# Summer Institute in Statistical Genetics

## Module1: Principles of Quantitative Genetics

*Week 1 – Session1*
*Monday 6 – Tuesday 7 February 2017*

***Instructors***
Professor Bruce Walsh – University of Arizona
Professor Steven Chenoweth – University of Queensland

The University of Queensland
Australia
IMB
Institute for Molecular Bioscience

# SYLLABUS
# PRINCIPLES OF QUANTITATIVE GENETICS

INSTRUCTORS:

Steve Chenoweth, School of Biological Sciences, UQ
s.chenoweth@uq.edu.au

Bruce Walsh, Department of Ecology & Evolutionary Biology, University of Arizona
jbwalsh@u.arizona.edu

## LECTURE SCHEDULE

**Monday, 6 Feb 2017**

| | |
|---|---|
| 8:30 10:00 am | 1. Introduction to complex traits (Walsh) |
| | Background reading: LW Chapter 4 |
| 10:00 10:30 am | Break |
| 10:30 12:00 | 2. Resemblance Between Relatives (Walsh) |
| | Background reading: LW Chapter 7 |
| 12:00 1:30 pm | Lunch |
| 1:30 3:00 pm | 3. Estimating variances (Walsh) |
| | Background reading: LW Chapter 7 |
| 3:00 3:30 pm | Break |
| 3:30 5:00 pm | 4. Artificial Selection (Walsh) |
| | Background reading: WL Chapter 13 |
| | Additional reading: WL Chapters 14-16 |

**Tuesday 7 Feb 2017**

| | |
|---|---|
| 8:30 10:00 am | 5. Inbreeding and Crossbreeding (Walsh) |
| | Background reading: LW Chapter 10 |
| 10:00 10:30 am | Break |
| 10:30 12:00 | 6. Correlated Characters (Chenoweth) |
| | Background reading: L&W chapter 21, chapter 3 |
| | W&L Chapter 34 (correlated responses) |
| 12:00 1:30 pm | Lunch |
| 1:30 3:00 pm | 7. Evolutionary Quantitative genetics (Chenoweth) |
| | Background reading: W&L v1. chapters 28, 29, 34 |
| | Additional reading: W&L v1. Chapter 27 |

| | | |
|---|---|---|
| 3:00 | 3:30 pm | Break |
| 3:30 | 5:00 pm | 8. QTL/Association Mapping (Walsh)<br>Background reading: LW Chapters 15, 16 |

Website for draft chapters from “Volume 2”: Walsh & Lynch: Evolution and Selection on Quantitative traits
http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html

## ADDITIONAL BOOKS ON QUANTITATIVE GENETICS

### General

Falconer, D. S. and T. F. C. Mackay. *Introduction to Quantitative Genetics*, 4th 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*. (3rd 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*. 5th 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* (4th 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. Konig. 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.

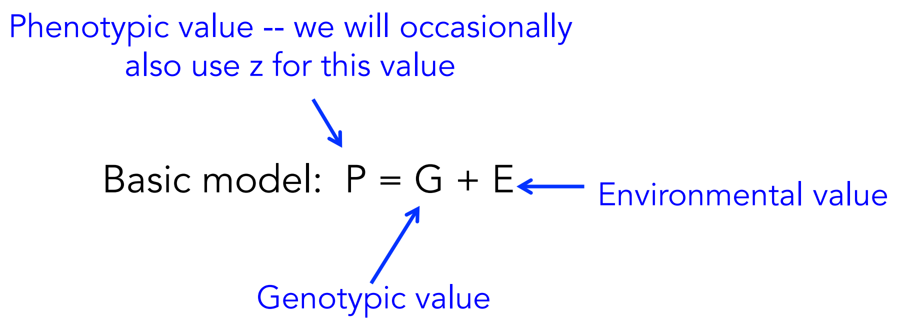
Saxton, A. M. (Ed). 2004*. Genetic Analysis of Complex Traits Using SAS*. SAS Press.

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

# Lecture 1:
# Introduction to Quantitative Genetics

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

## Basic model of Quantitative Genetics


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

Basic model: P = G + E

Environmental value

Genotypic value

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.

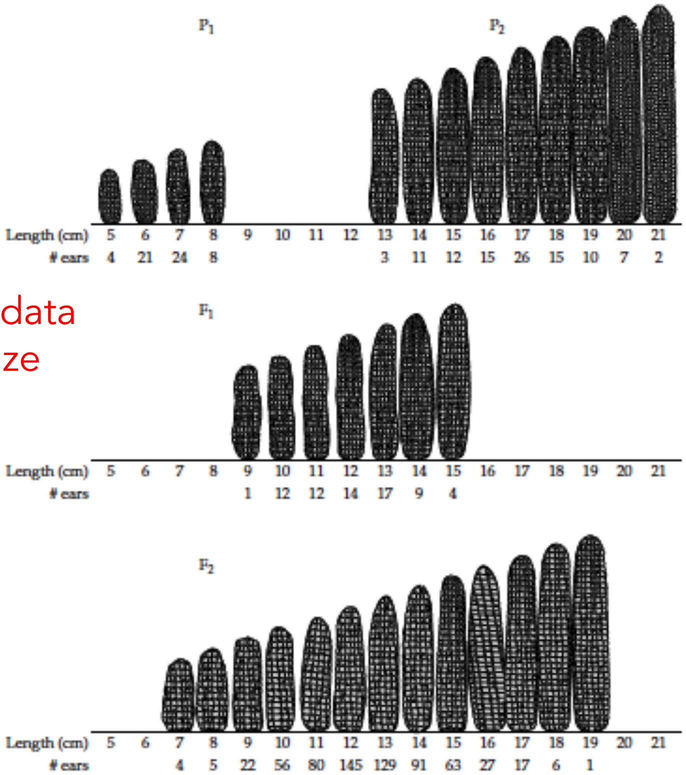
# Basic model of Quantitative Genetics

Basic model: P = G + E

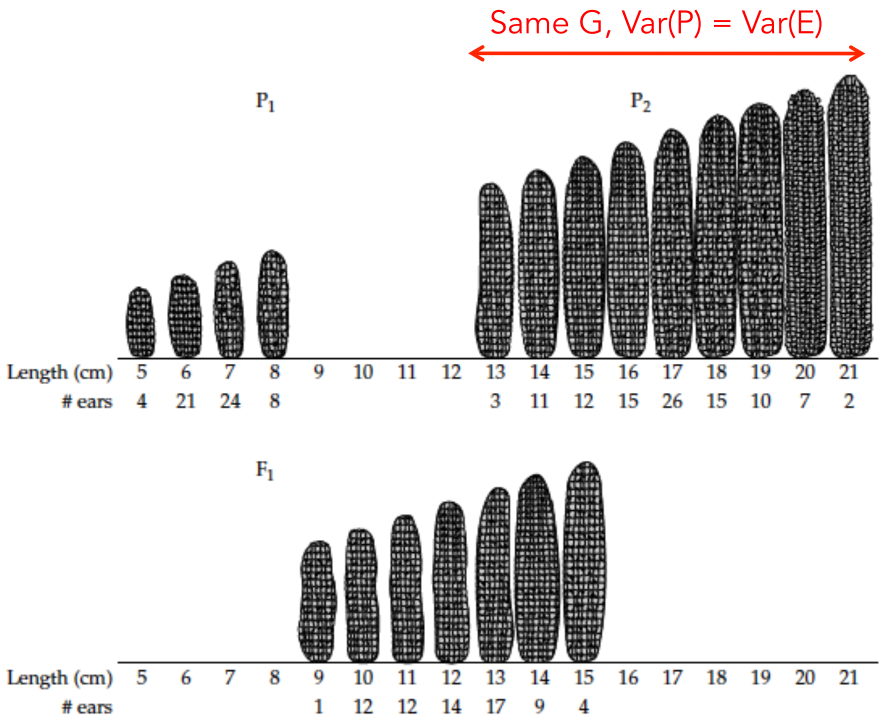
G = average phenotypic value for that genotype if we are able to replicate it over the universe of environmental values, G = E[P]

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

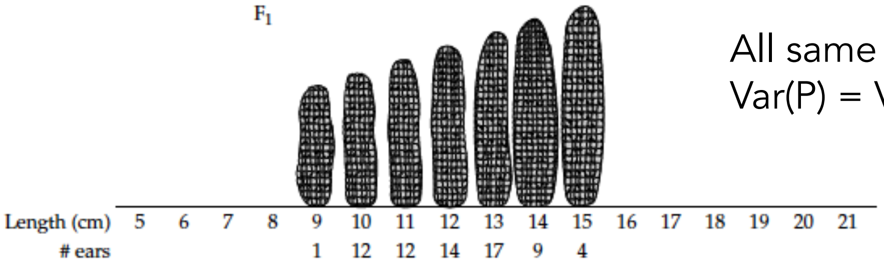


East (1911) data on US maize crosses

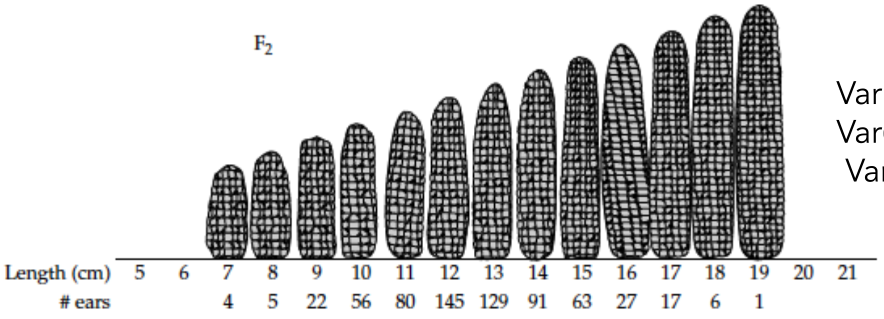


Same G, Var(P) = Var(E)

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



All same G, hence Var(P) = Var(E)



Variation in G
Var(P) = Var(G) + Var(E)

Var($F_2$) > Var($F_1$) due to Variation in G

# 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, <u>within</u> a given line, size of parental seed independent of size of offspring speed
  - No variation in G <u>within</u> a line

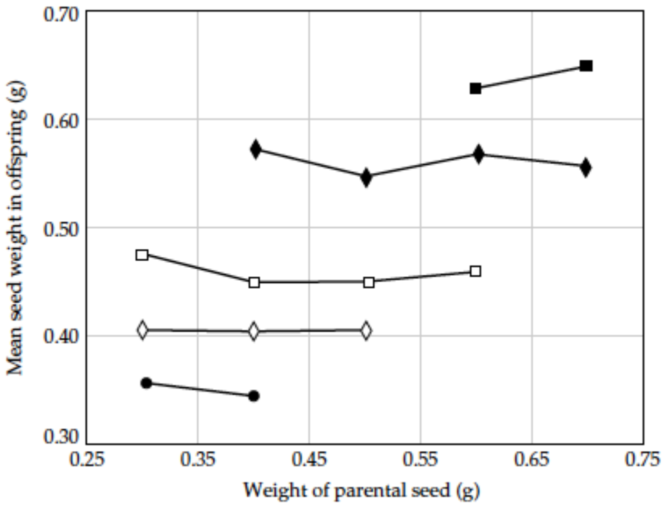


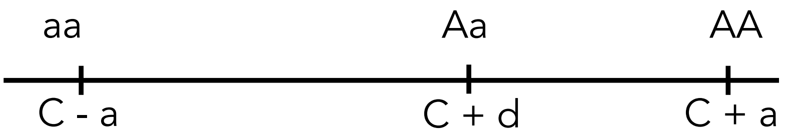
**Figure 1.4** Mean offspring seed size as a function of parental seed size for some of Johannsen's pure lines. The data for the different lines are denoted by different symbols. If there is a heritable component to seed weight within a pure line, a line with positive slope is expected — larger parents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

# 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)

# 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

# Computing a and d

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

| Genotype | aa | Aa | AA |
|---|---|---|---|
| Trait value | 10 | 15 | 16 |

$$
\begin{aligned}
C &= [G(AA) + G(aa)]/2 = (16+10)/2 = 13 \\
a &= [G(AA) - G(aa)]/2 = (16-10)/2 = 3 \\
d &= G(Aa)] - [G(AA) + G(aa)]/2 \\
&= G(Aa)] - C = 15 - 13 = 2
\end{aligned}
$$

# Population means: Random mating

Let p = freq(A), q = 1-p = freq(a). Assuming random-mating (Hardy-Weinberg frequencies),

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

$$
\begin{aligned}
\text{Mean} &= q^2(C - a) + 2pq(C + d) + p^2(C + a) \\
\mu_{RM} &= C + a(p-q) + d(2pq)
\end{aligned}
$$

Contribution from homozygotes

Contribution from heterozygotes

# 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 |

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

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

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 $F_2$.

# 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 an 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 |

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

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

# 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.
  - $\alpha_A$ = mean(offspring value given parent transmits A) - mean(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.

