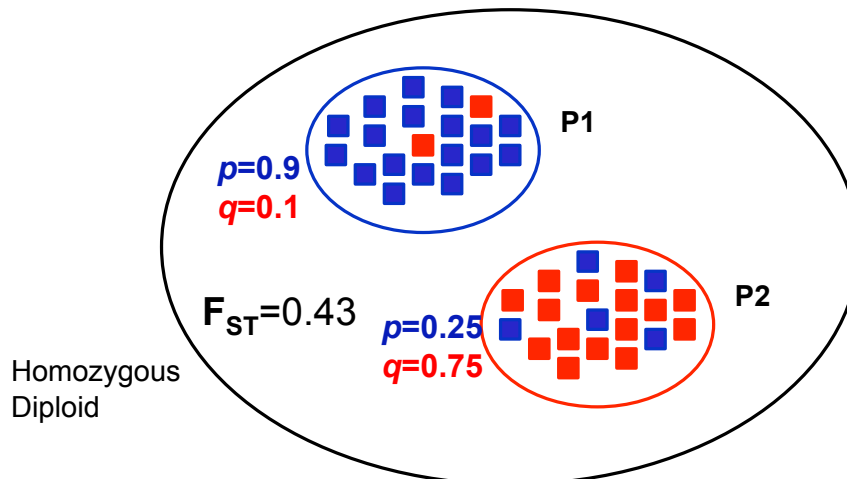


No population differentiation

41

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

## Graphical example of $F_{ST}$

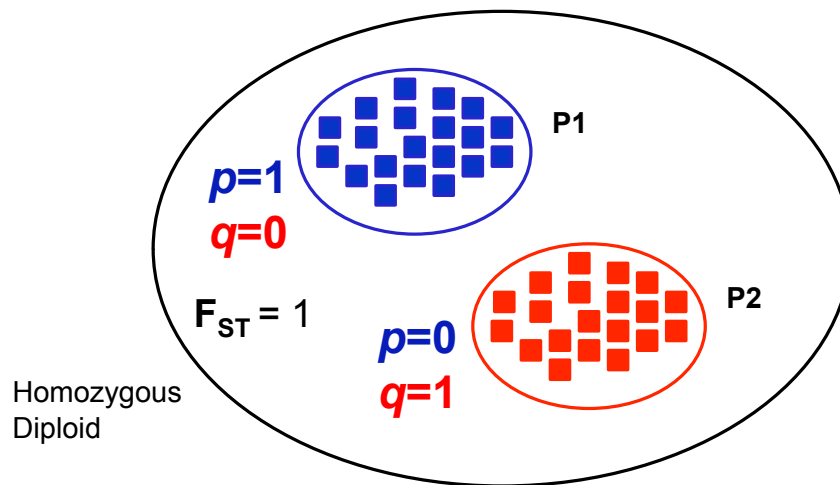


Strong population differentiation

42

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

## Graphical example of $F_{ST}$

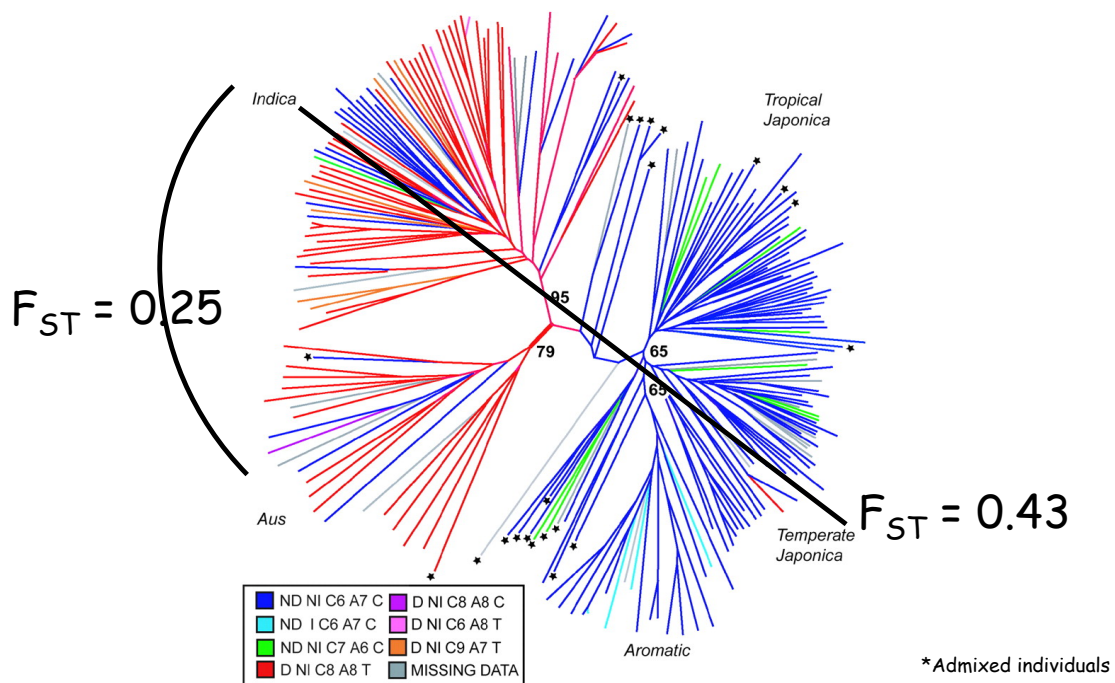


## Complete population differentiation

43

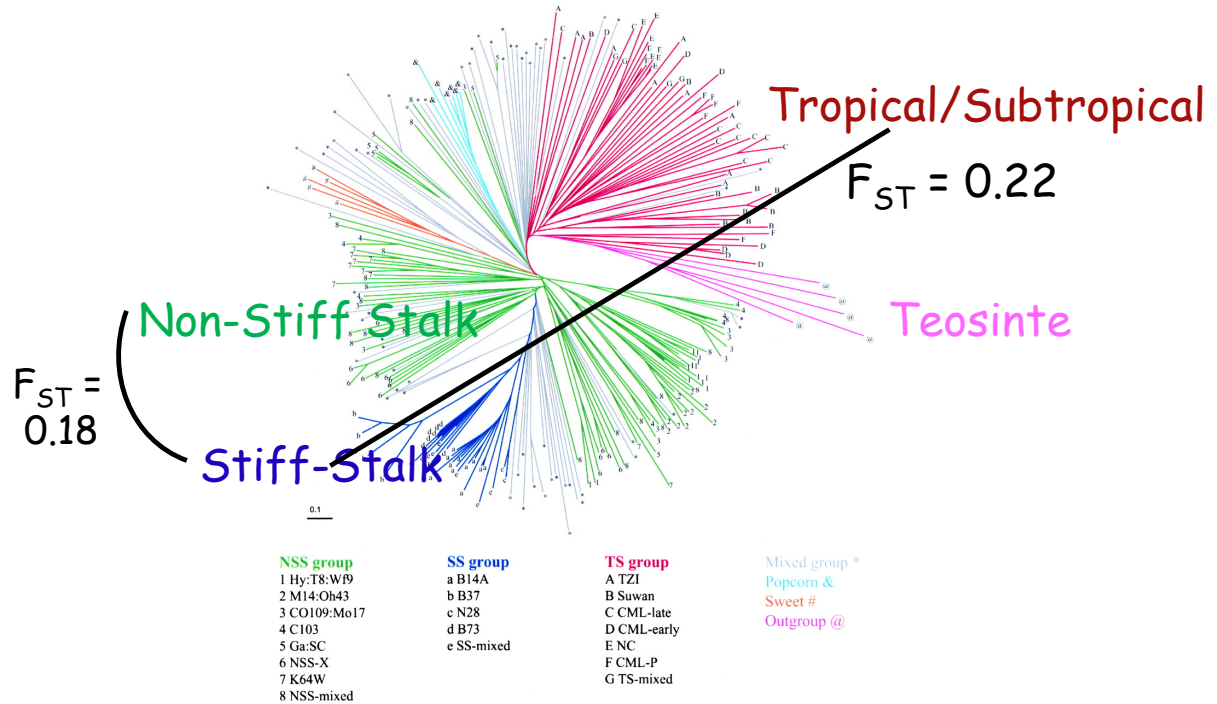
Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

## Rice population structure



Unrooted neighbor-joining tree based on C.S. Chord (Cavalli-Sforza and Edwards 1967) based on 169 nuclear SSRs. The key relates the color of the line to the chloroplast haplotype based on ORF100 and PS-1D sequences.

# Maize population structure



Phylogenetic tree for 260 inbred lines using the log-transformed proportion of shared alleles distance

Liu et al. 2003. Genetics 165:2117-2128

Flint-Garcia et al. 2005. Plant J. 144:1054-1064

## Selection

## One locus with two alleles

Genotype	AA	Aa	aa
Frequency (before selection)	$p^2$	$2p(1-p)$	$(1-p)^2$
Fitness	$W_{AA}$	$W_{Aa}$	$W_{aa}$
Frequency (after selection)	$\frac{p^2 W_{AA}}{\bar{W}}$	$\frac{2p(1-p) W_{Aa}}{\bar{W}}$	$\frac{(1-p)^2 W_{aa}}{\bar{W}}$

Where  $\bar{W} = p^2 W_{AA} + 2p(1-p) W_{Aa} + (1-p)^2 W_{aa}$

is the **mean population fitness**, the fitness of an random individual, e.g.  $\bar{W} = E[W]$

47

The new frequency  $p'$  of A is just  
 $\text{freq}(\text{AA after selection}) + (1/2) \text{freq}(\text{Aa after selection})$

$$p' = \frac{p^2 W_{AA} + p(1-p) W_{Aa}}{\bar{W}} = p \frac{p W_{AA} + (1-p) W_{Aa}}{\bar{W}}$$

The fitness rankings determine the ultimate fate of an allele

If  $W_{AA} \geq W_{Aa} > W_{aa}$ , allele **A** is fixed, a lost

If  $W_{Aa} > W_{AA}, W_{aa}$ , selection maintains both **A** & a

**Overdominant selection**

48

General expression for selection with n alleles

Let  $p_i = \text{freq}(A_i)$ ,  $W_{ij} = \text{fitness } A_i A_j$

$$p'_i = p_i \frac{W_i}{\bar{W}}, \quad W_i = \sum_{j=1}^n p_j W_{ij}, \quad \bar{W} = \sum_{i=1}^n p_i W_i$$

$W_i = \text{marginal fitness of allele } A_i$

$\bar{W} = \text{mean population fitness} = E[W_i] = E[W_{ij}]$

If  $W_i > \bar{W}$ , allele  $A_i$  increases in frequency

If a selective equilibrium exists, then  $W_i = \bar{W}$  for all segregating alleles.

49

- Suppose fitnesses are 1: 1.2:1.4 for the genotypes qq: Qq:QQ and  $p = \text{freq}(Q) = 0.2$

	qq	qQ	QQ
Freq	$0.8^2 = 0.64$	$2 \cdot 0.8 \cdot 0.2 = 0.32$	$0.2^2 = 0.04$
Fitness	1	1.2	1.4
Freq*fit	0.64	0.384	0.056

$$\text{Mean fitness} = 0.64 + 0.384 + 0.056 = 1.08$$

$$\text{Freq}(Qq \text{ after selection}) = 0.384 / 1.08 = 0.356$$

$$\text{Freq}(QQ \text{ after selection}) = 0.04 / 1.08 = 0.037$$

$$\text{New freq } (Q) = (1/2) \cdot 0.356 + 0.037 = 0.215$$

50



# Lecture 2: Introduction to Quantitative Genetics

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

1

## Basic model of Quantitative Genetics

Phenotypic value -- we will occasionally  
also use  $z$  for this value

Basic model:  $P = G + E$

The diagram shows the equation  $P = G + E$  with three blue arrows pointing to its components: one from 'Phenotypic value' to  $P$ , one from 'Genotypic value' to  $G$ , and one from 'Environmental value' to  $E$ .

$G$  = average phenotypic value for that genotype  
if we are able to replicate it over the **universe**  
of environmental values,  $G = E[P]$

Hence, genotypic values are **functions of the  
environments experienced**.

2

# Basic model of Quantitative Genetics

Basic model:  $P = G + E$

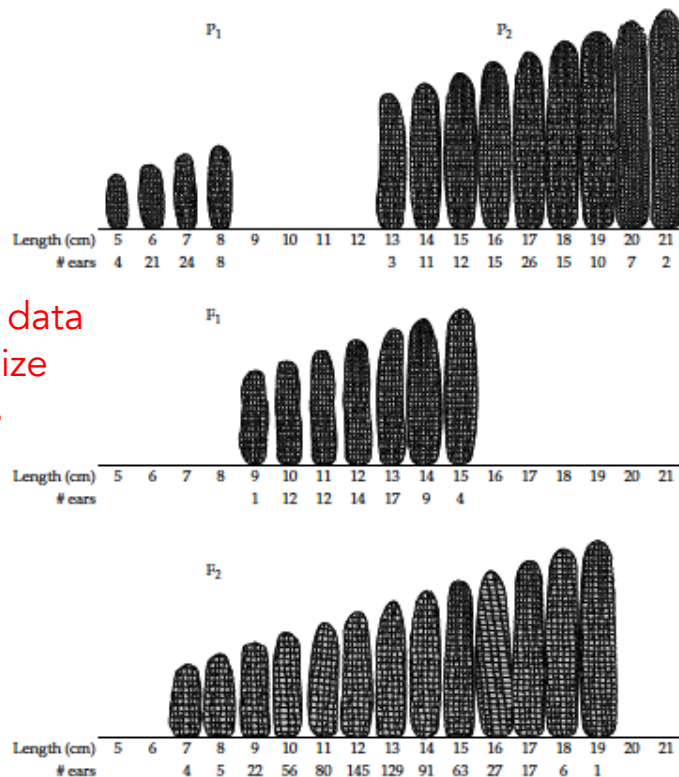
$G$  = average phenotypic value for that genotype if we are able to replicate it over the **universe** of environmental values,  $G = E[P]$

$G$  = average value of an inbred line over a series of environments

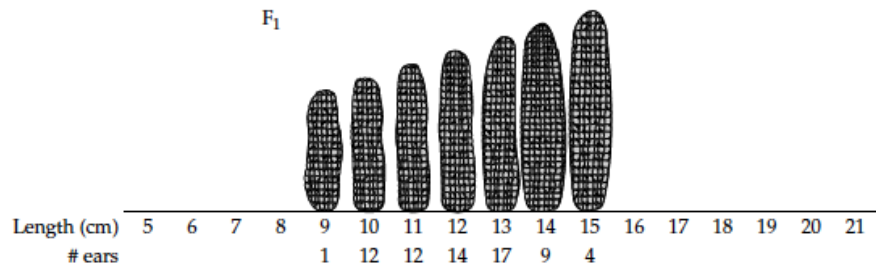
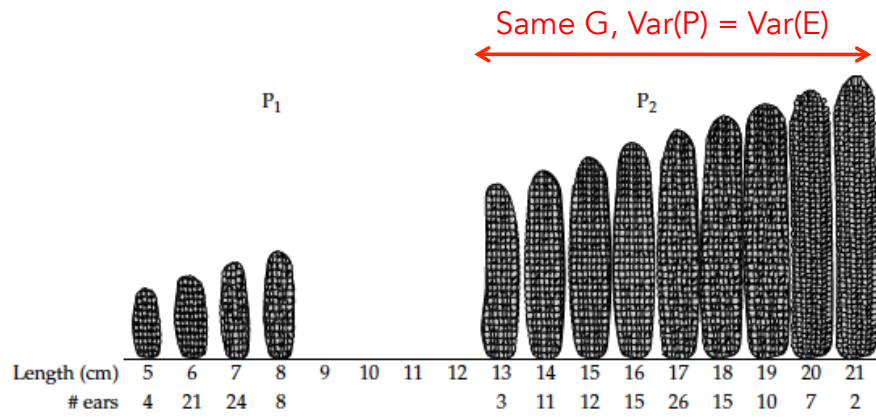
**G x E interaction** --- The performance of a particular genotype in a particular environment differs from the sum of the average performance of that genotype over all environments and the average performance of that environment over all genotypes. Basic model now becomes  $P = G + E + GE$

3

East (1911) data  
on US maize  
crosses

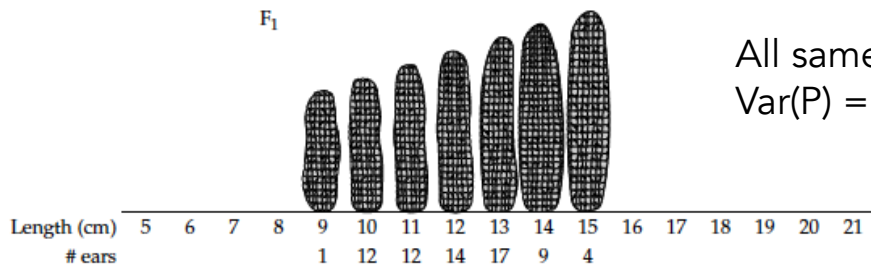


4

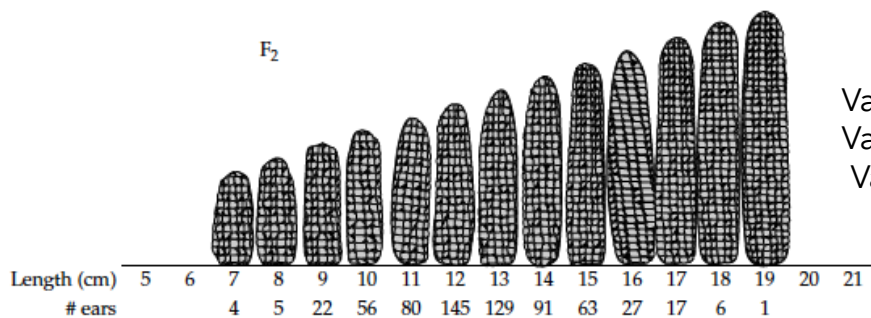


Each sample ( $P_1$ ,  $P_2$ ,  $F_1$ ) has same G, all variation in P is due to variation in E

5



All same G, hence  
 $\text{Var}(P) = \text{Var}(E)$



Variation in G  
 $\text{Var}(P) = \text{Var}(G) + \text{Var}(E)$

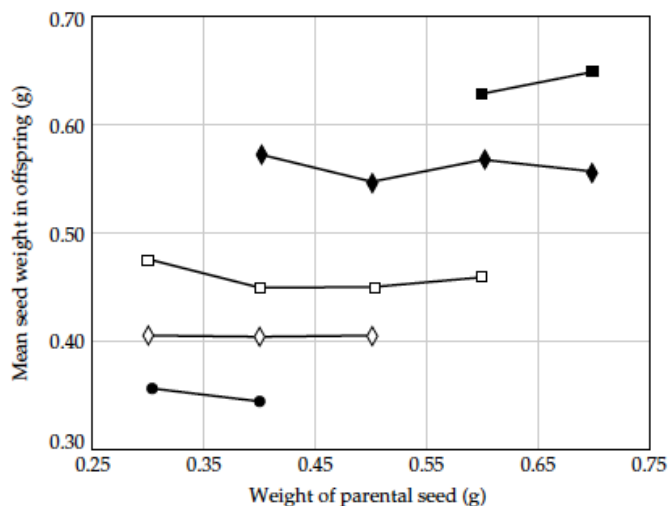
$\text{Var}(F_2) > \text{Var}(F_1)$  due to Variation in G

6

# Johannsen (1903) bean data

- Johannsen had a series of fully inbred (= pure) lines.
- There was a consistent between-line difference in the mean bean size
  - Differences in G across lines
- However, within a given line, size of parental seed independent of size of offspring seed
  - No variation in G within a line

7



**Figure 1.4** Mean offspring seed size as a function of parental seed size for some of Johannsen's pure lines. The data for the different lines are denoted by different symbols. If there is a heritable component to seed weight within a pure line, a line with positive slope is expected — larger parents should yield larger offspring. However, within each line, mean offspring size is essentially independent of the parental phenotype. (Data from Johannsen 1903.)

8

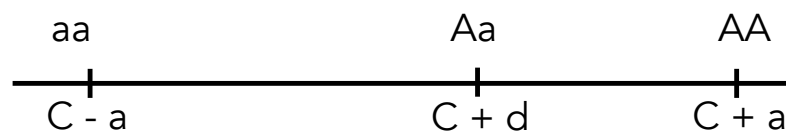
## The transmission of genotypes versus alleles

- With fully inbred lines, offspring have the same genotype as their parent, and hence the entire parental genotypic value  $G$  is passed along
  - Hence, favorable interactions between alleles (such as with dominance) are not lost by randomization under random mating but rather passed along.
- When offspring are generated by crossing (or random mating), each parent contributes a **single allele** at each locus to its offspring, and hence **only passes along a PART** of its genotypic value
- This part is determined by the **average effect of the allele**
  - Downside is that favorable interaction between alleles are NOT passed along to their offspring in a diploid (but, as we will see, are in an autoteraploid)

9

## Genotypic values

It will prove very useful to decompose the genotypic value into the difference between homozygotes ( $2a$ ) and a measure of dominance ( $d$  or  $k = d/a$ )



Note that the constant  $C$  is the average value of the two homozygotes.

If no dominance,  $d = 0$ , as heterozygote value equals the average of the two parents. Can also write  **$d = ka$** , so that  **$G(Aa) = C + ak$**

10

# Computing a and d

Suppose a major locus influences plant height, with the following values

Genotype	aa	Aa	AA
Trait value	10	15	16

$$C = [G(AA) + G(aa)]/2 = (16+10)/2 = 13$$

$$a = [G(AA) - G(aa)]/2 = (16-10)/2 = 3$$

$$d = G(Aa) - [G(AA) + G(aa)]/2 \\ = G(Aa) - C = 15 - 13 = 2$$

11

## Population means: Random mating

Let  $p = \text{freq}(A)$ ,  $q = 1-p = \text{freq}(a)$ . Assuming random-mating (Hardy-Weinberg frequencies),

Genotype	aa	Aa	AA
Value	$C - a$	$C + d$	$C + a$
Frequency	$q^2$	$2pq$	$p^2$

$$\text{Mean} = q^2(C - a) + 2pq(C + d) + p^2(C + a)$$

$$\mu_{\text{RM}} = C + a(p-q) + d(2pq)$$

Contribution from  
homozygotes

Contribution from  
heterozygotes

12

## Population means: Inbred cross $F_2$

Suppose two inbred lines are crossed. If A is fixed in one population and a in the other, then  $p = q = 1/2$

Genotype	aa	Aa	AA
Value	$C - a$	$C + d$	$C + a$
Frequency	$1/4$	$1/2$	$1/4$

$$\text{Mean} = (1/4)(C - a) + (1/2)(C + d) + (1/4)(C + a)$$

$$\mu_{RM} = C + d/2$$

Note that C is the average of the two parental lines, so when  $d > 0$ ,  $F_2$  exceeds this. Note also that the  $F_1$  exceeds this average by d, so only half of this passed onto  $F_2$ .

13

## Population means: RILs from an $F_2$

A large number of  $F_2$  individuals are fully inbred, either by selfing for many generations or by generating doubled haploids. If p and q denote the  $F_2$  frequencies of A and a, what is the expected mean over the set of resulting RILs?

Genotype	aa	Aa	AA
Value	$C - a$	$C + d$	$C + a$
Frequency	q	0	p

$$\mu_{RILs} = C + a(p-q)$$

Note this is independent of the amount of dominance (d)

14

# The average effect of an allele

- The average effect  $\alpha_A$  of an allele **A** is defined by the difference between offspring that get allele **A** and a random offspring.
  - $\alpha_A = \text{mean}(\text{offspring value given parent transmits A}) - \text{mean}(\text{all offspring})$
  - Similar definition for  $\alpha_a$ .
- Note that while  $C$ ,  $a$ , and  $d$  (the genotypic parameters) do not change with allele frequency,  $\alpha_x$  is clearly a function of the frequencies of alleles with which allele  $x$  combines.

15

## Random mating

Consider the average effect of allele **A** when a parent is randomly-mated to another individual from its population

Suppose parent contributes **A**

Allele from other parent	Probability	Genotype	Value
<b>A</b>	$p$	<b>AA</b>	$C + a$
<b>a</b>	$q$	<b>Aa</b>	$C + d$

$$\text{Mean(A transmitted)} = p(C + a) + q(C + d) = C + pa + qd$$

$$\alpha_A = \text{Mean(A transmitted)} - \mu = q[a + d(q-p)]$$

16



# Random mating

Now suppose parent contributes a

Allele from other parent	Probability	Genotype	Value
A	p	Aa	C + d
a	q	aa	C - a

$$\text{Mean}(a \text{ transmitted}) = p(C + d) + q(C - a) = C - qa + pd$$

$$\alpha_a = \text{Mean}(a \text{ transmitted}) - \mu = -p[a + d(q-p)]$$

17

## $\alpha$ , the average effect of an allelic substitution

- $\alpha = \alpha_A - \alpha_a$  is the average effect of an allelic substitution, the change in mean trait value when an a allele in a random individual is replaced by an A allele
  - $\alpha = a + d(q-p)$ . Note that
    - $\alpha_A = q\alpha$  and  $\alpha_a = -p\alpha$ .
    - $E(\alpha_X) = p\alpha_A + q\alpha_a = pq\alpha - qp\alpha = 0$ ,
    - The average effect of a random allele is zero, hence average effects are deviations from the mean

18

# Dominance deviations

- Fisher (1918) decomposed the contribution to the genotypic value from a single locus as  $G_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij}$ 
  - Here,  $\mu$  is the mean (a function of  $p$ )
  - $\alpha_i$  are the average effects
  - Hence,  $\mu + \alpha_i + \alpha_j$  is the **predicted genotypic value** given the average effect (over all genotypes) of alleles  $i$  and  $j$ .
  - The **dominance deviation** associated with genotype  $G_{ij}$  is the difference between its true value and its value predicted from the sum of average effects (essentially a residual)

19

## Fisher's (1918) Decomposition of G

One of Fisher's key insights was that the genotypic value consists of a **fraction that can be passed from parent to offspring** and a **fraction that cannot**.

In particular, under sexual reproduction, parents only pass along **SINGLE ALLELES** to their offspring

Consider the genotypic value  $G_{ij}$  resulting from an  $A_i A_j$  individual

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$$

Average contribution to genotypic value for allele  $i$

$$\text{Mean value } \mu_G = \sum G_{ij} \text{ Freq}(A_i A_j)$$

20

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$$

Since parents pass along single alleles to their offspring, the  $\alpha_i$  (the **average effect** of allele i) represent these contributions

The average effect for an allele is **POPULATION-SPECIFIC**, as it depends on the types and frequencies of alleles that it pairs with

The genotypic value predicted from the individual allelic effects is thus

$$\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$$

21

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$$

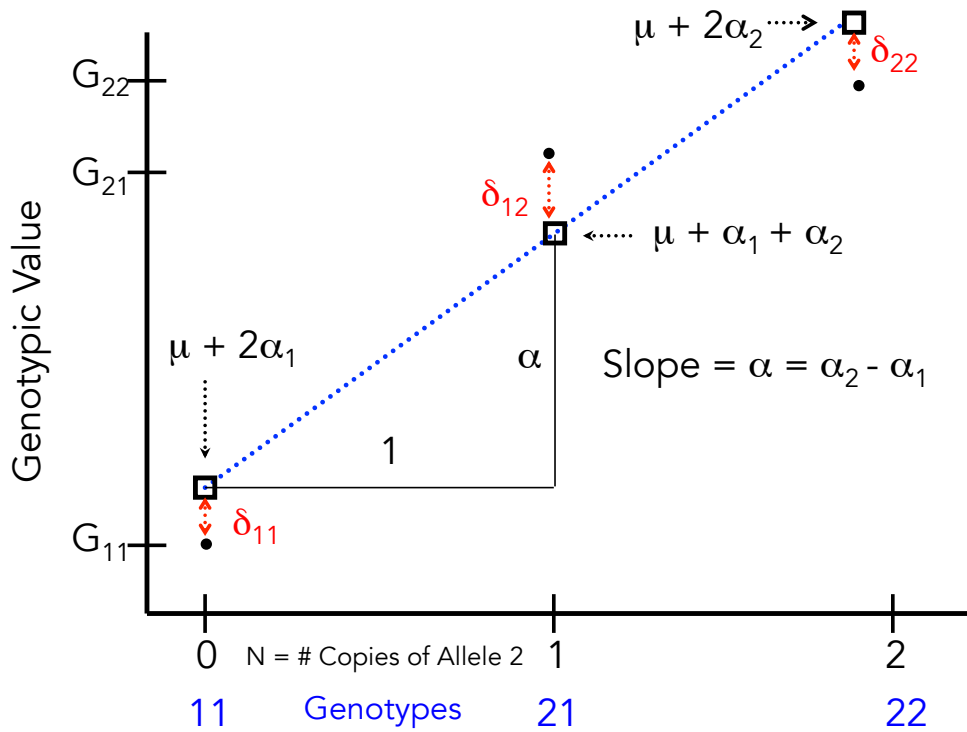
The genotypic value predicted from the individual allelic effects is thus

$$\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$$

**Dominance deviations** --- the difference (for genotype  $A_i A_j$ ) between the genotypic value predicted from the two single alleles and the actual genotypic value,

$$G_{ij} - \hat{G}_{ij} = \delta_{ij}$$

22



23

Fisher's decomposition is a Regression

$$G_{ij} = \underbrace{\mu_G + \alpha_i + \alpha_j}_{\text{Predicted value}} + \underbrace{\delta_{ij}}_{\text{Residual error}}$$

A notational change clearly shows this is a regression,

$$G_{ij} = \mu_G + 2\alpha_1 + (\alpha_2 - \alpha_1) N + \delta_{ij}$$

Independent (predictor) variable  $N = \# \text{ of } A_2 \text{ alleles}$

Note that the slope  $\alpha_2 - \alpha_1 = \alpha$ , the average effect of an allelic substitution

24

$$G_{ij} = \mu_G + 2\alpha_1 + (\alpha_2 - \alpha_1) N + \delta_{ij}$$

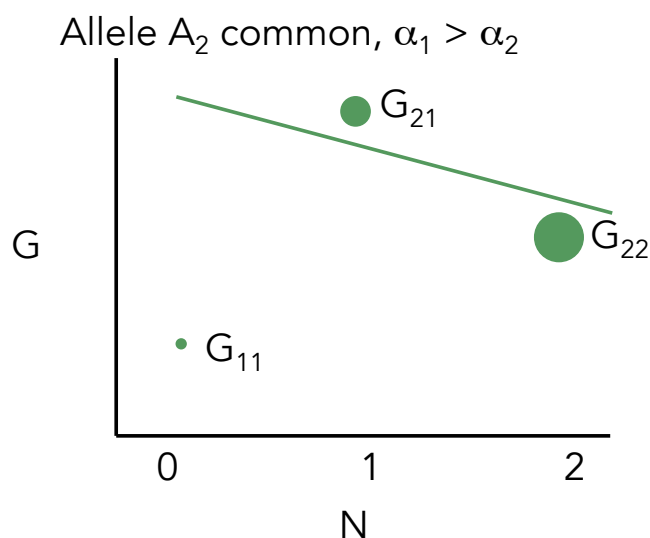
Intercept

Regression slope

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g., } A_1A_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g., } A_1A_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g., } A_2A_2 \end{cases}$$

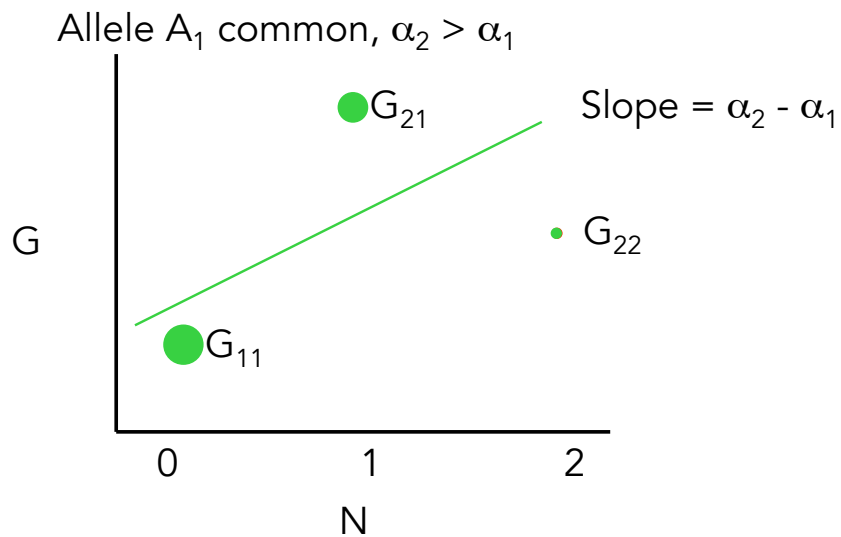
A key point is that the average effects change with allele frequencies. Indeed, if overdominance is present they can change sign with allele frequencies.

25



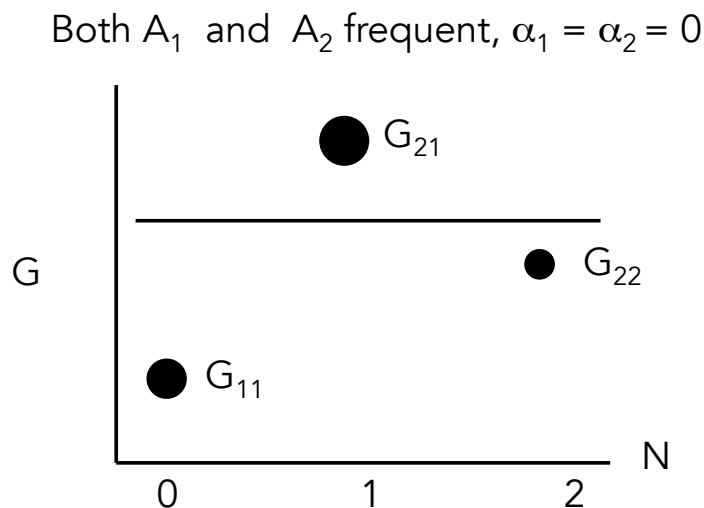
The size of the circle denotes the weight associated with that genotype. While the genotypic values do not change, their frequencies (and hence weights) do.

26



Again, same genotypic values as previous slide, but different weights, and hence a different slope (here a change in sign!)

27



With these allele frequencies, both alleles have the same mean value when transmitted, so that all parents have the same average offspring value -- no response to selection

28

## Average Effects and Additive Genetic Values

The  $\alpha$  values are the **average effects** of an allele

A key concept is the **Additive Genetic Value (A)** of an individual

$$A(G_{ij}) = \alpha_i + \alpha_j$$

$$A = \sum_{k=1}^n (\alpha_i^{(k)} + \alpha_j^{(k)})$$

$\alpha_i^{(k)}$  = effect of allele i at locus k

A is called the **Breeding value** or the **Additive genetic value**

29

$$A = \sum_{k=1}^n (\alpha_i^{(k)} + \alpha_j^{(k)})$$

Why all the fuss over A?

Suppose pollen parent has  $A = 10$  and seed parent has  $A = -2$  for plant height

Expected average offspring height is  $(10 - 2)/2$   
= 4 units above the population mean. Offspring A = average of parental A's

KEY: **parents only pass single alleles to their offspring.**  
**Hence, they only pass along the A part of their genotypic value G**

30

# Genetic Variances

Writing the genotypic value as

$$G_{ij} = \mu_G + (\alpha_i + \alpha_j) + \delta_{ij}$$

The genetic variance can be written as

$$\sigma^2(G) = \sum_{k=1}^n \sigma^2(\alpha_i^{(k)} + \alpha_j^{(k)}) + \sum_{k=1}^n \sigma^2(\delta_{ij}^{(k)})$$

This follows since

$$\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij})$$

$$\text{As Cov}(\alpha, \delta) = 0$$

31

# Genetic Variances

$$\sigma^2(G) = \sum_{k=1}^n \sigma^2(\alpha_i^{(k)} + \alpha_j^{(k)}) + \sum_{k=1}^n \sigma^2(\delta_{ij}^{(k)})$$

Additive Genetic Variance  
(or simply Additive Variance)

Dominance Genetic Variance  
(or simply dominance variance)

Hence, total genetic variance = additive + dominance variances,

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

32



## Key concepts (so far)

- $\alpha_i$  = average effect of allele i
  - Property of a single allele in a particular population (depends on genetic background)
- $A$  = Additive Genetic Value (A)
  - $A$  = sum (over all loci) of average effects
  - Fraction of  $G$  that parents pass along to their offspring
  - Property of an Individual in a particular population
- $\text{Var}(A)$  = additive genetic variance
  - Variance in additive genetic values
  - Property of a population
- Can estimate  $A$  or  $\text{Var}(A)$  without knowing any of the underlying genetical detail (forthcoming)

33

$$\sigma_A^2 = 2E[\alpha^2] = 2 \sum_{i=1}^m \alpha_i^2 p_i$$

$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
0	$a(1+k)$	$2a$

Since  $E[\alpha] = 0$ ,  
 $\text{Var}(\alpha) = E[(\alpha - \mu_\alpha)^2] = E[\alpha^2]$

One locus, 2 alleles:

$$\sigma_A^2 = 2p_1 p_2 a^2 [1 + k(p_1 - p_2)]^2$$

$\uparrow$   
 Dominance alters additive variance

When dominance present, Additive variance is an asymmetric function of allele frequencies

34

Dominance variance

$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
0	$a(1+k)$	$2a$

$$\sigma_D^2 = E[\delta^2] = \sum_{i=1}^m \sum_{j=1}^m \delta_{ij}^2 p_i p_j$$

Equals zero if  $k = 0$

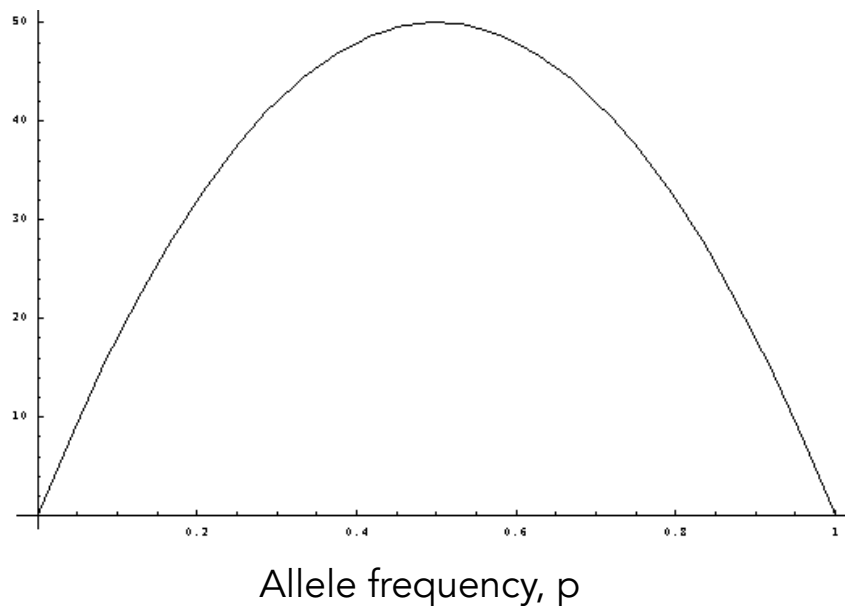
One locus, 2 alleles:  $\sigma_D^2 = (2p_1 p_2 a k)^2$

This is a symmetric function of allele frequencies

Can also be expressed in terms of  $d = ak$

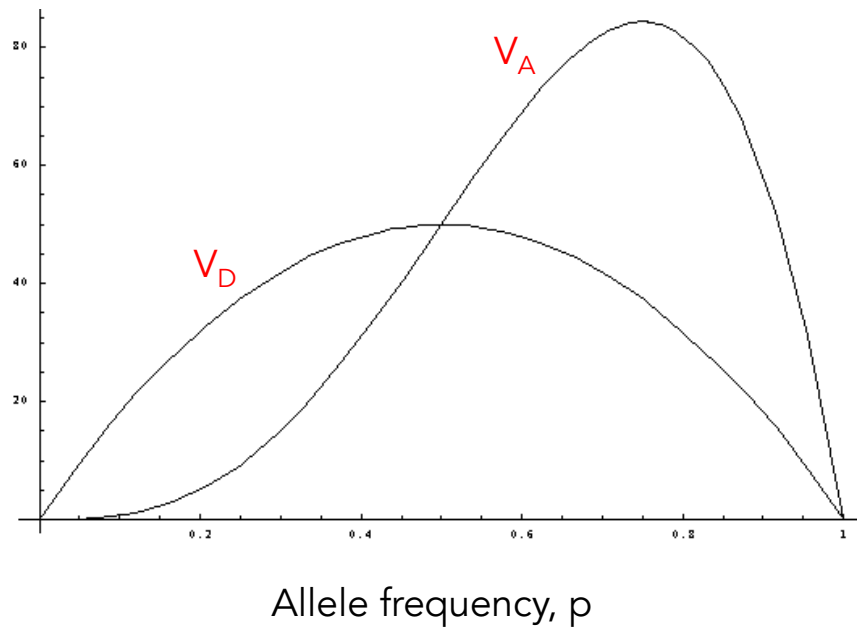
35

Additive variance,  $V_A$ , with no dominance ( $k = 0$ )



36

## Complete dominance ( $k = 1$ )



37

## Epistasis

$$\begin{aligned}
 G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kj}) \\
 &\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\
 &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\
 &\quad + (\delta\delta_{ijkl}) \\
 &= \mu_G + A + D + AA + AD + DD
 \end{aligned}$$

These components are defined to be uncorrelated, (or *orthogonal*), so that

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2$$

38

$$\begin{aligned}
G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kj}) \\
&\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\
&\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\
&\quad + (\delta\delta_{ijkl}) \\
&= \mu_G + A + D + AA + AD + DD
\end{aligned}$$

**Additive x Additive** interactions --  $\alpha\alpha$ , AA  
interactions between a single allele  
at one locus with a single allele at another

**Additive x Dominance** interactions --  $\alpha\delta$ , AD  
interactions between an allele at one  
locus with the genotype at another, e.g.  
allele  $A_i$  and genotype  $B_{kj}$

**Dominance x dominance** interaction ---  $\delta\delta$ , DD  
the interaction between the dominance  
deviation at one locus with the dominance  
deviation at another.

# Lecture 3:

## Resemblance and relatedness

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

1

## Heritability

- Central concept in quantitative genetics
- Fraction of phenotypic variance due to additive genetic values (Breeding values)
  - $h^2 = V_A/V_P$
  - This is called the narrow-sense heritability
  - Phenotypes (and hence  $V_P$ ) can be directly measured
  - Breeding values (and hence  $V_A$ ) must be estimated
- Estimates of  $V_A$  require known collections of relatives

2

# Broad-sense heritability

- Narrow-sense heritability  $h^2$  applies when outcrossing,
  - $h^2 = \text{Var}(A)/\text{Var}(P)$
  - = the fraction of all trait variation due to variation in breeding (additive genetic) values
- Broad-sense heritability  $H^2$  applies when selecting among a series of pure lines
  - $H^2 = \text{Var}(G)/\text{Var}(P)$
  - = the fraction of all trait variation due to variation in Genotypic values

3

## Defining $H^2$ for Plant Populations

Plant breeders often do not measure individual plants (especially with pure lines), but instead often measure a plot or a block of individuals.

This replication can result in inconsistent measures of  $H^2$  even for otherwise identical populations.

Let  $z_{ijkl}$  denote the value of the  $l$ -th replicate in plot  $k$  of genotype  $i$  in environment  $j$ . We can decompose this value as

$$z_{ijkl} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijkl}$$

Effect of the  $k$ -th plot      deviations of individual plants within this plot

4

Suppose we replicate the genotype over  $e$  environments, with  $r$  plots (replicates) per environment, and  $n$  individuals per plot.

If we set our unit of measurement as the average over all plots, the phenotypic variance for the mean of line  $i$  becomes

$$\sigma^2(\bar{z}_i) = \sigma_G^2 + \sigma_E^2 + \frac{\sigma_{GE}^2}{e} + \frac{\sigma_p^2}{er} + \frac{\sigma_e^2}{ern}$$

Thus,  $V_p$ , and  $H^2 = V_G/V_p$ , depend on our choice of  $e$ ,  $r$ , and  $n$

In order to compare broad-sense heritabilities we need to use a consistent design (same values of  $e$ ,  $r$ , and  $n$ )

5

## Key observations

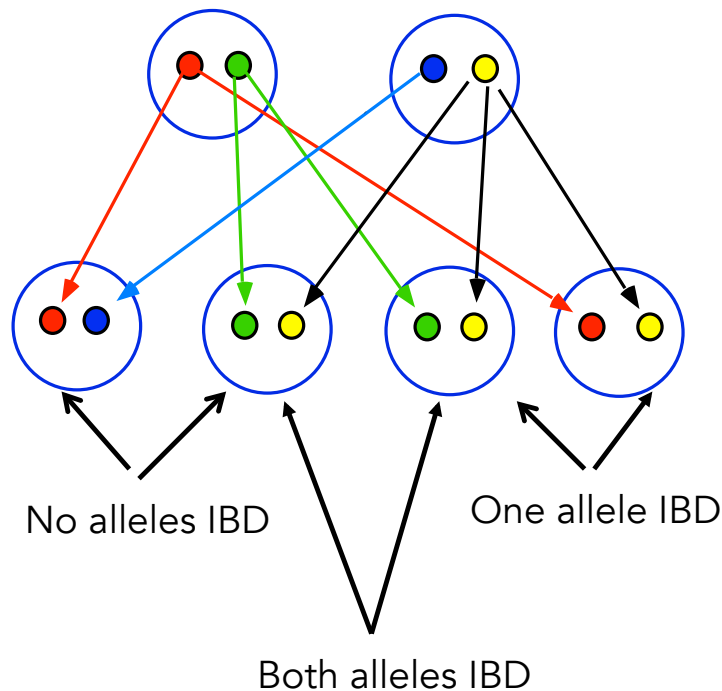
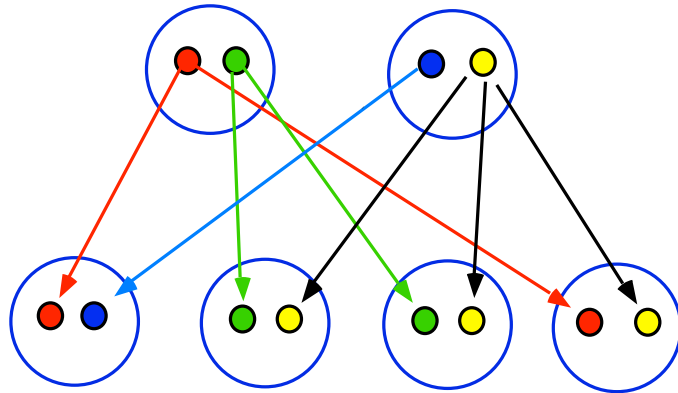
- The amount of **phenotypic resemblance** among relatives for the trait provides an indication of the amount of **genetic variation** for the trait.
- If trait variation has a significant genetic basis, the **closer the relatives**, the **more similar their appearance**
- The covariance between the phenotypic value of relatives measures the strength of this similarity, with larger Cov = more similarity

6

## Genetic Covariance between relatives

Sharing alleles means having alleles that are **identical by descent (IBD)**: both copies can be traced back to a single copy in a recent common ancestor.

Genetic covariances arise because two **related individuals are more likely to share alleles** than are two unrelated individuals.





## Resemblance between relatives and variance components

- The phenotypic variance between relatives can be expressed in terms of genetic variance components
  - $\text{Cov}(z_x, z_y) = a_{xy}V_A + b_{xy}V_D$ .
  - The weights  $a$  and  $b$  depend on the nature of the relatives  $x$  and  $y$ , and are measures of how often they are expected to share alleles identical by descent
  - These are critical in predicting selection response

9

## Parent-offspring genetic covariance

$\text{Cov}(G_p, G_o)$  --- Parents and offspring share  
EXACTLY one allele IBD

Denote this common allele by  $A_1$

$$\begin{aligned}
 G_p &= A_p + D_p = \alpha_1 + \alpha_x + D_{1x} \\
 G_o &= A_o + D_o = \alpha_1 + \alpha_y + D_{1y}
 \end{aligned}$$

10

$$\begin{aligned}
Cov(G_o, G_p) &= Cov(\alpha_1 + \alpha_x + D_{1x}, \alpha_1 + \alpha_y + D_{1y}) \\
&= \cancel{Cov(\alpha_1, \alpha_1)} + \cancel{Cov(\alpha_1, \alpha_y)} + \cancel{Cov(\alpha_1, D_{1y})} \\
&\quad + \cancel{Cov(\alpha_x, \alpha_1)} + \cancel{Cov(\alpha_x, \alpha_y)} + \cancel{Cov(\alpha_x, D_{1y})} \\
&\quad + \cancel{Cov(D_{1x}, \alpha_1)} + \cancel{Cov(D_{1x}, \alpha_y)} + \cancel{Cov(D_{1x}, D_{1y})}
\end{aligned}$$

All blue covariance terms are zero.

- By construction,  $\alpha$  and  $D$  are uncorrelated
- By construction,  $\alpha$  from non-IBD alleles are uncorrelated
- By construction,  $D$  values are uncorrelated unless both alleles are IBD

14

$$Cov(\alpha_x, \alpha_y) = \begin{cases} 0 & \text{if } x \neq y, \text{ i.e., not IBD} \\ Var(A)/2 & \text{if } x = y, \text{ i.e., IBD} \end{cases}$$

$$Var(A) = Var(\alpha_1 + \alpha_2) = 2Var(\alpha_1)$$

so that

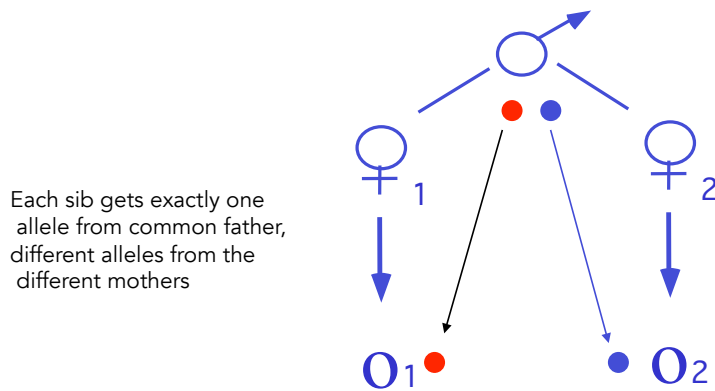
$$Var(\alpha_1) = Cov(\alpha_1, \alpha_1) = Var(A)/2$$

Hence, relatives sharing one allele IBD have a genetic covariance of  $Var(A)/2$

The resulting parent-offspring genetic covariance becomes  $Cov(G_p, G_o) = Var(A)/2$

12

## Half-sibs



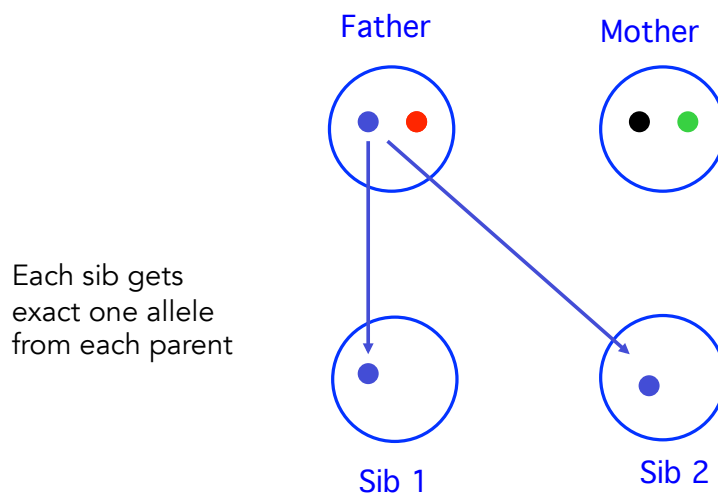
The half-sibs share no alleles IBD

- occurs with probability 1/2

Hence, the genetic covariance of half-sibs is just  $(1/2)\text{Var}(A)/2 = \text{Var}(A)/4$

13

## Full-sibs

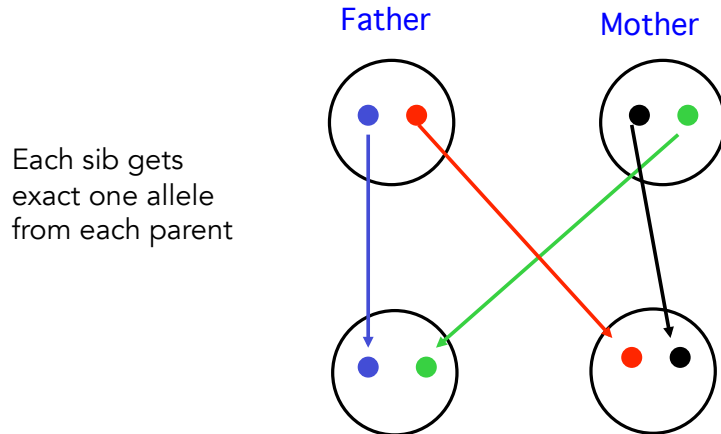


$\text{Prob}(\text{Allele from father IBD}) = 1/2$ . Given the allele in parent one,  $\text{prob} = 1/2$  that sib 2 gets same allele

$\text{Prob}(\text{Allele from father not IBD}) = 1/2$ . Given the allele in parent one,  $\text{prob} = 1/2$  that sib 2 gets different allele

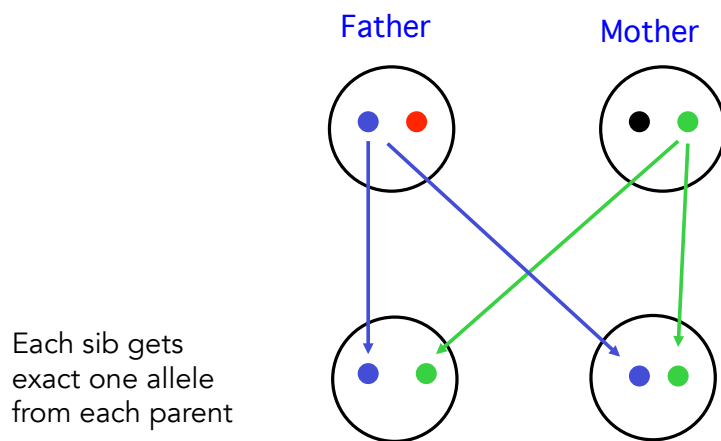
14

# Full-sibs



Paternal allele not IBD [ Prob =  $1/2$  ]  
 Maternal allele not IBD [ Prob =  $1/2$  ]  
 Prob(sibs share 0 alleles IBD) =  $1/2 * 1/2 = 1/4$

15



Paternal allele IBD [ Prob =  $1/2$  ]  
 Maternal allele IBD [ Prob =  $1/2$  ]  
 Prob(sibs share 2 alleles IBD) =  $1/2 * 1/2 = 1/4$

Prob(share 1 allele IBD) =  $1 - \text{Pr}(0) - \text{Pr}(2) = 1/2$

16

## Resulting Genetic Covariance between full-sibs

IBD alleles	Probability	Contribution
0	1/4	0
1	1/2	$\text{Var}(A)/2$
2	1/4	$\text{Var}(A) + \text{Var}(D)$
<hr/>		
$\text{Cov}(\text{Full-sibs}) = \text{Var}(A)/2 + \text{Var}(D)/4$		

17

## Genetic Covariances for General Relatives

Let  $r = (1/2)\text{Prob}(1 \text{ allele IBD}) + \text{Prob}(2 \text{ alleles IBD})$

Let  $u = \text{Prob}(\text{both alleles IBD})$

General genetic covariance between relatives

$$\text{Cov}(G) = r\text{Var}(A) + u\text{Var}(D)$$

When epistasis is present, additional terms appear

$$r^2\text{Var}(AA) + ru\text{Var}(AD) + u^2\text{Var}(DD) + r^3\text{Var}(AAA) +$$

18

# More general relationships

- To obtain the expected covariance for any set of relatives, we normally need only compute  $r$  and  $u$  for that set of relatives
- With general inbreeding, becomes more complex (as three other terms, in addition to  $V_A$  and  $V_D$  arise)
- With crosses involving inbred and/or related parents, values for  $r$  and  $u$  are different from those presented above.

19

## Coefficients of Coancestry

Suppose we pick a single allele each at random from two relatives. The probability that these are IBD is called  $\Theta$ , the **coefficient of coancestry**. In terms of our previous notation,  **$2\Theta = r = \text{the coeff on Var}(A)$**

$\Theta_{xy}$  denotes the coefficient for relatives  $x$  and  $y$

Consider an offspring  $z$  from a (hypothetical) cross of  $x$  and  $y$ .  $\Theta_{xy} = f_z$ , the inbreeding coefficient of  $z$ . Why? Because the offspring of  $x$  and  $y$  each get a randomly-chosen allele from each parent. The probability  $f_z$  that both alleles are IBD (the probability of inbreeding) is thus just  $\Theta_{xy}$ .

20

## $\theta$ and the coefficient on $V_A$

- The coefficient on the additive variance for the relatives  $x$  and  $y$  is just  $2\theta_{xy}$ .
- To see this,
  - let  $A_i A_j$  denote the two alleles in  $x$  and  $A_k A_l$  those in  $y$ .
  - $\text{Cov}(\text{breeding values}) = \Pr(A_i \text{ ibd } A_k) \text{cov}(\alpha_i, \alpha_k) + \Pr(A_i \text{ ibd } A_l) \text{cov}(\alpha_i, \alpha_l) + \Pr(A_j \text{ ibd } A_k) \text{cov}(\alpha_j, \alpha_k) + \Pr(A_j \text{ ibd } A_l) \text{cov}(\alpha_j, \alpha_l) = 4 \theta_{xy} \text{Var}(\alpha)$
  - Since  $\text{Var}(A) = 2\text{Var}(\alpha)$ ,  $\text{Cov} = 2 \theta_{xy} \text{Var}(A)$

21

## $\Theta_{xx}$ : The Coancestry of an individual with itself

Self  $x$ , what is the inbreeding coefficient of its offspring?

To compute  $\Theta_{xx}$ , denote the two alleles in  $x$  by  $A_1$  and  $A_2$

	Draw $A_1$	Draw $A_2$
Draw $A_1$	IBD	$f_x$
Draw $A_2$	$f_x$	IBD

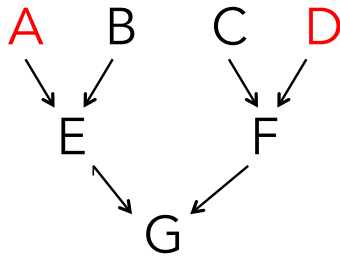
Hence, for a non-inbred individual,  $\Theta_{xx} = 2/4 = 1/2$

If  $x$  is inbred,  $f_x = \text{prob } A_1 \text{ and } A_2 \text{ IBD}$ ,

$$\Theta_{xx} = (1 + f_x)/2$$

22

# Example



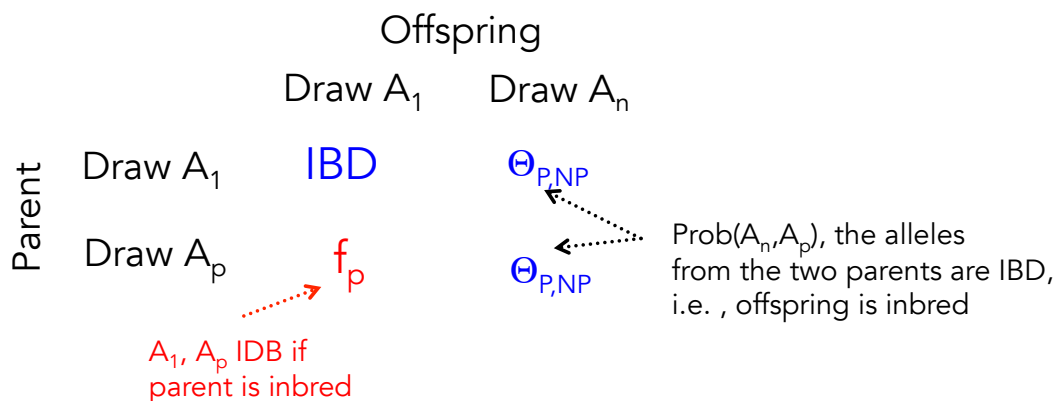
Consider the following pedigree  
Suppose **A** and **D** are **fully-inbred**,  
and related, lines with  $\theta_{AD} = 0.5$ .  
Further, B and C are unrelated and  
outcrossed individuals

Individual	A	B	C	D
$F_x$	1	0	0	1
$\theta_{xx} = (1 + F_x)/2$	1	1/2	1/2	1

23

## The Parent-offspring Coancestry

Let  $A_1, A_n$  denote the two alleles in the offspring, where  $A_n$  is the allele from the nonfocal parent (NP), while  $A_1, A_p$  are the two alleles in the focal parent (P)



For a non-inbred individual,  $\theta_{P0} = 1/4$

General:

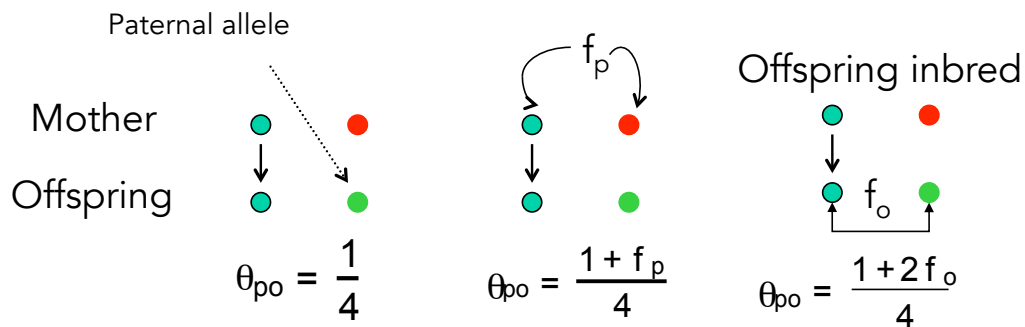
$$\theta_{PO} = (1 + f_p + 2\theta_{P,NP})/4 = (1 + f_p + 2f_o)/4$$

24



# $\Theta_{op}$ = Parent & Offspring

Parent inbred



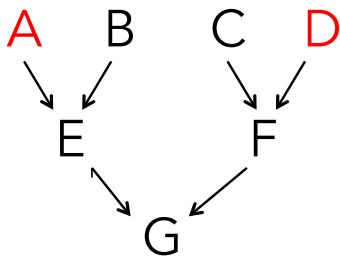
$1/2$  = Prob random offspring allele from father. Prob =  $\theta_{mf} = f_o$  that this allele is IBD to mother giving a contribution of  $f_o/2$

$$\theta_{po} = \frac{1}{4}(1 + f_p + 2\theta_{mf})$$

This is just  $2f_o$

25

From before



$$\begin{aligned} \theta_{AA} &= \theta_{DD} = 1; \theta_{BB} = \theta_{CC} = 1/2; \\ \theta_{AD} &= 1/2, \\ \theta_{AB} &= \theta_{AC} = \theta_{BC} = \theta_{BD} = \theta_{CD} = 0 \end{aligned}$$

Consider A - E (inbred parent - offspring)

$$\theta_{AE} = (1+f_A)/4 = (1+1)/4 = 1/2. \text{ Same value for } \theta_{DF}$$

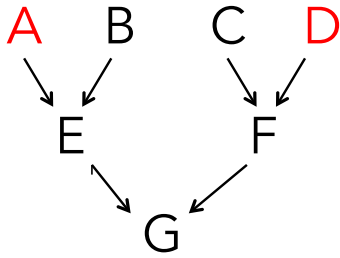
Consider B - E (outbred parent - offspring)

$$\theta_{BE} = (1+f_B)/4 = (1+0)/4 = 1/4. \text{ Same value for } \theta_{CF}$$

Consider E - G (outbred parent - offspring)

$$\theta_{EG} = (1+f_E)/4 = (1+0)/4 = 1/4. \text{ Same value for } \theta_{FG}$$

26



From before

$$\theta_{AA} = \theta_{DD} = 1; \theta_{BB} = \theta_{CC} = 1/2;$$

$$\theta_{AD} = 1/2,$$

$$\theta_{AB} = \theta_{AC} = \theta_{BC} = \theta_{BD} = \theta_{CD} = 0$$

What about  $\theta_{EF}$ ?

The randomly-chosen allele from E has equal chance of being from A or B. Likewise for F (from C or D)

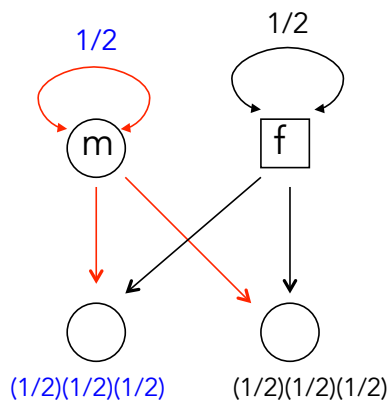
Of these four possible combinations (A&C, A&D, B&C, B&D), only an allele from A and an allele from D have a chance of being IBD, which is  $\theta_{AD} = 1/2$ .

Hence,  $\theta_{EF} = \theta_{AD}/4 = 1/8$

27

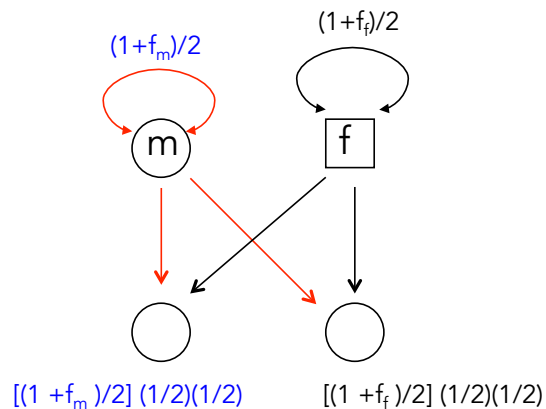
Full sibs (x and y) from parents m and f

$$\Theta = 1/8 + 1/8 = 1/4$$



Unrelated, non-inbred  
parents

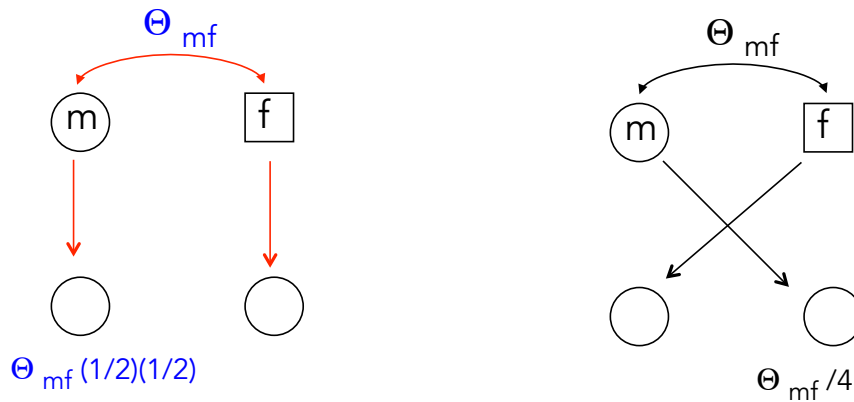
$$\Theta = (2 + f_m + f_f)/8$$



Unrelated, inbred  
parents

28

Full sibs (x and y) from parents m and f



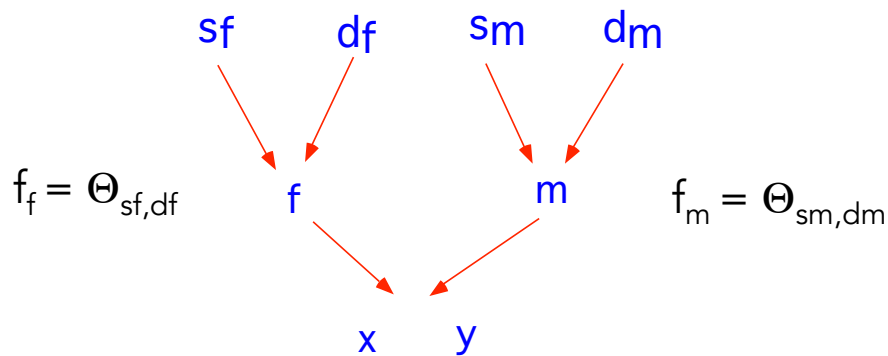
Parents inbred & related.  
Two additional paths to add  
to  $\Theta = (2 + f_m + f_f)/8$

This gives  $\Theta = (2 + f_m + f_f + 4 \Theta_{mf})/8$

29

Full sibs (x and y) from parents m and f

$$\Theta_{xy} = (2 + f_m + f_f + 4\Theta_{mf})/8$$

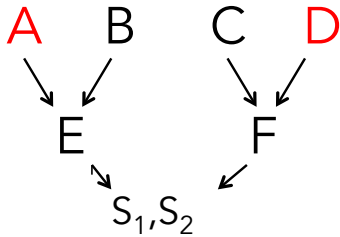


Putting all this together gives

$$\Theta_{xy} = (2 + \Theta_{sm,dm} + \Theta_{sf,df} + 4\Theta_{mf})/8$$

30

# Example



From before

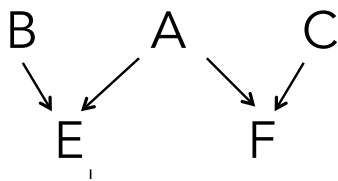
$$\begin{aligned} \theta_{AA} = \theta_{DD} = 1; \theta_{BB} = \theta_{CC} = 1/2; \\ \theta_{AD} = 1/2, \theta_{EF} = 1/8, \\ \theta_{AB} = \theta_{AC} = \theta_{BC} = \theta_{BD} = \theta_{CD} = 0 \end{aligned}$$

$$\Theta_{xy} = (2 + \Theta_{AB} + \Theta_{CD} + 4\Theta_{EF})/8$$

$$\theta_{S1S2} = (2 + 0 + 0 + 4[1/8])/8 = (4 + 1)/16 = 5/16$$

31

## Half-sibs



A is the common parent

- Using the same arguments as above,

$$\begin{aligned} \theta_{EF} &= (\theta_{AA} + \theta_{AB} + \theta_{AC} + \theta_{BC})/4 \\ &= ([1 + f_A]/2 + \theta_{AB} + \theta_{AC} + \theta_{BC})/4 \end{aligned}$$

Hence, if B and C unrelated,

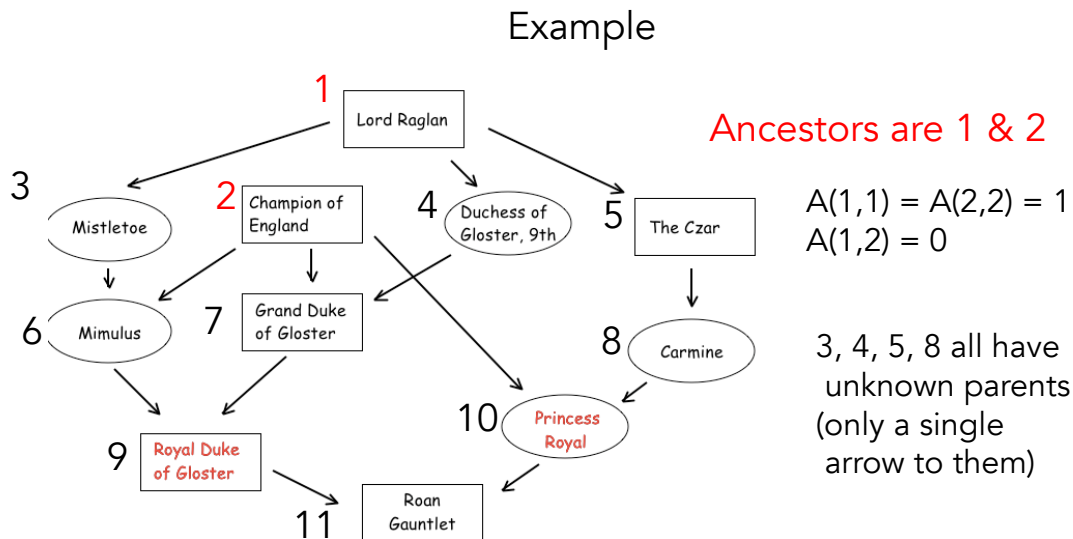
$$\theta_{EF} = (1 + f_A)/8$$

32

## Computing $\theta_{xy}$ -- The Recursive Method

- There is a simple recursive method for generating the elements  $A_{ij} = 2\theta_{ij}$  of a relationship matrix (used for BLUP selection). For ease of reading, we use the notation  $A(i,j) = A_{ij}$ 
  - Basic idea is that the founding individuals of the pedigree are assumed to be unrelated and not inbred (although this can also be accommodated). These founders are assigned values of  $A(i,i) = 1$ .
  - Likewise, any unknown parent of any future individual is assumed to be unrelated to all others in the pedigree and not inbred, and they are also assigned a value of  $A(i,i) = 1$ .
  - Let  $S_i$  and  $D_i$  denote the sire and dam (father and mother) of individual  $i$ . For this offspring  $A(i,i) = 1 + A(S_i, D_i)/2$
  - $A(i,j) = A(j,i) = [A(j,S_i) + A(j,D_i)]/2 = [A(i,S_i) + A(i,D_i)]/2$
  - The recursive (or tabular) method starts with the founding parents and then proceeds down the pedigree in a recursive fashion to fill out  $A$  for the desired pedigree.

33



3:  $S_3 = 1$ ,  $D_3 = \text{Unknown}$ ,  $A(3,3) = 1 + A(S_3, D_3)/2 = 1 + A(1, \text{unk})/2 = 1$

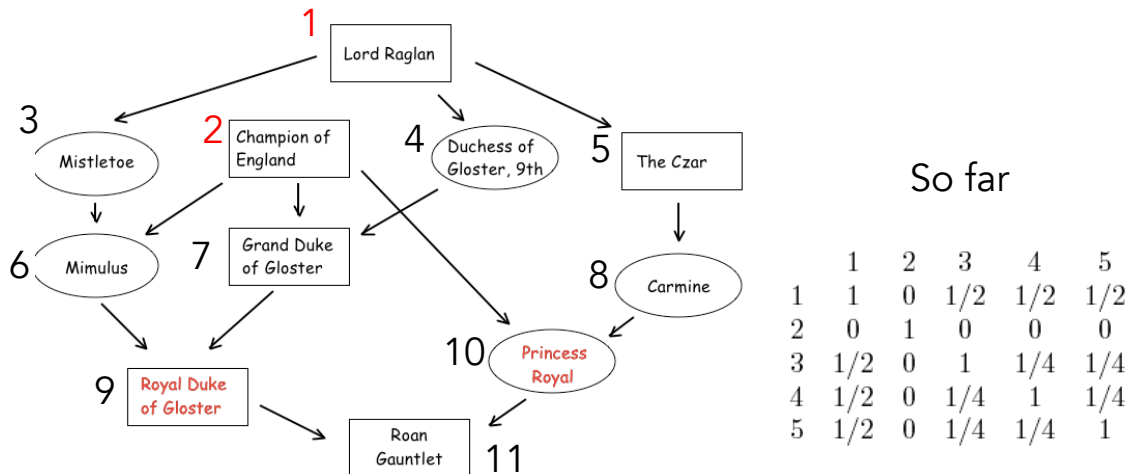
$A(1,3) = [A(1, S_3) + A(1, D_3)]/2 = [A(1,1) + A(1, \text{unk})]/2 = 1/2$ .

Note also that  $A(1,4) = A(1,5) = 1/2$ ,  $A(4,4) = A(5,5) = 1$ .

$A(3,4) = [A(3, S_4) + A(3, D_4)]/2 = [A(3,1) + A(3, \text{unk})]/2 = (1/2 + 0)/2 = 1/4$ .

Same for  $A(3,5) = 1/4$ . 2 is unrelated to 3, 4, 5, giving  $A(2,3) = A(2,4) = A(2,5) = 0$ .

34



6:  $S_6 = 2, D_6 = 3$ .  $A(6,6) = 1 + A(S_6, D_6)/2 = 1 + A(2,3)/2 = 1$   
 $A(6,1) = [A(1, S_6) + A(1, D_6)]/2 = [A(1,2) + A(1,3)]/2 = [0 + 1/2]/2 = 1/4$   
 $A(6,2) = [A(2, S_6) + A(2, D_6)]/2 = [A(2,2) + A(2,3)]/2 = [1 + 0]/2 = 1/2$   
 $A(6,3) = [A(3, S_6) + A(3, D_6)]/2 = [A(3,2) + A(3,3)]/2 = [0 + 1]/2 = 1/2$   
 $A(6,4) = [A(4, S_6) + A(4, D_6)]/2 = [A(4,2) + A(4,3)]/2 = [0 + 1/4]/2 = 1/8$   
 $A(6,5) = [A(5, S_6) + A(5, D_6)]/2 = [A(5,2) + A(5,3)]/2 = (0 + 1/4)/2 = 1/8$   
7:  $S_7 = 2, D_7 = 4$ .  $A(7,7) = 1 + A(S_7, D_7)/2 = 1 + A(2,4)/2 = 1 + 0/2 = 1$   
 $A(6,7) = [A(6, S_7) + A(6, D_7)]/2 = [A(6,2) + A(6,4)]/2 = (1/2 + 1/8)/2 = 5/16$   
8:  $S_8 = 5, D_8 = \text{unk}$ .  $A(8,8) = 1 + A(S_8, D_8)/2 = 1 + A(5, \text{unk})/2 = 1$ .  
 $A(6,8) = [A(6, S_8) + A(6, D_8)]/2 = [A(6,5) + A(6, \text{unk})]/2 = (1/8)/2 = 1/16$   
9:  $S_9 = 7, D_9 = 6$ .  $A(9,9) = 1 + A(S_9, D_9)/2 = 1 + A(6,7)/2 = 1 + 5/32 = 1.156 < \text{inbred!}$

35

## Actual relatedness versus expected values from pedigrees

Values for the coefficient of coancestry ( $\theta$ ) and the coefficient of fraternity ( $\Delta$ ) obtained from pedigrees are expected values. Due to random segregation of genes from parents, The actual value (or realization) can be different.

For example, we expect  $2\theta$  to be  $1/2$  for full subs. However, one pair of sibs may actually be more similar (0.6) and another less similar (say 0.35). On average,  $2\theta$  is  $1/2$  for pairs of full sibs, but if we knew the actual value of  $\theta$ , we have more information. With sufficient dense genetic markers, we can estimate these relationships directly.