# 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 |
|---|---|---|---|
| A | p | AA | C + a |
| a | q | Aa | C + d |

Mean(A transmitted) = p(C + a) + q(C + d) = C + pa + qd

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

# Random mating

Now suppose parent contributes a

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

Mean(a transmitted) = p(C + d) + q(C - a) = C - qa + pd

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

# $\alpha$, 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

# 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)

# 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 $A_iA_j$ individual

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

Average contribution to genotypic value for allele i

Mean value $\mu_G = \Sigma\, G_{ij}\, \text{Freq}(A_iA_j)$

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

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

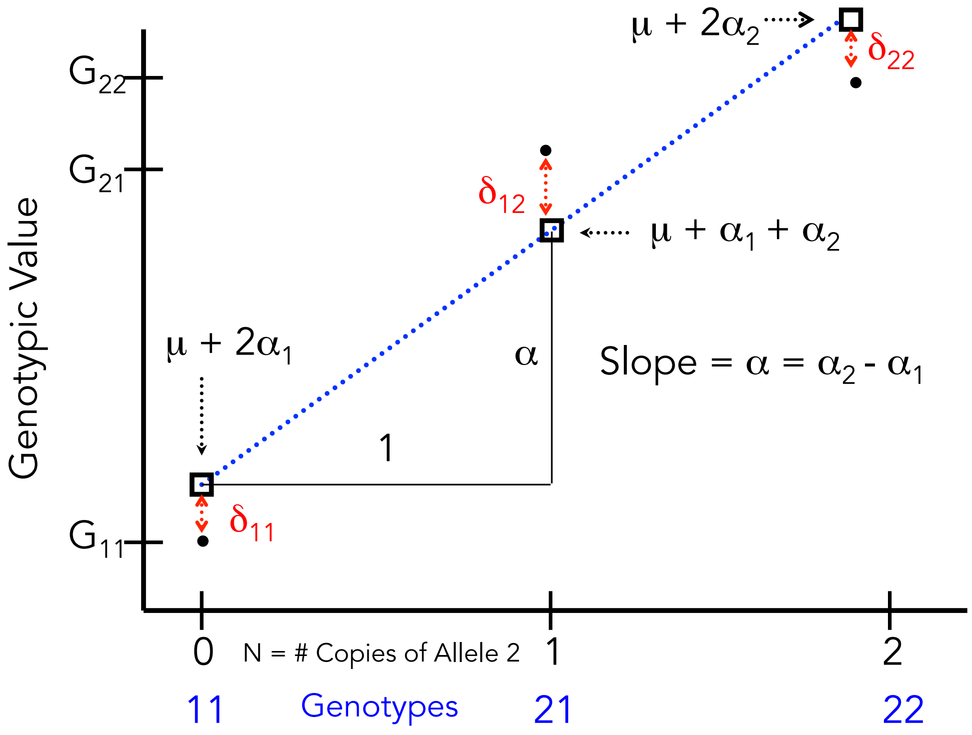
The genotypic value predicted from the individual allelic effects is thus $\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$

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

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 $A_iA_j$) between the genotypic value predicted from the two single alleles and the actual genotypic value,

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



# Fisher’s decomposition is a Regression

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

Predicted value

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 = # of $A_2$ alleles

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

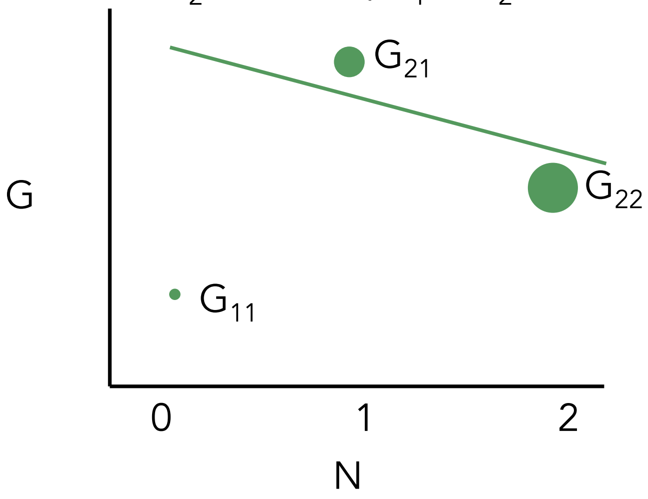
$$G_{ij} = \mu_G + 2\alpha_1 + (\alpha_2 - \alpha_1)\,N + \delta_{ij}$$

Intercept Regression slope

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

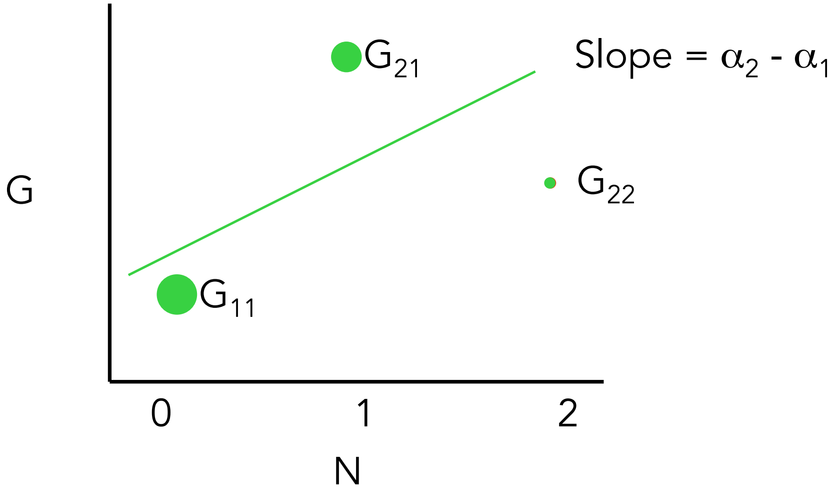
A key point is that the average effects change with allele frequencies. Indeed, if overdominance is present they can change sign with allele frequencies.

Allele $A_2$ common, $\alpha_1 > \alpha_2$



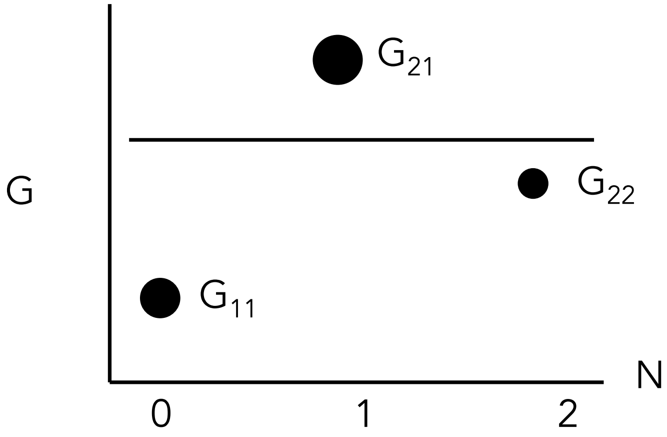
The size of the circle denotes the weight associated with that genotype. While the genotypic values do not change, their frequencies (and hence weights) do.

Allele $A_1$ common, $\alpha_2 > \alpha_1$



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

Both $A_1$ and $A_2$ frequent, $\alpha_1 = \alpha_2 = 0$



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

## 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

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

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

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

A is called the Breeding value or the Additive genetic value

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

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

# 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

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

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

# 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,

$$\sigma^2_G = \sigma^2_A + \sigma^2_D$$

# 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
- Var(A) = additive genetic variance
  - Variance in additive genetic values
  - Property of a population
- Can estimate A or Var(A) without knowing any of the underlying genetical detail (forthcoming)

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

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

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

One locus, 2 alleles: $\sigma^2_A = 2p_1 p_2 a^2 [1 + \underset{\uparrow}{k} (p_1 - p_2)]^2$

Dominance alters additive variance

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

Dominance variance

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

$$\sigma^2_D = E[\delta^2] = \sum_{i=1}^{m}\sum_{j=1}^{m}\delta^2_{ij}\,p_i\,p_j$$

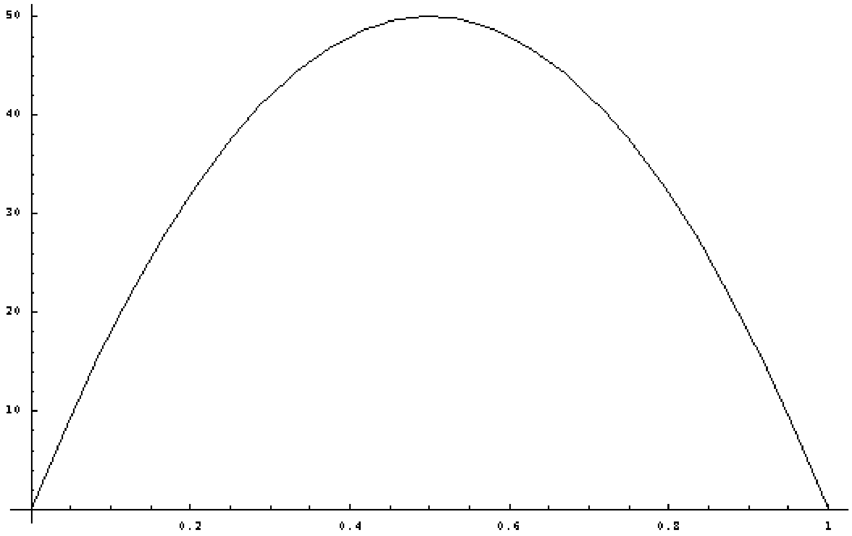
Equals zero if k = 0

**One locus, 2 alleles:** $\sigma^2_D = (2p_1\,p_2\,ak)^2$

This is a symmetric function of allele frequencies

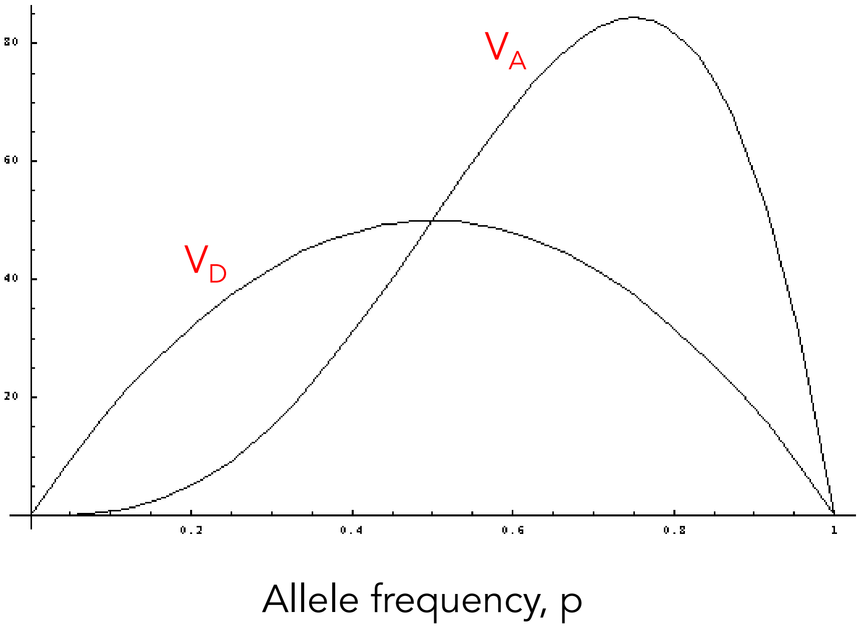
Can also be expressed in terms of d = ak

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



Allele frequency, p

Complete dominance (k = 1)



# 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

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD}$$

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

Additive x Additive interactions -- αα, AA
interactions between a single allele
at one locus with a single allele at another

Additive x Dominance interactions -- αδ, 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 --- δδ, DD
the interaction between the dominance
deviation at one locus with the dominance
deviation at another.

# Lecture 2: Resemblance and relatedness

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

# 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

# Broad-sense heritability

- Narrow-sense heritability $h^2$ applies when outcrossing,
  - $h^2 = Var(A)/Var(P)$
  - = 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
  - $H^2 = Var(G)/Var(P)$
  - = the fraction of all trait variation due to variation in Genotypic values

# 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

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

$$\sigma^2(\bar{z}_i) = \sigma^2_G + \sigma^2_E + \frac{\sigma^2_{GE}}{e} + \frac{\sigma^2_p}{er} + \frac{\sigma^2_e}{er\,n}$$

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

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

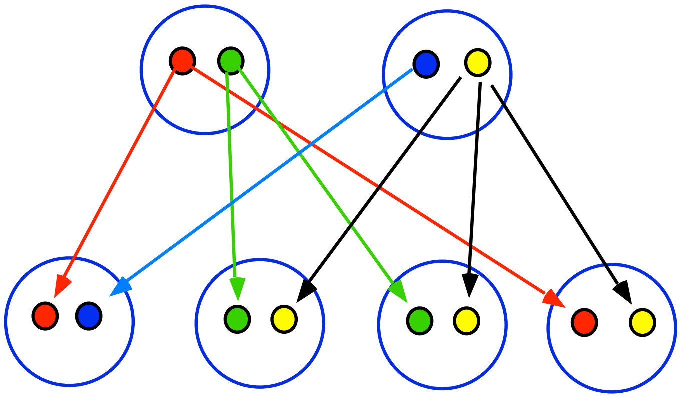
# Key observations

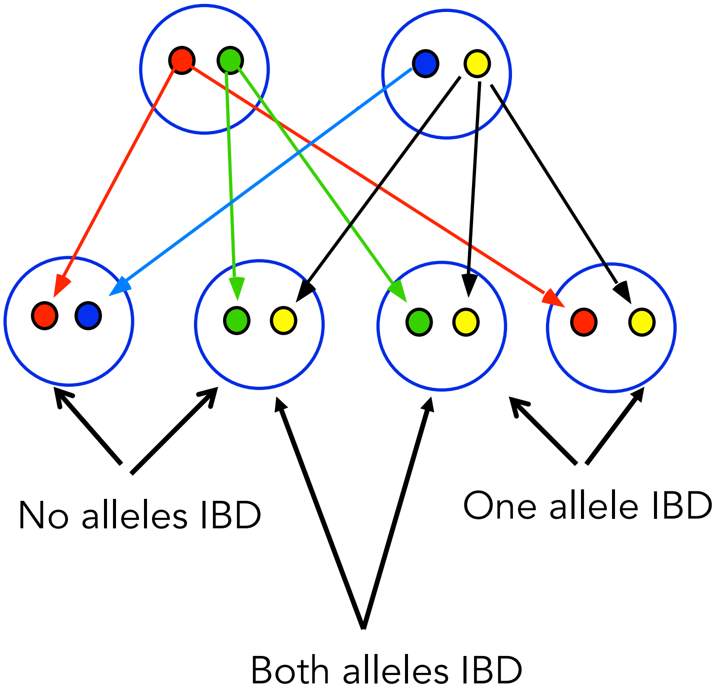
- The amount of phenotypic resemblance among relatives for the trait provides an indication of the amount of genetic variation for the trait.
- If trait variation has a significant genetic basis, the closer the relatives, the more similar their appearance
- The covariance between the phenotypic value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

# 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
  - $\mathrm{Cov}(z_x,z_y) = a_{xy}V_A + b_{xy}V_D$.
  - 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