**Genomic selection** uses this extra information.

36

## What about coefficient of coancestry $\theta$ ?

Genotype of $j$	Genotype of $i$		
	11	10	00
11	1	0.5	0
10	0.5	0.5	0.5
00	0	0.5	1

One computes the coefficient of coancestry for each SNP, taking the average value over all loci as the coefficient of coancestry for that pair of individuals. Toro et al. (2002) refer to this as **molecular coancestry**. Note that we can compare an individual with itself ( $i = j$ ), which returns 1 for each homozygous locus and 1/2 for each heterozygous loci.

37

Genotype of $j$	Genotype of $i$		
	11	10	00
11	1	0.5	0
10	0.5	0.5	0.5
00	0	0.5	1

Indiv x:	00	00	10	10	00	10	11	00	11	00
Indiv y:	10	00	11	11	10	11	11	10	11	10
Locus-specific $\theta$	0.5	1.0	0.5	0.5	0.5	0.5	1.0	0.5	1.0	0.5

Estimated  $\theta$  is the average over all ten loci, = 0.65

38

## The coefficient of fraternity

- While (twice) the coefficient of coancestry gives the weight on the additive variance for two relatives, a related measure of IDB status among relatives gives the weight on the dominance variance
- The probability that the two alleles in individual x are IBD to two alleles in individual y is denoted  $\Delta_{xy}$ , and is called the **coefficient of fraternity**.
- This can be expressed as a function of the coefficients of coancestry for the parents of (mx and fx) of x and the parents (my and fy) of y.
  - $\Delta_{xy} = \theta_{mxmy} \theta_{fxfy} + \theta_{mxfy} \theta_{fxmy}$

39

## The coefficient of fraternity (cont)

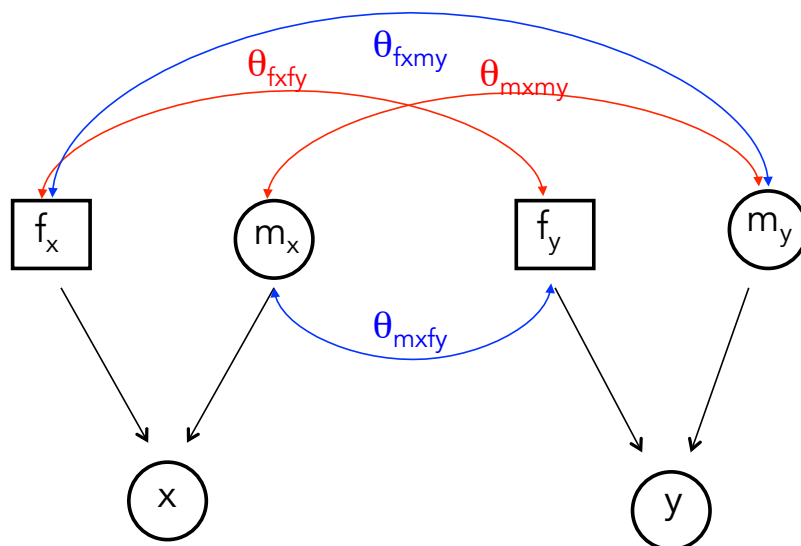
- x and y can have both alleles IBD if
  - The allele from the father (fx) of x and the father (fy) of y are IBD (probability  $\theta_{fxfy}$ ) AND the allele from the mother (mx) of x and the mother (my) of y are IBD (probability  $\theta_{mxmy}$ ) , or  $\theta_{fxfy} \theta_{mxmy}$
  - OR the allele from the mother (mx) of x and the father (fy) of y are IBD (probability  $\theta_{mxfy}$ ) AND the allele from the father (fx) of x and the mother (my) of y are IBD (probability  $\theta_{fxmy}$ ) , or  $\theta_{mxfy} \theta_{fxmy}$
  - Putting these together gives
    - $\Delta_{xy} = \theta_{mxmy} \theta_{fxfy} + \theta_{mxfy} \theta_{fxmy}$

40



## $\Delta_{xy}$ , The Coefficient of Fraternity

$\Delta_{xy} = \text{Prob}(\text{both alleles in } x \text{ \& } y \text{ IBD})$



$$\Delta_{xy} = \theta_{mxmy} \theta_{fxfy} + \theta_{mxfy} \theta_{fxmy}$$

41

## Examples of $\Delta_{xy}$ : Full sibs

- Full sibs share same mom, dad
  - $m_x = m_y = m$ ,  $f_x = f_y = f$
  - $\Delta_{xy} = \theta_{mxmy} \theta_{fxfy} + \theta_{mxfy} \theta_{fxmy} = \theta_{mm} \theta_{ff} + \theta_{mf}^2$
  - $\Delta_{xy} = (1+f_m)(1+f_f)/4 + \theta_{mf}^2$
- If parents unrelated,  $\theta_{fm} = 0$ , giving
  - $\Delta_{xy} = (1+f_m)(1+f_f)/4$
- If parents are unrelated and not inbred,
  - $\Delta_{xy} = 1/4$

42

## Examples of $\Delta_{xy}$ : Half sibs

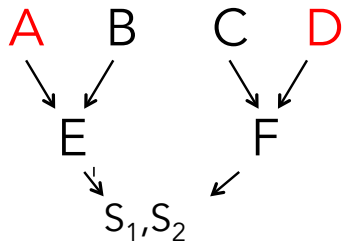
- Paternal half sibs share same dad, different moms
  - $f_x = f_y = f$ ;  $m_x$  and  $m_y$
  - $\Delta_{xy} = \theta_{mxmy}\theta_{fxfy} + \theta_{mxfy}\theta_{fxmy} = \theta_{mxmy}\theta_{ff} + \theta_{mxf}\theta_{myf}$
  - $\Delta_{xy} = \theta_{mxmy} (1+f_m)/2 + \theta_{mxf}\theta_{myf}$
- If mothers are unrelated to each other and to the common father,  $\theta_{mxmy} = \theta_{mxf} = \theta_{myf} = 0$ , giving
  - $\Delta_{xy} = 0$

43

## When is $\Delta$ non-zero?

- Since  $\Delta_{xy} = \theta_{mxmy}\theta_{fxfy} + \theta_{mxfy}\theta_{fxmy}$
- A nonzero value for  $\Delta$  requires either
  - That the fathers of both x and y are related AND the mothers of both x and y are related
  - OR that the father of x is related to the mother of y AND the mother of x is related to the father of y

44



From before

$$\begin{aligned}\theta_{AA} &= \theta_{DD} = 1; \theta_{BB} = \theta_{CC} = 1/2; \\ \theta_{AD} &= 1/2, \theta_{EF} = 1/8, \\ \theta_{AB} &= \theta_{AC} = \theta_{BC} = \theta_{BD} = \theta_{CD} = 0\end{aligned}$$

What is  $\Delta$  for the full sibs ( $S_1$  and  $S_2$ )?

$$\Delta_{xy} = \theta_{mxmy} \theta_{fxfy} + \theta_{mxfy} \theta_{fxmy} = \theta_{EE} \theta_{FF} + \theta_{EF}^2$$

$$\begin{aligned}\text{Giving } \Delta_{xy} &= \theta_{EE} \theta_{FF} + \theta_{EF}^2 \\ &= (1/2)(1/2) + (1/8)^2 \\ &= 1/4 + 1/64 = 17/64 = 0.266\end{aligned}$$

45

## $\Delta_{xy}$ and the coefficient on $V_D$

- The coefficient on the dominance variance for the relatives  $x$  and  $y$  is just  $\Delta_{xy}$ .
- To see this,
  - let  $A_i A_j$  denote the two alleles in  $x$  and  $A_k A_l$  those in  $y$ .
  - Suppose that alleles  $i$  and  $k$  come from the mothers of these two relatives and alleles  $j$  and  $l$  from their fathers.
  - $\text{Cov}(\text{dominance values}) = \Pr(A_i \text{ ibd } A_k, A_j \text{ ibd } A_l) \text{cov}(\delta_{ij}, \delta_{kl}) + \Pr(A_i \text{ ibd } A_l, A_j \text{ ibd } A_k) \text{cov}(\delta_{ij}, \delta_{kl})$
  - $= (\theta_{fxjy} \theta_{mxmy} + \theta_{mxfy} \theta_{jxmy}) \text{Var}(D) = \Delta_{xy} \text{Var}(D)$

46

# Estimating relationships using molecular data

With SNP data, treat **identity in state** (also called **alike in state**, AIS) as IBD

Suppose the genotypes of two individual at 10 SNPs are

Indiv x: 00 00 10 10 00 10 11 00 11 00

Indiv y: 10 00 11 11 10 11 11 10 11 10



3/10 loci have  $\Delta_{xy} = 1$ , so average  $\Delta_{xy}$  over all loci is  $0.3 \times 1 = 0.3$

47

## General Resemblance between relatives

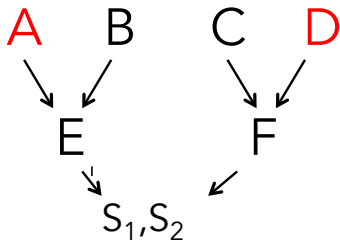
$$2\theta_{xy} = r_{xy}, \quad u_{xy} = \Delta_{xy}$$

$$\text{Cov}(G_x, G_y) = 2\theta_{xy}V_A + \Delta_{xy}V_D$$

$$\begin{aligned} \text{Cov}(G_x, G_y) = & 2\theta_{xy}V_A + \Delta_{xy}V_D \\ & + (2\theta_{xy})^2V_{AA} + 2\theta_{xy}\Delta_{xy}V_{AD} + \Delta_{xy}^2V_{DD} + \dots \end{aligned}$$

48

## Example



We found for full sibs  $S_1, S_2$  that  $\theta = 5/16$ , hence  $2\theta = 5/8$ ;  $\Delta = 17/64$

Expected genetic covariance between this sibs is

$$(5/8)\text{Var}(A) + (17/64)\text{Var}(D) + (5/8)^2\text{Var}(AA) + (5/8)(17/64)\text{Var}(AD) + (17/64)^2\text{Var}(DD) + \dots$$

49

## Autotetraploids

- Peanut, Potato, alfalfa, soybeans all examples of crops with at least some autotetraploid lines
- With autotetraploid, four alleles per locus, with a parent passing along two alleles to an offspring
- As a result, a parent can pass along the dominance contribution in G to an offspring
- Further, now there are four variance components associated with each locus

50

# Genetic variances for autotetraploids

- $G = A + D + T + Q$ 
  - A (additive) and D (dominance, or **digenic effects**) as with diploids
  - T (**trigenic effects**) are the three-way interactions among alleles at a locus
  - Q (**quadrigenic effects**) are the four-way interactions at a locus
- Total genetic variance becomes
  - $V_G = V_A + V_D + V_T + V_Q$

51

## Resemblance between autotetraploid relatives

Relatives	$V_A$	$V_D$	$V_T$	$V_Q$
Half-sibs	1/4	1/36		
Full-sibs	1/2	2/9	1/12	1/36
Parent -offspring	1/2	1/6		

Assumes unrelated, non-inbred parents

52

# Lecture 4

## Short-Term Selection

### Response: Breeder's equation

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

1

## Response to Selection

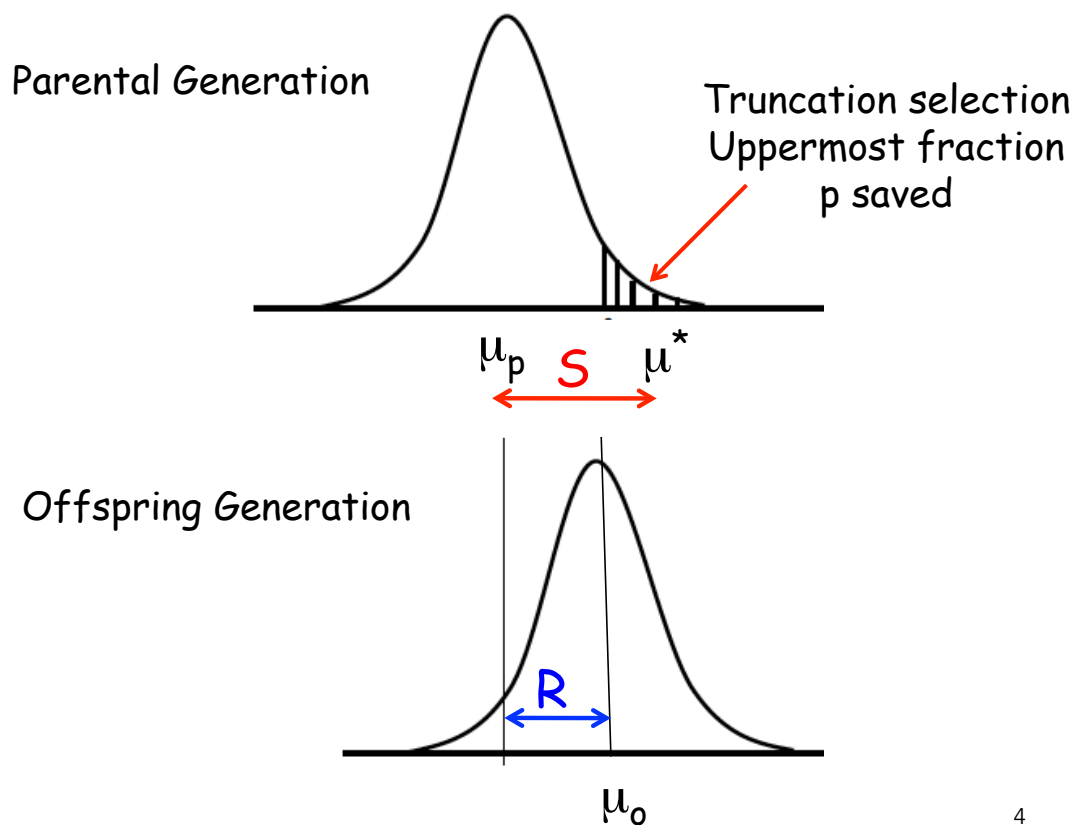
- Selection can change the distribution of phenotypes, and we typically measure this by changes in mean
  - This is a **within-generation change**
- Selection can also change the **distribution of breeding values**
  - This is the **response to selection**, the change in the trait in the next generation (the between-generation change)

2

# The Selection Differential and the Response to Selection

- The **selection differential  $S$**  measures the within-generation change in the mean
  - $S = \mu^* - \mu$
- The **response  $R$**  is the between-generation change in the mean
  - $R(t) = \mu(t+1) - \mu(t)$

3



4



## The Breeders' Equation: Translating S into R

Recall the regression of offspring value on midparent value

$$y_O = \mu_P + h^2 \left( \frac{P_f + P_m}{2} - \mu_P \right)$$

Averaging over the selected midparents,

$$E[(P_f + P_m)/2] = \mu^*,$$

Likewise, averaging over the regression gives

$$E[y_O - \mu] = h^2 (\mu^* - \mu) = h^2 S$$

Since  $E[y_O - \mu]$  is the change in the offspring mean, it represents the response to selection, giving:

$$R = h^2 S$$

The Breeders' Equation (Jay Lush)

- Note that no matter how strong S, if  $h^2$  is small, the response is small
- S is a measure of selection, R the actual response. One can get lots of selection but no response
- If offspring are asexual clones of their parents, the breeders' equation becomes
  - $R = H^2 S$
- If males and females subjected to differing amounts of selection,
  - $S = (S_f + S_m)/2$
  - Example: Selection on seed number in plants -- pollination (males) is random, so that  $S = S_f/2$

# Pollen control

- Recall that  $S = (S_f + S_m)/2$
- An issue that arises in plant breeding is **pollen control** --- is the pollen from plants that have also been selected?
- Not the case for traits (i.e., yield) scored after pollination. In this case,  $S_m = 0$ , so response only half that with pollen control
- Tradeoff: with an additional generation, a number of schemes can give pollen control, and hence twice the response
  - However, takes twice as many generations, so response per generation the same

7

# Selection on clones

- Although we have framed response in an outcrossed population, we can also consider selecting the best individual clones from a large population of different clones (e.g., inbred lines)
- $R = H^2S$ , now a function of the broad sense heritability. Since  $H^2 \geq h^2$ , the single-generation response using clones exceeds that using outcrossed individuals
- However, the genetic variation in the next generation is significantly reduced, reducing response in subsequent generations
  - In contrast, expect an almost continual response for several generations in an outcrossed population.

8

# Price-Robertson identity

- $S = \text{cov}(w, z)$
- The covariance between trait value  $z$  and relative fitness ( $w = W/\bar{W}$ , scaled to have mean fitness = 1)
- VERY! Useful result
- $R = \text{cov}(w, A_z)$ , as response = within generation change in BV
  - This is called [Robertson's secondary theorem of natural selection](#)

9

## Correcting for Reproductive Differences: Effective Selection Differentials

In artificial selection experiments,  $S$  is usually estimated as the difference between the mean of the selected adults and the sample mean of the population before selection. Selection need not stop at this stage. For example, strong artificial selection to increase a character might be countered by natural selection due to a decrease in the fertility of individuals with extreme character values. Biases introduced by such differential fertility can be removed by randomly choosing the same number of offspring from each selected parent, ensuring equal fertility.

Alternatively, biases introduced by differential fertility can be accounted for by using **effective selection differentials**,  $S_e$ .

$$S_e = \frac{1}{n_p} \sum_{i=1}^{n_p} \left( \frac{n_i}{\bar{n}} \right) (z_i - \mu_z) \quad (10.8)$$

where  $z_i$  and  $n_i$  are the phenotypic value and total number of offspring of the  $i$ th parent,  $n_p$  the number of parents selected to reproduce,  $\bar{n}$  the average number of offspring for selected parents, and  $\mu_z$  is the mean before selection. If all selected parents have the same number of offspring ( $n_i = \bar{n}$  for all  $i$ ), then  $S_e$  reduces to  $S$ . However, if there is variation in the number of offspring  $n_i$  among selected parents,  $S_e$  can be considerably different from  $S$ . This corrected differential is also referred to as the **realized selection differential**.

10

Suppose pre-selection mean = 30, and we select top 5. In the table  $z_i$  = trait value,  $n_i$  = number of offspring

$i$	$z_i$	$n_i$	$n_i/\bar{n}$
1	45	1	0.3125
2	40	2	0.6250
3	35	3	0.9375
4	33	5	1.563
5	32	5	1.563

$$\frac{1}{n_p} \sum_{i=1}^{n_p} \left( \frac{n_i}{\bar{n}} \right) z_i = 34.69$$

Hence,  $S_e = 4.69$ , for an expected response of  $R = 0.3 \cdot 4.69 = 1.4$ . In this case, not using the effective differential results in an overestimation of the expected response.

Unweighted  $S = 7$ , predicted response =  $0.3 \cdot 7 = 2.1$   
offspring-weighted  $S = 4.69$ , pred resp = 1.4

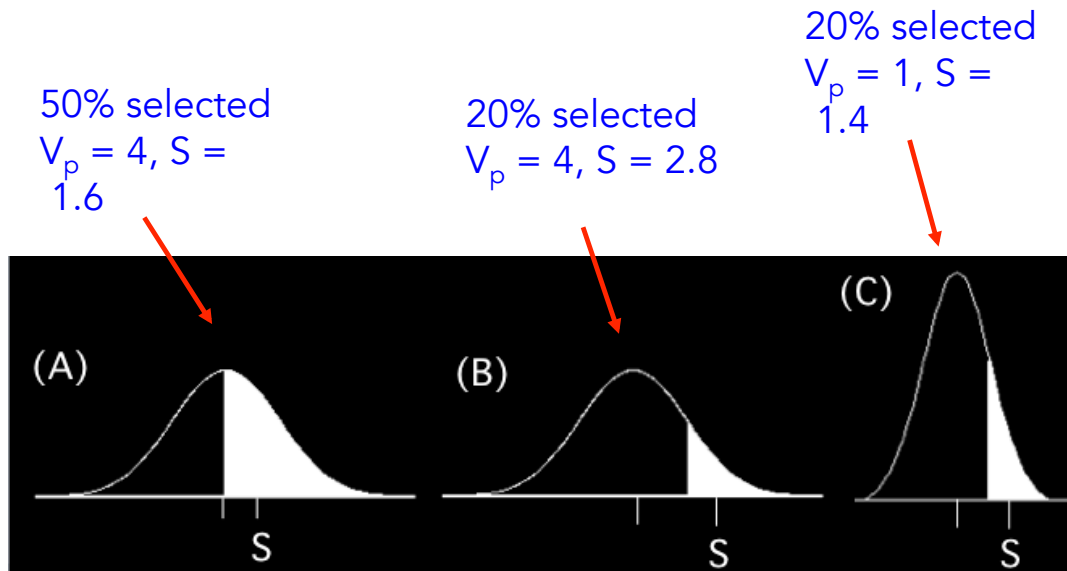
11

## Response over multiple generations

- Strictly speaking, the breeders' equation only holds for predicting a **single generation** of response from an **unselected base population**
- Practically speaking, the breeders' equation is usually pretty good for 5-10 generations
- The validity for an initial  $h^2$  predicting response over several generations depends on:
  - The reliability of the initial  $h^2$  estimate
  - Absence of environmental change between generations
  - The absence of genetic change between the generation in which  $h^2$  was estimated and the generation in which selection is applied

12

The selection differential is a function of both the phenotypic variance and the fraction selected



13

## The Selection Intensity, $i$

As the previous example shows, populations with the same selection differential ( $S$ ) may experience very different amounts of selection

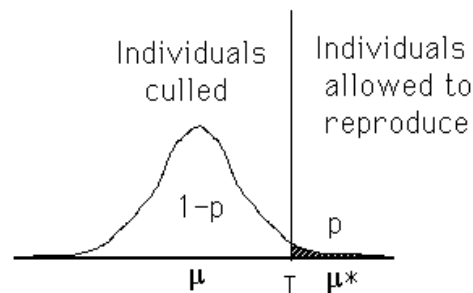
The **selection intensity**  $i$  provides a suitable measure for comparisons between populations,

$$i = \frac{S}{\sqrt{V_P}} = \frac{S}{\sigma_p}$$

14

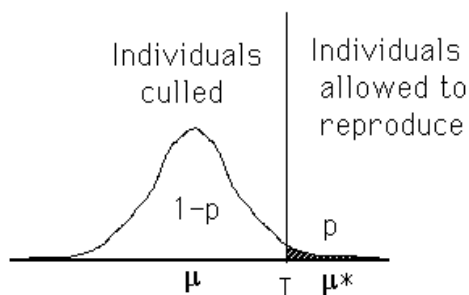
# Truncation selection

- A common method of artificial selection is truncation selection --- all individuals whose trait value is above some threshold (T) are chosen.
- Equivalent to only choosing the uppermost fraction p of the population



15

## Selection Differential Under Truncation Selection



$$S = \mu^* - \mu$$

$$S = \varphi\left(\frac{T - \mu}{\sigma}\right) \frac{\sigma}{p}$$

Likewise, 
$$\bar{i} = \frac{S}{\sigma} = \frac{\varphi(z_{[1-p]})}{p}$$

R code for i: `dnorm(qnorm(1-p)) / p`

16

# Truncation selection

- The fraction  $p$  saved can be translated into an expected selection intensity (assuming the trait is normally distributed),
  - allows a breeder (by setting  $p$  in advance) to choose an expected value of  $i$  before selection, and hence set the expected response

$$\bar{i} = \frac{S}{\sigma} = \frac{\varphi(z_{[1-p]})}{p}$$

Height of a unit normal at the threshold value corresponding to  $p$

$p$	0.5	0.2	0.1	0.05	0.01	0.005
$i$	0.798	1.400	1.755	2.063	2.665	2.892

R code for  $i$ : `dnorm(qnorm(1-p))/p`

17

## Selection Intensity Version of the Breeders' Equation

$$R = h^2 S = h^2 \frac{S}{\sigma_p} \sigma_p = i h^2 \sigma_p$$

$$\text{Since } h^2 \sigma_p = (\sigma_A^2 / \sigma_p^2) \sigma_p = \sigma_A (\sigma_A / \sigma_p) = h \sigma_A$$

$$R = i h \sigma_A$$

Since  $h$  = correlation between phenotypic and breeding values,  $h = r_{PA}$

$$R = i r_{PA} \sigma_A$$

$$\text{Response} = \text{Intensity} * \text{Accuracy} * \text{spread in Va}$$

When we select an individual solely on their phenotype, the accuracy (correlation) between BV and phenotype is  $h$

# Accuracy of selection

More generally, we can express the breeders equation as

$$R = i r_{uA} \sigma_A$$

Where we select individuals based on the index  $u$  (for example, the mean of  $n$  of their sibs).

$r_{uA}$  = the accuracy of using the measure  $u$  to predict an individual's breeding value = correlation between  $u$  and an individual's BV,  $A$

19

**Example 10.4. Progeny testing**, using the mean of a parent's offspring to predict the parent's breeding value, is an alternative predictor of an individual's breeding value. In this case, the correlation between the mean  $x$  of  $n$  offspring and the breeding value  $A$  of the parent is

$$\rho(x, A) = \sqrt{\frac{n}{n+a}}, \quad \text{where } a = \frac{4-h^2}{h^2}$$

From Equation 10.11, the response to selection under progeny testing is

$$R = i\sigma_A \sqrt{\frac{n}{n+a}} = i\sigma_A \sqrt{\frac{h^2 n}{4+h^2(n-1)}}$$

Note that for very large  $n$  that the accuracy approaches one. Progeny testing gives a larger response than simple selection on the phenotypes of the parents (**mass selection**) when

$$\sqrt{\frac{n}{4+h^2(n-1)}} > 1, \quad \text{or } n > \frac{4-h^2}{1-h^2}$$

In particular,  $n > 4, 5$ , and  $7$ , for  $h^2 = 0.1, 0.25$ , and  $0.5$ . Also note that the ratio of response for progeny testing ( $R_{pt}$ ) to mass selection ( $R_{ms}$ ) is just

$$\frac{R_{pt}}{R_{ms}} = \frac{1}{h} \sqrt{\frac{h^2 n}{4+h^2(n-1)}} = \sqrt{\frac{n}{4+h^2(n-1)}}$$

which approaches  $1/h$  for large  $n$ .



# Improving accuracy

- Predicting either the breeding or genotypic value from a single individual often has low accuracy ---  $h^2$  and/or  $H^2$  (based on a single individuals) is small
  - Especially true for many plant traits with high  $G \times E$
  - Need to replicate either clones or relatives (such as sibs) over regions and years to reduce the impact of  $G \times E$
  - Likewise, information from a set of relatives can give much higher accuracy than the measurement of a single individual

21

# Stratified mass selection

- In order to accommodate the high environmental variance with individual plant values, Gardner (1961) proposed the method of **stratified mass selection**
  - Population stratified into a number of different blocks (i.e., sections within a field)
  - The best fraction  $p$  within each block are chosen
  - Idea is that environmental values are more similar among individuals within each block, increasing trait heritability.

22

# Overlapping Generations

$L_x$  = **Generation interval** for sex x  
 = Average age of parents when progeny are born

The yearly rate of response is

$$R_y = \frac{i_m + i_f}{L_m + L_f} h^2 \sigma_p$$

Trade-offs: **Generation interval** vs. **selection intensity**:  
 If younger animals are used (decreasing L), i is also lower,  
 as more of the newborn animals are needed as replacements

23

## Computing generation intervals

OFFSPRING	Year 2	Year 3	Year 4	Year 5	total
Number (sires)	60	30	0	0	90
Number (dams)	400	600	100	40	1140

$$L_s = \frac{2 \cdot 60 + 3 \cdot 30}{60 + 30} = 2.33,$$

$$L_d = \frac{2 \cdot 400 + 3 \cdot 600 + 4 \cdot 100 + 5 \cdot 40}{400 + 600 + 100 + 40} = 2.81$$

24

# Generalized Breeder's Equation

$$R_y = \frac{i_m + i_f}{L_m + L_f} r_{uA} \sigma_A$$

Tradeoff between generation length  $L$  and accuracy  $r$

The longer we wait to replace an individual, the more accurate the selection (i.e., we have time for progeny testing and using the values of its relatives)

25

**Example 10.8.** As an example of the tradeoff between accuracy and generation intervals, consider a trait with  $h^2 = 0.25$  and selection only on sires. One scheme is to simply select on the sire's phenotype, which results in a sire generation interval of 1.5 years. Alternatively, one might perform progeny testing to improve the accuracy of the selected sires. This results in an increase of the sire generation interval to (say) 2.5 years. Suppose in both cases, the dam interval is steady at 1.5 years.

Since the intensity of selection and additive genetic variation are the same in both schemes, the ratio of response under mass selection to response under progeny testing is just

$$\frac{R(\text{Sire phenotype})}{R(\text{progeny mean})} = \frac{\rho(A, \text{Sire phenotype})/(L_s + L_d)}{\rho(A, \text{progeny mean})/(L_s + L_d)}$$

Here,  $\rho(A, \text{Sire phenotype}) = h = \sqrt{0.25} = 0.5$ , with generation intervals  $L_s + L_d = 1.5 + 1.5 = 3$ . With progeny testing, (Example 10.4)

$$\rho(A, \text{progeny mean}) = \sqrt{\frac{n}{n+a}} = \sqrt{\frac{n}{n+15}}$$

as  $a = (4 - h^2)/(h^2) = 15$ , with a total generation interval of  $L_s + L_d = 2.5 + 1.5 = 4$ . Hence,

$$\frac{R(\text{Sire phenotype})}{R(\text{progeny mean})} = \frac{0.5/3.0}{\sqrt{\frac{n}{n+15}}/4} = \frac{2}{3} \cdot \sqrt{\frac{n+15}{n}}$$

If (say)  $n = 2$  progeny are tested per sire, this ratio is 1.95, giving a much larger rate of response under sire-only selection. For  $n = 12$ , the ratio is exactly one, while for a very large number of offspring tested per sire, the ratio approaches  $2/3$ , or a 1.5-fold increase in the rate of response under progeny testing, despite the increase in sire generation interval.

# Permanent Versus Transient Response

Considering epistasis and shared environmental values, the single-generation response follows from the midparent-offspring regression

$$R = h^2 S + \frac{S}{\sigma_z^2} \left( \frac{\sigma_{AA}^2}{2} + \frac{\sigma_{AAA}^2}{4} + \dots + \sigma(E_{sire}, E_o) + \sigma(E_{dam}, E_o) \right)$$

Breeder's Equation
Response from epistasis
Response from shared environmental effects

Permanent component of response

Transient component of response --- contributes to short-term response. Decays away to zero over the long-term

27

# Permanent Versus Transient Response

The reason for the focus on  $h^2S$  is that this component is permanent in a random-mating population, while the other components are transient, initially contributing to response, but this contribution decays away under random mating

Why? Under HW, changes in allele frequencies are permanent (don't decay under random-mating), while LD (epistasis) does, and environmental values also become randomized

28

# Response with Epistasis

The response after one generation of selection from an unselected base population with A x A epistasis is

$$R = S \left( h^2 + \frac{\sigma_{AA}^2}{2\sigma_z^2} \right)$$

The contribution to response from this single generation after  $\tau$  generations of no selection is

$$R(1 + \tau) = S \left( h^2 + (1 - c)^\tau \frac{\sigma_{AA}^2}{2\sigma_z^2} \right)$$

$c$  is the average (pairwise) recombination between loci involved in A x A

29

# Response with Epistasis

$$R(1 + \tau) = S \left( h^2 + (1 - c)^\tau \frac{\sigma_{AA}^2}{2\sigma_z^2} \right)$$

Response from additive effects ( $h^2 S$ ) is due to changes in allele frequencies and hence is permanent. Contribution from A x A due to linkage disequilibrium

Contribution to response from epistasis decays to zero as linkage disequilibrium decays to zero

30

Why breeder's equation assumption of an unselected base population?  
 If history of previous selection, linkage disequilibrium may be present  
 and the mean can change as the disequilibrium decays

For  $t$  generation of selection followed by  
 $\tau$  generations of no selection (but recombination)

$$R(t + \tau) = t h^2 S + (1 - c)^\tau R_{AA}(t)$$

$R_{AA}$  has a limiting  
 value given by

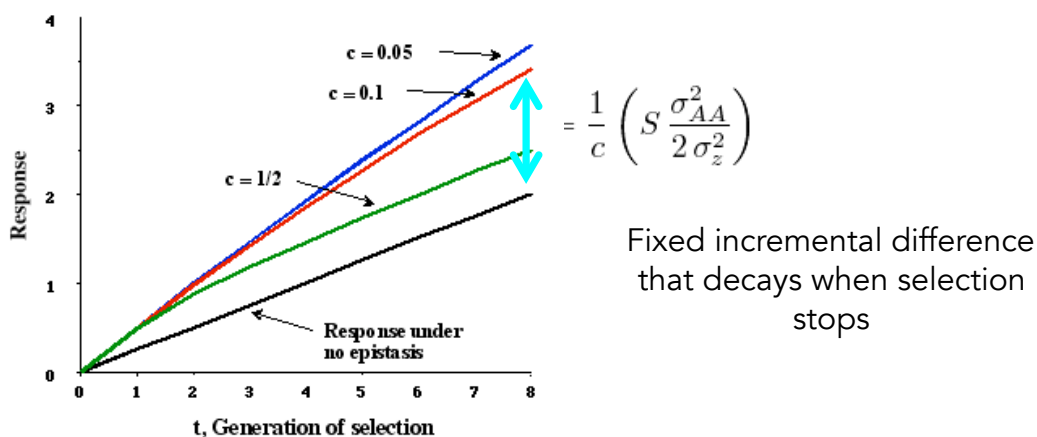
$$\tilde{R}_{AA} = \lim_{t \rightarrow \infty} R_{AA}(t) = \frac{1}{c} \left( S \frac{\sigma_{AA}^2}{2 \sigma_z^2} \right)$$

Time to equilibrium a  
 function of  $c$

$$t_{1/2} = \frac{-\ln(2)}{\ln(1 - c)}$$

Decay half-life

31



What about response with higher-order epistasis?

$S\sigma^2(A^i)/\sigma_z^2$	AA	AAA	AAAA	AAAAA
$R(1)$	0.500	0.250	0.125	0.063
Limit	1.000	0.333	0.143	0.067
% $R(1)/\text{limit}$	50.0	75.0	87.5	93.8

# Response in autotetraploids

- Autotetraploids pass along two alleles at each locus to their offspring
- Hence, dominance variance is passed along
- However, as with A x A, this depends upon favorable **combinations** of alleles, and these are randomized over time by transmission, so D component of response is transient.

33

## Autotetraploids

P-O covariance

Single-generation response

$$\sigma(z_p, z_o) = \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{6}, \quad R = S \left( h^2 + \frac{\sigma_D^2}{3\sigma_z^2} \right)$$

Response to t generations of selection with constant selection differential S

$$R(t) = th^2S + R_D(t)$$

$$R_D(t) = S \frac{3}{2} \left[ 1 - \left( \frac{1}{3} \right)^t \right] \frac{\sigma_D^2}{3\sigma_z^2}$$

Response remaining after t generations of selection followed by  $\tau$  generations of random mating

$$th^2S + (1/3)^\tau R_D(t)$$

Contribution from dominance quickly decays to zero

34

# General responses

- For both individual and family selection, the response can be thought of as a regression of some phenotypic measurement (such as the individual itself or its corresponding selection unit value  $x$ ) on either the offspring value ( $y$ ) or the breeding value  $R_A$  of an individual who will be a parent of the next generation (the recombination group).
- The regression slope for predicting
  - $y$  from  $x$  is  $\sigma(x,y)/\sigma^2(x)$
  - BV  $R_A$  from  $x$   $\sigma(x,R_A)/\sigma^2(x)$
- With transient components of response, these covariances now also become functions of time --- e.g. the covariance between  $x$  in one generation and  $y$  several generations later

35

## Maternal Effects:

Falconer's **dilution model**

$$z = G + m z_{\text{dam}} + e$$

$G$  = Direct genetic effect on character

$$G = A + D + I. \quad E[A] = (A_{\text{sire}} + A_{\text{dam}})/2$$

maternal effect passed from dam to offspring  $m z_{\text{dam}}$  is just a fraction  $m$  of the dam's phenotypic value

The presence of the maternal effects means that response is not necessarily linear and time lags can occur in response

$m$  can be negative --- results in the potential for a **reversed response**

36



Parent-offspring regression under the dilution model

In terms of parental breeding values,

$$E(z_o | A_{dam}, A_{sire}, z_{dam}) = \frac{A_{dam}}{2} + \frac{A_{sire}}{2} + m z_{dam}$$

Regression of BV on phenotype

$$A = \mu_A + b_{Az} (z - \mu_z) + e$$

The resulting slope becomes  $b_{Az} = h^2 / (2 - m)$

With no maternal effects,  $b_{az} = h^2$

37

Parent-offspring regression under the dilution model

With maternal effects, a covariance between BV and maternal effect arises, with  $\sigma_{A,M} = m \sigma_A^2 / (2 - m)$

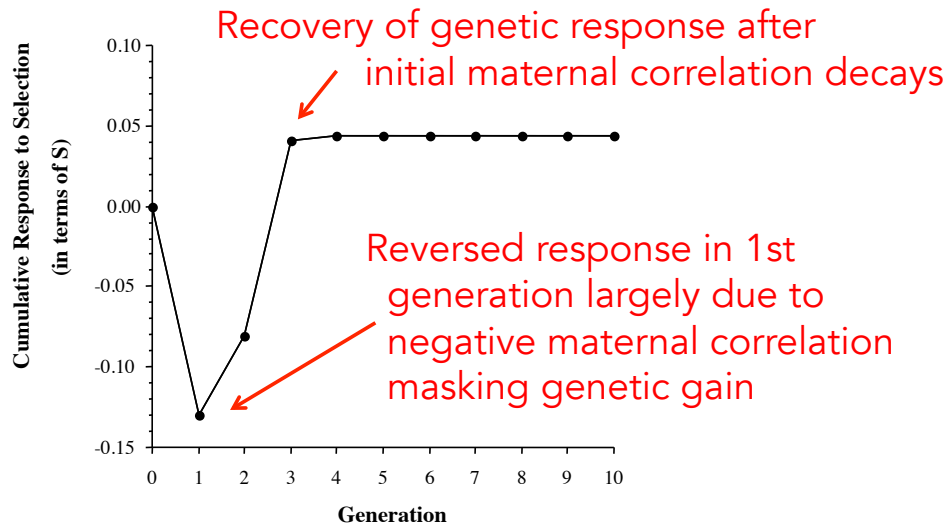
The response thus becomes

$$\Delta\mu_z = S_{dam} \left( \frac{h^2}{2 - m} + m \right) + S_{sire} \frac{h^2}{2 - m}$$

38

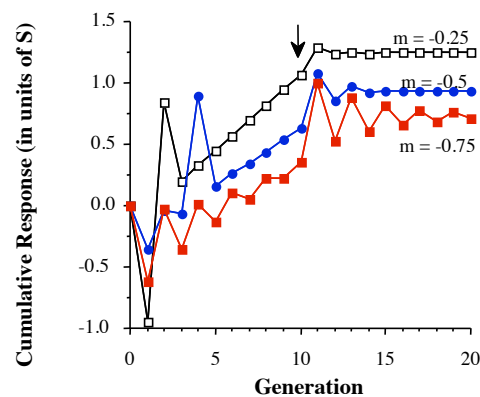
## Response to a single generation of selection

$h^2 = 0.11$ ,  $m = -0.13$  (litter size in mice)



39

## Selection occurs for 10 generations and then stops



$h^2 = 0.35$

40

# Additional material

Unlikely to be covered in class

41

## Selection on Threshold Traits

Response on a binary trait is a special case of response on a continuous trait

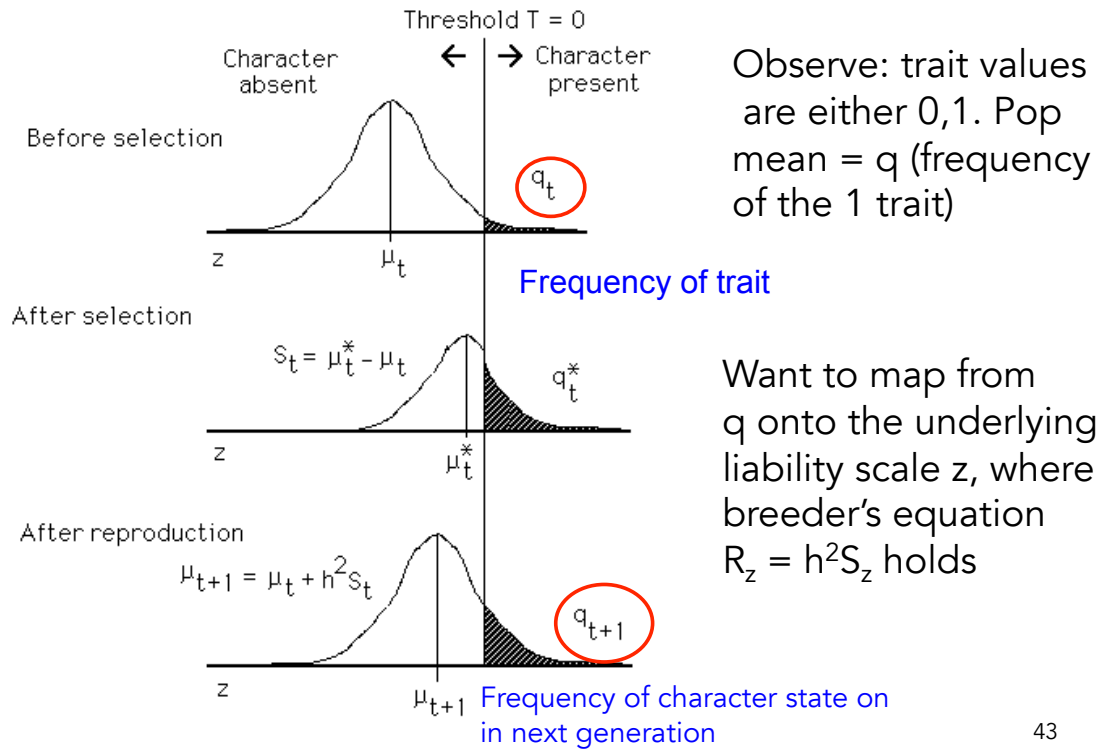
Assume some underlying continuous value  $z$ , the **liability**, maps to a discrete trait.

$z < T$     character state zero (i.e. no disease)

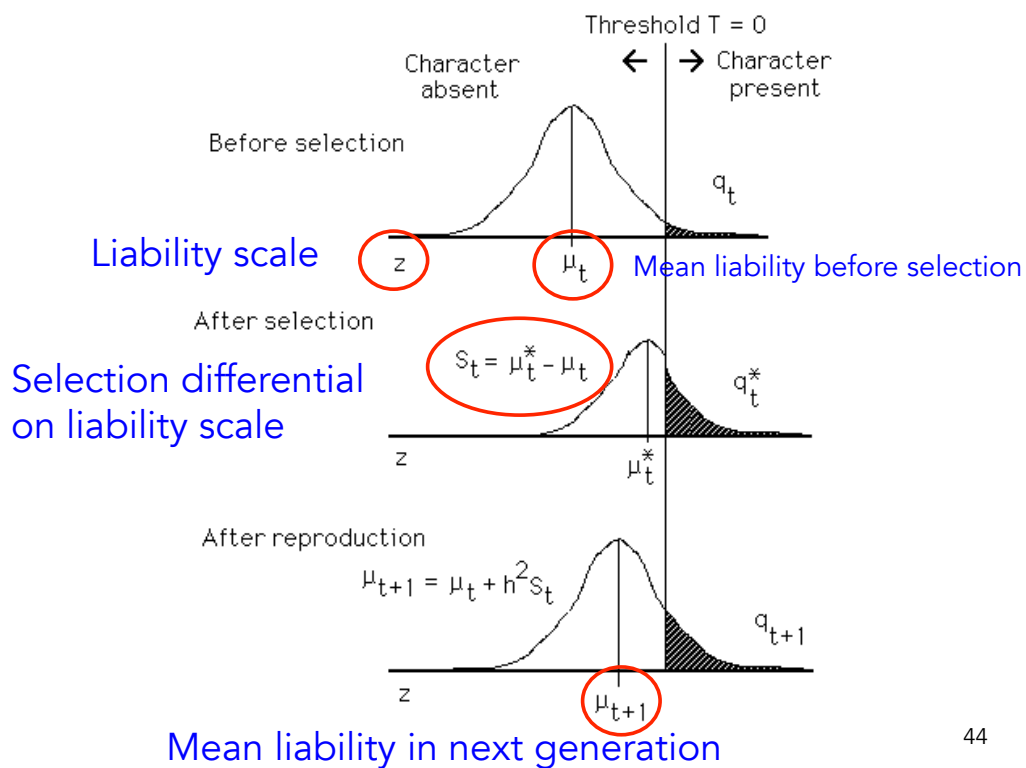
$z \geq T$     character state one (i.e. disease)

Alternative (but essentially equivalent model) is a **probit** (or **logistic**) model, when  $p(z) = \text{Prob}(\text{state one} \mid z)$ . Details in LW Chapter 14.

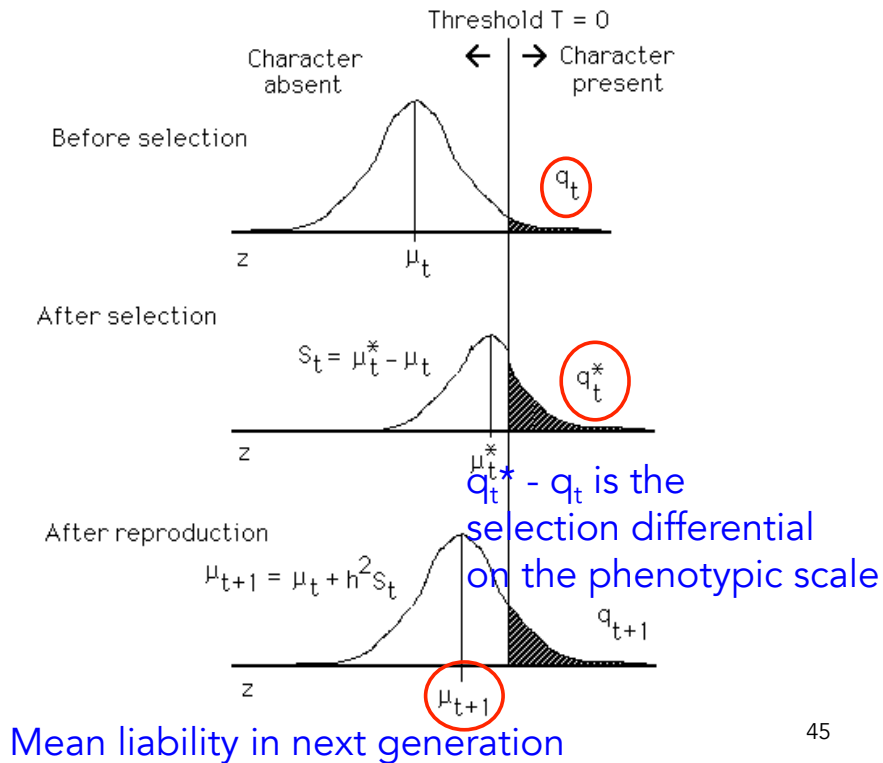
42



43



44



45

## Steps in Predicting Response to Threshold Selection

i) Compute initial mean  $\mu_0$

$$P(\text{trait}) = P(z \geq 0) = P(z - \mu \geq -\mu) = P(U \geq -\mu)$$

$U$  is a unit normal

Hence,  $z - \mu_0$  is a unit normal random variable

We can choose a scale where the liability  $z$  has variance of one and a threshold  $T = 0$

Define  $z_{[q]} = P(U < z_{[q]}) = q$ .  $P(U \geq z_{[1-q]}) = q$

General result:  $\mu = -z_{[1-q]}$

For example, suppose 5% of the pop shows the trait.  $P(U > 1.645) = 0.05$ , hence  $\mu = -1.645$ . Note: in R,  $z_{[1-q]} = \text{qnorm}(1-q)$ , with  $\text{qnorm}(0.95)$  returning 1.644854

46

## Steps in Predicting Response to Threshold Selection

ii) The frequency  $q_{t+1}$  of the trait in the next generation is just

$$q_{t+1} = P(U > -\mu_{t+1}) = P(U > -[h^2S + \mu_t]) \\ = P(U > -h^2S - z_{[1-q]})$$

iii) Hence, we need to compute  $S$ , the selection differential for the liability  $z$

Let  $p_t$  = fraction of individuals chosen in generation  $t$  that display the trait

$$\mu_t^* = (1 - p_t)E(z | z < 0, \mu_t) + p_tE(z | z \geq 0, \mu_t)$$

47

$$\mu_t^* = (1 - p_t)E(z | z < 0, \mu_t) + p_tE(z | z \geq 0, \mu_t)$$

↖  
This fraction does not display  
the trait, hence  $z < 0$

↖  
This fraction displays  
the trait, hence  $z \geq 0$

When  $z$  is normally distributed, this reduces to

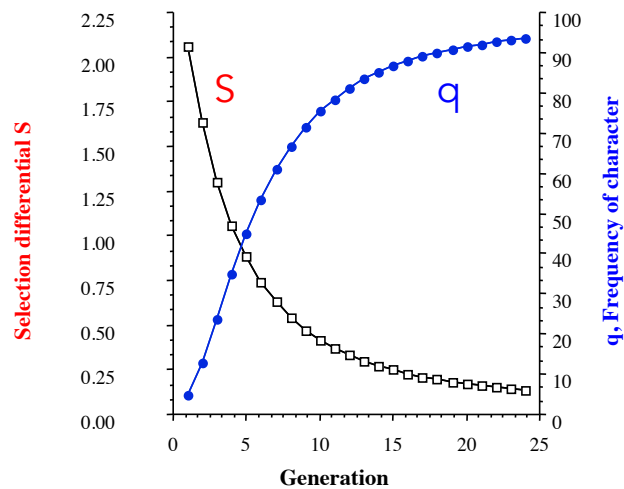
$$S_t = \pi_t^* - \pi_t = \frac{\phi(\pi_t)}{q_t} \frac{p_t - q_t}{1 - q_t}$$

↖  
Height of the unit normal density function  
at the point  $\mu_t$

Hence, we start at some initial value given  $h^2$  and  $\mu_0$ , and iterate to obtain selection response

48

Initial frequency of  $q = 0.05$ . Select only on adults showing the trait ( $p_t = 1$ )



49

## Ancestral Regressions

When regressions on relatives are linear, we can think of the response as the sum over all previous contributions

For example, consider the response after 3 gens:

$$R(3) = 8\beta_{3,0}S_0 + 4\beta_{3,1}S_1 + 2\beta_{3,2}S_2$$

8 great-grand parents

$S_0$  is there selection differential

$\beta_{3,0}$  is the regression coefficient for an offspring at time 3 on a great-grandparent From time 0

4 grandparents  
Selection diff  $S_1$

$\beta_{3,1}$  is the regression of relative in generation 3 on their gen 1 relatives

2 parents

50

# Ancestral Regressions

More generally,

$$R(T) = \sum_{t=0}^{T-1} 2^{T-t} \beta_{T,t} S_t \quad \beta_{T,t} = \text{cov}(z_T, z_t)$$

The general expression  $\text{cov}(z_T, z_t)$ , where we keep track of the actual generation, as oppose to  $\text{cov}(z, z_{T-t})$  -- how many generations separate the relatives, allows us to handle inbreeding, where the regression slope changes over generations of inbreeding.

Unless  $2^t \beta_{\tau+t, \tau}$  remains constant as  $t$  increases, the contribution to cumulative response from selection on adults in generation  $\tau$  changes over time. For example, when loci are strictly additive (no dominance or epistasis),  $\sigma_G(\tau + t, \tau) = 2^{-t} \sigma_A^2(\tau)$  and thus  $2^t \beta_{\tau+t, \tau} = h_{\tau}^2$ , the standard result from the breeders' equation. However, unless  $2^t \sigma_G(\tau + t, \tau)$  remains constant, any response contributed decays. Hence any term of  $\sigma_G(\tau + t, \tau)$  that decreases by more than 1/2 each generation contributes only to the transient response.

## Changes in the Variance under Selection

**The infinitesimal model** --- each locus has a very small effect on the trait.

Under the infinitesimal, require many generations for significant change in allele frequencies

However, can have significant change in genetic variances due to selection creating **linkage disequilibrium**

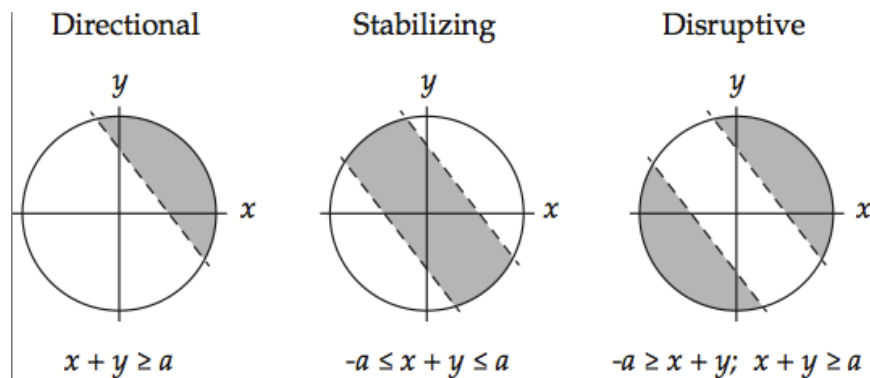
Under **linkage equilibrium**,  $\text{freq}(\text{AB gamete}) = \text{freq}(\text{A})\text{freq}(\text{B})$

With **positive linkage disequilibrium**,  $f(\text{AB}) > f(\text{A})f(\text{B})$ , so that AB gametes are more frequent

With **negative linkage disequilibrium**,  $f(\text{AB}) < f(\text{A})f(\text{B})$ , so that AB gametes are less frequent



Selection that reduces the variance generates negative  $d$ , selection that increases the variance generates positive  $d$



53

## Additive variance with LD:

Additive variance is the variance of the sum of allelic effects,

**Genic variance:** value of  $\text{Var}(A)$  in the absence of disequilibrium function of allele frequencies

$$\begin{aligned} \sigma^2 \left( \sum_{k=1}^n (a_1^{(k)} + a_2^{(k)}) \right) &= 2 \sum_{k=1}^n \sigma^2 (a^{(k)}) + 4 \sum_{k < j}^n \sigma (a^{(j)}, a^{(k)}) \\ &= 2 \sum_{k=1}^n C_{kk} + 4 \sum_{k < j}^n C_{jk} \\ &\rightarrow \sigma_A^2 = \sigma_a^2 + d \end{aligned}$$

Additive variance

Disequilibrium contribution. Requires covariances between allelic effects at different loci

54

Key: Under the infinitesimal model, no (selection-induced) changes in **genic variance**  $\sigma_a^2$

Selection-induced changes in  $d$  change  $\sigma_A^2, \sigma_z^2, h^2$

$$\sigma_z^2(t) = \sigma_E^2 + \sigma_D^2 + \sigma_A^2(t) = \sigma_z^2 + d(t)$$

$$h^2(t) = \frac{\sigma_A^2(t)}{\sigma_z^2(t)} = \frac{\sigma_a^2 + d(t)}{\sigma_z^2 + d(t)}$$

Dynamics of  $d$ : With unlinked loci,  $d$  loses half its value each generation (i.e,  $d$  in offspring is  $1/2$   $d$  of their parents,

$$d(t+1) = \frac{d(t)}{2}$$

55

Dynamics of  $d$ : Computing the effect of selection in generating  $d$

Consider the parent-offspring regression

$$z_o = \mu + \frac{h^2}{2}(z_m - \mu) + \frac{h^2}{2}(z_f - \mu) + e$$

$$\sigma_e^2 = \left(1 - \frac{h^4}{2}\right) \sigma_z^2$$

Taking the variance of the offspring given the selected parents gives

$$\begin{aligned} \sigma^2(z_o) &= \frac{h^4}{4} [\sigma^2(z_m^*) + \sigma^2(z_f^*)] + \sigma_e^2 \\ &= \frac{h^4}{2} [\sigma_z^2 + \delta(\sigma_z^2)] + \left(1 - \frac{h^4}{2}\right) \sigma_z^2 \\ &= \sigma_z^2 + \frac{h^4}{2} \delta(\sigma_z^2) \end{aligned}$$

Change in variance from selection

56

Change in d = change from recombination plus  
change from selection

$$d(t+1) = \frac{d(t)}{2} + \frac{h^4}{2} \delta(\sigma_z^2) = d(t+1) = \frac{d(t)}{2} + \frac{h^4(t)}{2} \delta(\sigma_{z(t)}^2)$$

Recombination                  Selection

In terms of change in d,

$$\Delta d(t) = \Delta \sigma_{z(t)}^2 = \Delta \sigma_A^2(t)$$

$$= -\frac{d(t)}{2} + \frac{h^4(t)}{2} \delta(\sigma_{z(t)}^2)$$

This is the [Bulmer Equation](#) (Michael Bulmer), and it is  
akin to a breeder's equation for [the change in variance](#)

At the selection-recombination  
equilibrium,

$$\tilde{d} = \tilde{h}^4 \tilde{\delta}(\sigma_z^2)$$

57

## Application: Egg Weight in Ducks

Rendel (1943) observed that while the change  
mean weight weight (in all vs. hatched) as  
negligible, but their was a significance decrease  
in the variance, suggesting stabilizing selection

Before selection, variance = 52.7, reducing to  
43.9 after selection. Heritability was  $h^2 = 0.6$

$$\tilde{d} = \tilde{h}^4 \tilde{\delta}(\sigma_z^2) = 0.6^2 (43.9 - 52.7) = -3.2$$

Var(A) =  $0.6 \times 52.7 = 31.6$ . If selection stops, Var(A)  
is expected to increase to  $31.6 + 3.2 = 34.8$

Var(z) should increase to 55.9, giving  $h^2 = 0.62$

58

# Specific models of selection-induced changes in variances

Proportional reduction model:

constant fraction  $\kappa$  of  
variance removed

$$\sigma_{z^*}^2 = (1 - \kappa) \sigma_z^2$$

$$\delta(\sigma_z^2) = \sigma_{z^*}^2 - \sigma_z^2 = -\kappa \sigma_z^2$$

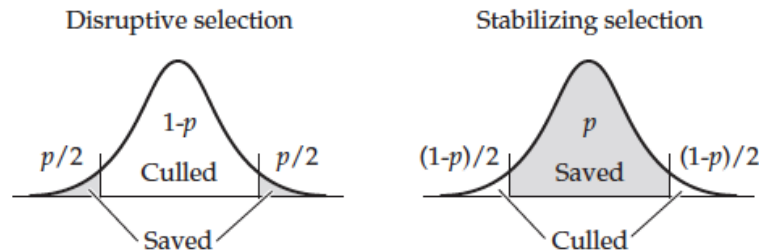
Bulmer equation simplifies  
to

$$\begin{aligned} d(t+1) &= \frac{d(t)}{2} - \frac{\kappa}{2} h^2(t) \sigma_A^2(t) \\ &= \frac{d(t)}{2} - \frac{\kappa}{2} \frac{[\sigma_a^2 + d(t)]^2}{\sigma_z^2 + d(t)} \end{aligned}$$

Closed-form solution  
to equilibrium  $h^2$

$$\tilde{h}^2 = \frac{-1 + \sqrt{1 + 4h^2(1 - h^2)\kappa}}{2\kappa(1 - h^2)}$$

59



**Directional Truncation Selection:** Uppermost (or lowermost)  $p$  saved

$$\kappa = \frac{\varphi(z_{[1-p]})}{p} \left( \frac{\varphi(z_{[1-p]})}{p} - z_{[1-p]} \right) = \bar{z} (\bar{z} - z_{[1-p]})$$

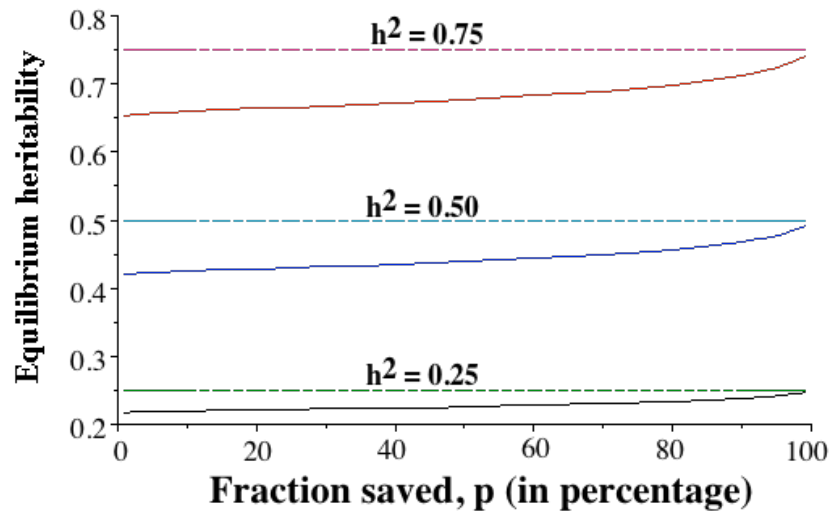
**Stabilizing Truncation Selection:** Middle fraction  $p$  of the distribution saved

$$\kappa = \frac{2 \varphi(z_{[1/2+p/2]})}{p} z_{[1/2+p/2]}$$

**Disruptive Truncation Selection:** Uppermost and lowermost  $p/2$  saved

$$\kappa = - \frac{2 \varphi(z_{[1-p/2]})}{p} z_{[1-p/2]}$$

## Equilibrium $h^2$ under direction truncation selection



61

## Directional truncation selection

$$\kappa = \bar{\tau} (\bar{\tau} - z_{[1-p]})$$

**Example 13.2.** Suppose directional truncation selection is performed (equally on both sexes) on a normally distributed character with  $\sigma_z^2 = 100$ ,  $h^2 = 0.5$ , and  $p = 0.20$  (the upper 20 percent of the population is saved). From normal distribution tables,

$$\Pr(U \leq 0.84) = 0.8, \quad \text{hence} \quad z_{[0.8]} = 0.84$$

Likewise, evaluating the unit normal gives  $\varphi(0.84) = 0.2803$ , so that (Equation 10.26a)

$$\bar{\tau} = \varphi(0.84)/p = 0.2803/0.20 = 1.402$$

From Equation 13.15b, the fraction of variance removed by selection is

$$\kappa = 1.402(1.402 - 0.84) = 0.787.$$

Hence, Equation 13.12 gives

$$d(t+1) = \frac{d(t)}{2} - 0.394 \frac{[50 + d(t)]^2}{100 + d(t)}$$

Generation	0	1	2	3	4	5	$\infty$
$d(t)$	0.00	-9.84	-11.96	-12.45	-12.56	-12.59	-12.59
$\sigma_A^2(t)$	50.00	40.16	38.04	37.55	37.44	37.41	37.41
$h^2(t)$	0.50	0.45	0.43	0.43	0.43	0.43	0.43

2

Changes in the variance = changes in  $h^2$   
and even  $S$  (under truncation selection)

$$R(t) = h^2(t) S(t)$$

How does this reduction in  $\sigma_A^2$  influence the per-generation change in mean,  $R(t)$ ? Since the selection  $\bar{i}$  is unchanged (being entirely a function of the fraction  $p$  of adults saved), but  $h^2$  and  $\sigma_z^2$  change over time, Equation 10.6b gives the response as

$$R(t) = h^2(t) \bar{i} \sigma_z(t) = 1.402 h^2(t) \sqrt{\sigma_z^2 + d(t)} = 1.402 h^2(t) \sqrt{100 + d(t)}$$

Response declines from an initial value of  $R = 1.4 \cdot 0.5 \cdot 10 = 7$  to an asymptotic per-generation value of  $\tilde{R} = 1.4 \cdot 0.43 \cdot \sqrt{87.41} = 5.6$ . Thus if we simply used the Breeders' equation to predict change in mean over several generations without accounting for the Bulmer effect, we would have *overestimated* the expected response by 25 percent.

# Lecture 5

## Inbreeding and Crossbreeding

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

1

## Inbreeding

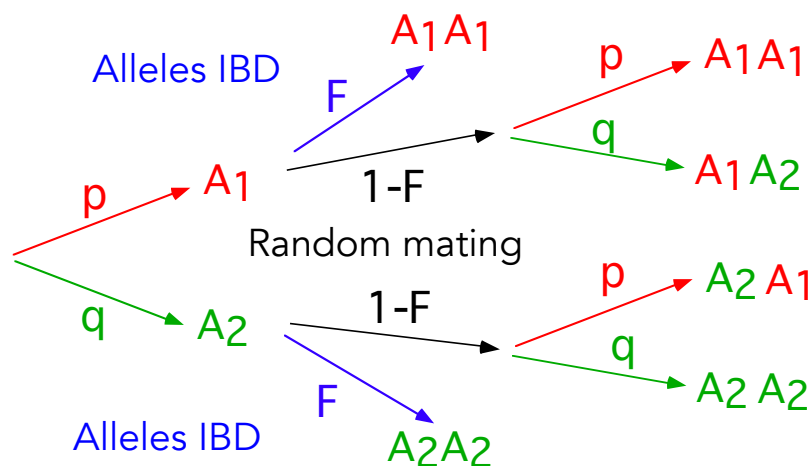
- Inbreeding = mating of related individuals
- Often results in a change in the mean of a trait
- Inbreeding is intentionally practiced to:
  - create genetic uniformity of laboratory stocks
  - produce stocks for crossing (animal and plant breeding)
- Inbreeding is unintentionally generated:
  - by keeping small populations (such as is found at zoos)
  - during selection

2

## Genotype frequencies under inbreeding

- The inbreeding coefficient,  $F$
- $F = \text{Prob}(\text{the two alleles within an individual are IBD})$  -- identical by descent
- Hence, with probability  $F$  both alleles in an individual are identical, and hence a homozygote
- With probability  $1-F$ , the alleles are combined at random

3



Genotype	Alleles IBD	Alleles not IBD	frequency
$A_1A_1$	$Fp$	$(1-F)p^2$	$p^2 + Fpq$
$A_2A_1$	0	$(1-F)2pq$	$(1-F)2pq$
$A_2A_2$	$Fq$	$(1-F)q^2$	$q^2 + Fpq$

4



# Changes in the mean under inbreeding

Genotypes	$A_1A_1$	$A_1A_2$	$A_2A_2$
	0	a+d	2a

$$\text{freq}(A_1) = p, \quad \text{freq}(A_2) = q$$

Using the genotypic frequencies under inbreeding, the population mean  $\mu_F$  under a level of inbreeding  $F$  is related to the mean  $\mu_0$  under random mating by

$$\mu_F = \mu_0 - 2Fpqd$$

5

**For k loci, the change in mean is**

$$\mu_F = \mu_0 - 2F \sum_{i=1}^k p_i q_i d_i = \mu_0 - BF$$

**Here B is the reduction in mean under complete inbreeding ( $F=1$ ), where**

$$B = 2 \sum p_i q_i d_i$$

- There will be a change of mean value if dominance is present ( $d \neq 0$ )
- For a single locus, if  $d > 0$ , inbreeding will decrease the mean value of the trait. If  $d < 0$ , inbreeding will increase the mean
- For multiple loci, a decrease ([inbreeding depression](#)) requires [directional dominance](#) --- dominance effects  $d_i$  tending to be positive.
- The magnitude of the change of mean on inbreeding depends on gene frequency, and is greatest when  $p = q = 0.5$

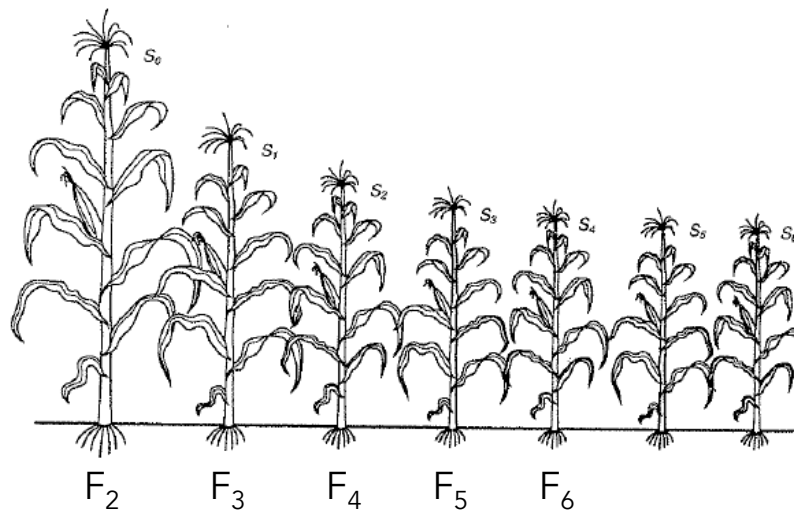
6

# Inbreeding Depression and Fitness traits



7

## Inbreeding depression



Example for maize height

8

## Fitness traits and inbreeding depression

- Often seen that inbreeding depression is strongest on fitness-related traits such as yield, height, etc.
- Traits less associated with fitness often show less inbreeding depression
- Selection on fitness-related traits may generate directional dominance

9

## Why do traits associated with fitness show inbreeding depression?

- Two competing hypotheses:
  - **Overdominance Hypothesis**: Genetic variance for fitness is caused by loci at which heterozygotes are more fit than both homozygotes. Inbreeding decreases the frequency of heterozygotes, increases the frequency of homozygotes, so fitness is reduced.
  - **Dominance Hypothesis**: Genetic variance for fitness is caused by rare deleterious alleles that are recessive or partly recessive; such alleles persist in populations because of recurrent mutation. Most copies of deleterious alleles in the base population are in heterozygotes. Inbreeding increases the frequency of homozygotes for deleterious alleles, so fitness is reduced.

10

# Inbred depression in largely selfing lineages

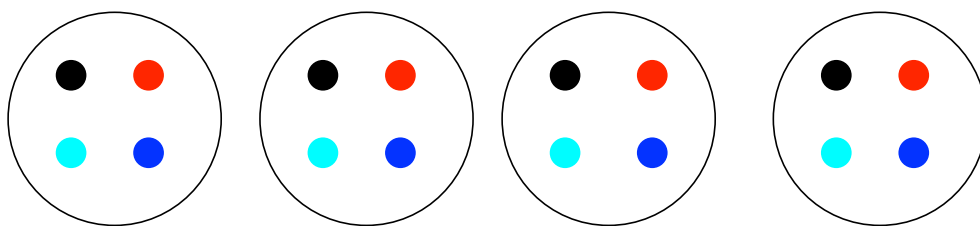
- Inbreeding depression is common in outcrossing species
- However, generally fairly uncommon in species with a high rate of selfing
- One idea is that the constant selfing have purged many of the deleterious alleles thought to cause inbreeding depression
- However, lack of inbreeding depression also means a lack of heterosis (a point returned to shortly)
  - Counterexample is Rice: Lots of heterosis but little inbreeding depression

11

## Variance Changes Under Inbreeding

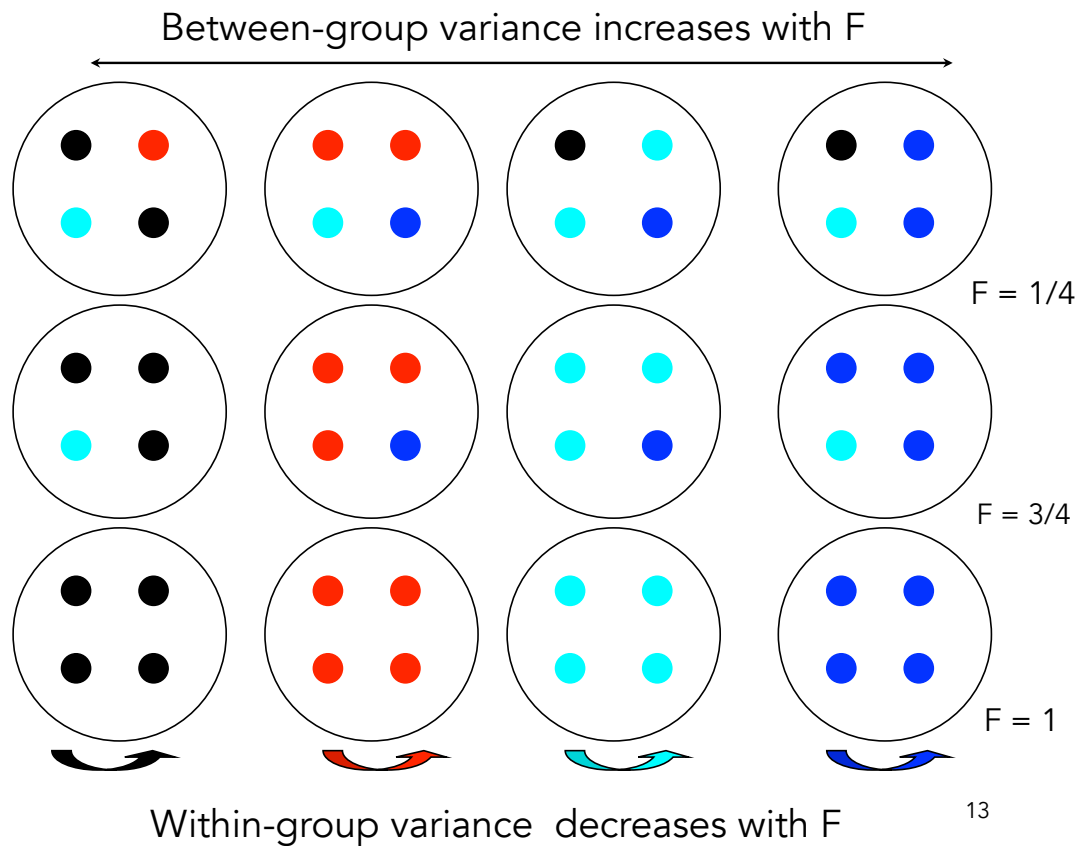
Inbreeding **reduces variation within each population**

Inbreeding **increases the variation between populations**  
(i.e., variation in the means of the populations)



$$F = 0$$

12



## Implications for traits

- A series of inbred lines from an  $F_2$  population are expected to show
  - **more within-line uniformity** (variance about the mean within a line)
    - Less within-family genetic variation for selection
  - **more between-line divergence** (variation in the mean value between lines)
    - More between-family genetic variation for selection

# Variance Changes Under Inbreeding

	General	F = 1	F = 0
Between lines	$2FV_A$	$2V_A$	0
Within Lines	$(1-F) V_A$	0	$V_A$
Total	$(1+F) V_A$	$2V_A$	$V_A$

The above results assume ONLY additive variance i.e., no dominance/epistasis. When nonadditive variance present, results very complex (see WL Chpt 11).

## Line Crosses: Heterosis

When inbred lines are crossed, the progeny show an increase in mean for characters that previously suffered a reduction from inbreeding.

This increase in the mean over the average value of the parents is called **hybrid vigor** or **heterosis**

$$H_{F_1} = \mu_{F_1} - \frac{\mu_{P_1} + \mu_{P_2}}{2}$$

A cross is said to show heterosis if  $H > 0$ , so that the  $F_1$  mean is larger than the average of both parents.

## Expected levels of heterosis

If  $p_i$  denotes the frequency of  $Q_i$  in line 1, let  $p_i + \delta p_i$  denote the frequency of  $Q_i$  in line 2.

The expected amount of heterosis becomes

$$H_{F_1} = \sum_{i=1}^n (\delta p_i)^2 d_i$$

- **Heterosis depends on dominance:**  $d = 0$  = no inbreeding depression and no Heterosis. As with inbreeding depression, directional dominance is required for heterosis.
- **H is proportional to the square of the difference in allele frequencies between populations** H is greatest when alleles are fixed in one population and lost in the other (so that  $|\delta p_i| = 1$ ).  $H = 0$  if  $\delta p = 0$ .
- **H is specific to each particular cross.** H must be determined empirically, since we do not know the relevant loci nor their gene frequencies.

17

## Heterosis declines in the $F_2$

In the  $F_1$ , all offspring are heterozygotes. In the  $F_2$ , random mating has occurred, reducing the frequency of heterozygotes.

As a result, there is a reduction of the amount of heterosis in the  $F_2$  relative to the  $F_1$ ,

$$H_{F_2} = \mu_{F_2} - \frac{\mu_{P_1} + \mu_{P_2}}{2} = \frac{(\delta p)^2 d}{2} = \frac{H_{F_1}}{2}$$

Since random mating occurs in the  $F_2$  and subsequent generations, the **level of heterosis stays at the  $F_2$  level.**

# Agricultural importance of heterosis

Crosses often show **high-parent heterosis**, wherein the  $F_1$  not only beats the average of the two parents (**mid-parent heterosis**), it exceeds the best parent.

Crop	% planted as hybrids	% yield advantage	Annual added yield: %	Annual added yield: tons	Annual land savings
Maize	65	15	10	55 x 10 <sup>6</sup>	13 x 10 <sup>6</sup> ha
Sorghum	48	40	19	13 x 10 <sup>6</sup>	9 x 10 <sup>6</sup> ha
Sunflower	60	50	30	7 x 10 <sup>6</sup>	6 x 10 <sup>6</sup> ha
Rice	12	30	4	15 x 10 <sup>6</sup>	6 x 10 <sup>6</sup> ha

19

## Hybrid Corn in the US

Shull (1908) suggested objective of corn breeders should be to find and maintain the best parental lines for crosses

Initial problem: early inbred lines had low seed set

Solution (Jones 1918): use a hybrid line as the seed parent, as it should show heterosis for seed set

1930's - 1960's: most corn produced by double crosses

Since 1970's most from single crosses

20



# A Cautionary Tale

1970-1971 the great Southern Corn Leaf Blight almost destroyed the whole US corn crop

Much larger (in terms of food energy) than the great potato blight of the 1840's

Cause: Corn can self-fertilize, so to make hybrids either have to manually detassel the pollen structures or use genetic tricks that cause male sterility.

Almost 85% of US corn in 1970 had Texas cytoplasm Tcms, a mtDNA encoded male sterility gene

Tcms turned out to be hyper-sensitive to the fungus *Helminthosporium maydis*. Resulted in over a billion dollars of crop loss

21

## Crossing Schemes to Reduce the Loss of Heterosis: Synthetics

Take  $n$  lines and construct an  $F_1$  population by making all pairwise crosses

Allow random mating from the  $F_2$  on to produce a synthetic population

$$F_2 = F_1 - \frac{F_1 - \bar{P}}{n} \quad H/n$$

$$H_{F_2} = H_{F_1} \left( 1 - \frac{1}{n} \right) \quad \text{Only } 1/n \text{ of heterosis lost vs. } 1/2$$

22

# Synthetics

- Major trade-off
  - As more lines are added, the  $F_2$  loss of heterosis declines
  - However, as more lines are added, the mean of the  $F_1$  also declines, as less elite lines are used
  - Bottom line: For some value of  $n$ ,  $F_1 - H/n$  reaches a maximum value and then starts to decline with  $n$

23

## Types of crosses

- The  $F_1$  from a cross of lines  $A \times B$  (typically inbreds) is called a **single cross**
- A **three-way cross** (also called a **modified single cross**) refers to the offspring of an  $A$  individual crossed to the  $F_1$  offspring of  $B \times C$ .
  - Denoted  $A \times (B \times C)$
- A **double** (or **four-way**) **cross** is  $(A \times B) \times (C \times D)$ , the offspring from crossing an  $A \times B$   $F_1$  with a  $C \times D$   $F_1$ .