# Parent-offspring genetic covariance

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

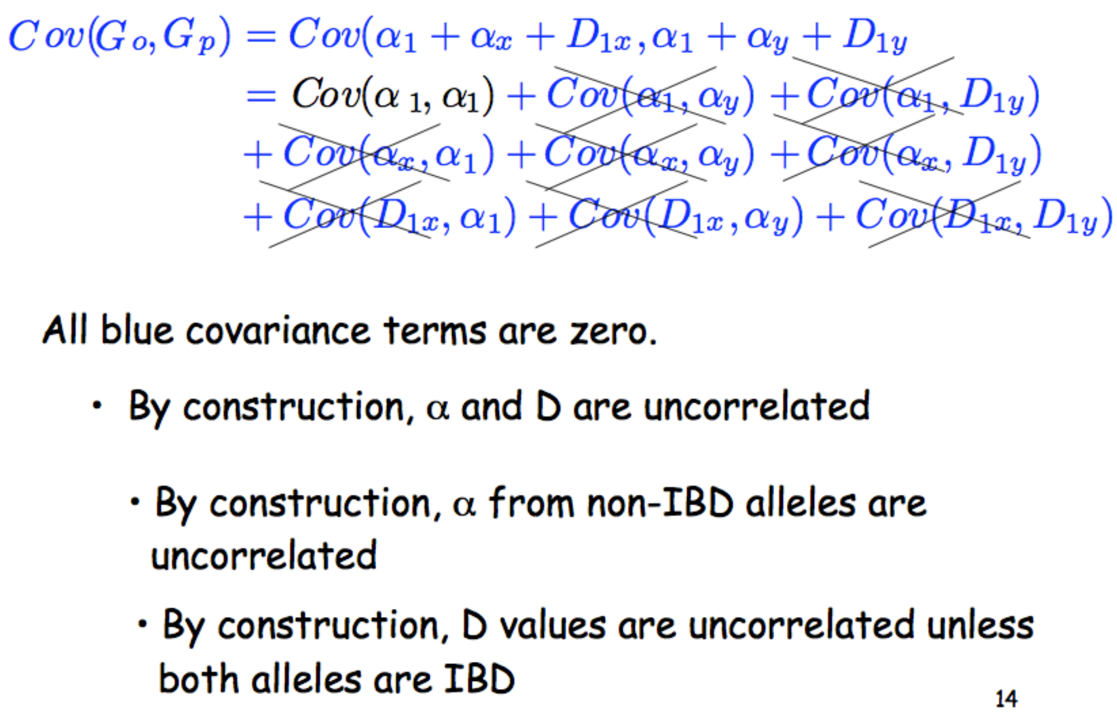
Denote this common allele by $A_1$

$$G_p = A_p + D_p = \alpha_1 + \alpha_x + D_{1x}$$

$$G_o = A_o + D_o = \alpha_1 + \alpha_y + D_{1y}$$

IBD allele

Non-IBD alleles



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

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

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

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

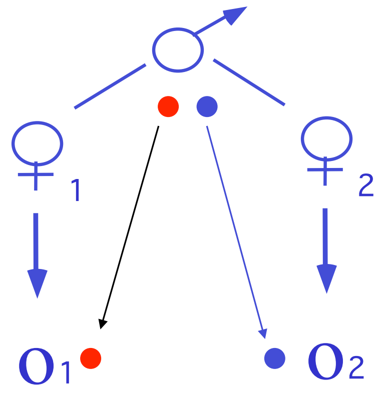
so that

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

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

# Half-sibs



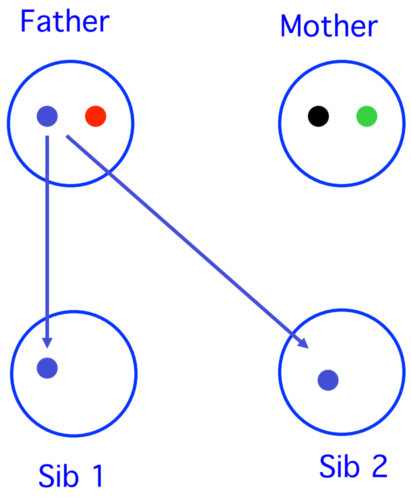
Each sib gets exactly one allele from common father, different alleles from the different mothers

The half-sibs share no alleles IBD

- occurs with probability 1/2

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

# Full-sibs

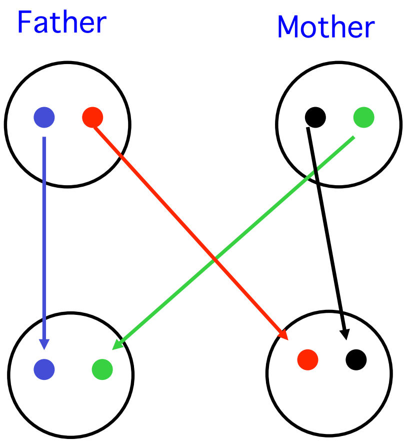


Each sib gets exact one allele from each parent

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

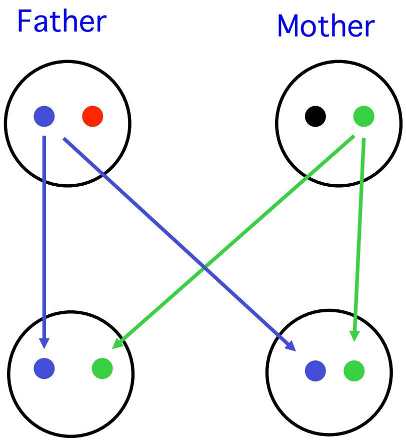
Prob(Allele from father not IBD) = 1/2. Given the allele in parent one, prob = 1/2 that sib 2 gets different allele

# Full-sibs



Each sib gets
exact one allele
from each parent

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



Each sib gets
exact one allele
from each parent

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-Pr(0) - Pr(2) = 1/2

## Resulting Genetic Covariance between full-sibs

| IBD alleles | Probability | Contribution |
|---|---|---|
| 0 | 1/4 | 0 |
| 1 | 1/2 | Var(A)/2 |
| 2 | 1/4 | Var(A) + Var(D) |

Cov(Full-sibs) = Var(A)/2 + Var(D)/4

## Genetic Covariances for General Relatives

Let r = (1/2)Prob(1 allele IBD) + Prob(2 alleles IBD)

Let u = Prob(both alleles IBD)

General genetic covariance between relatives
Cov(G) = rVar(A) + uVar(D)

When epistasis is present, additional terms appear
$r^2$Var(AA) + ruVar(AD) + $u^2$Var(DD) + $r^3$Var(AAA) +

# 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.

# 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, $2\Theta = r =$ 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 $\Theta_{xy}$.

# $\theta$ 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 $A_iA_j$ denote the two alleles in x and $A_kA_l$ those in y.
  - Cov(breeding values) = $\Pr(A_i \text{ ibd } A_k)\, \mathrm{cov}(\alpha_i, \alpha_k) + \Pr(A_i \text{ ibd } A_l)\, \mathrm{cov}(\alpha_i,\alpha_l) + \Pr(A_j \text{ ibd } A_k)\, \mathrm{cov}(\alpha_j, \alpha_k) + \Pr(A_j \text{ ibd } A_l)\, \mathrm{cov}(\alpha_j,\alpha_l) = 4\, \theta_{xy} \mathrm{Var}(\alpha)$
  - Since $\mathrm{Var}(A) = 2\mathrm{Var}(\alpha)$, $\mathrm{Cov} = 2\, \theta_{xy} \mathrm{Var}(A)$

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

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

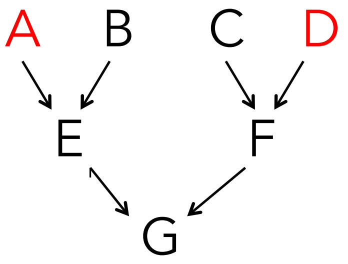
To compute $\Theta_{xx}$, denote the two alleles in x by $A_1$ and $A_2$

| | 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, $f_x$ = prob $A_1$ and $A_2$ IBD, $\boxed{\Theta_{xx} = (1+ f_x)/2}$

# 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 |

# 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)

| | | Offspring | |
|---|---|---|---|
| | | Draw $A_1$ | Draw $A_n$ |
| Parent | Draw $A_1$ | IBD | $\Theta_{P,NP}$ |
| | Draw $A_p$ | $f_p$ | $\Theta_{P,NP}$ |

Prob($A_n$,$A_p$), the alleles from the two parents are IBD, i.e. , offspring is inbred

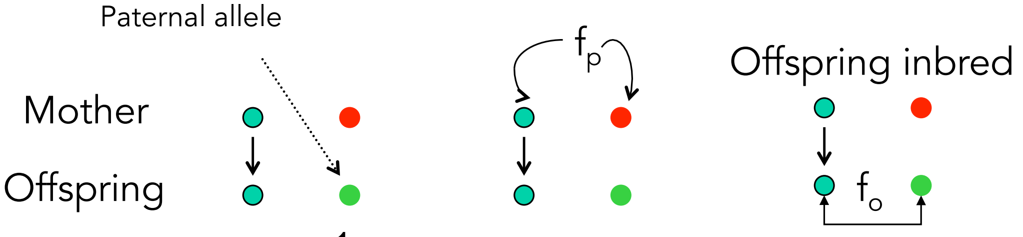
$A_1$, $A_p$ IDB if parent is inbred

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

General: $\boxed{\Theta_{PO} = (1 + f_p + 2\Theta_{P,NP})/4 = (1 + f_p + 2f_o)/4}$

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

Parent inbred

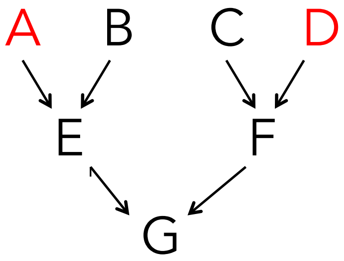


$\theta_{po} = \frac{1}{4}$ $\theta_{po} = \frac{1+f_p}{4}$ $\theta_{po} = \frac{1+2f_o}{4}$

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 $2f_0$



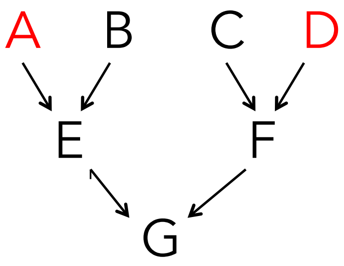
From before

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

Consider A - E (inbred parent - offspring)
$\theta_{AE} = (1+f_A)/4 = (1+1)/4 = 1/2$. Same value for $\theta_{DF}$

Consider B - E (outbred parent - offspring)
$\theta_{BE} = (1+f_B)/4 = (1+0)/4 = 1/4$. Same value for $\theta_{CF}$

Consider E - G (outbred parent - offspring)
$\theta_{EG} = (1+f_E)/4 = (1+0)/4 = 1/4$. Same value for $\theta_{FG}$



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 $\theta_{EF}$ ?

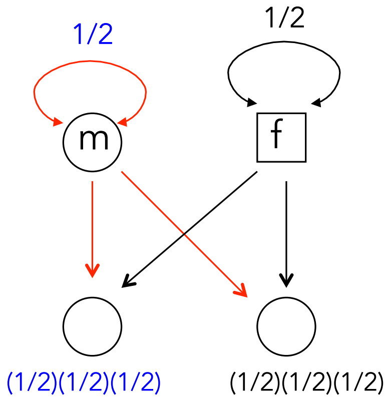
The randomly-chosen allele from E has equal chance of being from A or B. Likewise for F (from C or D)

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

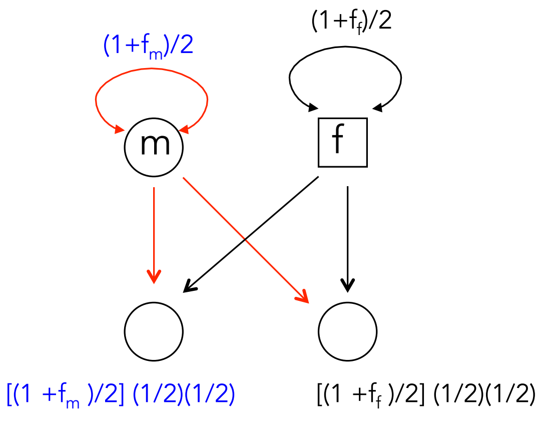
## Full sibs (x and y) from parents m and f

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



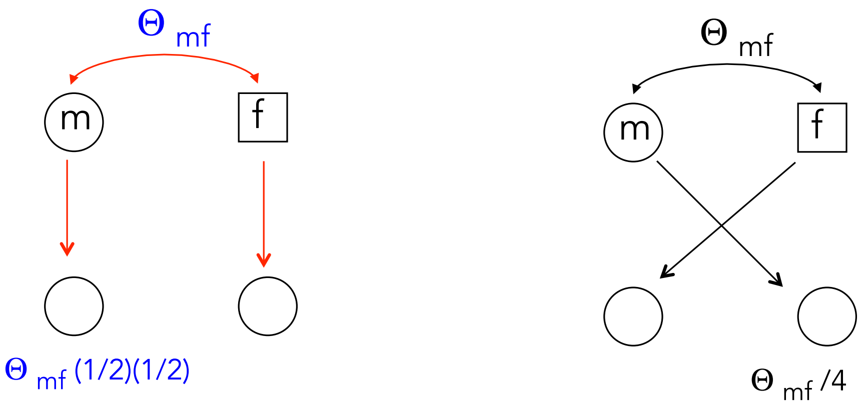
Unrelated, non-inbred
parents

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



Unrelated, inbred
parents

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

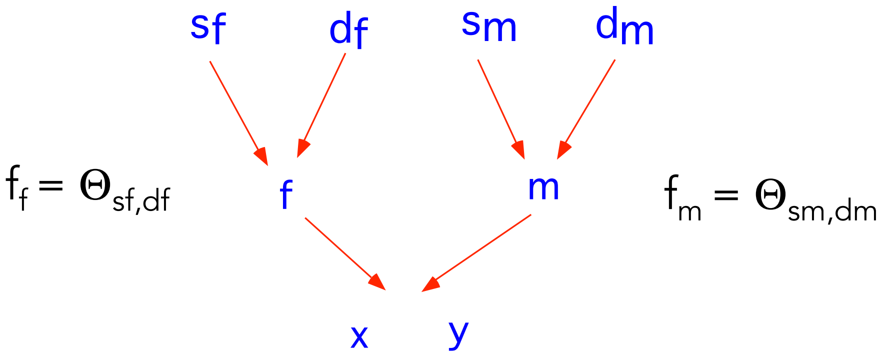


Parents inbred & related.
Two additional paths to add
to $\Theta = (2+f_m+f_f)/8$

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

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

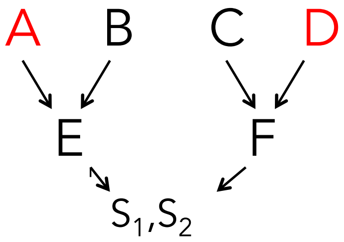
$$\Theta_{xy} = (2 + f_m + f_f + 4\Theta_{mf})/8$$



Putting all this together gives

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

# Example



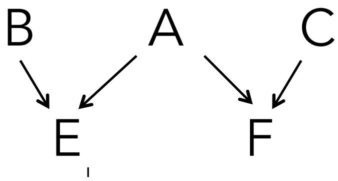
From before

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

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

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

# Half-sibs



A is the common parent

- Using the same arguments as above,

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

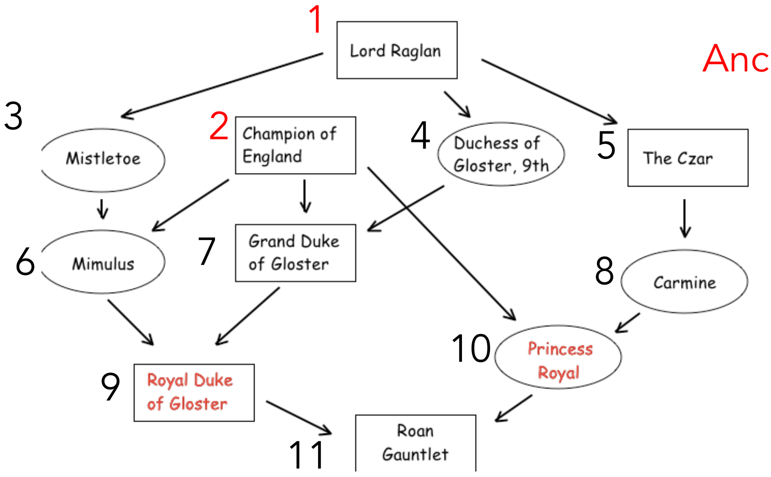
Hence, if B and C unrelated,

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

# 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 $A(i,i) = 1 + A(S_i, D_i)/2$
  - $A(i,j) = A(j,i) = [A(j,S_i) + A(j,D_i)]/2 = [A(i,S_j) + A(i,D_j)]/2$
  - The <u>recursive</u> (or <u>tabular</u>) method starts with the founding parents and then proceeds down the pedigree in a recursive fashion to fill out A for the desired pedigree.