24

# Predicting cross performance

- While single cross (offspring of A x B) hard to predict, three- and four-way crosses can be predicted if we know the means for single crosses involving these parents
- The three-way cross mean is the average mean of the two single crosses:
  - $\text{mean}(A \times \{B \times C\}) = [\text{mean}(A \times B) + \text{mean}(A \times C)]/2$
- The mean of a double (or four-way) cross is the average of all the single crosses,
  - $\text{mean}(\{A \times B\} \times \{C \times D\}) = [\text{mean}(A \times C) + \text{mean}(A \times D) + \text{mean}(B \times C) + \text{mean}(B \times D)]/4$

25

## Individual vs. Maternal Heterosis

- Individual heterosis
  - enhanced performance in a hybrid individual
- Maternal heterosis
  - enhanced maternal performance (such as increased litter size and higher survival rates of offspring)
  - Use of crossbred dams
  - Maternal heterosis is often comparable, and can be greater than, individual heterosis

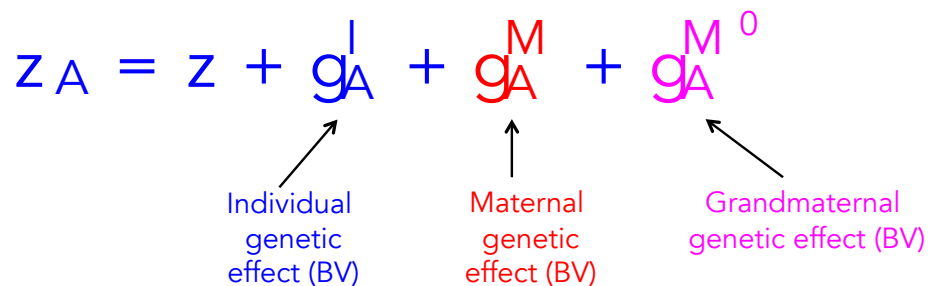
## Individual vs. Maternal Heterosis in Sheep traits

Trait	Individual H	Maternal H	total
Birth weight	3.2%	5.1%	8.3%
Weaning weight	5.0%	6.3%	11.3%
Birth-weaning survival	9.8%	2.7%	12.5%
Lambs reared per ewe	15.2%	14.7%	29.9%
Total weight lambs/ewe	17.8%	18.0%	35.8%
Prolificacy	2.5%	3.2%	5.7%

## Estimating the Amount of Heterosis in Maternal Effects

Contributions to mean value of line A

$$Z_A = Z + g_A^I + g_A^M + g_A^{M^0}$$

  
Individual genetic effect (BV)      Maternal genetic effect (BV)      Grandmaternal genetic effect (BV)

Consider the offspring of an A sire and a B dam

Individual genetic value is the average of both parental lines

Contribution from (individual) heterosis

$$Z_{AB} = Z + \frac{g_A^I + g_B^I}{2} + g_B^M + g_B^{M^0} + h_{AB}^I$$

Maternal and grandmaternal effects from the B mothers

$$Z_{AB} = Z + \frac{g_A^I + g_B^I}{2} + g_B^M + g_B^{M^0} + h_{AB}^I$$

Now consider the offspring of an B sire and a A dam

$$Z_{BA} = Z + \frac{g_A^I + g_B^I}{2} + g_A^M + g_A^{M^0} + h_{AB}^I$$

Maternal and grandmaternal genetic effects for B line

Difference between the two line means estimates difference in maternal + grandmaternal effects in A vs. B

Hence, an estimate of individual heterotic effects is

$$\frac{Z_{AB} + Z_{BA}}{2} - \frac{Z_{AA} + Z_{BB}}{2} = h_{AB}^I$$

The mean of offspring from a sire in line C crossed to a dam from a A X B cross (B = granddam, AB = dam)

Average individual genetic value  
(average of the line BV's)

Genetic maternal effect  
(average of maternal BV for both lines)

Grandmaternal genetic effect

New individual heterosis of C x AB cross

Maternal genetic heterotic effect

"Recombinational loss" --- decay of the F<sub>1</sub> heterosis in the F<sub>2</sub>

$$Z_{CAB} = \frac{2g_C^I + g_A^I + g_B^I}{4} + \frac{h_{CA}^I + h_{CB}^I}{2} + \frac{g_A^M + g_B^M}{2} + h_{AB}^M + g_B^{M^0} + \frac{r_{ab}^I}{2}$$

One estimate (confounded) of maternal heterosis

$$Z_{CAB} = \frac{Z_{CA} + Z_{CB}}{2} = h_{AB}^M + \frac{r_{ab}^I}{2}$$

# Lecture 6: Selection on Multiple Traits

Bruce Walsh lecture notes  
Introduction to Quantitative Genetics  
SISG, Seattle  
17 – 19 July 2017

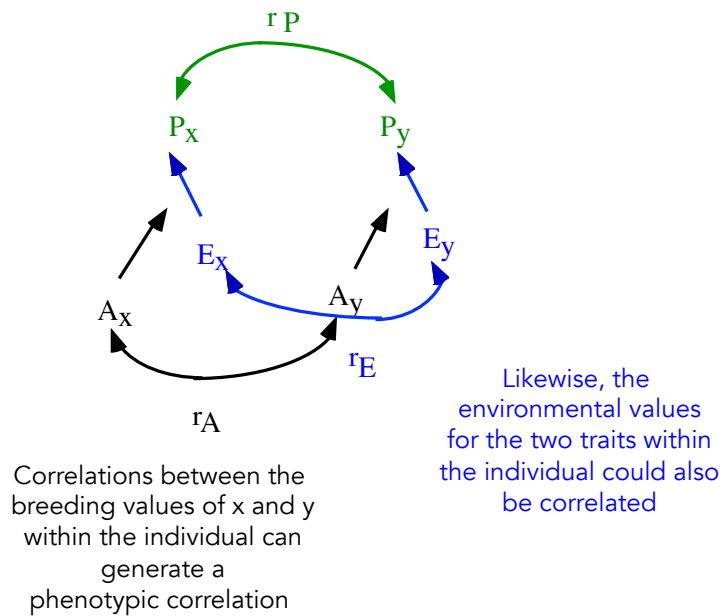
1

## Genetic vs. Phenotypic correlations

- Within an individual, trait values can be positively or negatively correlated,
  - height and weight -- positively correlated
  - Weight and lifespan -- negatively correlated
- Such phenotypic correlations can be directly measured,
  - $r_p$  denotes the phenotypic correlation
- Phenotypic correlations arise because genetic and/or environmental values within an individual are correlated.

2

The phenotypic values between traits x and y within an individual are correlated



3

## Genetic & Environmental Correlations

- $r_A$  = correlation in breeding values (the **genetic correlation**) can arise from
  - pleiotropic effects of loci on both traits
  - linkage disequilibrium, which decays over time
- $r_E$  = correlation in environmental values
  - includes non-additive genetic effects (e.g., D, I)
  - arises from exposure of the two traits to the same individual environment

4



The relative contributions of genetic and environmental correlations to the phenotypic correlation

$$r_P = r_A h_X h_Y + r_E \sqrt{(1 - h_x^2)(1 - h_Y^2)}$$

If heritability values are high for both traits, then the correlation in breeding values dominates the phenotypic correlation

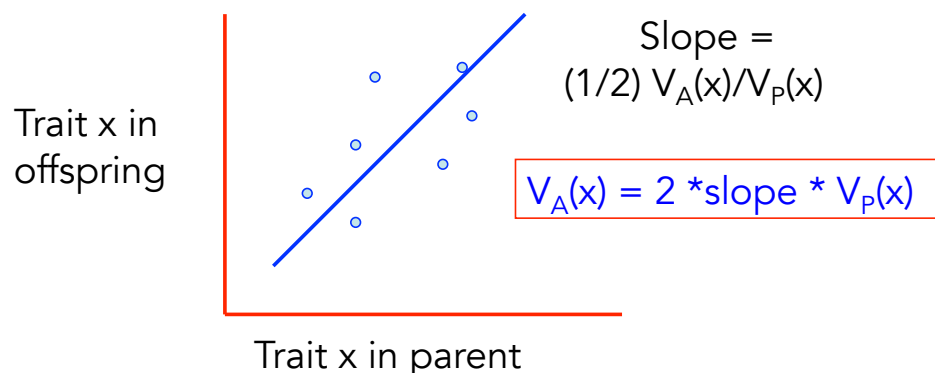
If heritability values in EITHER trait are low, then the correlation in environmental values dominates the phenotypic correlation

In practice, phenotypic and genetic correlations often have the same sign and are of similar magnitude, but this is not always the case

5

## Estimating Genetic Correlations

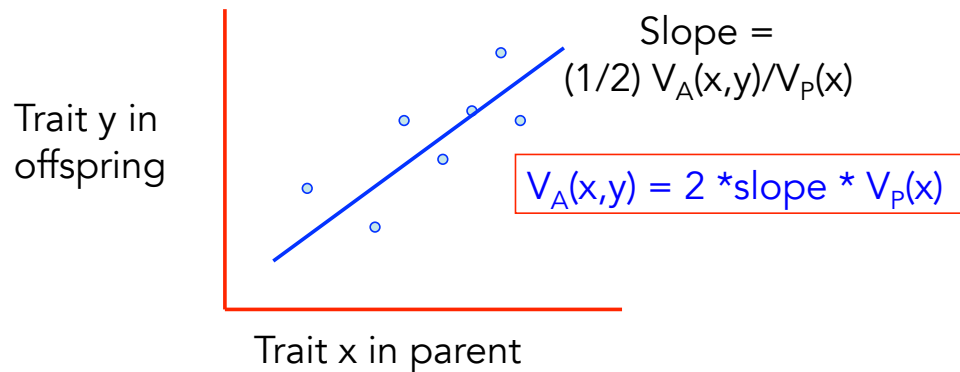
Recall that we estimated  $V_A$  from the regression of trait x in the parent on trait x in the offspring,



6

# Estimating Genetic Correlations

Similarly, we can estimate  $V_A(x,y)$ , the covariance in the breeding values for traits x and y, by the regression of trait x in the parent and trait y in the offspring



7

Thus, one estimator of  $V_A(x,y)$  is

$$V_A(x,y) = \frac{2 * b_{y|x} * V_P(x) + 2 * b_{x|y} * V_P(y)}{2}$$

giving

$$V_A(x,y) = b_{y|x} V_P(x) + b_{x|y} V_P(y)$$

Put another way,

$$\begin{aligned} \text{Cov}(x_O, y_P) &= \text{Cov}(y_O, x_P) = (1/2) \text{Cov}(A_x, A_y) \\ \text{Cov}(x_O, x_P) &= (1/2) V_A(x) = (1/2) \text{Cov}(A_x, A_x) \\ \text{Cov}(y_O, y_P) &= (1/2) V_A(y) = (1/2) \text{Cov}(A_y, A_y) \end{aligned}$$

Likewise, for half-sibs,

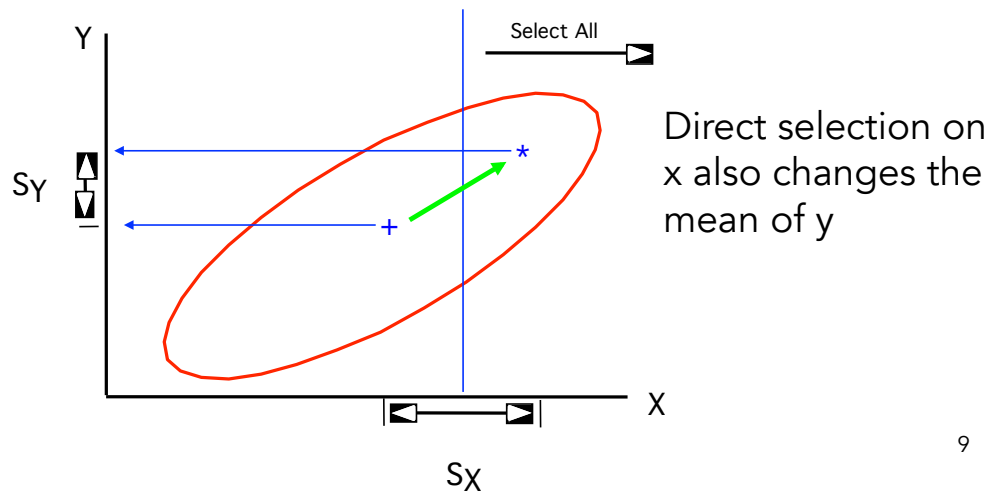
$$\begin{aligned} \text{Cov}(x_{HS}, y_{HS}) &= (1/4) \text{Cov}(A_x, A_y) \\ \text{Cov}(x_{HS}, x_{HS}) &= (1/4) \text{Cov}(A_x, A_x) = (1/4) V_A(x) \\ \text{Cov}(y_{HS}, y_{HS}) &= (1/4) \text{Cov}(A_y, A_y) = (1/4) V_A(y) \end{aligned}$$

General:  $\text{Cov}(x_i, y_j) = 2\theta_{ij} \text{Cov}(A_i, A_j)$

8

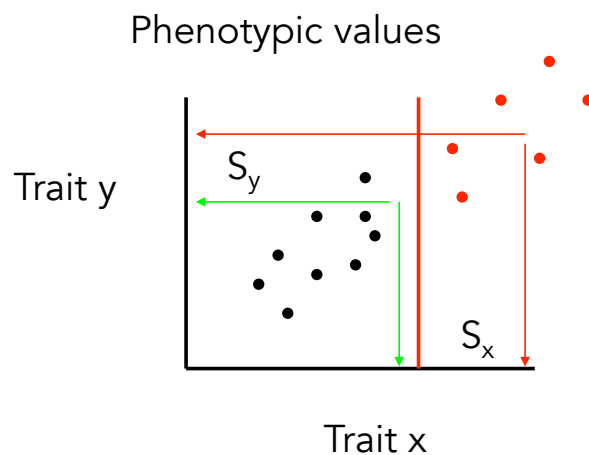
# Correlated Response to Selection

Direct selection of a character can cause a within-generation change in the mean of a phenotypically correlated character.



9

Phenotypic correlations induce **within-generation changes**



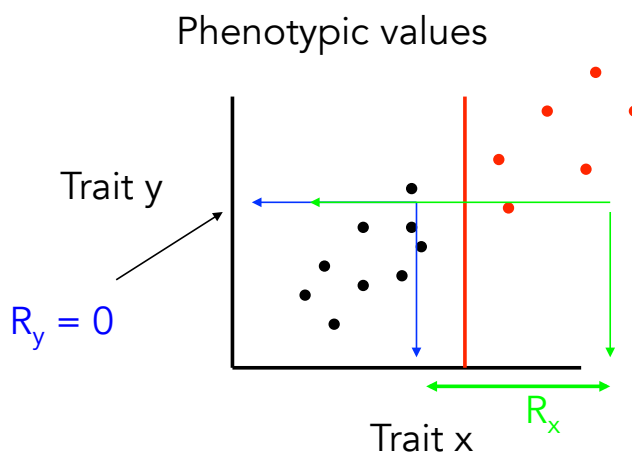
For there to be a **between-generation change**, the **breeding values must be correlated**. Such a change is called a **correlated response to selection**

10

# Example

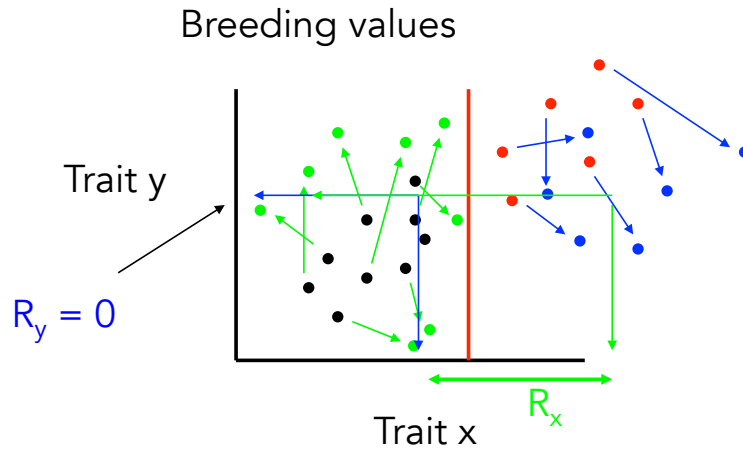
- Suppose  $h^2$  trait x = 0.5,  $h^2$  trait y = 0.3
- Select on trait one to give  $S_x = 10$ 
  - Expected response is  $R_x = 5$
- Suppose  $\text{Cov}(t_x, t_y) = 0.5$ , then  $S_y = 5$
- What is the response in trait 2?
  - is it  $CR_y = 0.3 \cdot 5 = 1.5$ . NO!
  - Could be positive, negative, or zero
  - Depends on the Genetic correlation between traits x and y. Why??

11

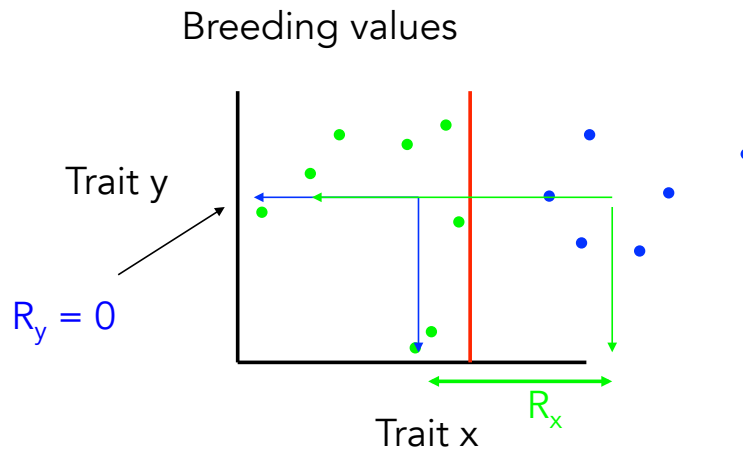


12

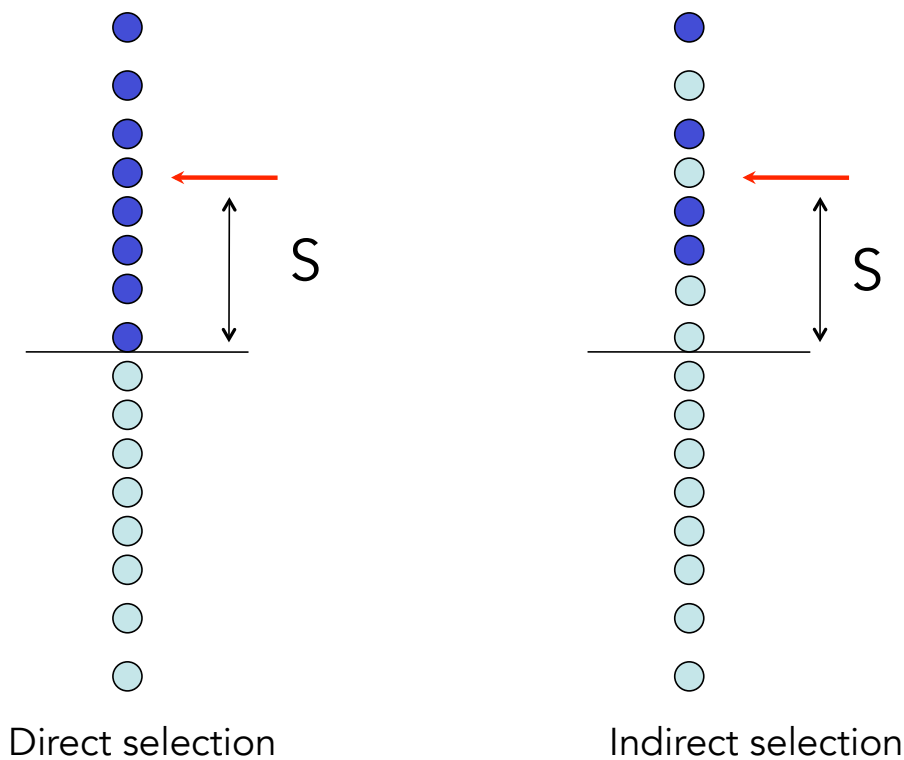
Phenotypic values are misleading, what we want are the breeding values for each of the selected individuals. Each arrow takes an individual's phenotypic value into its actual breeding value.



13



14



15

## Predicting the correlated response

The change in character  $y$  in response to selection on  $x$  is the regression of the breeding value of  $y$  on the breeding value of  $x$ ,

$$A_y = b_{A_y|A_x} A_x$$

where

$$b_{A_y|A_x} = \frac{\text{Cov}(A_x, A_y)}{\text{Var}(A_x)} = r_A \frac{\sigma(A_y)}{\sigma(A_x)}$$

If  $R_x$  denotes the direct response to selection on  $x$ ,  $CR_y$  denotes the correlated response in  $y$ , with

$$CR_y = b_{A_y|A_x} R_x$$

16

We can rewrite  $CR_y = b_{A_y|A_x} R_x$  as follows

First, note that  $R_x = h_x^2 S_x = i_x h_x \sigma_A(x)$

↑  
Recall that  $i_x = S_x / \sigma_P$   
(x) is the selection  
intensity on x

Since  $b_{A_y|A_x} = r_A \sigma_A(x) / \sigma_A(y)$ ,

We have  $CR_y = b_{A_y|A_x} R_x = r_A \sigma_A(y) h_x i_x$

Substituting  $\sigma_A(y) = h_y \sigma_P(y)$  gives our final result:

$$CR_y = i_x h_x h_y r_A \sigma_P(y)$$

17

$$CR_y = i_x h_x h_y r_A \sigma_P(y)$$

Noting that we can also express the direct response as  
 $R_x = i_x h_x^2 \sigma_P(x)$

shows that  $h_x h_y r_A$  in the corrected response plays the same role as  $h_x^2$  does in the direct response. As a result,  $h_x h_y r_A$  is often called the **co-heritability**

18

## Direct vs. Indirect Response

We can change the mean of x via a direct response  $R_x$  or an indirect response  $CR_x$  due to selection on y

$$\frac{CR_x}{R_x} = \frac{i_Y r_A \sigma_{AX} h_Y}{i_X h_X \sigma_{AX}} = \frac{i_Y r_A h_Y}{i_X h_X}$$

Hence, indirect selection gives a large response when

$$i_Y r_A h_Y > i_X h_X$$

- The selection intensity is much greater for y than x. This would be true if y were measurable in both sexes but x measurable in only one sex.
- Character y has a greater heritability than x, and the genetic correlation between x and y is high. This could occur if x is difficult to measure with precision but y is not.

19

## G x E

The same trait measured over two (or more) environments can be considered as two (or more) correlated traits.

If the genetic correlation  $|\rho| = 1$  across environments and the genetic variance of the trait is the same in both environments, then no G x E

However, if  $|\rho| < 1$ , and/or  $\text{Var}(A)$  of the trait varies over environments, then G x E present

Hence, dealing with G x E is a *multiple-trait problem*



# Participatory breeding

The environment where a crop line is developed may be different from where it is grown

An especially important example of this is **participatory breeding**, wherein subsistence farmers are involved in the field traits.

Here, the correlated response is the yield in subsistence environment given selection at a regional center, while direct response is yield when selection occurred in subsistence environment. Regional center selection works when

$$i_Y r_A h_Y > i_X h_X$$

21

## Matrices

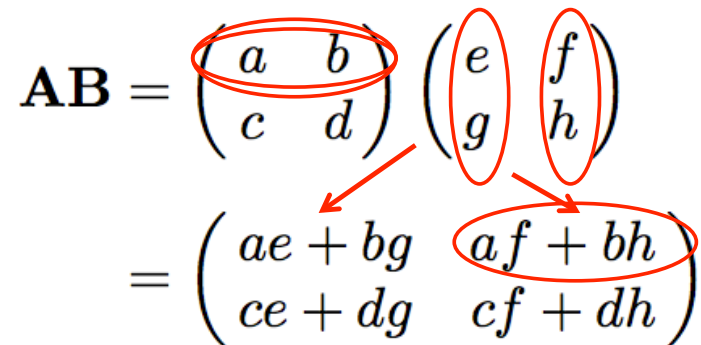
$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} i \\ j \end{pmatrix}$$

Dimensions given by rows x columns (r x c)

The identity matrix  $\mathbf{I}$ ,  $\mathbf{I}_{2 \times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

22

## Matrix Multiplication

$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix}$$

$$= \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}$$

In order to multiply two matrices, they must conform

$$A_{r \times c} B_{c \times k} = C_{r \times k}$$

23

## Matrix Multiplication

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} i \\ j \end{pmatrix}$$

$$\mathbf{BA} = \begin{pmatrix} ae + cf & eb + df \\ ga + ch & gd + dh \end{pmatrix} \quad \mathbf{AC} = \begin{pmatrix} ai + bj \\ ci + dj \end{pmatrix}$$

The **identity matrix I** serves the role of one in matrix multiplication:  $\mathbf{AI} = \mathbf{A}$ ,  $\mathbf{IA} = \mathbf{A}$


24

# The Inverse Matrix, $A^{-1}$

For a square matrix  $A$ , define the **Inverse** of  $A$ ,  $A^{-1}$ , as the matrix satisfying

$$A^{-1}A = AA^{-1} = I$$

For  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$   $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$



If this quantity (the determinant) is zero, the inverse does not exist.

The inverse serves the role of division in matrix multiplication

Suppose we are trying to solve the system  $Ax = c$  for  $x$ .

$$A^{-1}Ax = A^{-1}c. \text{ Note that } A^{-1}Ax = Ix = x, \text{ giving } x = A^{-1}c$$

# The Multivariate Breeders' Equation

Suppose we are interested in the vector  $\mathbf{R}$  of responses when selection occurs on  $n$  correlated traits

Let  $\mathbf{S}$  be the vector of selection differentials.

In the univariate case, the relationship between  $\mathbf{R}$  and  $\mathbf{S}$  was the Breeders' Equation,  $\mathbf{R} = h^2 \mathbf{S}$

What is the multivariate version of this?

27

$$\mathbf{S} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} \sigma^2(z_2) & \sigma(z_1, z_2) \\ \sigma(z_1, z_2) & \sigma^2(z_2) \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} \sigma^2(A_2) & \sigma(A_1, A_2) \\ \sigma(A_1, A_2) & \sigma^2(A_2) \end{pmatrix}$$

28

## The multivariate breeder's equation

$$R = G P^{-1} S$$

$$R = h^2 S = (V_A / V_P) S$$

Natural parallels with univariate breeder's equation

$P^{-1} S = \beta$  is called the **selection gradient** and measures the amount of direct selection on a character

The gradient version of the breeder's equation is given by  $R = G \beta$ . This is often called the Lande Equation (after Russ Lande)

29

## Sources of within-generation change in the mean

Since  $\beta = P^{-1} S$ ,  $S = P \beta$ ,  
giving the j-th element as

$$S_j = \underbrace{\sigma^2(P_j)}_{\text{Change in mean from direct selection on trait j}} \beta_j + \underbrace{\sum_{i \neq j} \sigma(P_j, P_i) \beta_i}_{\text{Change in mean from phenotypically correlated characters under direct selection}}$$

30

Within-generation change in the mean

$$S_j = \sigma^2(P_j) \beta_j + \sum_{i \neq j} \sigma(P_j, P_i) \beta_i$$

Response in the mean

Between-generation  
change (response)  
in trait j

Indirect response  
from genetically  
correlated  
characters under  
direct selection

$$R_j = \sigma^2(A_j) \beta_j + \sum_{i \neq j} \sigma(A_j, A_i) \beta_i$$

Response from direct  
selection on trait j

Correlated response

Direct response

31

## Example in R

Consider three of these traits,  $z_1$  = oil content,  $z_2$  = protein content, and  $z_3$  = yield. For these characters, Brim et al. estimated the covariance matrices as

$$\mathbf{P} = \begin{pmatrix} 287.5 & 477.4 & 1266 \\ 477.4 & 935 & 2303 \\ 1266 & 2303 & 5951 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 128.7 & 160.6 & 492.5 \\ 160.6 & 254.6 & 707.7 \\ 492.5 & 707.7 & 2103 \end{pmatrix}$$

Suppose you observed a within-generation change of -10 for oil, 10 for protein, and 100 for yield.

What is R? What is the nature of selection on each trait?

Enter G, P, and S

```
> P<-matrix(c(287.5,477.4,1266,477.4,935,2303,1266,2303,5951),nrow=3)
> P
      [,1] [,2] [,3]
[1,] 287.5 477.4 1266
[2,] 477.4 935.0 2303
[3,] 1266.0 2303.0 5951
> G<-matrix(c(128.7,160.6,492.5,160.6,254.6,707.7,492.5,707.7,2103),nrow=3)
> G
      [,1] [,2] [,3]
[1,] 128.7 160.6 492.5
[2,] 160.6 254.6 707.7
[3,] 492.5 707.7 2103.0
> S<-matrix(c(-10,10,100),nrow=3)
> S
      [,1]
[1,] -10
[2,] 10
[3,] 100
```

$$R = G P^{-1} S$$

```
> G %%% solve(P) %%% S
      [,1]
[1,] -13.57729
[2,] 12.28425
[3,] 65.14172
```

13.6 decrease in oil  
12.3 increase in protein  
65.1 increase in yield

33

S versus  $\beta$  : Observed change versus targets of Selection,  $\beta = P^{-1} S$ ,  $S = P \beta$ ,

$$S_j = \sigma^2(P_j) \beta_j + \sum_{i \neq j} \sigma(P_j, P_i) \beta_i$$

```
> solve(P) %%% S
      [,1]
[1,] -2.708160
[2,] -1.431750
[3,] 1.147009
```

←→

```
> S
      [,1]
[1,] -10
[2,] 10
[3,] 100
```

$\beta$ : targets of selection

S: observed within-generation change

Observe a within-generation increase in protein, but the actual selection was to *decrease* it.

34

## Quantifying Multivariate Constraints to Response

Is there genetic variation in the direction of selection?

Consider the following  $\mathbf{G}$  and  $\boldsymbol{\beta}$ :

$$\mathbf{G} = \begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

Taken one trait at a time, we might expect  $R_i = G_{ii}\beta_i$

Giving  $R_1 = 20$ ,  $R_2 = -40$ .

What is the actual response?

$$\mathbf{R} = \mathbf{G}\boldsymbol{\beta} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

35

## Constraints Imposed by Genetic Correlations

While  $\boldsymbol{\beta}$  is the directional optimally favored by selection, the actual response is dragged off this direction, with  $\mathbf{R} = \mathbf{G}\boldsymbol{\beta}$ .

**Example: Suppose**

$$\mathbf{S} = \begin{pmatrix} 10 \\ -10 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} 20 & -10 \\ -10 & 40 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 20 & 5 \\ 5 & 10 \end{pmatrix}$$

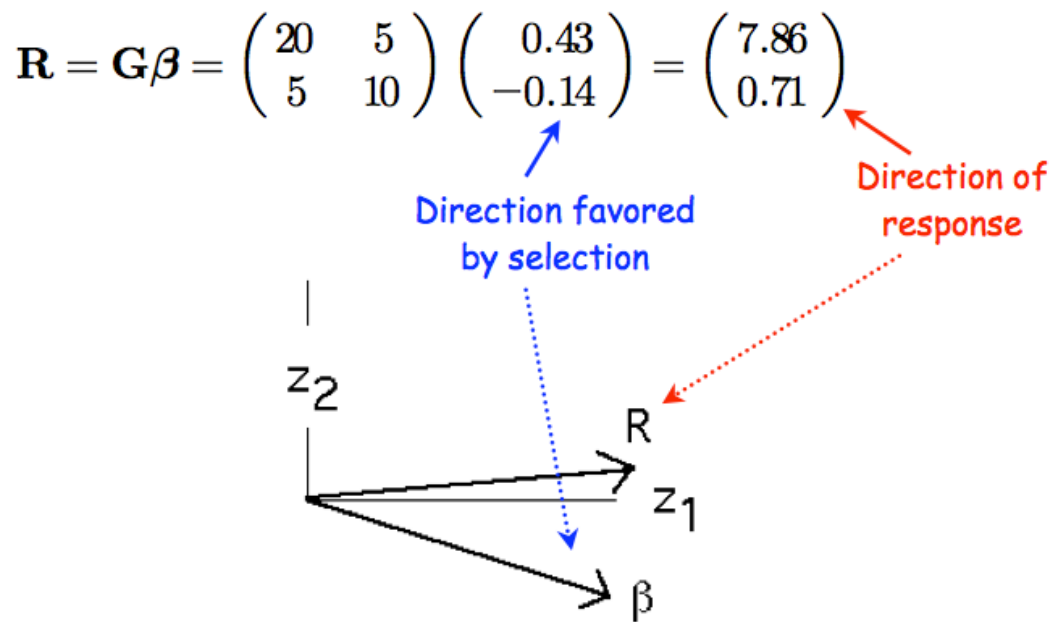
What is the true nature of selection on the two traits?

$$\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S} = \mathbf{P}^{-1} \begin{pmatrix} 10 \\ -10 \end{pmatrix} = \begin{pmatrix} 0.43 \\ -0.14 \end{pmatrix}$$

36



What does the actual response look like?



37

## Time for a short diversion: The Geometry of a matrix

A vector is a geometric object, leading from the origin to a specific point in  $n$ -space.

Hence, a vector has a length and a direction.

We can thus change a vector by both rotation and scaling

The length (or norm) of a vector  $\mathbf{x}$  is denoted by  $\|\mathbf{x}\|$

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T \mathbf{x}}$$

38

The (Euclidean) distance between two vectors  $\mathbf{x}$  and  $\mathbf{y}$  (of the same dimension) is

$$\|\mathbf{x}-\mathbf{y}\|^2 = \sum_{i=1}^n (x_i - y_i)^2 = (\mathbf{x}-\mathbf{y})^T (\mathbf{x}-\mathbf{y}) = (\mathbf{y}-\mathbf{x})^T (\mathbf{y}-\mathbf{x})$$

The angle  $\theta$  between two vectors provides a measure for how they differ.

If two vectors satisfy  $\mathbf{x} = a\mathbf{y}$  (for a constant  $a$ ), then they point in the same direction, i.e.,  $\theta = 0$  (Note that  $a < 0$  simply reflects the vector about the origin)

Vectors at right angles to each other,  $\theta = 90^\circ$  or  $270^\circ$  are said to be orthogonal. If they have unit length as well, they are further said to be orthonormal.

39

## Matrices Describe Vector transformations

Matrix multiplication results in a **rotation** and a **scaling** of a vector

The action of multiplying a vector  $\mathbf{x}$  by a matrix  $\mathbf{A}$  generates a new vector  $\mathbf{y} = \mathbf{A}\mathbf{x}$ , that has different dimension from  $\mathbf{x}$  unless  $\mathbf{A}$  is square.

Thus  $\mathbf{A}$  describes a **transformation** of the original coordinate system of  $\mathbf{x}$  into a new coordinate system.

Example: Consider the following  $\mathbf{G}$  and  $\boldsymbol{\beta}$ :

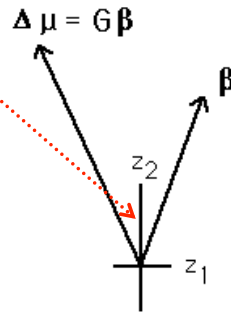
$$\mathbf{G} = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \quad \mathbf{R} = \mathbf{G}\boldsymbol{\beta} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$$

40

The resulting angle between  $\mathbf{R}$  and  $\beta$  is given by

$$\cos \theta = \frac{\beta^T \mathbf{R}}{|\mathbf{R}| |\beta|} = \frac{1}{\sqrt{2}}$$

For an angle of  $\theta = 45^\circ$



41

## Eigenvalues and Eigenvectors

The **eigenvalues** and their associated **eigenvectors** fully describe the geometry of a matrix.

Eigenvalues describe how the original coordinate axes are **scaled** in the new coordinate systems

Eigenvectors describe how the original coordinate axes are **rotated** in the new coordinate systems

For a square matrix  $A$ , any vector  $y$  that satisfies  $Ay = \lambda y$  for some scalar  $\lambda$  is said to be an **eigenvector** of  $A$  and  $\lambda$  its associated **eigenvalue**.

42

Note that if  $y$  is an eigenvector, then so is  $a*y$  for any scalar  $a$ , as  $Ay = \lambda y$ .

Because of this, we typically take eigenvectors to be scaled to have unit length (their norm = 1)

An **eigenvalue**  $\lambda$  of  $A$  satisfies the equation  $\det(A - \lambda I) = 0$ , where  $\det$  = determinant

For an  $n$ -dimensional square matrix, this yields an  $n$ -degree polynomial in  $\lambda$  and hence up to  $n$  unique roots.

Two nice features:

$\det(A) = \prod_i \lambda_i$  The determinant is the product of the eigenvalues

$\text{trace}(A) = \sum_i \lambda_i$ . The **trace** (sum of the diagonal elements) is the sum of the eigenvalues

43

Note that  $\det(A) = 0$  if and only if at least one eigenvalue = 0

For symmetric matrices (such as covariance matrices) the resulting  $n$  eigenvectors are mutually orthogonal, and we can factor  $A$  into its **spectral decomposition**,

$$A = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T + \cdots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T$$

Hence, we can write the product of any vector  $x$  and  $A$  as

$$\begin{aligned} Ax &= \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T x + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T x + \cdots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T x \\ &= \lambda_1 \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_1) + \lambda_2 \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_2) + \cdots + \lambda_n \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_n) \end{aligned}$$

44

Example: Let's reconsider a previous G matrix

$$|\mathbf{G} - \lambda \mathbf{I}| = \left| \begin{pmatrix} 4 - \lambda & -2 \\ -2 & 2 - \lambda \end{pmatrix} \right|$$

$$= (4 - \lambda)(2 - \lambda) - (-2)^2 = \lambda^2 - 6\lambda + 4 = 0$$

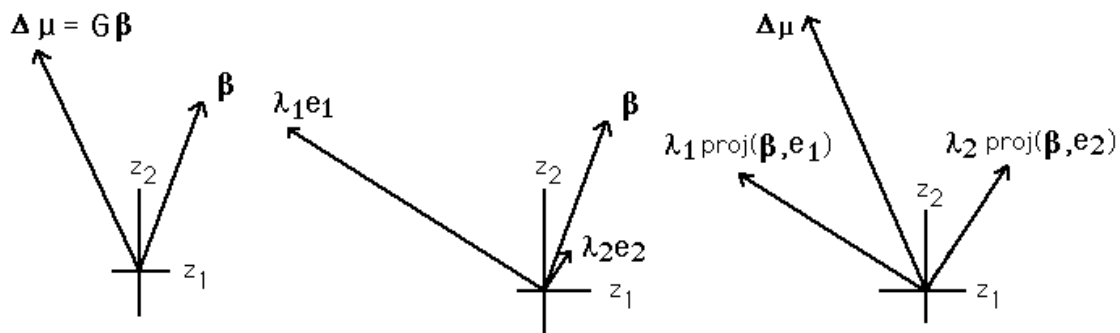
The solutions are

$$\lambda_1 = 3 + \sqrt{5} \simeq 5.236 \quad \lambda_2 = 3 - \sqrt{5} \simeq 0.764$$

The corresponding eigenvectors become

$$\mathbf{e}_1 \simeq \begin{pmatrix} -0.851 \\ 0.526 \end{pmatrix} \quad \mathbf{e}_2 \simeq \begin{pmatrix} 0.526 \\ 0.851 \end{pmatrix}$$

45



Even though  $\beta$  points in a direction very close of  $\mathbf{e}_2$ , because most of the variation is accounted for by  $\mathbf{e}_1$ , **its projection is this dimension yields a much longer vector**. The sum of these two projections yields the selection response  $R$ .