# Example

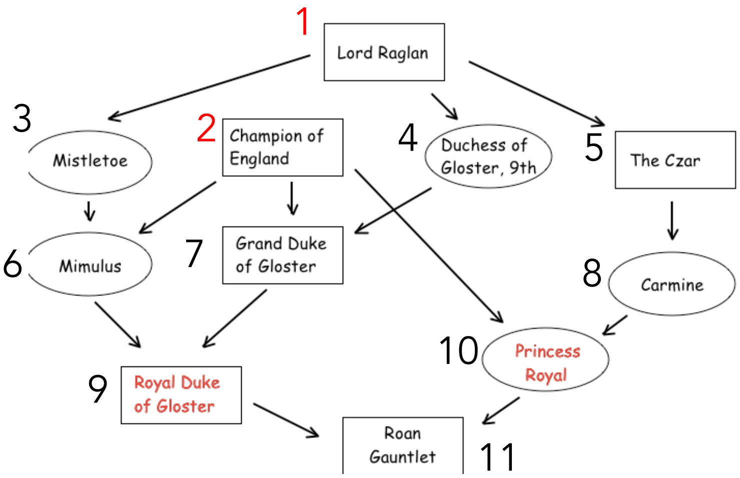


Ancestors are 1 & 2

$A(1,1) = A(2,2) = 1$
$A(1,2) = 0$

3, 4, 5, 8 all have unknown parents (only a single arrow to them)

3: $S_3 = 1$, $D_3$ = 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 $A(2,3) = A(2,4) = A(2,5) = 0$.



So far

|  | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 1/2 | 1/2 | 1/2 |
| 2 | 0 | 1 | 0 | 0 | 0 |
| 3 | 1/2 | 0 | 1 | 1/4 | 1/4 |
| 4 | 1/2 | 0 | 1/4 | 1 | 1/4 |
| 5 | 1/2 | 0 | 1/4 | 1/4 | 1 |

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: $S_9 = 7, D_9 = 6$. $A(9,9) = 1 + A(S_9, D_9)/2 = 1 + A(6,7)/2 = 1 + 5/32 = 1.156$ <- inbred!

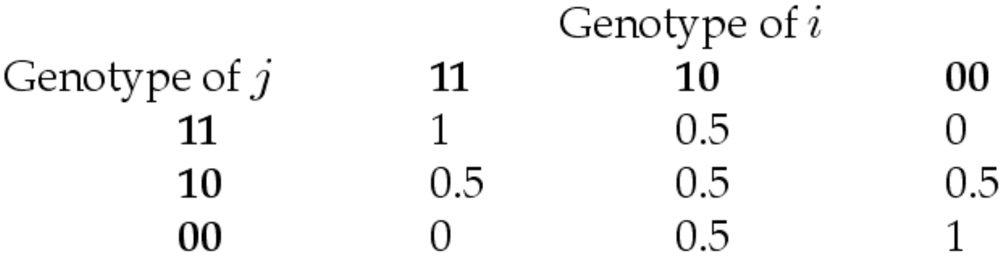
# Actual relatedness versus expected values from pedigrees

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

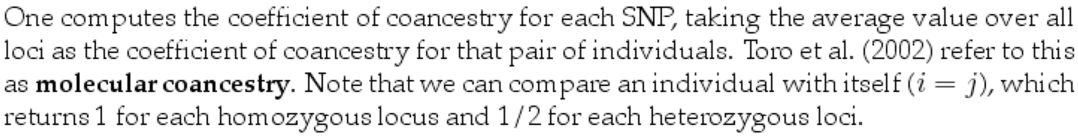
For example, we expect 2θ to be ½ for full subs. However, one pair of sibs may actually be more similar (0.6) and another less similar (say 0.35). On average, 2θ is ½ for pairs of full sibs, but if we knew the actual value of θ, we have more information. With sufficient dense genetic markers, we can estimate these relationships directly.

Genomic selection uses this extra information.

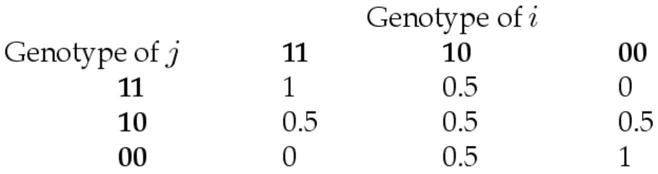
# What about coefficient of coancestry θ ?



|  | Genotype of $i$ |  |  |
|---|---|---|---|
| Genotype of $j$ | **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.



|  | Genotype of $i$ |  |  |
|---|---|---|---|
| Genotype of $j$ | **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 θ | 0.5 | 1.0 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 0.5 | 1.0 | 0.5 |

Estimated θ is the average over all ten loci, = 0.65

# The coefficient of fraternity

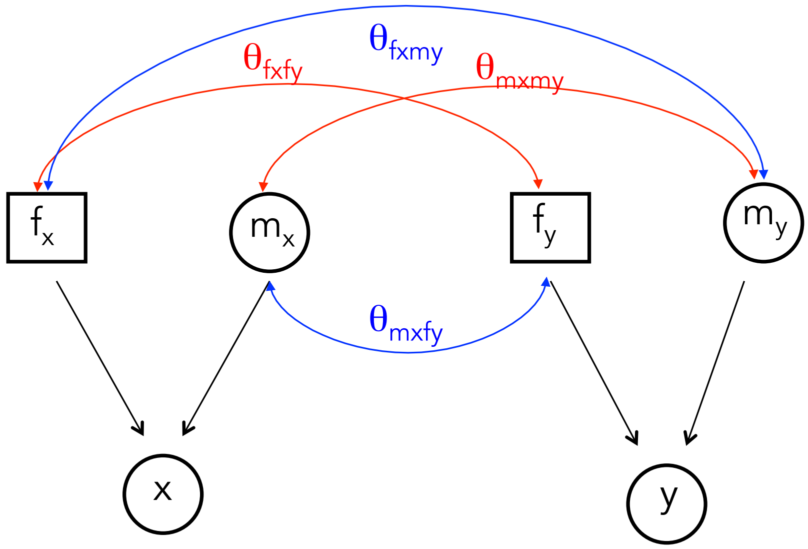
- While (twice) the coefficient of coancestry gives the weight on the additive variance for two relatives, a related measure of IDB status among relatives gives the weight on the dominance variance
- The probability that the two alleles in individual x are IBD to two alleles in individual y is denoted $\Delta_{xy}$, 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.
  - $\Delta_{xy} = \theta_{mxmy}\theta_{fxfy} + \theta_{mxfy}\theta_{fxmy}$

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

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

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



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

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

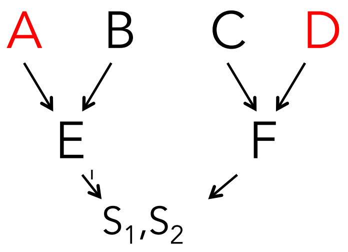
- Full sibs share same mon, dad
  - $m_x = m_y = m$, $f_x = f_y = f$
  - $\Delta_{xy} = \theta_{mxmy}\theta_{fxfy} + \theta_{mxfy}\theta_{fxmy} = \theta_{mm}\theta_{ff} + \theta_{mf}^2$
  - $\Delta_{xy} = (1+f_m)(1+f_f)/4 + \theta_{mf}^2$
- If parents unrelated, $\theta_{fm} = 0$, giving
  - $\Delta_{xy}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

# 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
  - $\Delta_{xy} = 0$

# 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



From before

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

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

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

Giving $\Delta_{xy} = \theta_{EE}\theta_{FF} + \theta_{EF}^2$

$$= (1/2)(1/2) + (1/8)^2$$

$$= 1/4 + 1/64 = 17/64 = 0.266$$

# $\Delta_{xy}$ 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 $A_iA_j$ denote the two alleles in x and $A_kA_l$ 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.
  - Cov(dominance values) = Pr($A_i$ ibd $A_{k;}$ $A_j$ ibd $A_l$ ) cov($\delta_{ij}$, $\delta_{kl}$) + Pr($A_i$ ibd $A_{l;}$ $A_j$ ibd $A_k$)cov($\delta_{ij}$, $\delta_{kl}$)
  - = $(\theta_{fxfy}\theta_{mxmy} + \theta_{mxfy}\theta_{jxmy})$ Var(D) = $\Delta_{xy}$ Var(D)

# 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* 1 = 0.3

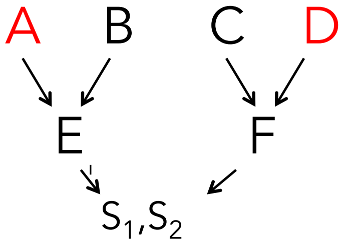
# General Resemblance between relatives

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

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

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

# 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

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

# 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 assocated with each locus

# 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
  - $V_G = V_A + V_D + V_T + V_Q$

# 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

# Lecture 3
# Estimation of genetic variances

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 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 content 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 $h^2$:

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

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

| People | | $h^s$ |
|---|---|---|
| | 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 | 0.3 |
| | Egg production | 0.20 |

Traits more closely associated with fitness tend to have lower heritabilities

# Basic approach to estimating genetic variances

Different crosses are made, which allow us to express the covariance between relatives (which are functions of the genetic variances) with the variance between measured groups. Between-group variances estimated by ANOVA

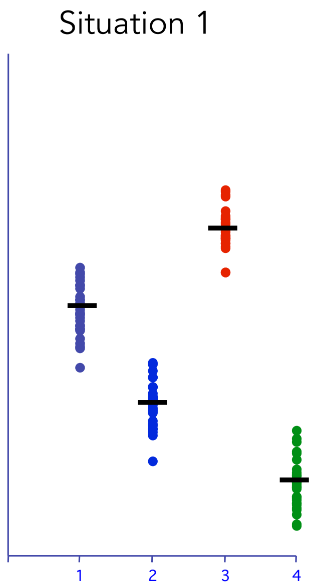
For example, variance between the means of full-sib families = cov(full sibs) = Var(A)/2 + Var(D)/4 + Var(Ec)

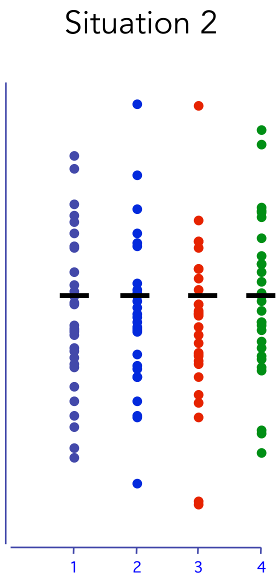
# Types of crosses (mating designs)

- Parent-offspring
- Full sib
- Half sib
- Nested full sib/half sib
  - North Carolina (NC) design one: all males crossed to same set of females
  - NC design two: males crossed to random (different) females
- dialleles

# ANOVA: Analysis of variation

- Partitioning of trait variance into within- and among -group components
- Two key ANOVA identities
  - Total variance = between-group variance + within-group variance
    - Var(T) = Var(B) + Var(W)
  - Variance(between groups) = covariance (within groups)
  - Intraclass correlation, t = Var(B)/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)





Var(B) = 2.5
Var(W) = 0.2
Var(T) = 2.7

t = 2.5/2.7 = 0.93

Var(B) = 0
Var(W) = 2.7
Var(T) = 2.7

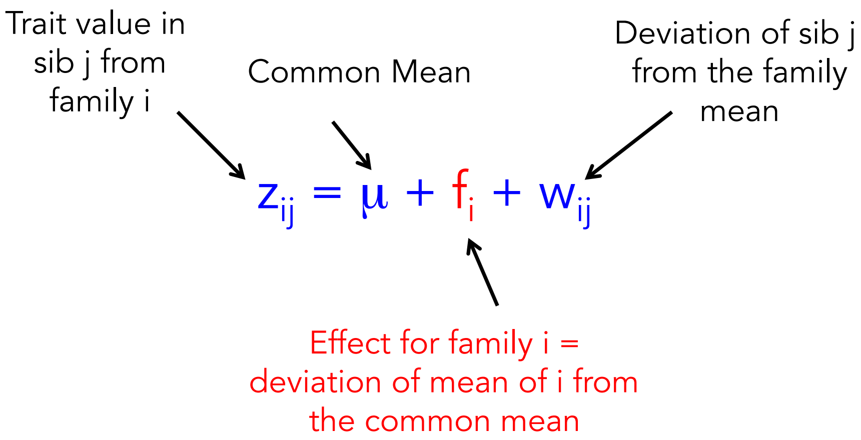
t = 0

# Why cov(within) = variance(among)?

- Let $z_{ij}$ denote the jth 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



Trait value in sib j from family i

Common Mean

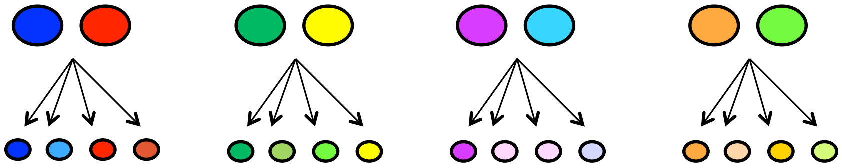
Deviation of sib j from the family mean

$$z_{ij} = \mu + f_i + w_{ij}$$

Effect for family i = deviation of mean of i from the common mean

## Mating Designs

FULL-SIB DESIGN: *N* full-sib families with *n* offspring each.



$$z_{ij} = \mu + \underline{f_i} + \underline{w_{ij}}$$

$z_{ij}$ = phenotype of the j-th offspring of the i-th family

$\mu$ = population mean

$f_i$ = effect of the i-th family

$w_{ij}$ = residual error (segregation, dominance, environmental contribution)

within-family variance

| SoV | df | SS | MS | EMS |
|---|---|---|---|---|
| Among-families | N-1 | $SS_f = n\sum_i \left(\bar{z}_{i.} - \bar{z}_{..}\right)^2$ | $SS_f/df_{(f)}$ | $\sigma^2_{w(FS)} + n\sigma^2_f$ |
| Within-families | n(N-1) | $SS_w = \sum_{i,j} (z_{ij} - \bar{z}_{i.})^2$ | $SS_w/df_{(w)}$ | $\sigma^2_{w(FS)}$ |

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

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

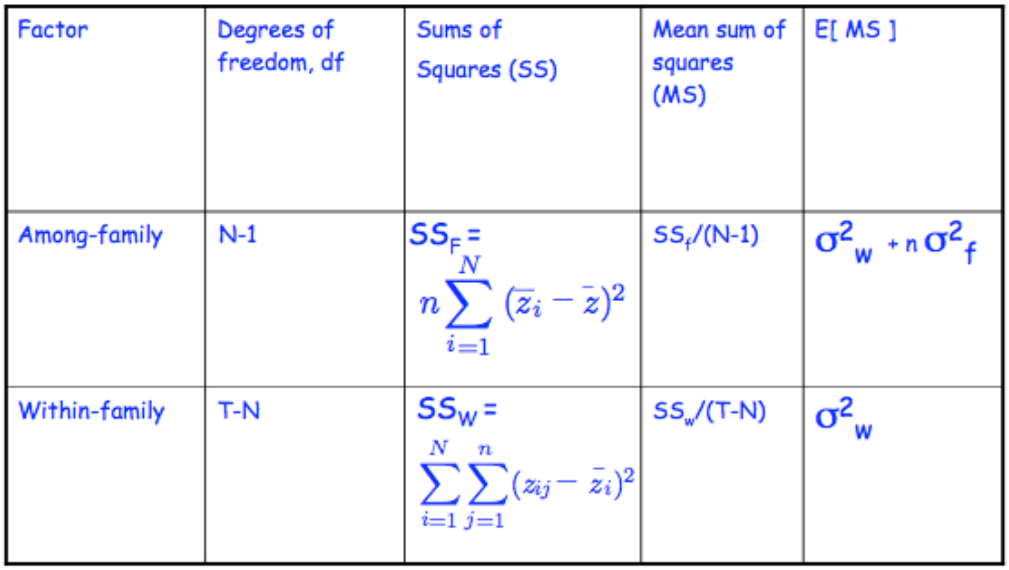
Hence, the variance among family effects equals the covariance between full sibs

$$\sigma^2_f = \sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}$$

The within-family variance $\sigma^2_w = \sigma^2_P - \sigma^2_f$,

$$\begin{aligned} \sigma^2_{w(FS)} &= \sigma^2_P - (\,\sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}\,) \\ &= \sigma^2_A + \sigma^2_D + \sigma^2_E - (\,\sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}\,) \\ &= (1/2)\sigma^2_A + (3/4)\sigma^2_D + \sigma^2_E - \sigma^2_{Ec} \end{aligned}$$

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



| Factor | Degrees of freedom, df | Sums of Squares (SS) | Mean sum of squares (MS) | E[ MS ] |
|---|---|---|---|---|
| Among-family | N-1 | $SS_F = n\sum_{i=1}^{N}(\overline{z}_i - \overline{z})^2$ | $SS_f/(N-1)$ | $\sigma^2_w + n\sigma^2_f$ |
| Within-family | T-N | $SS_W = \sum_{i=1}^{N}\sum_{j=1}^{n}(z_{ij} - \overline{z}_i)^2$ | $SS_w/(T-N)$ | $\sigma^2_w$ |

Estimating the variance components:

$$\mathrm{Var}(f) = \frac{\mathrm{MS}_f - \mathrm{MS}_w}{n}$$

$$\mathrm{Var}(w) = \mathrm{MS}_w$$

$$\mathrm{Var}(z) = \mathrm{Var}(f) + \mathrm{Var}(w)$$

**Since** $\sigma^2_f = \sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}$

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

$$\sigma^2(\mathrm{MS}_x) \simeq \frac{2\,(\mathrm{MS}_x)^2}{\mathrm{df}_x + 2}$$

$$\mathrm{Var}[\,\mathrm{Var}(w(FS))\,] = \mathrm{Var}(\mathrm{MS}_w) \simeq \frac{2(\mathrm{MS}_w)^2}{T - N + 2}$$

$$\begin{aligned}\mathrm{Var}[\,\mathrm{Var}(f)\,] &= \mathrm{Var}\left[\frac{\mathrm{MS}_f - \mathrm{MS}_w}{n}\right] \\ &\simeq \frac{2}{n^2}\left(\frac{(\mathrm{MS}_f)^2}{N+1} + \frac{(\mathrm{MS}_w)^2}{T-N+2}\right)\end{aligned}$$

## Estimating heritability

$$t_{\mathrm{FS}} = \frac{\mathrm{Var}(f)}{\mathrm{Var}(z)} = \frac{1}{2}h^2 + \frac{\sigma^2_D/4 + \sigma^2_{E_c}}{\sigma^2_z}$$

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

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

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

# Worked example

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

| Factor | Df | SS | MS | EMS |
|---|---|---|---|---|
| Among-familes | 9 | $SS_f$ = 405 | 45 | $\sigma^2_w + 5\,\sigma^2_f$ |
| Within-families | 40 | $SS_w$ = 800 | 20 | $\sigma^2_w$ |

$$\mathrm{Var}(f) = \frac{\mathrm{MS}_f - \mathrm{MS}_w}{n} = \frac{45 - 20}{5} = 5 \longrightarrow V_A < 10$$

$$\mathrm{Var}(w) = \mathrm{MS}_w = 20$$

$$\mathrm{Var}(z) = \mathrm{Var}(f) + \mathrm{Var}(w) \quad = 25 \longrightarrow$$

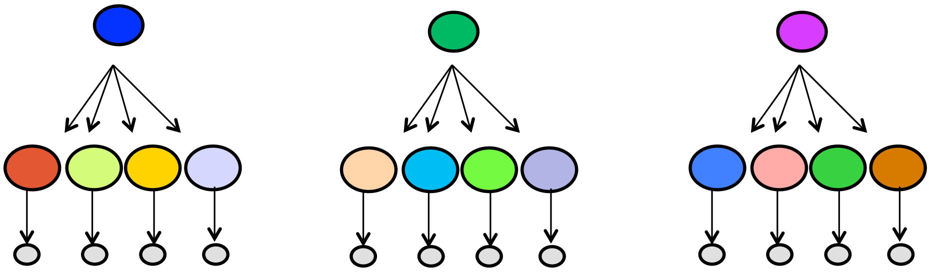
$$h^2 < 2\,(5/25) = 0.4$$

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

# Same approach works using half-sib families

# Mating Designs

## HALF-SIB DESIGN: *N* half-sib families with *n* offspring each.



$$z_{ij} = \mu + \underline{f_i} + \underline{w_{ij}}$$

$z_{ij}$ = phenotype of the j - th offspring of the i - th family

$\mu$ = population mean

$f_i$ = effect of the i - th family

$w_{ij}$ = residual error (segregation, dominance, environmental contribution)

within - family variance

| SoV | df | SS | MS | EMS |
|---|---|---|---|---|
| Among-families | N-1 | $SS_f = n\sum_i \left(\bar{z}_{i.} - \bar{z}_{..}\right)^2$ | $SS_f/df_{(f)}$ | $\sigma^2_{w(FS)} + n\sigma^2_f$ |
| Within-families | n(N-1) | $SS_w = \sum_{i,j} (z_{ij} - \bar{z}_{i.})^2$ | $SS_w/df_{(w)}$ | $\sigma^2_{w(FS)}$ |

# Mating Designs

## HALF-SIB DESIGN: *N* half-sib families with *n* offspring each.

$$z_{ij} = \mu + \underline{f_i} + \underline{w_{ij}}$$

$z_{ij}$ = phenotype of the j - th offspring of the i - th family

$\mu$ = population mean

$f_i$ = effect of the i - th family

$w_{ij}$ = residual error (segregation, dominance, environmental contribution)

within - family variance

| SoV | df | SS | MS | EMS |
|---|---|---|---|---|
| Among-families | N-1 | $SS_f = n\sum_i \left(\bar{z}_{i.} - \bar{z}_{..}\right)^2$ | $SS_f/df_{(f)}$ | $\sigma^2_{w(FS)} + n\sigma^2_f$ |
| Within-families | n(N-1) | $SS_w = \sum_{i,j} (z_{ij} - \bar{z}_{i.})^2$ | $SS_w/df_{(w)}$ | $\sigma^2_{w(FS)}$ |

$$\mathrm{Var}(f) = \frac{\mathrm{MS}_f - \mathrm{MS}_w}{\mathrm{n}}$$

$$\mathrm{Var}(w) = \mathrm{MS}_\mathrm{w}$$

$$\mathrm{Var}(z) = \mathrm{Var}(f) + \mathrm{Var}(w)$$

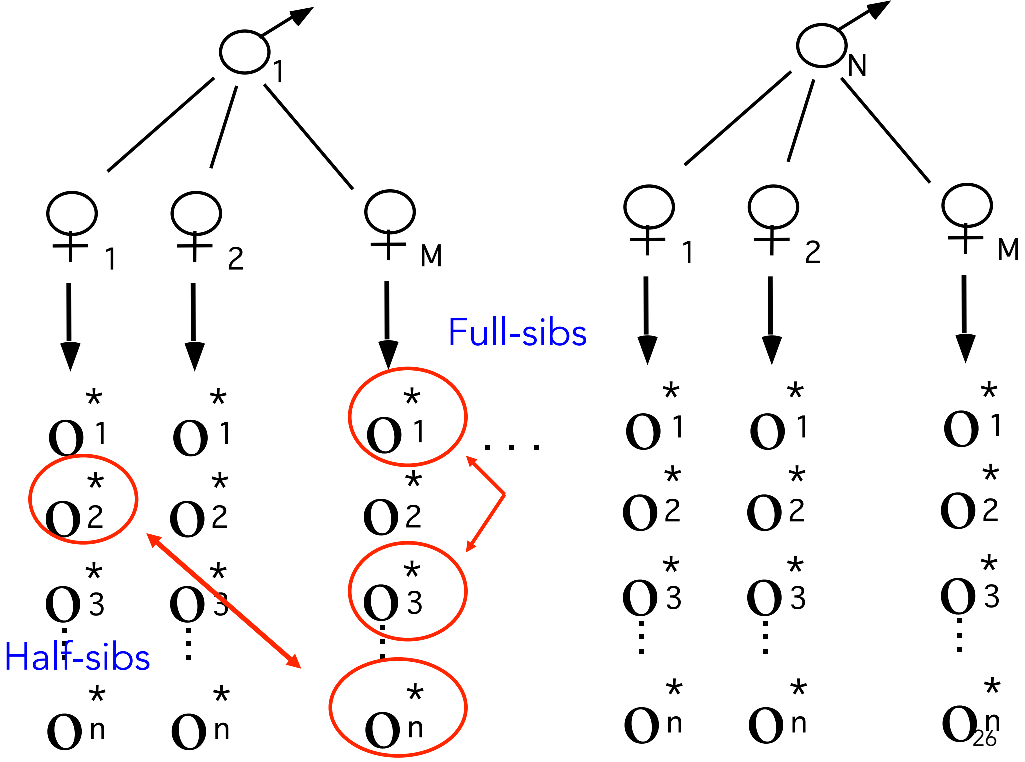
$$t_{HS} = \frac{\mathrm{Var}(f)}{\mathrm{Var}(z)} = \frac{\frac{1}{4}\sigma^2_A}{\sigma^2_z} = \frac{1}{4}h^2$$

$$h^2 \cong 4t_{HS}$$

# Nested designs

- Under a nested design, several types of relatives are jointly considered, typically full- vs. half-sibs
- Under the North Carolina Design one (NC I), males are crossed to a random series of unrelated females
  - No common females (each unique to a cross)
- Under NC II, males are crossed to a set of common (but unrelated) females
  - All males crossed to the same set of females
- Under a diallel, a (full or partial) set of all pairwise crosses is made.

## Full sib-half sib design: Nested ANOVA



# Estimation: Nested ANOVA (NC I)

Balanced full-sib / half-sib design: N males (sires) are crossed to M dams each of which has n offspring: Nested ANOVA model for NC I is

Effect of dam j of sire i = deviation of mean of dam j from sire and overall mean

Overall mean

Value of the kth offspring from the jth dam for sire i

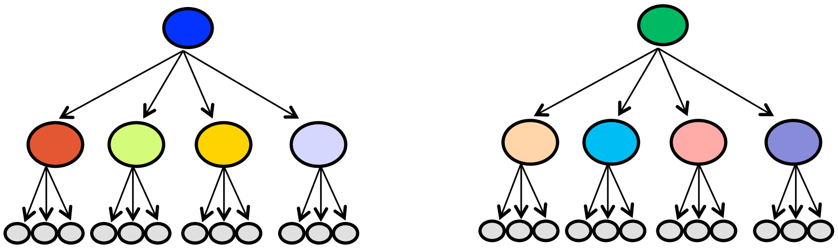
$$z_{ijk} = \mu + s_i + d_{ij} + w_{ijk}$$

Effect of sire i = deviation of mean of i's family from overall mean

Within-family deviation of kth offspring from the mean of the ij-th family

## Mating Designs

NORTH CAROLINA DESIGN I: Each male (N sire) is mated to several unrelated females (M dams) to produce n offspring per dam.