46

## Realized Selection Gradients

Suppose we observe a difference in the vector of means for two populations,  $\mathbf{R} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ .

*If* we are willing to assume they both have a common  $\mathbf{G}$  matrix that has remained constant over time, then we can estimate the nature and amount of selection generating this difference by

$$\boldsymbol{\beta} = \mathbf{G}^{-1} \mathbf{R}$$

Example: You are looking at oil content ( $z_1$ ) and yield ( $z_2$ ) in two populations of soybeans. Population a has  $\mu_1 = 20$  and  $\mu_2 = 30$ , while for Pop 2,  $\mu_1 = 10$  and  $\mu_2 = 35$ .

47

Here

$$\mathbf{R} = \begin{pmatrix} 20 - 10 \\ 30 - 35 \end{pmatrix} = \begin{pmatrix} 10 \\ -5 \end{pmatrix}$$

Suppose the variance-covariance matrix has been stable and equal in both populations, with

$$\mathbf{G} = \begin{pmatrix} 20 & -10 \\ -10 & 40 \end{pmatrix}$$

The amount of selection on both traits to obtain this response is

$$\boldsymbol{\beta} = \begin{pmatrix} 20 & -10 \\ -10 & 40 \end{pmatrix}^{-1} \begin{pmatrix} 10 \\ -5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$$

48

# Lecture 7

## Estimation of Basic Genetic Parameters

Guilherme J. M. Rosa

University of Wisconsin-Madison

Introduction to Quantitative Genetics  
SISG, Seattle  
17 - 19 July 2017

## Heritability

### Narrow vs. broad sense

Narrow sense:  $h^2 = V_A/V_P$

Slope of midparent - offspring regression  
(sexual reproduction)

Broad sense:  $H^2 = V_G/V_P$

Slope of a parent - cloned offspring regression  
(asexual reproduction)

When one refers to heritability, the default is narrow-sense,  $h^2$

$h^2$  is the measure of (easily) usable genetic variation under sexual reproduction

## Why $h^2$ instead of $h$ ?

Blame Sewall Wright, who used  $h$  to denote the correlation between phenotype and breeding value. Hence,  $h^2$  is the total fraction of phenotypic variance due to breeding values

$$r(A, P) = \frac{\sigma(A, P)}{\sigma_A \sigma_P} = \frac{\sigma_A^2}{\sigma_A \sigma_P} = \frac{\sigma_A}{\sigma_P} = h$$

## Heritabilities are functions of populations

Heritability values only make sense in the context of the population for which it was measured

Heritability measures the *standing genetic variation* of a population

A zero heritability DOES NOT imply that the trait is not genetically determined

Heritabilities are functions of the distribution of environmental values (i.e., the *universe* of E values)

Decreasing  $V_p$  increases  $h^2$ .

Heritability values measured in one environment (or distribution of environments) may not be valid under another

Measures of heritability for lab-reared individuals may be very different from heritability in nature



## Heritability and the Prediction of Breeding Values

If  $P$  denotes an individual's phenotype, then best linear predictor of their breeding value  $A$  is

$$A = \frac{\sigma(P, A)}{\sigma_P^2} (P - \mu_P) + e = h^2 (P - \mu_P) + e$$

The residual variance is also a function of  $h^2$ :

$$\sigma_e^2 = (1 - h^2) \sigma_P^2$$

The larger the heritability, the tighter the distribution of true breeding values around the value  $h^2(P - \mu_P)$  predicted by an individual's phenotype.

## Heritability and Population Divergence


*Heritability is a **completely unreliable predictor** of long-term response*

Measuring heritability values in two populations that show a difference in their means provides no information on whether the underlying difference is genetic

## Sample Heritabilities

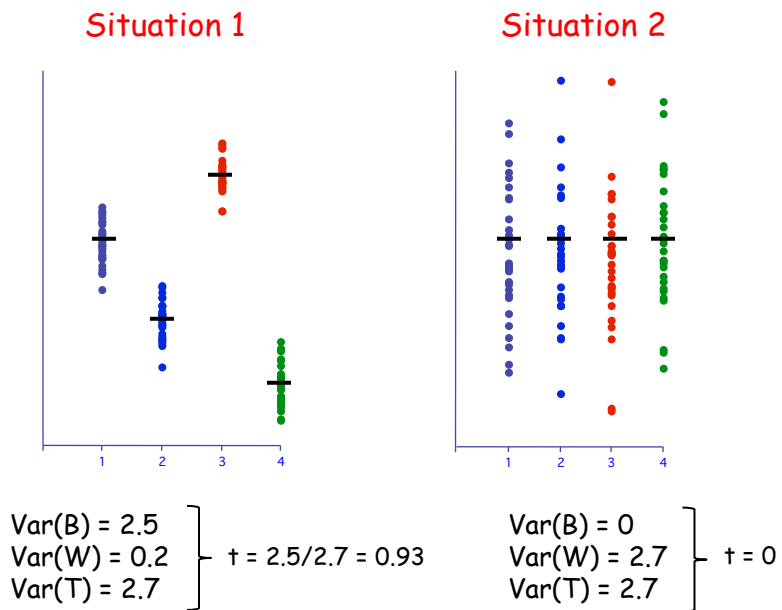
		$h^2$
People	Height	0.80
	Serum IG	0.45
Pigs	Back-fat	0.70
	Weight gain	0.30
	Litter size	0.05
Fruit Flies	Abdominal Bristles	0.50
	Body size	0.40
	Ovary size	0.30
	Egg production	0.20

Traits more closely associated with fitness tend to have lower heritabilities



## ANOVA: Analysis of Variance

- Partitioning of trait variance into within- and among-group components
- Two key ANOVA identities
  - Total variance = between-group variance + within-group variance
    - $\text{Var}(T) = \text{Var}(B) + \text{Var}(W)$
  - Variance(between groups) = covariance (within groups)
  - Intraclass correlation,  $t = \text{Var}(B)/\text{Var}(T)$
- The more similar individuals are within a group (higher within-group covariance), the larger their between-group differences (variance in the group means)



## Phenotypic Resemblance Between Relatives

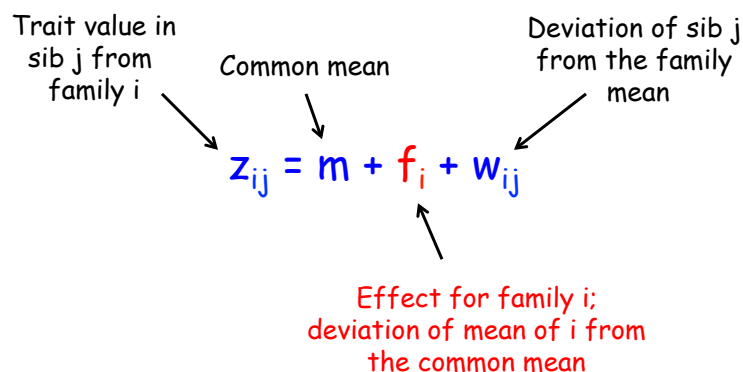
Relatives	Covariance	Regression (b) or correlation (t)
Offspring and one parent	$\frac{1}{2}V_A$	$b = \frac{1}{2} \frac{V_A}{V_P}$
Offspring and mid-parent	$\frac{1}{2}V_A$	$b = \frac{V_A}{V_P}$
Half sibs	$\frac{1}{4}V_A$	$t = \frac{1}{4} \frac{V_A}{V_P}$
Full sibs	$\frac{1}{2}V_A + \frac{1}{4}V_D + V_{E_c}$	$t = \frac{\frac{1}{2}V_A + \frac{1}{4}V_D + V_{E_c}}{V_P}$

## Why $\text{cov}(\text{within}) = \text{variance}(\text{among})$ ?

- Let  $z_{ij}$  denote the  $j$ th member of group  $i$ .
  - Here  $z_{ij} = u + g_i + e_{ij}$
  - $g_i$  is the group effect
  - $e_{ij}$  the residual error
- Covariance within a group  $\text{Cov}(z_{ij}, z_{ik})$ 
  - $= \text{Cov}(u + g_i + e_{ij}, u + g_i + e_{ik})$
  - $= \text{Cov}(g_i, g_i)$  as all other terms are uncorrelated
  - $\text{Cov}(g_i, g_i) = \text{Var}(g)$  is the among-group variance

## Estimation: One-way ANOVA

Simple (balanced) full-sib design:  $N$  full-sib families, each with  $n$  offspring: One-way ANOVA model



Covariance between members of the same group  
equals the variance among (between) groups

$$\begin{aligned}
 \text{Cov}(\text{Full Sibs}) &= \sigma(z_{ij}, z_{ik}) \\
 &= \sigma[(\mu + f_i + w_{ij}), (\mu + f_i + w_{ik})] \\
 &= \sigma(f_i, f_i) + \sigma(f_i, w_{ik}) + \sigma(w_{ij}, f_i) + \sigma(w_{ij}, w_{ik}) \\
 &= \sigma_f^2
 \end{aligned}$$

Hence, the variance among family effects equals the  
covariance between full sibs

$$\sigma_f^2 = \sigma_A^2 / 2 + \sigma_D^2 / 4 + \sigma_{Ec}^2$$

The within-family variance  $\sigma_w^2 = \sigma_p^2 - \sigma_f^2$ ,

$$\begin{aligned}
 \sigma_{w(FS)}^2 &= \sigma_P^2 - (\sigma_A^2 / 2 + \sigma_D^2 / 4 + \sigma_{Ec}^2) \\
 &= \sigma_A^2 + \sigma_D^2 + \sigma_E^2 - (\sigma_A^2 / 2 + \sigma_D^2 / 4 + \sigma_{Ec}^2) \\
 &= (1/2)\sigma_A^2 + (3/4)\sigma_D^2 + \sigma_E^2 - \sigma_{Ec}^2
 \end{aligned}$$

## One-way ANOVA: N families with n sibs, T = Nn

Factor	Degrees of freedom, df	Sum of squares (SS)	Mean squares (MS)	E[MS]
Among family	N-1	$SS_f = n \sum_{i=1}^N (\bar{z}_i - \bar{z})^2$	$SS_f/(N-1)$	$\sigma_w^2 + n \sigma_f^2$
Within family	T-N	$SS_w = \sum_{i=1}^N \sum_{j=1}^n (\bar{z}_{ij} - \bar{z}_i)^2$	$SS_w/(T-N)$	$\sigma_w^2$

Estimating the variance components:

$$Var(f) = \frac{MS_f - MS_w}{n}$$

$$Var(w) = MS_w$$

$$Var(z) = Var(f) + Var(w)$$

Since  $\sigma_f^2 = \sigma_A^2 / 2 + \sigma_D^2 / 4 + \sigma_{Ec}^2$

$2Var(f)$  is an upper bound for the additive variance

### Assigning standard errors (= square root of Var)

**Fun fact:** Under normality, the (large-sample) variance for a mean-square is given by

$$\sigma^2(MS_x) \cong \frac{2(MS_x)^2}{df_x + 2}$$

$$Var[Var(w(FS))] = Var(MS_w) \cong \frac{2(MS_x)^2}{T - N + 2}$$

$$\begin{aligned} Var[Var(f)] &= Var\left[\frac{MS_f - MS_w}{n}\right] \\ &\cong \frac{2}{n^2} \left[ \frac{(MS_f)^2}{N+1} + \frac{(MS_w)^2}{T-N+2} \right] \end{aligned}$$

### Estimating heritability

$$t_{FS} = \frac{Var(f)}{Var(z)} = \frac{1}{2}h^2 + \frac{\sigma_D^2 / 4 + \sigma_{Ec}^2}{\sigma_z^2}$$

Hence,  $h^2 \leq 2 t_{FS}$

**An approximate large-sample standard error for  $h^2$  is given by**

$$SE(h^2) \cong 2(1 - t_{FS})[1 + (n-1)t_{FS}] \sqrt{2 / [Nn(n-1)]}$$

## Worked Example

10 full-sib families, each with 5 offspring are measured

Factor	df	SS	MS	EMS
Among-families	9	$SS_f = 405$	45	$\sigma_w^2 + 5 \sigma_f^2$
Within-families	40	$SS_w = 800$	20	$\sigma_w^2$

$$Var(f) = \frac{MS_f - MS_w}{n} = \frac{45 - 20}{5} = 5 \quad \xrightarrow{\text{green}} \quad V_A < 10$$

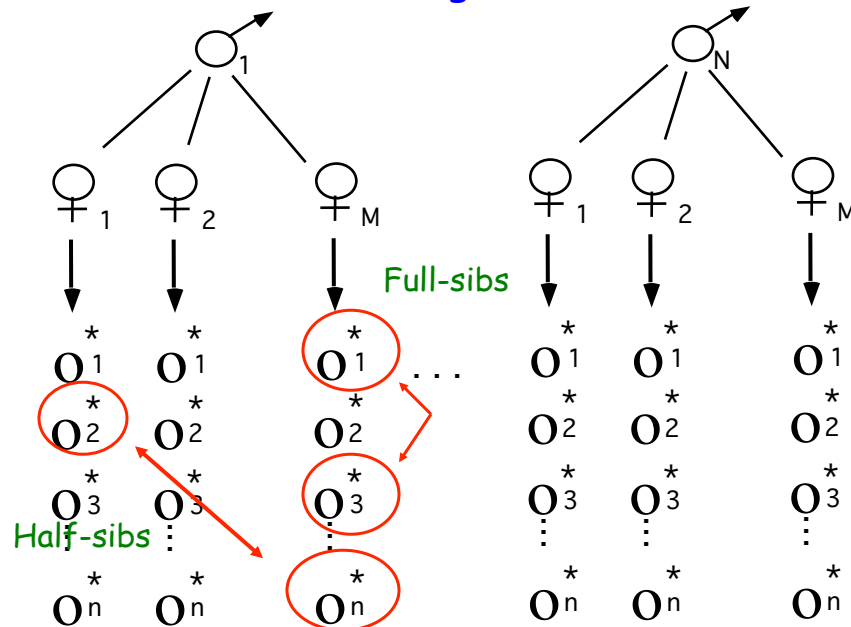
$$Var(w) = MS_w = 20$$

$$Var(z) = Var(f) + Var(w) = 25$$

$$h^2 < 2 (5/25) = 0.4$$

$$SE(h^2) \cong 2(1 - 0.4)[1 + (5 - 1)0.4] \sqrt{2 / [50(5 - 1)]} = 0.312$$

## Full sib-half sib design: Nested ANOVA

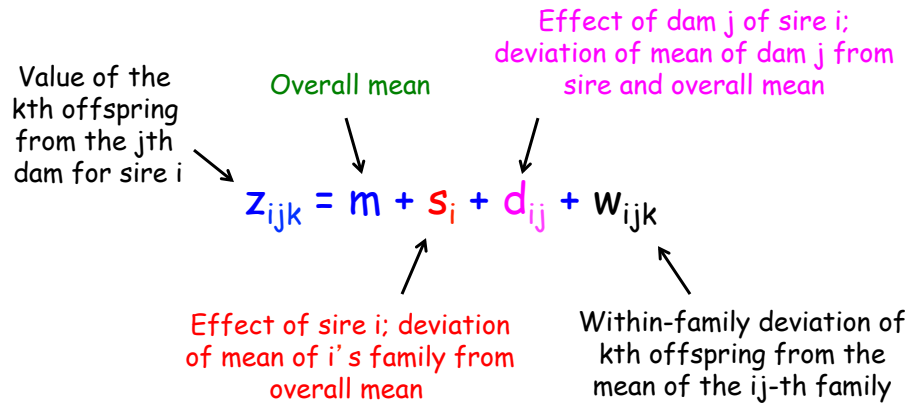




## Estimation: Nested ANOVA

Balanced full-sib / half-sib design: N males (**sires**) are crossed to M **dams** each of which has n offspring:

Nested ANOVA model



## Nested ANOVA Model

$$z_{ijk} = m + s_i + d_{ij} + w_{ijk}$$

$\sigma_s^2$  = between-sire variance = variance in sire family means

$\sigma_d^2$  = variance among dams within sires = variance of dam means for the same sire

$\sigma_w^2$  = within-family variance

$$\sigma_T^2 = \sigma_s^2 + \sigma_d^2 + \sigma_w^2$$

**Nested ANOVA:** N sires crossed to  
M dams, each with n sibs, T = NMn

Factor	df	SS	MS	E[MS]
Sires	N-1	$SS_s$	$SS_s/(N-1)$	$\sigma_w^2 + n\sigma_d^2 + Mn\sigma_s^2$
Dams(Sires)	N(M-1)	$SS_d$	$SS_d/[N(M-1)]$	$\sigma_w^2 + n\sigma_d^2$
Sibs(Dams)	T-NM	$SS_w$	$SS_w/(T-NM)$	$\sigma_w^2$

where:  $SS_s = Mn \sum_{i=1}^N (\bar{z}_i - \bar{z})^2$

$$SS_d = n \sum_{i=1}^N \sum_{j=1}^M (\bar{z}_{ij} - \bar{z}_i)^2 \quad \text{and} \quad SS_w = n \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^n (z_{ijk} - \bar{z}_{ij})^2$$

Estimation of sire, dam, and family variances:

$$Var(s) = \frac{MS_s - MS_d}{Mn}$$

$$Var(d) = \frac{MS_d - MS_w}{n}$$

$$Var(e) = MS_w$$

Translating these into the desired variance components:

- $Var(\text{Total}) = Var(\text{between FS families}) + Var(\text{within FS})$

$$\rightarrow \sigma_w^2 = \sigma_z^2 - Cov(FS)$$

- $Var(\text{Sires}) = Cov(\text{Paternal half-sibs})$

$$\sigma_d^2 = \sigma_z^2 - \sigma_s^2 - \sigma_w^2 = \sigma(FS) - \sigma(PHS)$$

Summarizing:

$$\begin{aligned}\sigma_s^2 &= \sigma(PHS) & \sigma_d^2 &= \sigma_z^2 - \sigma_s^2 - \sigma_w^2 \\ \sigma_w^2 &= \sigma_z^2 - \sigma(FS) & &= \sigma(FS) - \sigma(PHS)\end{aligned}$$

Expressing these in terms of the genetic and environmental variances:

$$\begin{aligned}\sigma_s^2 &\cong \frac{\sigma_A^2}{4} & \sigma_d^2 &\cong \frac{\sigma_A^2}{4} + \frac{\sigma_D^2}{4} + \sigma_{Ec}^2 \\ \sigma_w^2 &\cong \frac{\sigma_A^2}{2} + \frac{3\sigma_D^2}{4} + \sigma_{Es}^2\end{aligned}$$

Intraclass correlations and estimating heritability

$$\begin{aligned}t_{PHS} &= \frac{Cov(PHS)}{Var(z)} = \frac{Var(s)}{Var(z)} \rightarrow 4t_{PHS} = h^2 \\ t_{FS} &= \frac{Cov(FS)}{Var(z)} = \frac{Var(s) + Var(d)}{Var(z)} \rightarrow h^2 \leq 2t_{FS}\end{aligned}$$

Note that  $4t_{PHS} = 2t_{FS}$  implies no dominance or shared family environmental effects

Worked Example: N = 10 sires, M = 3 dams, n = 10 sibs/dam

Factor	df	SS	MS	E[MS]
Sires	9	4,230	470	$\sigma_w^2 + 10\sigma_d^2 + 30\sigma_s^2$
Dams(Sires)	20	3,400	170	$\sigma_w^2 + 10\sigma_d^2$
Within Dams	270	5,400	20	$\sigma_w^2$

$$\sigma_w^2 = MS_w = 20$$

$$\sigma_d^2 = \frac{MS_d - MS_w}{n} = \frac{170 - 20}{10} = 15$$

$$\sigma_s^2 = \frac{MS_s - MS_d}{Nn} = \frac{470 - 170}{30} = 10$$

$$\sigma_P^2 = \sigma_s^2 + \sigma_d^2 + \sigma_w^2 = 45$$

$$\begin{aligned}\sigma_d^2 &= 15 = (1/4)\sigma_A^2 + (1/4)\sigma_D^2 + \sigma_{Ec}^2 \\ &= 10 + (1/4)\sigma_D^2 + \sigma_{Ec}^2\end{aligned}$$

$$\sigma_A^2 = 4\sigma_s^2 = 40$$

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} = \frac{40}{45} = 0.89$$

$$\sigma_D^2 + 4\sigma_{Ec}^2 = 20$$

## Beetle Example

Messina and Fry (2003): 24 males each mated to 4 or 5 dams (different for each sire), and 5 female progeny from each dam were measured for two traits, mass eclosion and lifetime fecundity

### ANOVA for fecundity

Factor	df	SS	MS
Sires	23	33,983	1,477.5
Dams(Sires)	86	64,441	749.3
Sibs(Dams)	439	77,924	177.5



beetle example

## Beetle Example

### Expected Mean Squares (EMS)

$$\begin{array}{ll} \text{Sires:} & \sigma_R^2 + n\sigma_D^2 + nq\sigma_S^2 \\ \text{Dams(Sires):} & \sigma_R^2 + n\sigma_D^2 \\ \text{Sibs(Dams):} & \sigma_R^2 \end{array}$$

Approximately  $n = 5$  progeny by mating, and an average of  $q = 4.58$  dams per sire, so:

$$\begin{aligned} \sigma_R^2 &= 177.5 \\ \sigma_D^2 &= (749.3 - 177.5)/5 = 114.36 \\ \sigma_S^2 &= (1,477.5 - 749.3)/22.9 = 31.80 \end{aligned}$$

- Note: ANOVA method works only with balanced or slightly unbalanced data sets; otherwise ML or REML should be preferred

## Beetle Example

### Estimation of genetic (causal) parameters:

$$\begin{aligned} \sigma_S^2 &= V_A/4 \\ \sigma_D^2 &= V_A/4 + V_D/4 + V_{Ec} \\ \sigma_R^2 &= V_A/2 + 3V_D/4 + V_{Es} \end{aligned}$$

For simplicity, assuming  $V_D = 0$ , the following estimates are obtained for the causal components:

$$\begin{aligned} V_A &= 4\sigma_S^2 = 127.2 \\ V_{Ec} &= \sigma_D^2 - \sigma_S^2 = 82.56 \\ V_{Es} &= \sigma_R^2 - 2\sigma_S^2 = 113.9 \end{aligned}$$

Heritability:  $h^2 = V_A/(\sigma_R^2 + \sigma_D^2 + \sigma_S^2) = 0.393$

## Parent-offspring Regression

Single parent - offspring regression

$$z_{o_i} = \mu + b_{olp}(z_{p_i} - \mu) + e_i$$

The expected slope of this regression is:

$$E(b_{olp}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \cong \frac{(\sigma_A^2 / 2) + \sigma(E_o, E_p)}{\sigma_z^2} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma_z^2}$$

Residual error variance (spread around expected values)

$$\sigma_e^2 = \left(1 - \frac{h^2}{2}\right) \sigma_z^2$$

The expected slope of this regression is:

$$E(b_{olp}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \cong \frac{(\sigma_A^2 / 2) + \sigma(E_o, E_p)}{\sigma_z^2} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma_z^2}$$

Shared environmental values

To avoid this term, typically regressions are male-offspring, as female-offspring more likely to share environmental values

Midparent-offspring  
regression:

$$z_{oi} = \mu + b_{olMP} \left( \frac{z_{mi} + z_{fi}}{2} - \mu \right) + e_i$$

$$\begin{aligned} b_{olMP} &= \frac{\text{Cov}[z_o, (z_m + z_f) / 2]}{\text{Var}[(z_m + z_f) / 2]} \\ &= \frac{[\text{Cov}(z_o, z_m) + \text{Cov}(z_o, z_f)] / 2}{[\text{Var}(z) + \text{Var}(z)] / 4} \\ &= \frac{2\text{Cov}(z_o, z_p)}{\text{Var}(z)} = 2b_{olp} \end{aligned}$$

The expected slope of this regression is  $h^2$

Residual error variance (spread around expected values)

$$\sigma_e^2 = \left( 1 - \frac{h^2}{2} \right) \sigma_z^2$$

## Standard Errors

Single parent-offspring regression, N parents, each with n offspring

$$\text{Var}(b_{olp}) \cong \frac{n(t - b_{plp}^2) + (1 - t)}{Nn}$$

Square regression slope

Total number of offspring

Sib correlation  $t = \begin{cases} t_{HS} = h^2 / 4 & \text{for half-sibs} \\ t_{FS} = h^2 / 2 + \frac{\sigma_D^2 + \sigma_{Ec}^2}{\sigma_z^2} & \text{for full-sibs} \end{cases}$

$$\text{Var}(h^2) = \text{Var}(2b_{olp}) = 4\text{Var}(b_{olp})$$

Midparent-offspring regression,  
N sets of parents, each with n offspring

$$\text{Var}(h^2) = \text{Var}(b_{olMP}) \cong \frac{2[n(t_{FS} - b_{olMP}^2 / 2) + (1 - t_{FS})]}{Nn}$$

- Midparent-offspring variance half that of single parent-offspring variance

$$\text{Var}(h^2) = \text{Var}(2b_{olp}) = 4\text{Var}(b_{olp})$$



### Estimating Heritability in Natural Populations

Often, sibs are reared in a laboratory environment, making parent-offspring regressions and sib ANOVA problematic for estimating heritability

Let  $b'$  be the slope of the regression of the values of lab-raised offspring regressed in the trait values of their parents in the wild

A lower bound can be placed of heritability using parents from nature and their lab-reared offspring,

$$h_{min}^2 = (b'_{olMP})^2 \frac{\text{Var}_n(z)}{\text{Var}_l(A)}$$

 Trait variance in nature  
 Additive variance in lab



Why is this a lower bound?

Covariance between  
breeding value in nature  
and BV in lab

$$(b'_{olMP})^2 \frac{Var_n(z)}{Var_l(A)} = \left[ \frac{Cov_{l,n}(A)}{Var_n(z)} \right]^2 \frac{Var_n(z)}{Var_l(A)} = \gamma^2 h_n^2$$

where  $\gamma = \frac{Cov_{l,n}(A)}{\sqrt{Var_n(A)Var_l(A)}}$

is the additive genetic covariance  
between environments and hence  $\gamma^2 \leq 1$

## Defining $H^2$ for Plant Populations

Plant breeders often do not measure individual plants (especially with pure lines), but instead measure a **plot** or a **block** of individuals. This can result in inconsistent measures of  $H^2$  even for otherwise identical populations

$$z_{ijkl} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijkl}$$

Genotype i      Environment j      Effect of plot k for genotype i in environment j

Interaction between genotype i and environment j      Deviations of individual plants within plots

$$z_{ijkl} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijkl}$$

$$\sigma^2(z_i) = \sigma_G^2 + \sigma_E^2 + \frac{\sigma_{GE}^2}{e} + \frac{\sigma_p^2}{er} + \frac{\sigma_e^2}{ern}$$

$$\left\{ \begin{array}{l} e = \text{number of environments} \\ r = (\text{replicates}) \text{ number of plots/environment} \\ n = \text{number of individuals per plot} \end{array} \right.$$

Hence,  $V_p$ , and hence  $H^2$ , depends  
on our choice of  $e$ ,  $r$ , and  $n$

# Lecture 8

## Mixed Models, BLUP Breeding Values

Guilherme J. M. Rosa

University of Wisconsin-Madison

Introduction to Quantitative Genetics  
SISG, Seattle  
17 - 19 July 2017

## OUTLINE

- The General Linear Model
- Linear Mixed Models
- The 'Animal Model'
- EBV and Prediction Accuracy
- Multiple Random Effects

## General Linear Model (Fixed Effects Model)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

responses
residuals

design/incidence matrix (known)
overall mean + fixed effects parameters

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n \sigma^2) \rightarrow \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

⇒ **Fixed effect:** levels included in the study represent all levels about which inference is to be made. **Fixed effects models:** models containing only fixed effects

### Example 1

Experiment to compare growth performance of pigs under two experimental groups (Control and Treatment), with three replications each.

**Model:**

Control	Treatment
53	61
46	66
58	57

$$y_{ij} = \mu + \delta_i + e_{ij}$$

$y_{ij}$ : weight gain of pig j of group i  
 $\mu$ : constant; general mean  
 $\delta_i$ : effect of group i  
 $e_{ij}$ : residual term

4

## Matrix Notation

Control	Treatment
53	61
46	66
58	57

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 53 \\ 46 \\ 58 \\ 61 \\ 66 \\ 57 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}$$

5

## Example 2

Flowering time (days, log scale) of *Brassica napus* according to genotype in specific locus, such as a candidate gene

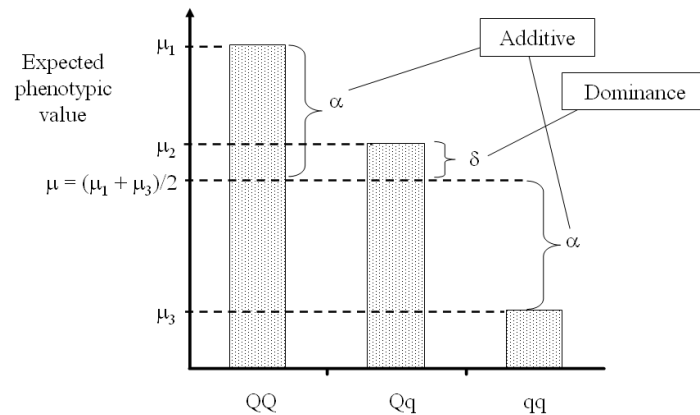
Genotype		
qq	Qq	QQ
3.4	2.9	3.1
3.7	2.5	2.6
3.2		

**Model:**  $y_{ij} = \mu_i + e_{ij}$

$$\left\{ \begin{array}{l} y_{ij}: \text{flowering time of replication } j \text{ (} j = 1, \dots, n_i \text{) of} \\ \text{genotype } i \text{ (} i = qq, Qq \text{ and } QQ \text{)} \\ \mu_i: \text{expected flowering time of plants of genotype } i \\ e_{ij}: \text{residual (environment and polygenic effects)} \end{array} \right.$$

6

⇒ The expected phenotypic values  $\mu_i$ , however, can be expressed as a function of the additive and dominant effects



Expected phenotypic value according to the genotype on a specific locus.

7

The model can be written then as:

$$y_{ij} = \mu + x_{ij}\alpha + (1 - |x_{ij}|)\delta + e_{ij}$$

- $\mu$ : constant (mid-point flowering time between homozygous genotypes)
- $x_{ij}$ : indicator variable (genotype), coded as -1, 0 and 1 for genotypes qq, Qq and QQ
- $\alpha$  and  $\beta$ : additive and dominance effects

In matrix notation:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 3.4 \\ 3.7 \\ 3.2 \\ 2.9 \\ 2.5 \\ 3.1 \\ 2.6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \delta \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

## Least-Squares Estimation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{I}_n \sigma^2) \rightarrow \varepsilon_i \stackrel{\text{iid}}{\sim} (0, \sigma^2)$$

An estimate ( $\hat{\boldsymbol{\beta}}$ ) of the vector  $\boldsymbol{\beta}$  can be obtained by the method of least-squares, which aims to minimize the residual sum of squares, given (in matrix notation) by:

$$\text{RSS} = \sum_{i=1}^n (\hat{\varepsilon}_i)^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Taking the derivatives and equating to zero, it can be shown that the least-squares estimator of  $\boldsymbol{\beta}$  is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

➡ It is shown that  $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$  and  $\text{Var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$

9

## Least-Squares Estimation

The estimator  $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  is called **ordinary least squares (OLS)** estimator, and it is indicated only in situations with homoscedastic and uncorrelated residuals

If the residual variance is heterogeneous (i.e.,  $\text{Var}(\varepsilon_i) = \sigma_i^2 = w_i \sigma^2$ ), the residual variance matrix can be expressed as  $\text{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\sigma^2$ , where  $\mathbf{W}$  is a diagonal matrix with the elements  $w_i$ , a better estimator of  $\boldsymbol{\beta}$  is given by:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{y}$$

which is generally referred to as **weighted least squares (WLS)** estimator.

Furthermore, in situations with a general residual variance-covariance matrix  $\mathbf{V}$ , including correlated residuals, a **generalized least squares (GLS)** estimator  $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$  is obtained by minimizing the generalized sum of squares, given by:

$$\text{GSS} = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

10

## Maximum Likelihood Estimation

**Likelihood Function:** any function of the model parameters that is proportional to the density function of the data

Hence, to use a likelihood-based approach for estimating model parameters, some extra assumptions must be made regarding the distribution of the data

In the case of the linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , if the residuals are assumed normally distributed with mean vector zero and variance-covariance matrix  $\mathbf{V}$ , i.e.  $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{V})$ , the response vector  $\mathbf{y}$  is also normally distributed, with expectation  $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$  and variance  $\text{Var}[\mathbf{y}] = \mathbf{V}$

11

## Maximum Likelihood Estimation

The distribution of  $\mathbf{y}$  has a density function given by:

$$p(\mathbf{y} | \boldsymbol{\beta}, \mathbf{V}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

so that the **likelihood** and the **log-likelihood** functions can be expressed respectively as:

$$L(\boldsymbol{\beta}, \mathbf{V}) \propto |\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

and

$$l(\boldsymbol{\beta}, \mathbf{V}) = \log[L(\boldsymbol{\beta}, \mathbf{V})] \propto -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

12



## Maximum Likelihood Estimation

Assuming  $V$  known, the likelihood equations for  $\beta$  are given by taking the first derivatives of  $l(\beta, V)$  with respect to  $\beta$  and equating it to zero

The maximum likelihood estimator (MLE) for  $\beta$  is then shown to be:

$$\text{MLE}(\beta) = \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

Note: Under normality the MLE coincides with the GLS estimator discussed previously

In addition, it is shown that:  $\hat{\beta} \sim \text{MVN}(\beta, (X^T V^{-1} X)^{-1})$

13

## Two-stage Analysis of Longitudinal Data Step 1

Supposed a series of **longitudinal data** (e.g., repeated measurements on time) on  $n$  individuals. Let  $y_{ij}$  represent the observation  $j$  ( $j = 1, 2, \dots, n_i$ ) on individual  $i$  ( $i = 1, 2, \dots, n$ ), and the following quadratic regression of measurements on time ( $z_{ij}$ ) for each individual:

$$y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \beta_{2i} z_{ij}^2 + \epsilon_{ij}$$

where  $\beta_{0i}$ ,  $\beta_{1i}$  and  $\beta_{2i}$  are **subject-specific regression** parameters, and  $\epsilon_{ij}$  are residual terms, assumed normally distributed with mean zero and variance  $\sigma_\epsilon^2$

14

In matrix notation such [subject-specific regressions](#) can be expressed as:

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

where  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^T$ ,  $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})^T$ ,

$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{in_i})^T \sim N(\mathbf{0}, \mathbf{I}\sigma_{\varepsilon}^2)$  and

$$\mathbf{Z}_i = \begin{bmatrix} 1 & z_{i1} & z_{i1}^2 \\ 1 & z_{i2} & z_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & z_{in_i} & z_{in_i}^2 \end{bmatrix}$$

15

Under these specifications, it is shown that the least-squares estimate of  $\beta_i$  is:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \mathbf{y}_i$$

Note that this is also the maximum likelihood estimate of  $\beta_i$

Such estimates can be viewed as [summary statistics](#) for the longitudinal data, the same way one could use area under the curve (AUC), or peak (maximum value of  $y_{ij}$ ), or mean response.