Note no common females between crosses

$$z_{ijk} = \mu + \underline{s_i} + \underline{d_{j(i)}} + \underline{w_{ijk}}$$

$z_{ijk}$ = phenotype of the k - th offspring from the family of the i - th sire and j - th dam

$\mu$ = population mean

$s_i$ = effect of the i - th sire

$d_{ij}$ = effect of the j - th dam mated to the i - th sire

$w_{ijk}$ = residual error (within - family variance deviations)

Nested ANOVA model (for NC I):

$$z_{ijk} = \mu + s_i + d_{ij} + w_{ijk}$$

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

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

$\sigma^2_w$ = within-family variance

$\sigma^2_T = \sigma^2_s + \sigma^2_d + \sigma^2_w$

## Mating Designs

NORTH CAROLINA DESIGN I: Each male (N sire) is mated to several unrelated females (M dams) to produce n offspring per dam.

$$z_{ijk} = \mu + \underline{s_i} + \underline{d_{j(i)}} + \underline{w_{ijk}}$$

| SoV | df | SS | MS | EMS |
|---|---|---|---|---|
| Sires | N-1 | $SS_s = Mn\sum_{i,j}\left(\bar{z}_{i.} - \bar{z}_{..}\right)^2$ | $MS_s/df_{(s)}$ | $\sigma^2_w + n\sigma^2_d + Mn\sigma^2_s$ |
| Dams(Sire) | N(M-1) | $SS_d = \sum_{i,j}(z_{ij} - \bar{z}_{i.})^2$ | $MS_d/df_{(d)}$ | $\sigma^2_w + n\sigma^2_d$ |
| Sibs(dams) | T-NM | $SS_w = \sum_{i,j,k}(z_{ijk} - \bar{z}_{ij.})^2$ | $MS_w/df_{(w)}$ | $\sigma^2_w$ |

$$\mathrm{Var}(s) = \frac{\mathrm{MS}_S - \mathrm{MS}_d}{\mathrm{Mn}}$$

$$\mathrm{Var}(d) = \frac{\mathrm{MS}_d - \mathrm{MS}_W}{\mathrm{n}}$$

$$\mathrm{Var}(w) = \mathrm{MS}_w$$

$$\mathrm{Var}(z) = \mathrm{Var}(s) + \mathrm{Var}(d) + \mathrm{Var}(w)$$

$$t_{PHS} = \frac{\mathrm{Var}(s)}{\mathrm{Var}(z)} = \frac{\tfrac{1}{4}\sigma^2_A}{\sigma^2_z} = \frac{1}{4}h^2$$

$$t_{FS} = \frac{\mathrm{Var}(s) + \mathrm{Var}(d)}{\mathrm{Var}(z)} = \frac{\tfrac{1}{2}\sigma^2_A + \tfrac{1}{4}\sigma^2_D + \sigma^2_{Ec}}{\sigma^2_z} = \frac{1}{2}h^2 + \frac{\tfrac{1}{4}\sigma^2_D + \sigma^2_{Ec}}{\sigma^2_z}$$

$$h^2 \cong 4t_{PHS}$$

Estimation of sire, dam, and family variances:

$$\mathrm{Var}(s) = \frac{\mathrm{MS}_s - \mathrm{MS}_d}{Mn}$$

$$\mathrm{Var}(d) = \frac{\mathrm{MS}_d - \mathrm{MS}_w}{n}$$

$$\mathrm{Var}(e) = \mathrm{MS}_w$$

Translating these into the desired variance components

- Var(Total) = Var(between FS families) + Var(Within FS)

⟶ $\sigma^2_w = \sigma^2_z$ - Cov(FS)

- Var(Sires) = Cov(Paternal half-sibs)

$$\begin{aligned}\sigma_d^2 &= \sigma_z^2 - \sigma_s^2 - \sigma_w^2 \\ &= \sigma(\mathrm{FS}) - \sigma(\mathrm{PHS})\end{aligned}$$

Summarizing,

$$\sigma_s^2 = \sigma(\mathrm{PHS}) \qquad \sigma_d^2 = \sigma_z^2 - \sigma_s^2 - \sigma_w^2$$

$$\sigma_w^2 = \sigma_z^2 - \sigma(\mathrm{FS}) \qquad = \sigma(\mathrm{FS}) - \sigma(\mathrm{PHS})$$

Expressing these in terms of the genetic and environmental variances,

$$\sigma_s^2 \simeq \frac{\sigma_A^2}{4}$$

$$\sigma_d^2 \simeq \frac{\sigma_A^2}{4} + \frac{\sigma_D^2}{4} + \sigma_{E_c}^2$$

$$\sigma_w^2 \simeq \frac{\sigma_A^2}{2} + \frac{3\sigma_D^2}{4} + \sigma_{E_s}^2$$

## Intraclass correlations and estimating heritability

$$t_{PHS} = \frac{\mathrm{Cov(PHS)}}{\mathrm{Var(z)}} = \frac{\mathrm{Var(s)}}{\mathrm{Var(z)}} \longrightarrow 4t_{PHS} = h^2$$

$$t_{FS} = \frac{\mathrm{Cov(FS)}}{\mathrm{Var(z)}} = \frac{\mathrm{Var(s)} + \mathrm{Var(d)}}{\mathrm{Var(z)}} \longrightarrow h^2 \leq 2t_{FS}$$

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 | EMS |
|---|---|---|---|---|
| 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 | $\sigma_w^2$ |

$$\sigma_w^2 = MS_w = 20$$

$$\sigma_d^2 = \frac{MS_d - MS_w}{n} = \frac{170 - 20}{10} = 15$$

$$\sigma_s^2 = \frac{MS_s - MS_d}{Nn} = \frac{470 - 170}{30} = 10 \longrightarrow \boxed{\sigma_A^2 = 4\sigma_s^2 = 40}$$

$$\sigma_P^2 = \sigma_s^2 + \sigma_d^2 + \sigma_w^2 = 45 \longrightarrow \boxed{h^2 = \frac{\sigma_A^2}{\sigma_P^2} = \frac{40}{45} = 0.89}$$

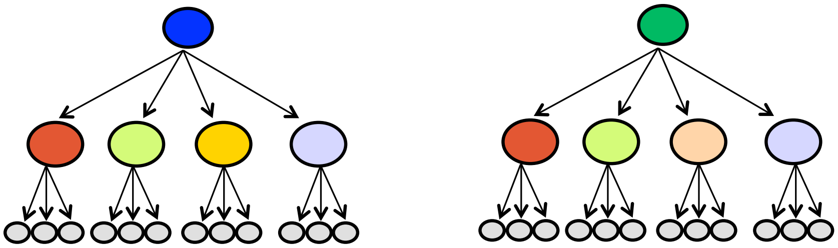
$$\sigma_d^2 = 15 = (1/4)\sigma_A^2 + (1/4)\sigma_D^2 + \sigma_{E_c}^2$$

$$= 10 + (1/4)\sigma_D^2 + \sigma_{E_c}^2 \longrightarrow \boxed{\sigma_D^2 + 4\sigma_{E_c}^2 = 20}$$

# Mating Designs

NORTH CAROLINA DESIGN II: A group of sires ($N_S$ sires) are mated to an independent group of dams ($N_D$ dams) to produce n offspring

.



*Note* ***<u>same set</u>*** *of females in all crosses*

$$z_{ijk} = \mu + \underline{s_i} + \underline{d_j} + \underline{I_{ij}} + \underline{w_{ijk}}$$

$z_{ijk}$ = phenotype of the k-th offspring from the family of the i-th sire and j-th dam

$\mu$ = population mean

$s_i$ = effectof the i-th sire

$d_i$ = effectof the j-th dam

Iij = effectof the interaction between the i-th sire and the j-th dam

$w_{ijk}$ = residual error (within-family variance deviations)

# Estimation: Nested ANOVA (NC II)

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

Interaction between effects of sire i and dam j

Overall mean

Value of the kth offspring from the jth dam for sire i

$$z_{ijk} = \mu + s_i + d_j + I_{ij} \;\; w_{ijk}$$

Effect of sire i = deviation of mean of i's family from overall mean

Effect of dam j

Within-family deviation of kth offspring from the mean of the ij-th family

The $d_{ij}$ term under NC I is replaced in NC II by $d_j + I_{ij}$

# Mating Designs

NORTH CAROLINA DESIGN II: A group of sires ($N_S$ sires) are mated to an independent group of dams ($N_D$ dams) to produce n offspring

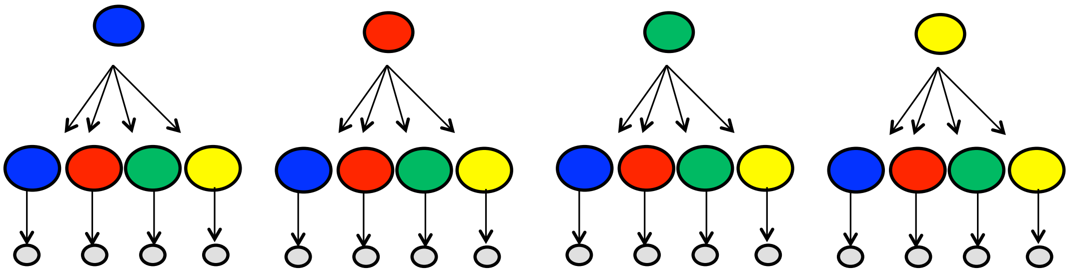
$$z_{ijk} = \mu + \underline{s_i} + \underline{d_j} + \underline{I_{ij}} + \underline{w_{ijk}}$$

| SoV | df | SS | EMS |
|---|---|---|---|
| Sires | $N_s$-1 | $SS_s = nN_d \sum_i (\bar{z}_{i..} - \bar{z}_{...})^2$ | $\sigma_w^2 + n\sigma_I^2 + nN_d\sigma_s^2$ |
| Dams | $N_d$-1 | $SS_d = nN_s \sum_j (\bar{z}_{.j.} - \bar{z}_{...})^2$ | $\sigma_w^2 + n\sigma_I^2 + nN_s\sigma_d^2$ |
| Interaction | ($N_s$-1)($N_d$-1) | $SS_I = \sum_{i,j} (\bar{z}_{ij.} - \bar{z}_{i..} - \bar{z}_{.j.} - \bar{z}_{...})^2$ | $\sigma_w^2 + n\sigma_I^2$ |
| Sibs | $N_sN_d$(n-1) | $SS_w = \sum_{i,j,k} (z_{ijk} - \bar{z}_{ij.})^2$ | $\sigma_w^2$ |

$$t_{PHS} = \frac{\frac{1}{4}\sigma_A^2}{\sigma_z^2} \qquad t_{MHS} = \frac{\frac{1}{4}\sigma_A^2 + \sigma_{Gm}^2 + \sigma_{Ec}^2}{\sigma_z^2} \qquad t_I = \frac{\frac{1}{4}\sigma_D^2}{\sigma_z^2}$$

# Mating Designs

DIALLELS: A group of individuals (N) are mated to the same set of individuals (N) to produce n offspring



Full Diallele (all selfed and reciprocal crosses are made)
Incomplete Diallele – no selfed crosses
Incomplete Diallele – no selfed, no reciprocal crosses

$$z_{ijk} = \mu + \underline{g_i} + \underline{g_j} + \underline{s_{ij}} + \underline{w_{ijk}}$$

$z_{ijk}$ = phenotype of the k-th offspring from the i-th and j-th parents and j-th dam
$\mu$ = population mean
$g_i$ = general combining ability of parent i-th
$g_j$ = general combining ability of parent j-th
$s_{ij}$ = specific combining ability of parents i-th and j-th
$w_{ijk}$ = residual error (within-family variance deviations)

## Mating Designs

DIALLELS: A group of individuals (N) are mated to the same set of individuals (N) to produce n offspring. Analysis for incomplete diallele without selfed or reciprocal crosses.

$$z_{ijk} = \mu + \underline{g_i} + \underline{g_j} + \underline{s_{ij}} + \underline{w_{ijk}}$$

| SoV | df | SS | EMS |
|---|---|---|---|
| GCA | N-1 | $SS_{GCA} = \frac{n(N-1)^2}{N-2} \sum_i (\bar{z}_{i..} - \bar{z}_{...})^2$ | $\sigma_w^2 + n\sigma_{SGA}^2 + n(N-2)\sigma_{GCA}^2$ |
| SCA | N(N-3)/2 | $SS_{SCA} = n\sum_{i<j} (\bar{z}_{ij.} - \bar{z}_{...})^2 - SS_{GCA}$ | $\sigma_w^2 + n\sigma_{SCA}^2$ |
| Sibs | (n-1)[N(N-1)/2-1] | $SS_w = \sum_{i<j,k} (z_{ijk} - \bar{z}_{ij.})^2$ | $\sigma_w^2$ |

$$t_{GCA} = \frac{\frac{1}{4}\sigma_A^2}{\sigma_z^2} \qquad t_{SCA} = \frac{\frac{1}{4}\sigma_D^2}{\sigma_z^2}$$

## Parent-offspring regression

Single parent - offspring regression

$$z_{o_i} = \mu + b_{o|p}(z_{p_i} - \mu) + e_i$$

The expected slope of this regression is:

$$E(b_{o|p}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \simeq \frac{(\sigma_A^2/2) + \sigma(E_o, E_p)}{\sigma_z^2} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma_z^2}$$

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:

$$E(b_{o|p}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \simeq \frac{(\sigma_A^2/2) + \sigma(E_o, E_p)}{\sigma_z^2} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma_z^2}$$

Shared environmental values

To avoid this term, typically regressions are male-offspring, as female-offspring more likely to share environmental values

Midparent - offspring regression

$$z_{oi} = \mu + b_{o|MP}\left(\frac{z_{mi} + z_{fi}}{2} - \mu\right) + e_i$$

$$\begin{aligned} b_{o||MP} &= \frac{\text{Cov}[z_o, (z_m + z_f)/2]}{\text{Var}[(z_m + z_f)/2]} \\ &= \frac{[\text{Cov}(z_o, z_m) + \text{Cov}(z_o, z_f)]/2}{[\text{Var}(z) + \text{Var}(z)]/4} \\ &= \frac{2\text{Cov}(z_o, z_p)}{\text{Var}(z)} = 2b_{o|p} \end{aligned}$$

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

Squared regression slope

$$\mathrm{Var}(b_{o|p}) \simeq \frac{n(t - b^2_{o|p}) + (1-t)}{Nn}$$

Total number of offspring

Sib correlation $$t = \begin{cases} t_{HS} = h^2/4 & \text{for half-sibs} \\ t_{FS} = h^2/2 + \dfrac{\sigma^2_D + \sigma^2_{E_c}}{\sigma^2_z} & \text{for full sibs} \end{cases}$$

$$\mathrm{Var}(h^2) = \mathrm{Var}(2b_{o|p}) = 4\mathrm{Var}(b_{o|p})$$

## Midparent-offspring regression, N sets of parents, each with n offspring

$$\mathrm{Var}(h^2) = \mathrm{Var}(b_{o|MP}) \simeq \frac{2[n(t_{FS} - b^2_{o|MP}/2) + (1 - t_{FS})])}{Nn}$$

- Midparent-offspring variance half that of single parent-offspring variance

$$\mathrm{Var}(h^2) = \mathrm{Var}(2b_{o|p}) = 4\mathrm{Var}(b_{o|p})$$

## Parent-Offspring Regression

Regression one parent – offspring (one offspring or the mean of multiple offspring).

$$E\left(\hat{b}_{o|p}\right) = \frac{\sigma\left(\mathrm{z_o}, \mathrm{z_p}\right)}{\sigma^2\left(\mathrm{z_p}\right)} \cong \frac{\tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_{AA}^2 + \sigma\left(Eo, Ep\right)}{\sigma_z^2} \cong \frac{1}{2}h^2, \quad h^2 \cong 2b_{o|p}$$

Regression one parent on offspring – no environment correlation among parent and offspring.

$$E\left(\hat{b}_{o|p}\right) = \frac{\sigma\left(\mathrm{z_o}, \mathrm{z_p}\right)}{\sigma^2\left(\mathrm{z_p}\right)} \cong \frac{\tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_{AA}^2}{\sigma_z^2} \cong \frac{1}{2}h^2, \quad h^2 \cong 2b_{o|p}$$

Regression mid parent on offspring – no environment correlation among parent and offspring.

$$E\left(\hat{b}_{o|\bar{p}}\right) = \frac{\sigma\left(\mathrm{z_o}, \bar{\mathrm{z}}_\mathrm{p}\right)}{\sigma^2\left(\bar{\mathrm{z}}_\mathrm{p}\right)} = \frac{\sigma\left(\mathrm{z_o}, \tfrac{1}{2}\mathrm{z_{P1}} + \tfrac{1}{2}\mathrm{z_{P2}}\right)}{\sigma^2\left(\tfrac{1}{2}\mathrm{z_{P1}} + \tfrac{1}{2}\mathrm{z_{P2}}\right)} = \frac{\tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_{AA}^2}{\tfrac{1}{2}\sigma_z^2} = h^2, \quad h^2 \cong b_{o|\bar{p}}$$

Regression parent -offspring inbreeding – no environment correlation.

$$E\left(\hat{b}_{S0:1|S0}\right) = \frac{\sigma\left(\mathrm{z_{S0}}, \mathrm{z_{S0:1}}\right)}{\sigma^2\left(\mathrm{z_{S0}}\right)} = \frac{\sigma_A^2 + \tfrac{1}{2}\sigma_D^2 + \tfrac{1}{2}\sigma_{D1}^2 + \sigma_{AA}^2}{\sigma_z^2} \cong h^2, \quad h^2 \cong b_{S0:1|S0}$$

## 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

$$h^2_{min} = (b'_{o|MP})^2 \frac{\mathrm{Var}_n(z)}{\mathrm{Var}_l(A)}$$

Trait variance in nature

Additive variance in lab

Why is this a lower bound?

Covariance between
breeding value in nature
and BV in lab

$$(b'_{o|MP})^2 \frac{\mathrm{Var}_n(z)}{\mathrm{Var}_l(A)} = \left[\frac{\mathrm{Cov}_{l,n}(A)}{\mathrm{Var}_n(z)}\right]^2 \frac{\mathrm{Var}_n(z)}{\mathrm{Var}_l(A)} = \boxed{\gamma^2 h_n^2}$$

where $$\gamma = \frac{\mathrm{Cov}_{l,n}(A)}{\sqrt{\mathrm{Var}_n(A)\mathrm{Var}_l(A)}}$$

is the additive genetic covariance between environments and hence $\gamma^2 < 1$

# 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

Genotype i

Interaction
between Genotype i
and environment j

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

Environment j

Effect of plot k for
Genotype i
in environment j

deviations of individual
plants within this plot

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

$$\sigma^2(z_i) = \sigma^2_G + \sigma^2_E + \frac{\sigma^2_{GE}}{e} + \frac{\sigma^2_p}{e\,r} + \frac{\sigma^2_e}{e\,r\,n}$$

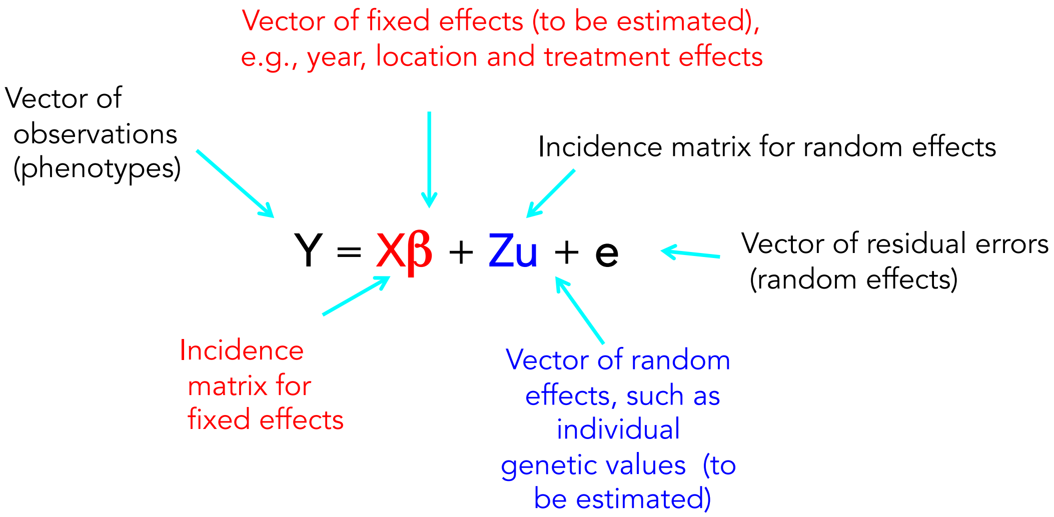
e = number of environments
r = (replicates) number of plots/environment
n = number of individuals per plot

Hence, $V_P$, and hence $H^2$, depends on our choice of e, r, and n

# Mixed Models

- The above designs only compare a small set of relatives (e.g., sibs, parent-offspring). More generally, esp. in plant breeding, we may have much richer sets of relatedness. Further, designs are usually unbalanced, unequal numbers of relatives
- The framework of mixed models (BLUP for estimation of genetic effects, REML for estimation of genetic variances) handles such completely general designs.
  - A relationship matrix **A** for the $\theta$ values for all individuals is used to allow us to extract the maximal amount of information.
  - Easily handles unbalanced designs
  - Mixed Models covered later in the course.

# The general mixed model



Vector of fixed effects (to be estimated), e.g., year, location and treatment effects

Vector of observations (phenotypes)

Incidence matrix for random effects

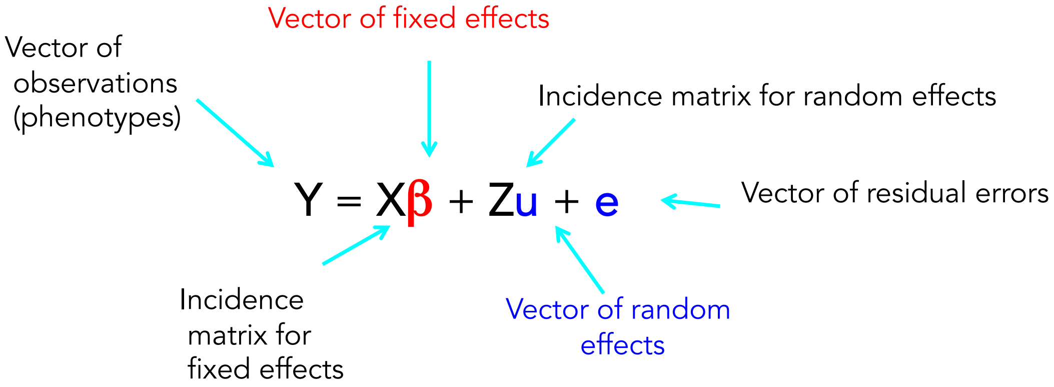
$$Y = X\beta + Zu + e$$

Vector of residual errors (random effects)

Incidence matrix for fixed effects

Vector of random effects, such as individual genetic values (to be estimated)

# The general mixed model



Vector of fixed effects

Vector of observations (phenotypes)

Incidence matrix for random effects

$$Y = X\beta + Zu + e$$

Vector of residual errors

Incidence matrix for fixed effects

Vector of random effects

Observe **y**, X, Z.

Assume Cov(**u**) = Var(A)*A

Estimate fixed effects **β**

Estimate random effects **u**, **e**

# Lecture 4
# Short-Term Selection Response: Breeder's equation

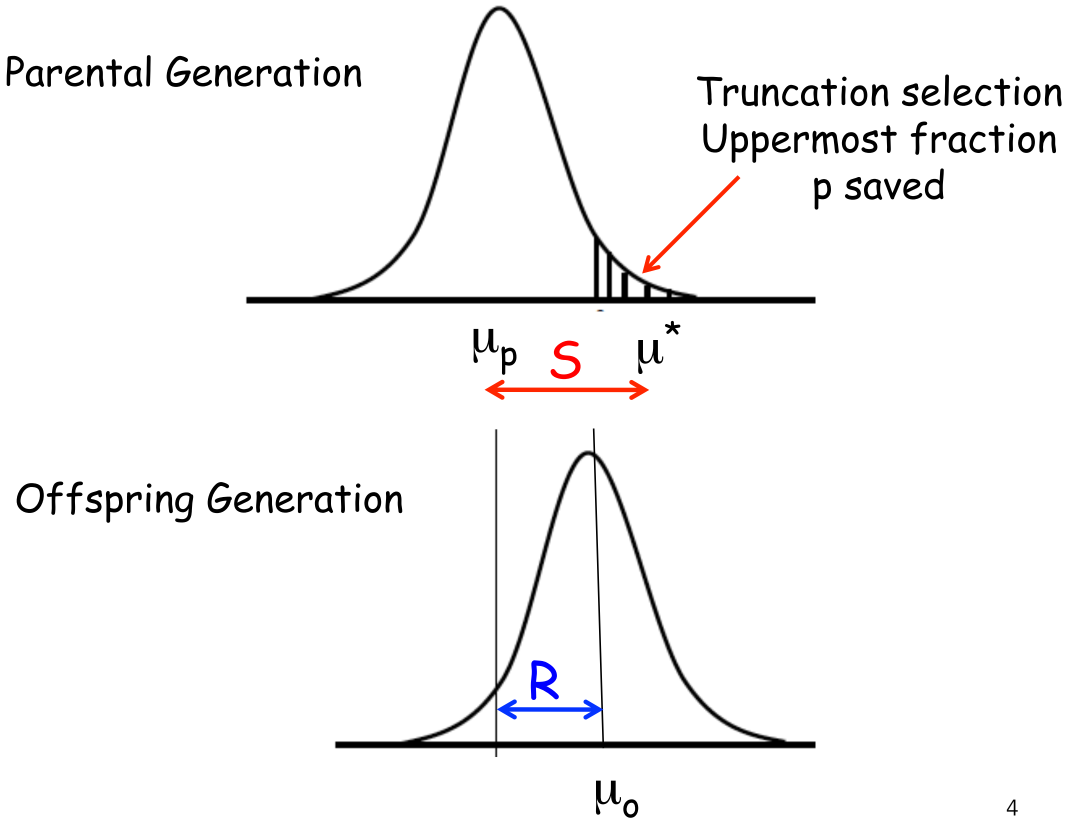
Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

# 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)

# 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
  - $R(t) = \mu(t+1) - \mu(t)$



## 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:

$\boxed{R = h^2 S}$ 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

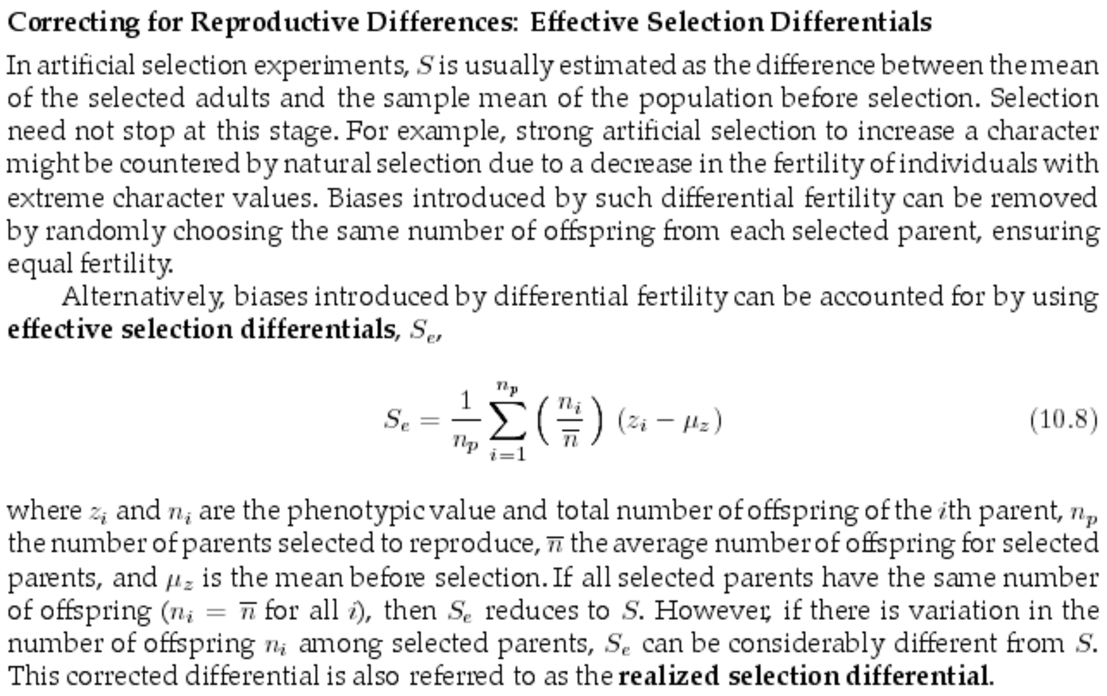
- Recall that $S = (S_f + S_m)/2$
- An issue that arises in plant breeding is pollen control --- is the pollen from plants that have also been selected?
- Not the case for traits (i.e., yield) scored after pollination. In this case, $S_m = 0$, so response only half that with pollen control
- Tradeoff: 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

# 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 board 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.

# Price-Robertson identity

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



**Correcting for Reproductive Differences: Effective Selection Differentials**

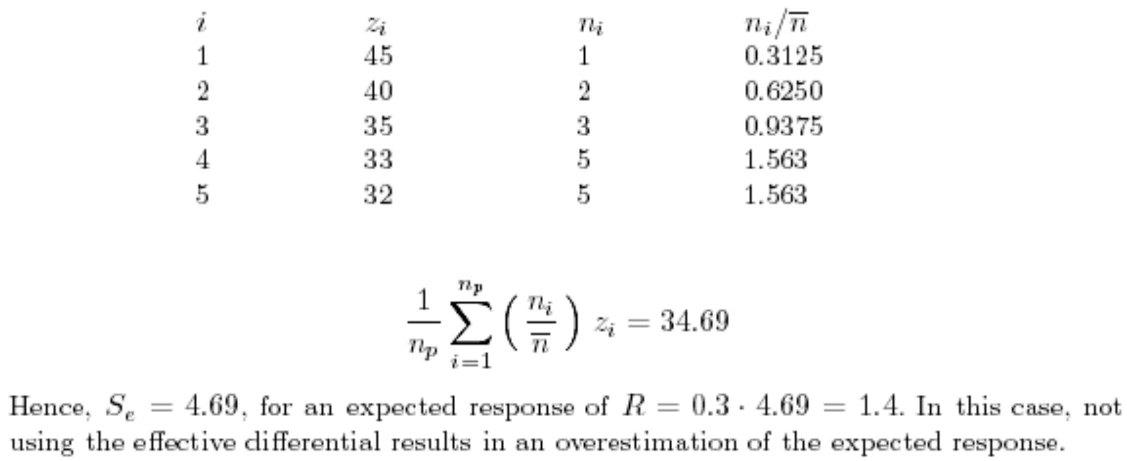
In artificial selection experiments, $S$ is usually estimated as the difference between the mean of the selected adults and the sample mean of the population before selection. Selection need not stop at this stage. For example, strong artificial selection to increase a character might be countered by 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**, $S_e$,

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

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, $\overline{n}$ 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 ($n_i = \overline{n}$ 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**.