16

## Two-stage Analysis of Longitudinal Data

### Step 2

Supposed now we are interested on the **effect of some other variables** (such as gender, treatment, year, etc.) on the values of  $\beta_i$

Such effects could be studied using a model as:

$$\hat{\beta}_i = \mathbf{W}_i \boldsymbol{\beta} + \mathbf{u}_i$$

where  $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{D})$ , which is an approximation for the model:

$$\beta_i = \mathbf{W}_i \boldsymbol{\beta} + \mathbf{u}_i \quad (2)$$

17

## Single-stage Analysis of Longitudinal Data

The two step-analysis described here can be merged into a single stage approach by substituting (2) in (1):

$$\mathbf{y}_i = \mathbf{Z}_i [\mathbf{W}_i \boldsymbol{\beta} + \mathbf{u}_i] + \boldsymbol{\varepsilon}_i$$

which can be expressed as:

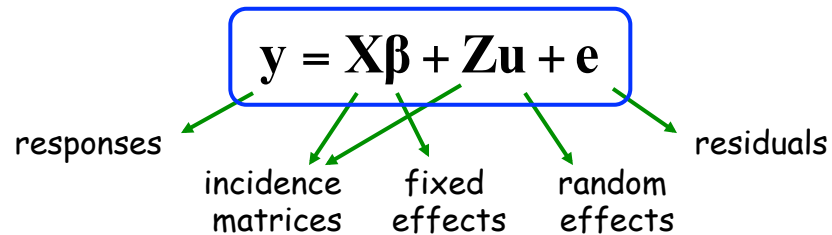
$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

where  $\mathbf{X}_i = \mathbf{Z}_i \mathbf{W}_i$ . By concatenating observations from multiple individuals, we have the following **mixed model**:

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \boldsymbol{\varepsilon}$$

18

## Linear Mixed Effects Model



$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \Sigma \end{bmatrix} \right)$$

19

## Estimation of Fixed Effects

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

with  $\varepsilon = \mathbf{Z}\mathbf{u} + \mathbf{e}$ , such that  $\text{Var}[\varepsilon] = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \Sigma$

→ MLE of  $\beta$ :

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \sim \text{MVN}(\beta, (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1})$$

where  $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \Sigma$

20

## Prediction of Random Effects

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{ZG} \\ \mathbf{GZ}^T & \mathbf{G} \end{bmatrix} \right)$$

$$\begin{aligned} E[\mathbf{u} | \mathbf{y}] &= E[\mathbf{u}] + \text{Cov}[\mathbf{u}, \mathbf{y}^T] \text{Var}^{-1}[\mathbf{y}] (\mathbf{y} - E[\mathbf{y}]) \\ &= \mathbf{GZ}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{GZ}^T (\mathbf{ZGZ}^T + \boldsymbol{\Sigma})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Replacing  $\boldsymbol{\beta}$  by its estimate:

$$\hat{\mathbf{u}} = \mathbf{GZ}^T (\mathbf{ZGZ}^T + \boldsymbol{\Sigma})^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

21

## Mixed Model Equations

$$\begin{bmatrix} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \\ \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{bmatrix}$$

BLUP and BLUE:

$$\left\{ \begin{aligned} \hat{\mathbf{u}} &= (\mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\boldsymbol{\beta}} &= \{ \mathbf{X}^T [\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{Z} (\mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}^{-1}] \mathbf{X} \}^{-1} \\ &\quad \times \mathbf{X}^T [\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{Z} (\mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{G}^{-1})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}^{-1}] \mathbf{y} \end{aligned} \right.$$

22

## Estimation of Variance Components

BLUE and BLUP require knowledge of  $\mathbf{G}$  and  $\mathbf{\Sigma}$

These matrices, however, are rarely known and must be estimated

Variance and covariance components estimation:

- Analysis of Variance (ANOVA)
- Maximum Likelihood
- Restricted Maximum Likelihood (REML)
- Bayesian Inference

23

## Mixed Models in Animal and Plant Breeding

Animal/plant breeding programs are based on the principle that phenotypic observations on related individuals can provide information about their underlying genotypic values

The additive component of genetic variation is the primary determinant of the degree to which offspring resemble their parents, and therefore this is usually the component of interest in artificial selection programs

24

## Mixed Models in Animal and Plant Breeding

Many statistical methods for analysis of genetic data are specific (or more appropriate) for phenotypic measurements obtained from planned experimental designs and with balanced data sets

While such situations may be possible within laboratory or greenhouse experimental settings, data from natural populations and agricultural species are generally highly unbalanced and fragmented by numerous kinds of relationships

25

## Animal Model

Culling of data to accommodate conventional statistical techniques (e.g. ANOVA) may introduce bias and/or lead to a substantial loss of information

The mixed model methodology allows efficient estimation of genetic parameters (such as variance components and heritability) and breeding values while accommodating extended pedigrees, unequal family sizes, overlapping generations, sex-limited traits, assortative mating, and natural or artificial selection

To illustrate such application of mixed models in breeding programs, we consider here the so-called **Animal Model** in situations with a single trait and a single observation (including missing values) per individual

26

## Animal Model

The animal model can be described as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- $\mathbf{y}$  is an  $(n \times 1)$  vector of observations (phenotypic scores)
- $\boldsymbol{\beta}$  is a  $(p \times 1)$  vector of fixed effects (e.g. herd-year-season effects)
- $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$  is a  $(q \times 1)$  vector of breeding values (relative to all individuals with record or in the pedigree file, such that  $q$  is in general bigger than  $n$ )
- $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$  represents residual effects, where  $\sigma_e^2$  is the residual variance

27

## The Matrix $\mathbf{A}$

The matrix  $\mathbf{G}$  describing the covariances among the random effects (here the breeding values) follows from standard results for the covariances between relatives

It is seen that the additive genetic covariance between two relatives  $i$  and  $i'$  is given by  $2\theta_{ii'}\sigma_a^2$ , where  $\theta_{ii'}$  is the coefficient of coancestry between individuals  $i$  and  $i'$ , and  $\sigma_a^2$  is the additive genetic variance in the base population

Hence, under the animal model,  $\mathbf{G} = \mathbf{A}\sigma_a^2$ , where  $\mathbf{A}$  is the additive genetic (or numerator) relationship matrix, having elements given by  $a_{ii'} = 2\theta_{ii'}$

28



## The Matrix **A**

For each animal  $i$  in the pedigree ( $i = 1, 2, \dots, n$ ), going from older to younger animals, compute  $a_{ii}$  and  $a_{ij}$  ( $j = 1, 2, \dots, i-1$ ) as follows:

If both parents ( $s$  and  $d$ ) of animal  $i$  are known:

$$a_{ij} = a_{ji} = (a_{js} + a_{jd})/2 \text{ and } a_{ii} = 1 + a_{sd}/2$$

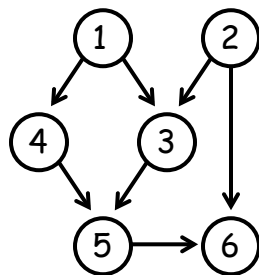
If only one parent (e.g.  $d$ ) of animal  $i$  is known:

$$a_{ij} = a_{ji} = a_{jd}/2 \text{ and } a_{ii} = 1$$

If parents unknown:

$$a_{ij} = a_{ji} = 0 \text{ and } a_{ii} = 1$$

### Example



Animal	Sire	Dam
1	-	-
2	-	-
3	1	2
4	1	-
5	4	3
6	5	2



pedigree matrix **A**

$$A = \begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .25 \\ 0 & 1 & .5 & 0 & .25 & .625 \\ .5 & .5 & 1 & .25 & .625 & .563 \\ .5 & 0 & .25 & 1 & .625 & .313 \\ .5 & .25 & .625 & .625 & 1.125 & .688 \\ .25 & .625 & .563 & .313 & .688 & 1.125 \end{bmatrix}$$

## Animal Model

In general, in animal/plant breeding interest is on prediction of breeding values (for selection of superior individuals), and on estimation of variance components and functions thereof, such as heritability

The fixed effects are, in some sense, nuisance factors with no central interest in terms of inferences, but which need to be taken into account (i.e., they need to be corrected for when inferring breeding values)

31

## Animal Model

Since under the animal model  $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2}$  and  $\mathbf{R}^{-1} = \mathbf{I}_n\sigma_e^{-2}$ , the mixed model equations can be expressed as:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$$

where  $\lambda = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$ , such that:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$$

32

Conditional on the variance components ratio  $\lambda$ , the BLUP of the breeding values are given then by:

$$\hat{\mathbf{u}} = (\mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{A}^{-1})^{-1} \mathbf{Z}^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

These are generally referred to as **Estimated Breeding Values (EBV)**

Alternatively, some breeders associations express their results as Predicted Transmitting Abilities (PTA) (or Estimated Transmitting Abilities (ETA) or Expected Progeny Difference (EPD)), which are equal to half the EBV, representing the portion of an animal's breeding values that is passed to its offspring

33

The amount of information contained in an animal's genetic evaluation depends on the availability of its own record, as well as how many (and how close) relatives it has with phenotypic information

As a measure of amount of information in livestock genetic evaluations, EBVs are typically reported with its associated accuracies

**Accuracy** of predictions is defined as the correlation between true and estimated breeding values, i.e.,  $r_i = \rho(\hat{u}_i, u_i)$

Instead of accuracy, some livestock species genetic evaluations use **reliability**, which is the squared correlation of accuracy ( $r_i^2$ )

34

## Prediction Accuracy

The calculation of  $\rho(\hat{u}_i, u_i)$  requires the diagonal elements of the inverse of the **MME coefficient matrix**, represented as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{A}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix}$$

It is shown that the **prediction error variance** of EBV  $\hat{u}_i$  is given by:

$$\text{PEV} = \text{Var}(\hat{u}_i - u_i) = c_i^{uu} \sigma_e^2$$

where  $c_i^{uu}$  is the  $i$ -th diagonal element of  $\mathbf{C}^{uu}$ , relative to animal  $i$ .

35

## Prediction Accuracy

The PEV can be interpreted as the fraction of additive genetic variance not accounted for by the prediction

Therefore, PEV can be expressed also as:

$$\text{PEV} = (1 - r_i^2) \sigma_a^2$$

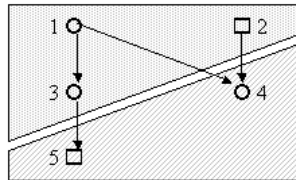
such that  $c_i^{uu} \sigma_e^2 = (1 - r_i^2) \sigma_a^2$ , from which the reliability is obtained as:

$$r_i^2 = 1 - c_i^{uu} \sigma_e^2 / \sigma_a^2 = 1 - \lambda c_i^{uu}$$

36

## Animal Model

herd 1



Animal	Sire	Dam	Herd	Observation
1	–	–	h1	310
2	–	–	h1	–
3	–	1	h1	270
4	2	1	h2	350
5	–	3	h2	–

herd 2

$$\begin{bmatrix} 310 \\ 270 \\ 350 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}}_{\mathbf{X} \boldsymbol{\beta}} + \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix}}_{\mathbf{Z} \mathbf{u}} + \begin{bmatrix} e_1 \\ e_3 \\ e_4 \end{bmatrix}$$

$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \mathbf{e}$

37

## Animal Model

Breeding values:  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$ , with

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 & 0.25 \\ 0 & 1 & 0 & 0.5 & 0 \\ 0.5 & 0 & 1 & 0.25 & 0.5 \\ 0.5 & 0.5 & 0.25 & 1 & 0.125 \\ 0.25 & 0 & 0.5 & 0.125 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{X} & \mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T \mathbf{y} \\ \mathbf{Z}^T \mathbf{y} \end{bmatrix}$$

$\lambda = \frac{\sigma_e^2}{\sigma_u^2} = \frac{1-h^2}{h^2}$

38

## R Code



animal model  
toy example

```

y<-matrix(c(310,270,350),nrow=3)
X<-matrix(c(1,1,0,0,0,1),nrow=3)
Z<-matrix(c(1,0,0,0,0,0,0,1,0,0,0,0,1,0),nrow=3, byrow = TRUE)
A<-matrix(c(1,0,0.5,0.5,0.25,
            0,1,0,0.5,0,
            0.5,0,1,0.25,0.5,
            0.5,0.5,0.25,1,0.125,
            0.25,0,0.5,0.125,1),nrow=5)

h2<-1/3 # heritability
a=(1-h2)/h2

# crossproducts
XX<-crossprod(X,X)
XZ<-t(X) %*% Z
ZX<-t(Z) %*% X
ZZ<-crossprod(Z,Z)+a*solve(A)

# mixed model equations
# coefficient matrix and right hand side
C<-rbind(cbind(XX,XZ),cbind(ZX,ZZ))
rhs<-rbind(t(X) %*% y,t(Z) %*% y)

#solution
theta.hat <- solve(C) %*% rhs

```

$$h^2 = \frac{1}{3} \rightarrow \alpha = 2 \Rightarrow \begin{cases} \hat{h}_1 = 290 \\ \hat{h}_2 = 348 \\ \hat{u}_1 = 4.0 \\ \hat{u}_2 = 0.0 \\ \hat{u}_3 = -4.0 \\ \hat{u}_4 = 2.0 \\ \hat{u}_5 = -2.0 \end{cases}$$

39

## Animal Model

The animal model can be extended to model multiple (correlated) traits, multiple random effects (such as maternal effects and common environmental effects), repeated records (e.g. test day models), and so on

**Example (Mrode 1996, pp74-76):** Weaning weight (kg) of piglets, progeny of three sows mated to two boars:

Piglet	Sire	Dam	Sex	Weight
6	1	2	1	90
7	1	2	2	70
8	1	2	2	65
9	3	4	2	98
10	3	4	1	106
11	3	4	2	60
12	3	4	2	80
13	1	5	1	100
14	1	5	2	85
15	1	5	1	68

40

A linear model with the (fixed) effect of sex, and the (random) effects of common environment (related to each litter) and breeding values can be expressed as  $\mathbf{X}$ :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{c} + \mathbf{e}$$

Weight → ↗ Sex → ↗ Breeding values → ↗ Common environment → ↗ Residual

Assuming that  $\sigma_u^2 = 20$ ,  $\sigma_c^2 = 15$  and  $\sigma_e^2 = 65$ , the MME are as follows:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{W} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \mathbf{A}^{-1}\lambda_1 & \mathbf{Z}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T\mathbf{W} + \mathbf{I}\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{bmatrix}$$

where  $\lambda_1 = \frac{\sigma_e^2}{\sigma_u^2} = 3.25$  and  $\lambda_2 = \frac{\sigma_e^2}{\sigma_c^2} = 4.3$

41

The BLUEs and BLUPs (inverting the numerator relationship matrix) are:



Mrode example

Effects	Solutions
<i>Sex</i>	
1	91.493
2	75.764
<i>Animals</i>	
1	-1.441
2	-1.175
3	1.441
4	1.441
5	-0.266
6	-1.098
7	-1.667
8	-2.334
9	3.925
10	2.895
11	-1.141
12	1.525
13	0.448
14	0.545
15	-3.819
<i>Environ.</i>	
2	-1.762
4	2.161
5	-0.399

42

# Lecture 9

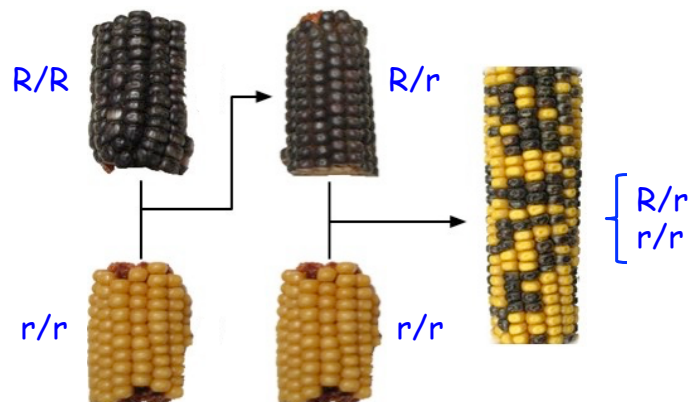
## QTL and Association Mapping

Guilherme J. M. Rosa

University of Wisconsin-Madison

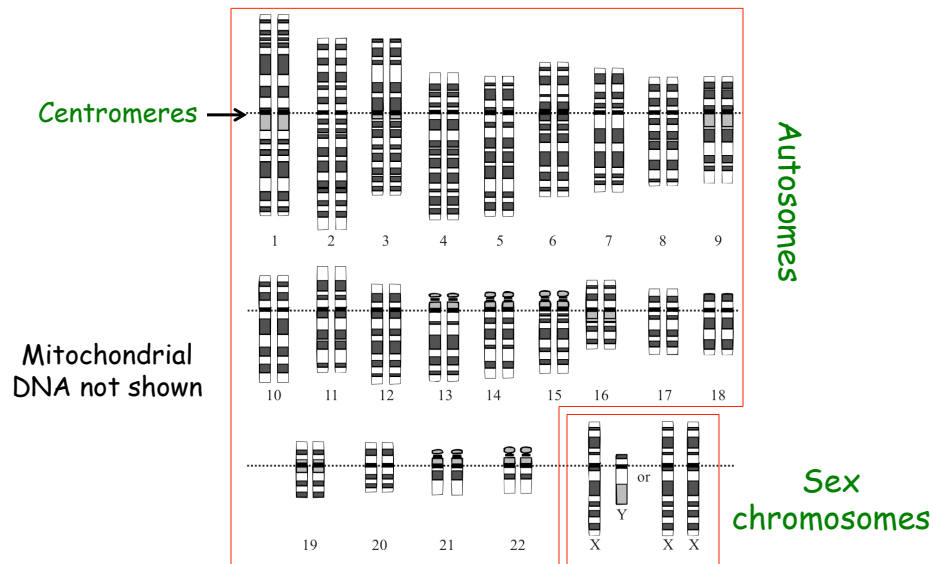
Introduction to Quantitative Genetics  
SISG, Seattle  
17 - 19 July 2017

## Linkage Analysis and QTL Mapping



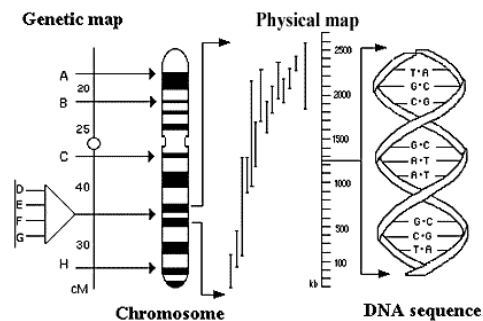


## Human Genome, Chromosomes



Graphical representation of the idealized human diploid karyotype

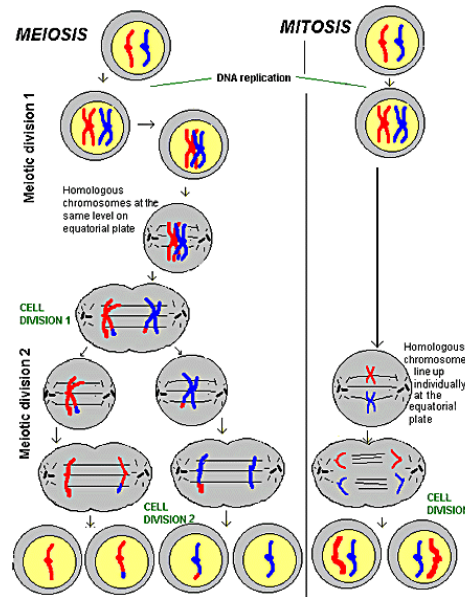
## Sequences of Base Pairs Mapping



**Genetic maps:** relative positions of loci in chromosomes or linkage groups. Distances in genetic maps are measured in centimorgans (cM, about 1 million base pairs)

**Physical maps:** overlapping collections of DNA fragments (measured in kilobases, kb) which are assembled together to build the base-by-base sequence of DNA

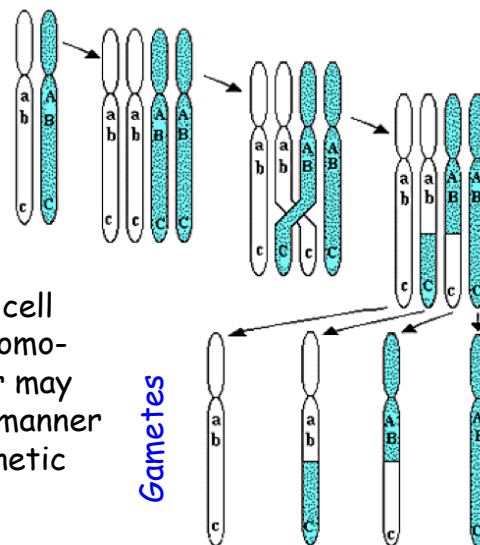
## Comparison of Meiosis and Mitosis



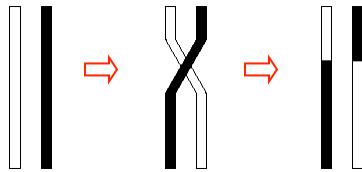
## Crossing-Over and Recombination During Meiosis

In meiosis, the precursor cells of the sperm or ova must multiply and at the same time reduce the number of chromosomes to one full set.

During the early stages of cell division in meiosis, two chromosomes of a homologous pair may exchange segments in the manner shown above, producing genetic variations in germ cells.



## Crossing Over and Recombination



An odd number of **crossovers** between two loci results in a **recombination** between them

Because crossing over takes place at random, the probability of recombination ( $r$ ) is higher for loci that are farther apart than for loci that are closer to each other

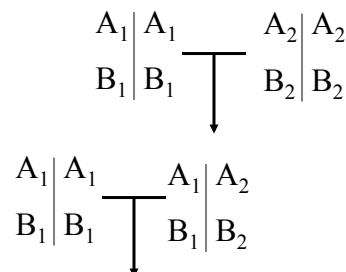
$$0 \leq r \leq 0.5$$

completely linked loci  $\swarrow$   $\searrow$  unlinked loci

## Two Point Linkage Analysis

- ⇒ Backcross experiment
- ⇒ Genotypic information for two loci (A and B)
- ⇒ Estimate the recombination rate  $r_{AB}$
- ⇒ Are these two loci linked?

Individual	A	B
1	0	0
2	0	1
⋮	⋮	⋮
n	1	1



Four possible genotypes

## Two Point Linkage Analysis

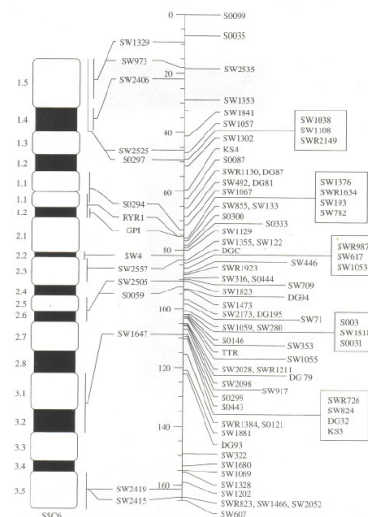
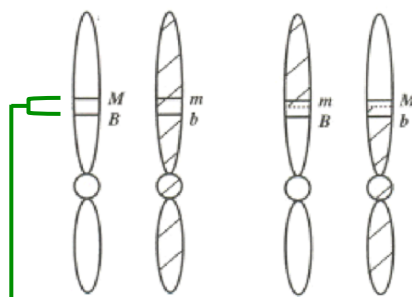
⇒ Suppose  $n = 80$  and  $y = 16$  (recombinants)

⇒ Point estimate of  $r_{AB}$  :  $\hat{r}_{AB} = \frac{y}{n} = 0.20$

⇒ Confidence interval (95%) of  $r_{AB}$  :

$$CI(r_{AB}; 95\%) = [0.1189; 0.3044]$$

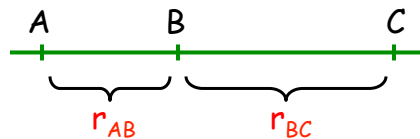
## Recombination Rate and Linkage Map



Estimates of recombination rates between pairs of markers are used to order markers and to infer their genetic distances (centimorgans; cM)

## Interference

⇒ Lack of independence in recombinations at different intervals on a chromosome



- If  $r_{AB}$  and  $r_{BC}$  are independent, the probability of double recombination is  $\Pr(\text{DR}) = r_{AB} \times r_{BC}$
- If  $r_{AB}$  and  $r_{BC}$  are not independent, the above probability is given by  $\Pr(\text{DR}) = c \times r_{AB} \times r_{BC}$  where  $c$  is called "coefficient of coincidence"
- **Interference:**  $I = 1 - c$

## Map Distance

The map distance  $x$  between two loci, in **Morgan** units, is defined as the expected number of crossovers between them

Unlike recombination rates, map distances are additive

The relationship between map distances and recombination rates is discussed next

## Map Functions

Map functions provide a transformation from map distance to recombination rate. Two approaches have been used to derive map functions:

In the first case, a probability model is assumed for the number of crossovers in an interval of length  $x$ . Then, recombination rate is calculated as the probability of an odd number of crossovers in the interval

In the second approach, recombination events in two adjacent intervals are modeled, allowing for interference

Examples of map functions: [Haldane](#), [Binomial](#), [Kosambi](#)

## Haldane Map Function

Haldane (1919) suggested that the number of crossovers in any chromosomal interval follows a Poisson distribution, with no interference

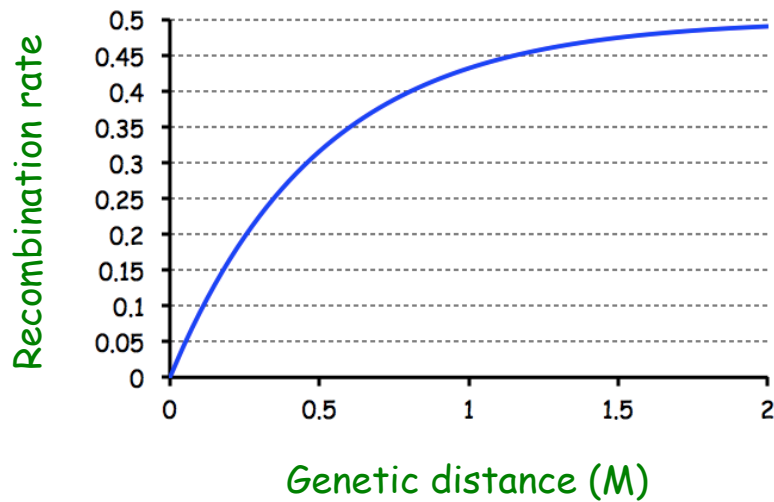
If  $P_k$  is the probability of  $k$  crossovers, then the probability of recombination ( $r$ ) is  $r = P_1 + P_3 + P_5 + \dots$

This leads to the [Haldane's map function](#):

$$r = \frac{1}{2}(1 - e^{-2x})$$

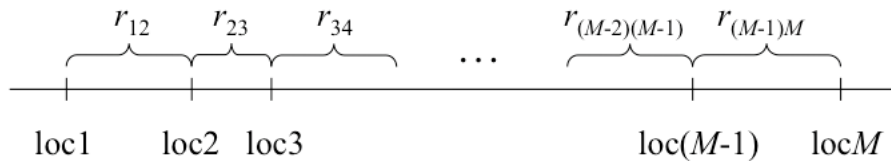
The inverse of which is:  $x = \begin{cases} -\frac{1}{2}\ln(1-2r) & , \text{ if } 0 \leq r < 0.5 \\ \infty & , \text{ if } r = 0.5 \end{cases}$

## Haldane Map Function



## Multipoint Point Linkage Analysis

- ⇒ Instead of two loci, suppose there are  $M$  loci
- ⇒ If order is unknown:  $M!/2$  alternatives



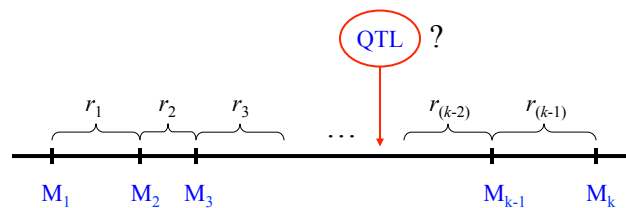
**Goal:** Determine the order of the loci and estimate recombination fractions between neighboring loci, i.e. “Map Construction”

## Methods for Mapping QTL

- ⇒ Single Marker Analysis
- ⇒ Interval Mapping
- ⇒ Composite Interval Mapping
- ⇒ Bayesian Methods

## QTL Mapping

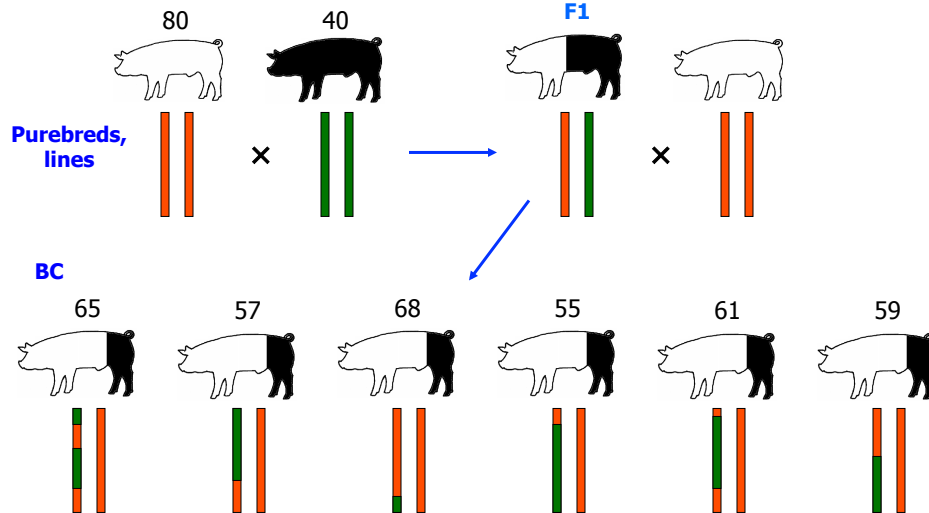
- ⇒ Methods based on linkage disequilibrium between markers and QTL (line crossing or segregating population)
- ⇒ Requirements:
  - ① Linkage (marker) maps
  - ② Variation for the quantitative trait





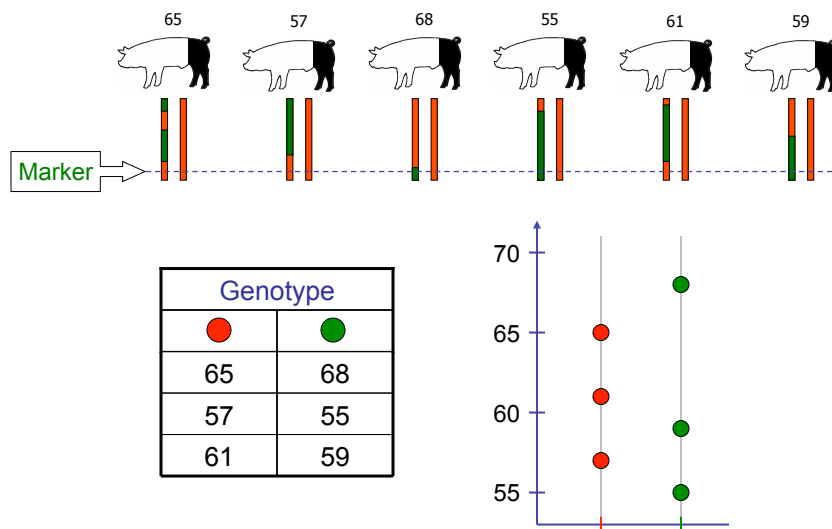
## QTL Mapping

### Single Marker Analysis; Example with Backcross



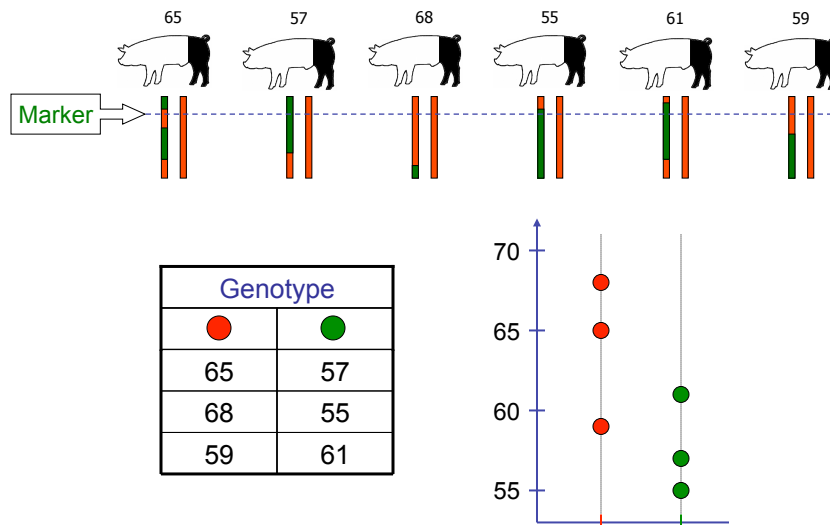
## QTL Mapping

### Single Marker Analysis; Example with Backcross



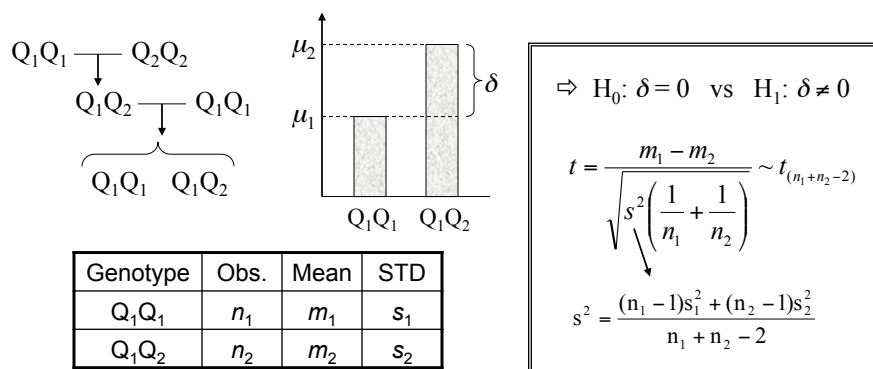
# QTL Mapping

## Single Marker Analysis; Example with Backcross



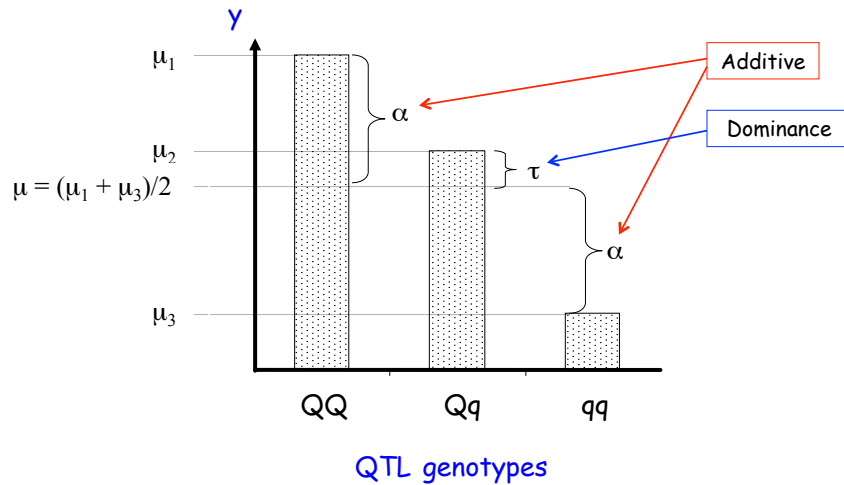
## Single Marker Analysis

Simple example with candidate gene and BC population



$$CI[\delta; (1 - \alpha)]: (m_2 - m_1) \pm t_{(n_1 + n_2 - 2; \alpha/2)} \sqrt{\frac{s^2}{n_1 + n_2 - 2}}$$

## Example with F2 Population



## Example with F2 Population

Information on phenotypes and genotypes for a specific marker

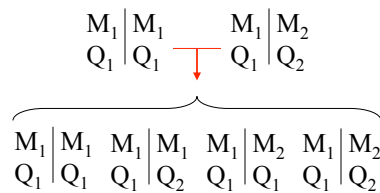


Marker Genotype	Phenotype (8 individuals per group)
MM	95.9, 108.0, 96.5, 92.9 101.0, 94.5, 93.7, 89.8
Mm	105.2, 107.9, 89.9, 113.4 109.7, 102.4, 97.1, 107.1
mm	117.1, 95.2, 106.4, 104.7 92.5, 123.9, 97.8, 100.5

## Single Marker Analysis

👉 QTL and marker (M); recombination frequency =  $r$

Genotype	Freq.	E[y]	Marker group	Freq.	E[y]
$M_1M_1Q_1Q_1$	$(1-r)/2$	$\mu_1$	$M_1M_1$	$\frac{1}{2}$	$(1-r)\mu_1 + r\mu_2$
$M_1M_1Q_1Q_2$	$r/2$	$\mu_2$			
$M_1M_2Q_1Q_1$	$r/2$	$\mu_1$	$M_1M_2$	$\frac{1}{2}$	$r\mu_1 + (1-r)\mu_2$
$M_1M_2Q_1Q_2$	$(1-r)/2$	$\mu_2$			



Difference between marker group expected values

$$r\mu_1 + (1-r)\mu_2 - (1-r)\mu_1 - r\mu_2 \\ = (1-2r)(\mu_2 - \mu_1) = (1-2r)\delta$$

## Single Marker Analysis

(EXAMPLE)

- ⇒ *Brassica napus*; Flowering time
- ⇒ 10 Markers  
(positions: 0, 8.8, 20.6, 27.4, 34.2, 42.9, 53.6, 64.1, 69.2, 83.9 cM)
- ⇒ 104 individuals; Double haploid

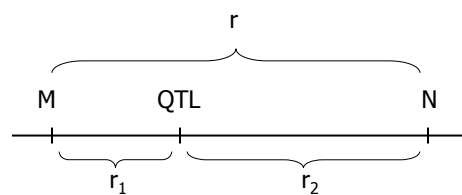
3.0204	-1	-1	-1	-1	-1	-1	-1	-99	-1
2.9704	-1	-1	-1	-1	-99	-1	-1	-1	1
2.7408	-1	-1	1	1	1	1	1	1	1
:	:	:	:	:	:	:	:	:	:
3.3673	1	1	1	1	-1	-1	-1	-1	1
3.0681	1	1	1	1	-99	1	1	1	-1
3.2771	-1	-99	-1	-1	-1	-1	-1	-1	-1

(Satagopan et al. *Genetics* 144: 805-816, 1996)

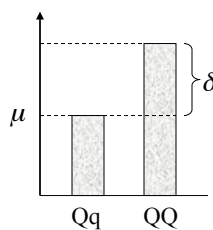
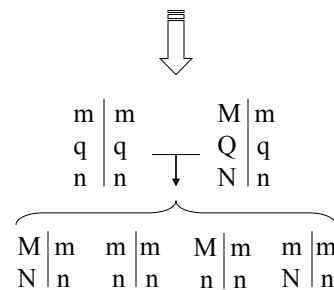
Chrom.	Marker	$\mu$	$\tau$	LRT	F	p-value
1	1	3.184	-0.202	9.379	9.624	0.002 **
1	2	3.204	-0.230	11.378	11.789	0.001 ***
1	3	3.232	-0.266	14.706	15.485	0.000 ***
1	4	3.229	-0.259	13.885	14.562	0.000 ***
1	5	3.240	-0.276	15.554	16.446	0.000 ****
1	6	3.259	-0.307	19.518	21.041	0.000 ****
1	7	3.252	-0.302	19.747	21.312	0.000 ****
1	8	3.257	-0.318	23.450	25.775	0.000 ****
1	9	3.258	-0.330	25.156	27.884	0.000 ****
1	10	3.252	-0.362	31.518	36.059	0.000 ****

## Interval Mapping

(Lander & Botstein, 1989)



Backcross



phenotype

$$y_i = \mu + q_i \delta + \varepsilon_i$$

QTL  
genotype

residual

$$q_i = \begin{cases} 0, & \text{if } qq \\ 1, & \text{if } Qq \end{cases}$$

## Interval Mapping

If  $\varepsilon_i \sim N(0, \sigma^2) \rightarrow y_i | q_i \sim N(\mu + q_i \delta, \sigma^2)$

$$p(y_i | q_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - \mu - q_i \delta)^2\right\}$$

$$L(\mu, \delta, \sigma^2, \lambda, \mathbf{q} | \mathbf{y}) \propto \prod_{i=1}^N [f(y_i | q_i = 0) \Pr(q_i = 0) + f(y_i | q_i = 1) \Pr(q_i = 1)]$$

$$L(\mu, \delta, \sigma^2, \lambda, \mathbf{q} | \mathbf{y}) \propto \prod_{i=1}^N \left[ \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - \mu)^2\right\} \Pr(q_i = 0 | \lambda) \right. \\ \left. + \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - \mu - \delta)^2\right\} \Pr(q_i = 1 | \lambda) \right]$$

QTL position ↗

## Interval Mapping

$\Pr(q_i | \lambda)$  is modeled in terms of recombinations between flanking markers and QTL:

Marker Genotypes	$\Pr(q_i = QQ)$	$\Pr(q_i = Qq)$
M,N	$(1 - r_1)(1 - r_2)/(1 - r)$	$r_1 r_2 / (1 - r)$
M,n	$(1 - r_1) r_2 / r$	$r_1 (1 - r_2) / r$
m,N	$r_1 (1 - r_2) / r$	$(1 - r_1) r_2 / r$
m,n	$r_1 r_2 / (1 - r)$	$(1 - r_1)(1 - r_2)/(1 - r)$

Approximation:  
(no double recombination)

Markers	$\Pr(q_i = QQ)$	$\Pr(q_i = Qq)$
M,N	1	0
M,n	$(1 - p)$	$p$
m,N	$p$	$(1 - p)$
m,n	0	1

$$p = \frac{r_1}{r}$$

## Interval Mapping

- ⇒ **Likelihood estimation:** EM algorithm to estimate parameters, including  $\lambda$  (position of QTL)
- ⇒ **Alternatively:** Fix  $\lambda$  (grid search) and evaluate LOD

$$\text{LOD}_\lambda = \log_{10} \left[ \frac{L(\hat{\mu}, \hat{\delta}, \hat{\sigma}^2, \hat{q} | y)}{L(\hat{\mu}, \hat{\sigma}^2, \hat{q} | y, \delta = 0)} \right]$$

- ⇒ A QTL is detected whenever the LOD score gets larger than a threshold; estimated position of the QTL maximizes LOD

## Interval Mapping

### REGRESSION APPROACH

(Haley & Knott, 1992)

$$y = X\beta + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ \vdots & \vdots \\ p_{N1} & p_{N2} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

alternatively  $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & p_{12} \\ 1 & p_{22} \\ \vdots & \vdots \\ 1 & p_{N2} \end{bmatrix} \begin{bmatrix} \mu \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$

$$\hat{\beta} = (X'X)^{-1}X'y$$

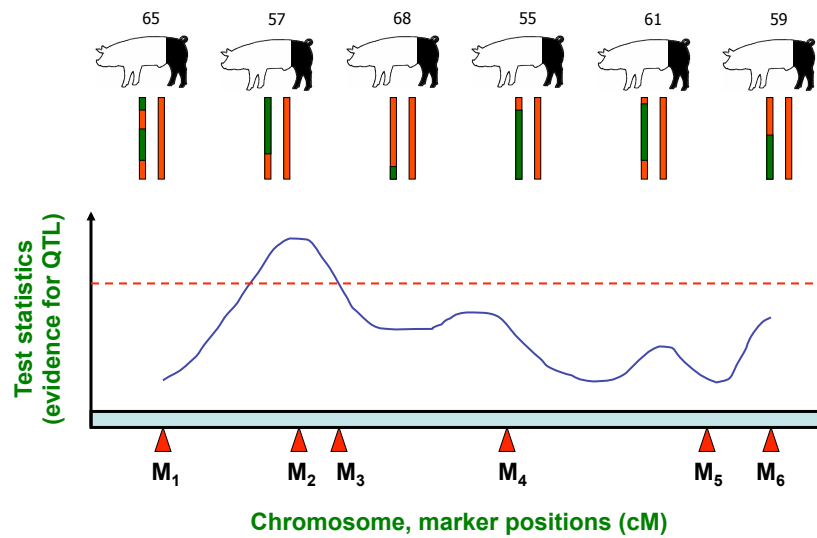
**Residual Sum of Squares:**

$$\text{RSS} = y'y - \hat{\beta}'X'y$$

Estimated position of the QTL minimizes RSS.

# QTL Mapping

## Interval Mapping; Example with Backcross



## Interval Mapping

### ⇒ COMMENTS:

- ① Backcross to both parental lines, or use F<sub>2</sub> design, to estimate additive and dominance effects
- ② Threshold; multiple testing; false positives
- ③ Confidence intervals
- ④ Multiple QTL, ghost QTL



## Interval Mapping Example

R/QTL package in R: Simulated backcross data (Broman and Saunak, 2009) with 400 individuals (200 males and 200 females; sex == 1 and 0, respectively) with a single quantitative phenotype.

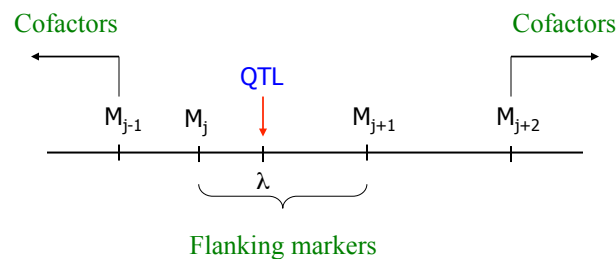
Interval mapping with sex as an additive covariate and sex as an interactive covariate, and also with males and females separately. Detection of regions of the genome affecting the phenotype, and also QTL  $\times$  sex interactions?



## Composite Interval Mapping

(Zeng, 1993, 1994)

- ⇒ Interval analysis adding marker cofactors (to account for the effects of unlinked QTLs); combination of single interval mapping and multiple linear regression



# Composite Interval Mapping

(Zeng, 1993, 1994)

$$y = X\beta + \varepsilon$$

↓

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$y_i = \beta_0 + \beta^* x_{ij} + \sum_{k \neq j, j+1} \beta_k w_{ik} + \varepsilon_i$$

Intercept

Dummy variables

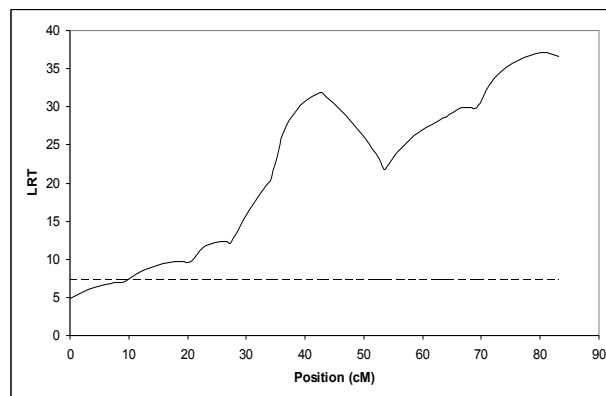
Genetic effect of the putative QTL (between markers j and j+1)

$$X = \begin{bmatrix} 1 & x_{1j} & w_{11} & \cdots & w_{1p} \\ 1 & x_{2j} & w_{21} & \cdots & w_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{Nj} & w_{N1} & \cdots & w_{Np} \end{bmatrix}$$

# Interval Mapping

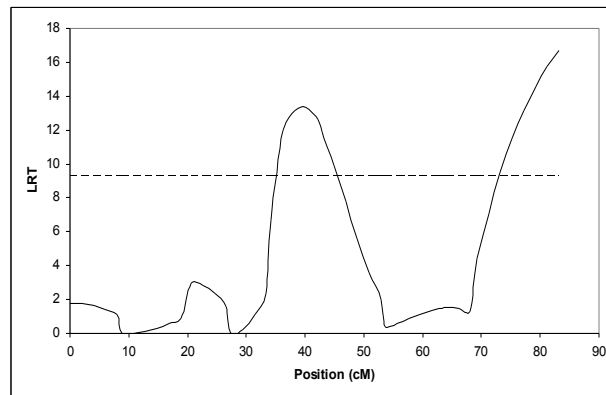
(Example)

⇒ *Brassica napus*; Flowering time (Satagopan et al., 1996)



## Composite Interval Mapping (Example)

→ *Brassica napus*; Flowering time (Satagopan et al., 1996)



## QTL Database (Livestock)

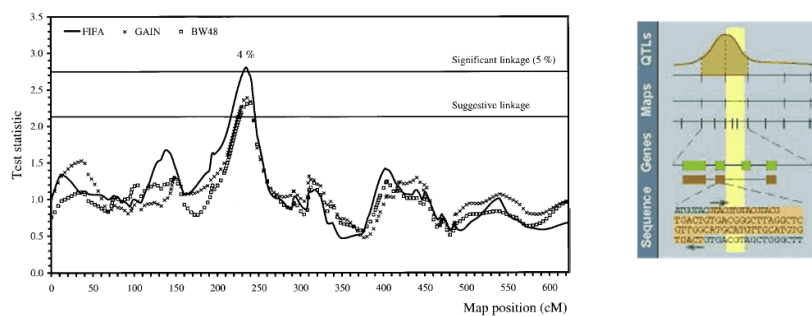
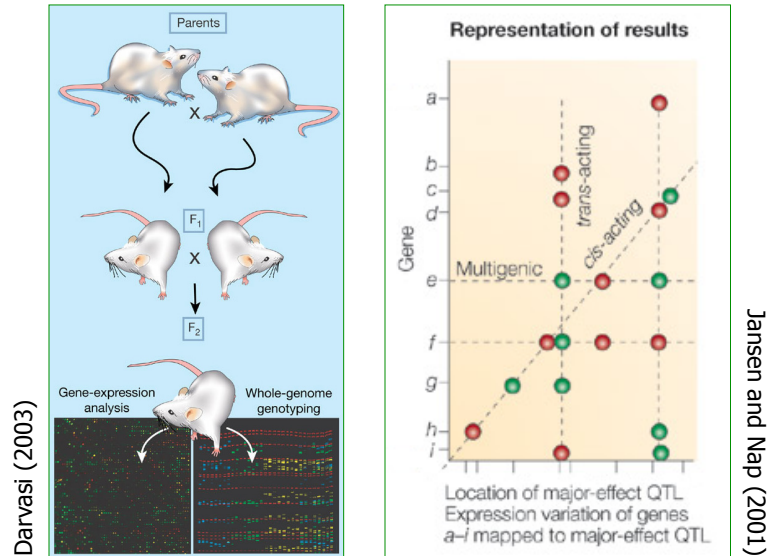


FIGURE 1. Test statistic values from the analysis of body weight at 48 d (BW48), growth between 23 and 48 d (GAIN), and feed intake between 23 and 48 d (FIFA) for quantitative trait loci on Chromosome 1. Significant and suggestive linkage thresholds of FIFA are included. The thresholds for BW48 and GAIN were slightly higher. Map positions are given using the Haldane scale.

## EXPRESSION QTL (eQTL)



## Genome-Wide Association Analysis (GWAS)

Guilherme J. M. Rosa

University of Wisconsin-Madison

## Gene Mapping

- ⇒ Linkage Analysis (QTL Analysis)
- ⇒ Fine Mapping Strategies (LDLA approach, Selective Genotyping, etc.)
- ⇒ Association Analysis, Candidate Gene Approach
- ⇒ Genome-wide Association Analysis (GWAS)

## High Density SNP Panels

- ⇒ Many species: humans, plants, animals
- ⇒ Technology (Affymetrix, Illumina, etc.)
- ⇒ Genome-wide Association Analysis (GWAS),  
Genome-wide Marker Assisted Selection (GWMAS),  
Population Structure, Selection Signature, etc.

## Descriptive Statistics & Data Cleaning

- ⇒ Measurement/recording error
- ⇒ Genotyping error; Mendelian inconsistencies
- ⇒ Redundancies
- ⇒ Heterozygosity (H)  
Polymorphism Information Content (PIC)
- ⇒ Minor Allele Frequency (MAF)
- ⇒ Hardy-Weinberg equilibrium

## Single Marker Regression

- ⇒ Series of models, one for each marker  $j$  ( $j = 1, 2, \dots, k$ ):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{m}g_j + \mathbf{e}$$

where:

$\mathbf{y}$ : vector of phenotypic observations ( $n$  individuals)

$\boldsymbol{\beta}$ : environmental covariates, such as gender, age, etc.

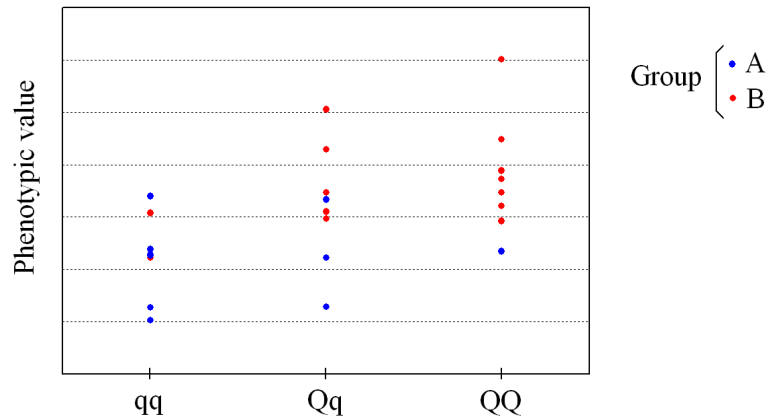
$\mathbf{X}$ : incidence matrix relating  $\boldsymbol{\beta}$  to  $\mathbf{y}$

$g_j$ : 'effect' of marker  $j$  ( $j = 1, 2, \dots, k$ )

$\mathbf{m} = [m_{1j}, m_{2j}, \dots, m_{nj}]^T$ : vector of genotypes for marker  $j$ , with  $m_{ij} = -1, 0$  or  $1$

$\mathbf{e}$ : residual vector

## Confounding



⇒ True model:  $y_{ij} = \mu + \text{Group}_i + e_{ij}$

## Accounting for Population Stratification

⇒ Series of models, one for each marker  $j$  ( $j = 1, 2, \dots, k$ ):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi} + \mathbf{m}g_j + \mathbf{e}$$

where:  $\boldsymbol{\Psi}$  is a population structure term (e.g. PC built from genotypes)



## Mixed Model Approach

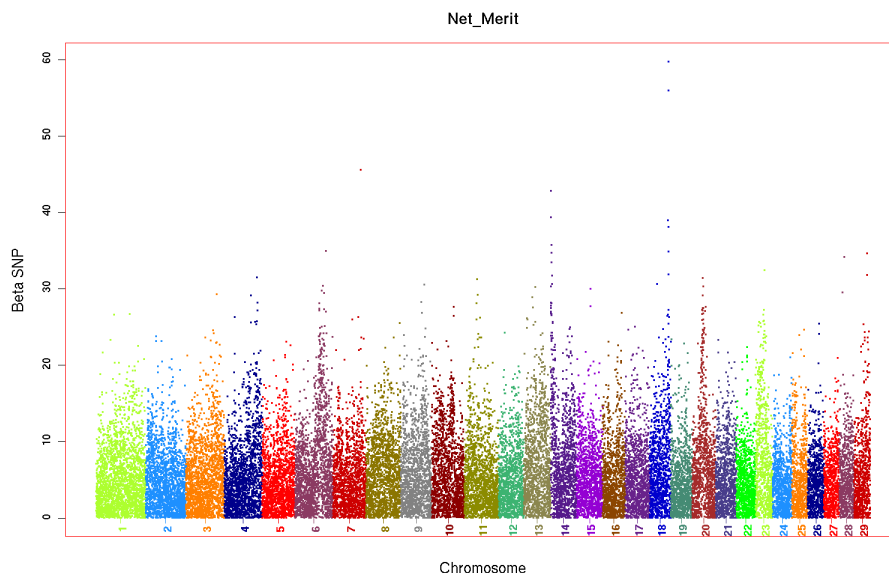
⇒ The model now is expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{m}g_j + \mathbf{e}$$

where all terms are as before, except that a polygenic (infinitesimal) term  $\mathbf{u}$  is included to account for population sub-structure, with  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{K}\sigma_u^2)$ ;  $\mathbf{K}$  is a kinship matrix built from pedigree information (e.g.  $\mathbf{A}$ ) or genotypic information (e.g.  $\mathbf{G}$ )

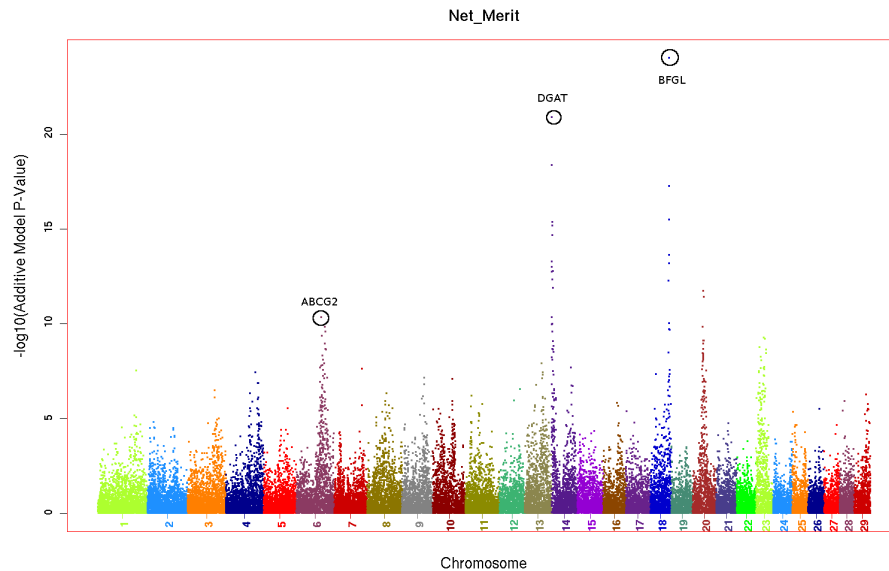
**Note:** Efficient computation, e.g. EMMA and GEMMA

## Manhattan Plot with Marker Effects





## Manhattan Plot with Significance Tests



## Statistical Power

⇒ Power is a function of:

- Significance level ( $\alpha$ )
- Sample size ( $n$ )
- Effect size ( $\delta$ ), expressed as a proportion of variance in measured phenotype, subsumes allele frequency, mode of inheritance, measurement reliability, degree of LD, and all other aspects of genetic model
- Test statistic ( $T$ )

# Hypothesis Testing

	$H_0$ is not rejected	$H_0$ is rejected
$H_0$ is true	No error ( $1-\alpha$ )	Type I error ( $\alpha$ )
$H_0$ is false	Type II error ( $\beta$ )	No error ( $1-\beta$ )

Significance level

Power

➔ Standard approach:

- ① Specify an acceptable type I error rate ( $\alpha$ )
- ② Seek tests that minimize the type II error rate ( $\beta$ ),  
i.e., maximize power ( $1 - \beta$ )

## The Multiple Testing Issue

Suppose you carry out 10 hypothesis tests at the 5% level  
(assume independent tests)

{ The probability of declaring a particular test significant under its null hypothesis is 0.05  
 But the probability of declaring at least 1 of the 10 tests significant is 0.401  
 If you perform 20 hypothesis tests, this probability increases to 0.642...

→  $1 - 0.95^{10}$

- ➔ Typically thousands of markers tested simultaneously
- ➔ Example: Suppose trait with  $H^2 = 0$  and association analysis considering 100 markers and  $\alpha = 5\%$  (for each test)
  - Expected  $100 \times 0.05 = 5$  false associations...

## The Multiple Testing Issue

	# $H_0$ not rejected	# $H_0$ rejected	
# true $H_0$	A	B	$m_0$
# false $H_0$	C	D	$m_1$
	$m - R$	R	m

Observable quantity (n° rejected  $H_0$ )      known quantity (number of tests)

## The Multiple Testing Issue

- Family-wise error rate (FWER):

$$\text{FWER} = \Pr(B \geq 1) = 1 - \Pr(B = 0)$$

- False discovery rate (FDR):

$$\text{FDR} = E[B/R \mid R > 0] \Pr(R > 0)$$

Positive FDR (pFDR); Storey (2002)

## ➡ Controlling the FWER at level $\alpha$ :

$$\Pr[V \geq 1]$$

- **Bonferroni**: Rejects any hypothesis  $H_j$  with p-value less than or equal to  $\alpha/m$ , i.e.:

$$\tilde{p}_j = \min[mp_j, 1]$$

adjusted p-value

unadjusted p-value

- **Sidak**: Rejects any hypothesis  $H_j$  with p-value less than or equal to  $1-(1-\alpha)^{1/g}$ , i.e.:

$$\tilde{p}_j = \min[1 - (1 - p_j)^g, 1]$$

- Very similar to Bonferroni adjustment.
- Both are too conservative...

## ➡ Controlling the FDR:

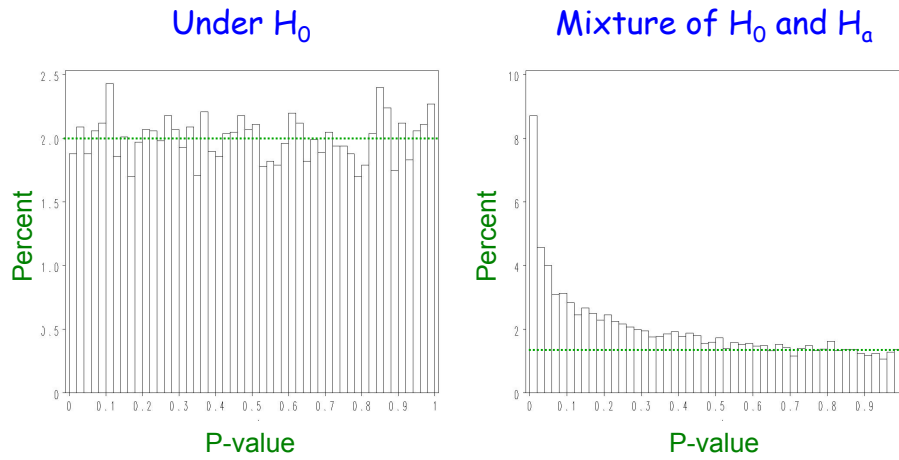
**Definition**:  $\text{FDR} = E[V/R \mid R > 0] \Pr[R > 0]$ ; expected proportion of false positive findings among all rejected hypotheses times the probability of making at least one rejection.

Positive FDR (pFDR); Storey (2002)

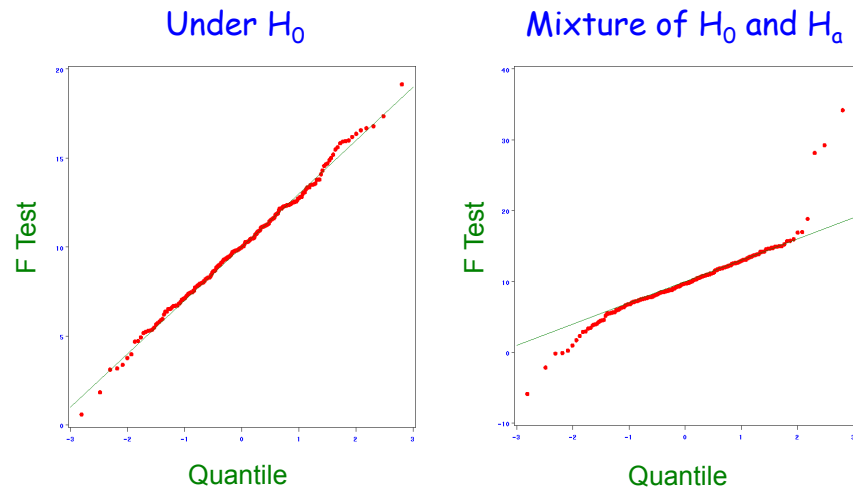
- **Benjamini and Hochberg (1995) algorithm**:

- Fix a value  $\alpha^* \in (0, 1)$
- Let  $p_{(1)}, p_{(2)}, \dots, p_{(m)}$  be the ordered observed p-values
- Let  $\hat{k} = \max\{k: p_{(k)} \leq \alpha^*(k/m)\}$   
(If  $p_{(k)} > \alpha^*(k/m)$  for all  $k = 1, \dots, m$ , let  $\hat{k} = 0$ )
- If  $\hat{k} \geq 1$ , reject the hypotheses corresponding to  $p_{(1)}, p_{(2)}, \dots, p_{(\hat{k})}$
- If  $\hat{k} = 0$ , do not reject any hypothesis

## Distribution of P-values (Histogram)



## Distribution of P-values (Q-Q Plot)



## Replication

- ⇒ Confounding factors, population structure and stratification, Type I error, etc.
- ⇒ Biased estimates of gene effects due to significance threshold
- ⇒ Multiple genes, with modest individual effects
- ⇒ Gene  $\times$  gene and gene  $\times$  environment interactions
- ⇒ Inter population heterogeneity
- ⇒ Low statistical power
- ⇒ Validation of association findings
- ⇒ But what constitutes a replication?

# Lecture 10

## Multi-Trait Models, Binary and Count Traits, Genome-enhanced prediction

Guilherme J. M. Rosa

University of Wisconsin-Madison

Introduction to Quantitative Genetics

SISG, Seattle

17 - 19 July 2017

## OUTLINE

- Animal Model
- Multiple-trait Model
- Repeatability Model
- Maternal Effects
- Generalized Linear Models
- Genome-enhanced Prediction

## Animal Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Diagram illustrating the Animal Model equation  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ . The components are labeled as follows:

- $\mathbf{y}$ : responses
- $\mathbf{X}$ : incidence matrices
- $\boldsymbol{\beta}$ : fixed effects
- $\mathbf{Z}$ : breeding values
- $\mathbf{u}$ : breeding values
- $\mathbf{e}$ : residuals

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix} \right)$$

3

## Mixed Model Equations

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$$

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1-h^2}{h^2}$$

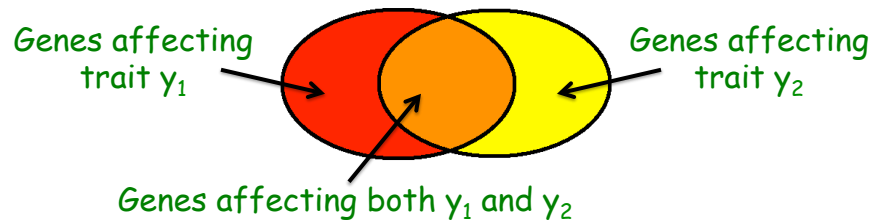
BLUP:  $\hat{\mathbf{u}} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$

4



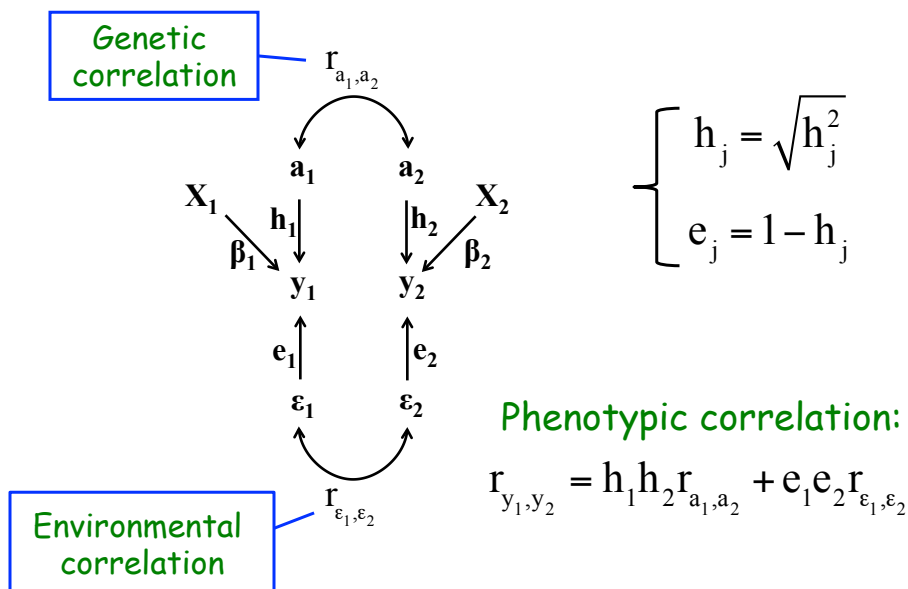
## Genetic Correlation

### Schematic representation of pleiotropy



- Pleiotropic genes affect both  $y_1$  and  $y_2$  resulting in a genetic correlation between the two traits
- In addition to pleiotropy, genetic correlations can be caused also by linkage disequilibrium (LD) between genes affecting the different traits. LD however is a 'temporary' cause of genetic correlation as recombination can breakdown LD over the generations

## Multiple (Correlated) Traits



## Multiple (Correlated) Traits

The animal model can be extended for the joint analysis of multiple traits

Let the model for each of k traits be:

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{a}_j + \boldsymbol{\varepsilon}_j$$

where j is an index to indicate the trait (j = 1, 2, ..., k).

For the joint analysis of the k trait, the model becomes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$$

with design matrices given by:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_k \end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_k \end{bmatrix}$$

## Multiple (Correlated) Traits

In this case it is assumed that:

$$\text{Var} \begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \otimes \mathbf{I} \end{bmatrix}$$

where  $\mathbf{G}$  and  $\boldsymbol{\Sigma}$  are the genetic and residual variance-covariance matrices, given by:

$$\mathbf{G} = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} & \cdots & \sigma_{a_1 a_k} \\ \sigma_{a_1 a_2} & \sigma_{a_2}^2 & \cdots & \sigma_{a_2 a_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1 a_k} & \sigma_{a_2 a_k} & \cdots & \sigma_{a_k}^2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_1 \varepsilon_2} & \cdots & \sigma_{\varepsilon_1 \varepsilon_k} \\ \sigma_{\varepsilon_1 \varepsilon_2} & \sigma_{\varepsilon_2}^2 & \cdots & \sigma_{\varepsilon_2 \varepsilon_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_1 \varepsilon_k} & \sigma_{\varepsilon_2 \varepsilon_k} & \cdots & \sigma_{\varepsilon_k}^2 \end{bmatrix}$$

**Note:**  $\otimes$  represents the direct (Kronecker) product

## Multiple (Correlated) Traits

The MME for multi-trait analyses are of the same form as before, i.e.:

$$\begin{bmatrix} X'(\Sigma^{-1} \otimes I)X & X'(\Sigma^{-1} \otimes I)Z \\ Z'(\Sigma^{-1} \otimes I)X & Z'(\Sigma^{-1} \otimes I)Z + G^{-1} \otimes A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'(\Sigma^{-1} \otimes I)y \\ Z'(\Sigma^{-1} \otimes I)y \end{bmatrix}$$

from which the BLUEs and BLUPs of  $\beta$  and  $a$  can be obtained.

## Multiple (Correlated) Traits

The dimensionality of multi-trait MME, however, can become a hurdle for solving it when more than two or three traits are considered

An alternative for the analysis of multiple traits is to use a [canonical transformation](#) of the traits, which consists of transforming the vectors of correlated traits into a new vector of uncorrelated variables

In such case, each transformed variable can be analyzed independently using standard single trait models, and subsequently the estimated breeding values are transformed back to the original scale of measurement

## Repeatability Model



## Repeatability Model

For the analysis of repeated measurements, environmental effects can be partitioned into **permanent** and **temporary effects**

In this case, the mixed model, usually called 'repeatability model', can be written as:

$$y = X\beta + Za + Wp + \epsilon$$

where  $p \sim N(0, I\sigma_p^2)$  is the vector of permanent environmental effects, with each level pertaining to a common effect to all observations of each animal

## Repeatability Model

It is often assumed that  $\mathbf{a}$ ,  $\mathbf{p}$ , and  $\boldsymbol{\varepsilon}$ , which are independent from each other

Under these assumptions, the MME becomes:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda_a \mathbf{A}^{-1} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} + \lambda_p \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

with  $\lambda_a = \sigma_{\boldsymbol{\varepsilon}}^2 / \sigma_a^2$  and  $\lambda_p = \sigma_{\boldsymbol{\varepsilon}}^2 / \sigma_p^2$

## Repeatability Model

An important definition related to repeated measurements refers to repeatability ( $r$ ), which is given by the intraclass correlation, i.e., the ratio of the within-individual (or between repeated measurements) to the phenotypic variances:

$$r = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_y^2} = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_a^2 + \sigma_p^2 + \sigma_{\boldsymbol{\varepsilon}}^2}$$

The repeatability coefficient measures the correlation between records on the same animal, and so it is useful for example in the estimation of **producing ability** of an animal

## Maternal Effects



## Maternal Effects

There are some traits of interest in livestock, such as weaning weight in beef cattle, in which progeny performance is affected by the dam's ability to affect the calf's environment, such as in the form of nourishment through her milk production, the quantity and quality of which is in part genetically determined

In such cases, dams contribute to the performance of their progeny not only through the genes passed to the progeny (the "direct genetic effects") but also through their ability to provide a suitable environment (the "indirect genetic effects")

## Maternal Effects

Maternally influenced traits can be analyzed by using a model as:

$$y = X\beta + Za + Km + Wp + \epsilon$$

where  $\mathbf{m}$  is a vector of random maternal genetic effects, and  $\mathbf{p}$  is a vector of random maternal permanent environmental effects

It is assumed that  $\mathbf{m} \sim N(\mathbf{0}, A\sigma_m^2)$  and  $\mathbf{p} \sim N(\mathbf{0}, I\sigma_p^2)$ , and quite often a covariance structure between direct and maternal additive genetic effects is considered, assumed equal to  $A\sigma_{a,m}$

## Computing Strategies

Solving the MME does not necessarily require the inversion of the coefficient matrix  $\mathbf{C}$

More computationally convenient alternatives for solving high dimensional systems of linear equations include methods based on iteration on the MME, such as the Jacobi or Gauss-Seidel iteration, and the "iteration on the data" strategy, which is commonly used methodology in national genetic evaluations involving millions of records

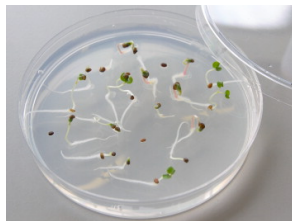
## Generalized Linear Mixed Models

The models discussed so far assumed a Gaussian (normal) distribution of the phenotypic traits

Often however phenotypic traits are expressed as a binary (e.g., pregnancy in dairy cattle, or germination in seeds) or count variable (e.g., litter size in swine, or fruits in trees)

In such cases the linear (Gaussian) model is not appropriate, and a generalized linear model (GLM) approach is necessary

## Generalized Linear Mixed Models





## Generalized Linear Mixed Models

GLM can actually model outcomes (response variables) generated from any distribution from the **exponential family**, which includes the normal, binomial, Poisson and gamma distributions, among others

The GLM consists of three elements:

1. **Probability distribution** from the exponential family.
2. **Linear predictor**  $\eta = X\beta$
3. **Link function**  $g$  such that  $E(Y) = \mu = g^{-1}(\eta)$ .

## Generalized Linear Mixed Models

Notice that the Gaussian model is a specific case of the GLM, with the normal distribution and an identity link function

In the case of Generalized Linear Mixed Models, including the applications in animal/plant breeding, the model is defined as:

1. **Probability distribution** from the exponential family.
2. **Linear predictor**  $\eta = X\beta + Zu$
3. **Link function**  $g$  such that  $E(Y|u) = \mu = g^{-1}(\eta)$

## GLMM in R

GLMM can be implemented in R using the package `lme4`

`lme4`, however, assumes independence between levels of random effects, and as such it is not suitable for many animal/plant breeding applications

`pedigreemm` is an R package that uses `lme4` with a Cholesky decomposition strategy to overcome this problem

## pedigreemm

An R package for fitting generalized linear mixed models in animal breeding

$$\begin{aligned} g(\mu_{Y|U}) &= Z\mathbf{u} + X\boldsymbol{\beta} \\ \mu_{Y|U} &= E[Y|U = \mathbf{u}] \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2) \\ \mathbf{u}^* &= \mathbf{L}^{-1}\mathbf{u} \longrightarrow g(\mu_{Y|U}) = Z\mathbf{L}(\mathbf{L}^{-1}\mathbf{u}) + X\boldsymbol{\beta} = \mathbf{Z}^*\mathbf{u}^* + X\boldsymbol{\beta} \\ \mathbf{A} &= \mathbf{L}\mathbf{L}' \quad \mathbf{u}^* \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2) \end{aligned}$$

(Harville and Callanan 1989)

## Technical note: An R package for fitting generalized linear mixed models in animal breeding<sup>1</sup>

A. I. Vazquez,<sup>\*2</sup> D. M. Bates,<sup>†</sup> G. J. M. Rosa,<sup>\*</sup> D. Gianola,<sup>\*‡</sup> and K. A. Weigel<sup>\*</sup>

<sup>\*</sup>Department of Dairy Science, <sup>†</sup>Department of Statistics, and <sup>‡</sup>Department of Animal Sciences, University of Wisconsin, Madison 53706

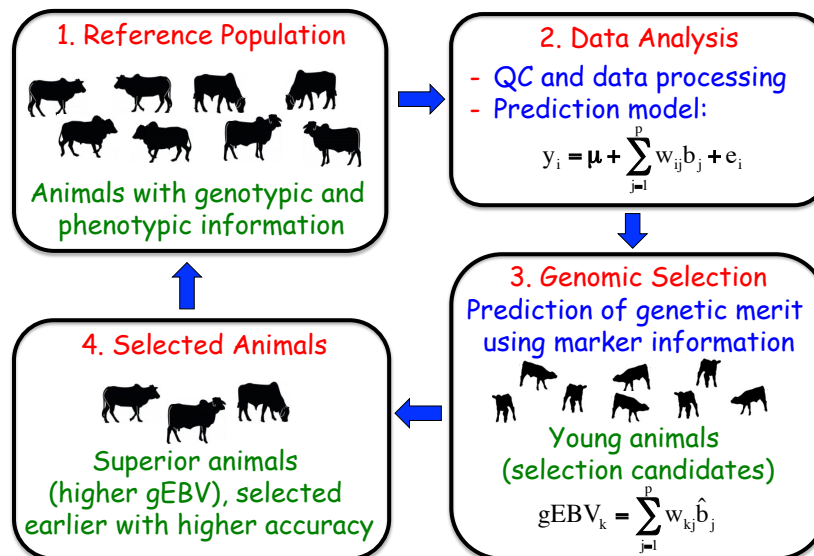
© 2010 American Society of Animal Science. All rights reserved.

J. Anim. Sci. 2010. 88:497–504  
doi:10.2527/jas.2009-1952

**Data Set 1.** Milk production records of 3,397 lactations from first- through fifth-parity Holsteins were available. These records were from 1,359 cows, daughters of 38 sires in 57 herds. Records are in the *milk* data set in the *pedigreemm* package. The data were downloaded from the USDA site (<http://www.aipl.arsusda.gov/>). All lactation records represent cows with at least 100 d in milk, with an average of 347 d. Milk yield ranged from 4,065 to 19,345 kg estimated for 305 d, averaging 11,636 kg. There were 1,314, 1,006, 640, 334, and 103 records for first-, second-, third-, fourth-, and fifth-lactation animals, respectively. A 5-generation pedigree of the cows with a total of 6,547 animals was used in the analysis (<http://www.aipl.arsusda.gov/>). The pedigree information is available in the *pedCows* and *pedCowsR* pedigree objects also included in the package; the second one is a lighter pedigree (with 70% of the information on *pedCows*). The milk production data used in the first 2 examples are described below.

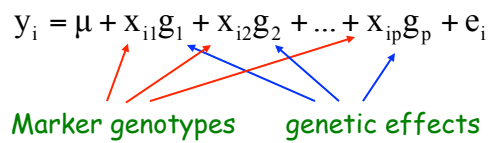


## Genome-enhanced Selection



# Genome-enhanced Selection

(Meuwissen et al., 2001)

$$y_i = \mu + x_{i1}g_1 + x_{i2}g_2 + \dots + x_{ip}g_p + e_i$$


Marker genotypes      genetic effects

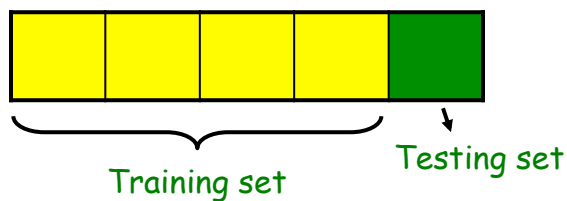
Genomic EBV:  $GEBV = x_{i1}\hat{g}_1 + x_{i2}\hat{g}_2 + \dots + x_{ip}\hat{g}_p = \sum_{j=1}^p x_{ij}\hat{g}_j$

- ⇒ 'big p small n paradigm'
- ⇒ Dimension reduction techniques (e.g. SVD and PLS), and stepwise strategies
- ⇒ Alternatively: penalized regression, shrinkage estimation

## Cross-validation

(Predictive Ability)

→ K-fold



$$\begin{cases} y = X\beta + e \\ \hat{\beta}: \text{estimate of } \beta \end{cases} \rightarrow \begin{cases} PMSE = \frac{1}{m} \sum_i (y_i - \hat{y}_i)^2 \\ \hat{y} = X\hat{\beta} \end{cases}$$

→ Leave-one-out ("n-fold")

## GBLUP

Regression with genetic effects with normal distribution with common variance

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^p \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad , \text{ with: } \mathbf{g}_j | \sigma_g^2 \sim N(0, \sigma_g^2)$$

### Equivalent Model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{b} + \mathbf{e} \quad , \text{ with: } \mathbf{b} | \sigma_b^2 \sim N(\mathbf{0}, \mathbf{G}\sigma_b^2)$$

⇒  $\mathbf{G}$  is the genomic relationship matrix (VanRaden 2008):

$$\mathbf{G} = \left( 2 \sum_{j=1}^p p_j (1 - p_j) \right)^{-1} (\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})'$$

## ssGBLUP

- Single-step GBLUP (Misztal et al. 2009)
- Single mixed model with all animals (genotyped and non-genotyped) included, with matrix  $\mathbf{A}$  replaced by  $\mathbf{H}$ :

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

## Preventive and Personalized Medicine