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



| $i$ | $z_i$ | $n_i$ | $n_i/\overline{n}$ |
|---|---|---|---|
| 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 |

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

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*7 = 2.1
offspring-weighted S = 4.69, pred resp = 1.4

# Response over multiple generations

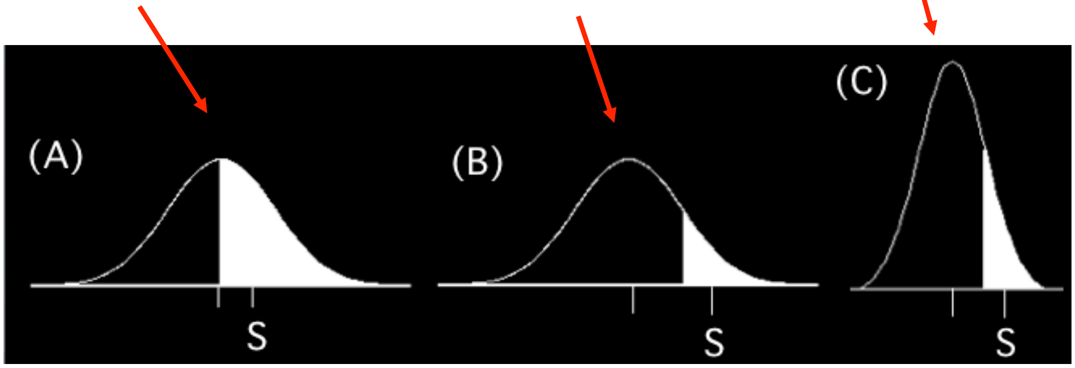
- Strictly speaking, the breeders' equation only holds for predicting a single generation of response from an unselected base population
- Practically speaking, the breeders' equation is usually pretty good for 5-10 generations
- The validity for an initial $h^2$ predicting response over several generations 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

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

50% selected
$V_p$ = 4, S = 1.6

20% selected
$V_p$ = 4, S = 2.8

20% selected
$V_p$ = 1, S = 1.4



## The Selection Intensity, i

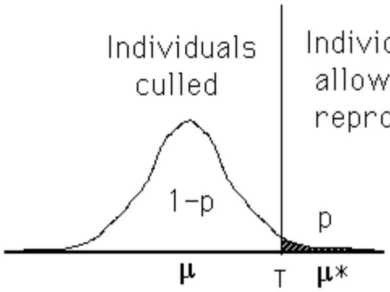
As the previous example shows, populations with the same selection differential (S) may experience very different amounts of selection

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

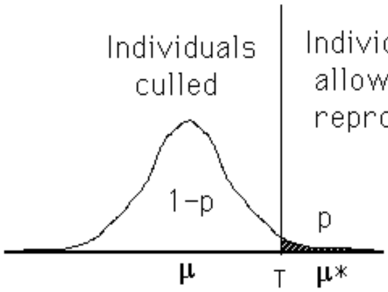
$$i = \frac{S}{\sqrt{V_P}} = \frac{S}{\sigma_p}$$

# 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



# Selection Differential Under Truncation Selection



S =μ*- μ

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

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

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

# 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 chose an expected value of i before selection, and hence set the expected response

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

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`

## Selection Intensity Version of the Breeders' Equation

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

Since $h^2\sigma_P = (\sigma^2_A/\sigma^2_P)\,\sigma_P = \sigma_A(\sigma_A/\sigma_P) = h\,\sigma_A$

$$R = i\, h\, \sigma_A$$

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

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

Response = Intensity * Accuracy * spread in Va

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

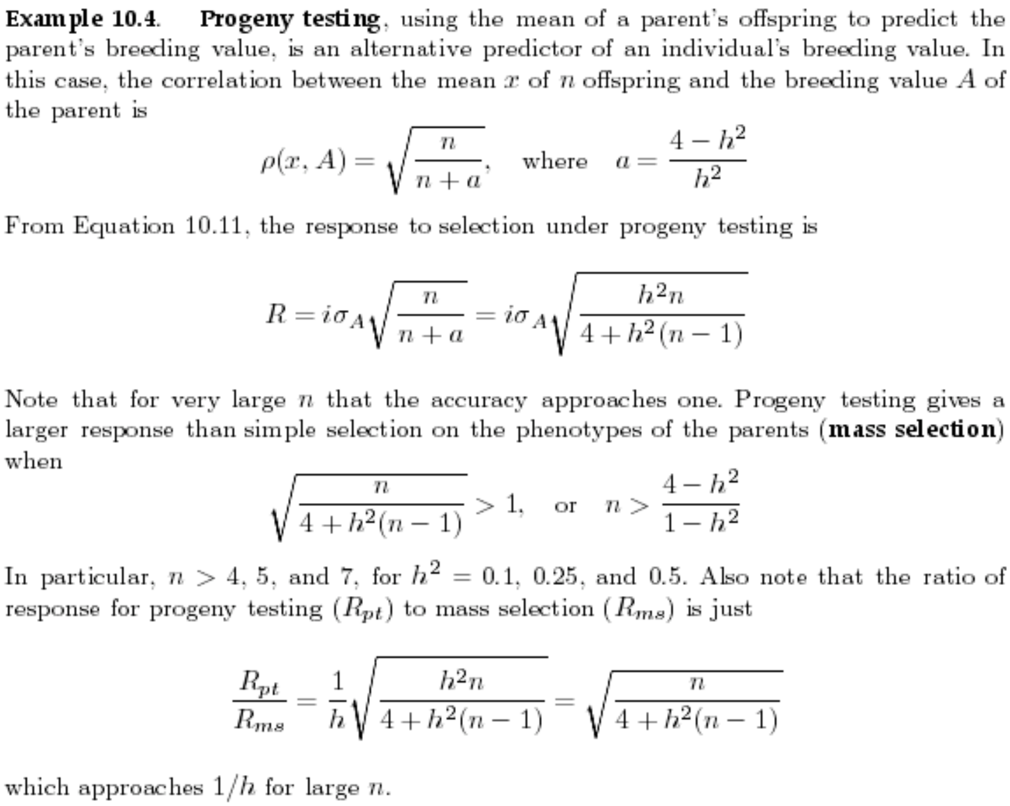
# Accuracy of selection

More generally, we can express the breeders equation as

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

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



**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

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

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

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

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 x E
  - Need to replicate either clones or relatives (such as sibs) over regions and years to reduce the impact of G x E
  - Likewise, information from a set of relatives can give much higher accuracy than the measurement of a single individual

# 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.

# 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

# 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 |

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

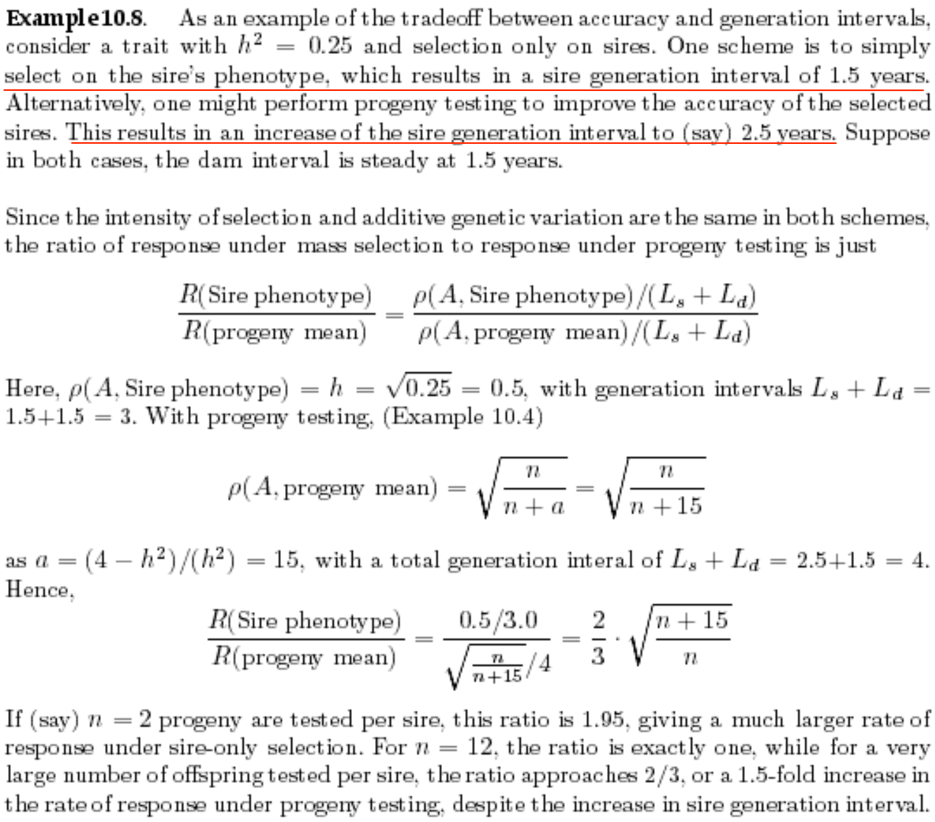
$$L_d = \frac{2 \cdot 400 + 3 \cdot 600 + 4 \cdot 100 + 5 \cdot 40}{400 + 600 + 100 + 40} = 2.81$$

# 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)



**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)

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

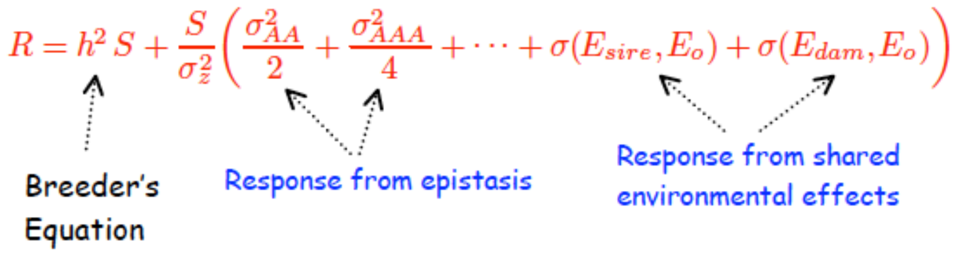
as $a = (4 - h^2)/(h^2) = 15$, with a total generation interal of $L_s + L_d = 2.5 + 1.5 = 4$. Hence,

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

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



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

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

## 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

# Response with Epistasis

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

$$R = S\left(h^2 + \frac{\sigma^2_{AA}}{2\,\sigma^2_z}\right)$$

The contribution to response from this single generation after τ generations of no selection is

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

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

# Response with Epistasis

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

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

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)

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

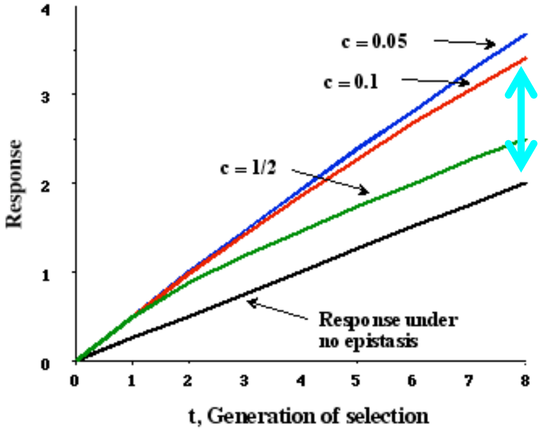
$R_{AA}$ has a limiting value given by

$$\widetilde{R}_{AA} = \lim_{t\to\infty} R_{AA}(t) = \frac{1}{c}\left(S\,\frac{\sigma^2_{AA}}{2\,\sigma^2_z}\right)$$

Time to equilibrium a function of c

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

Decay half-life



$$= \frac{1}{c}\left(S\,\frac{\sigma^2_{AA}}{2\,\sigma^2_z}\right)$$

Fixed incremental difference that decays when selection stops

What about response with higher-order epistasis?

| $S\sigma^2(A^i)/\sigma^2_z$, | AA | AAA | AAAA | AAAAA |
|---|---|---|---|---|
| $R(1)$ | 0.500 | 0.250 | 0.125 | 0.063 |
| Limit | 1.000 | 0.333 | 0.143 | 0.067 |
| % $R(1)$/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.

## Autotetraploids

P-O covariance

Single-generation response

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

Response to t generations of selection with constant selection differential S

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

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

Response remaining after t generations of selection followed by τ generations of random mating

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

Contribution from dominance quickly decays to zero

# 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

# Maternal Effects:

## Falconer's dilution model

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

G = Direct genetic effect on character

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

maternal effect passed from dam to offspring $m\, z_{dam}$ 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

## 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**

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

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

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

## Parent-offspring regression under the dilution model

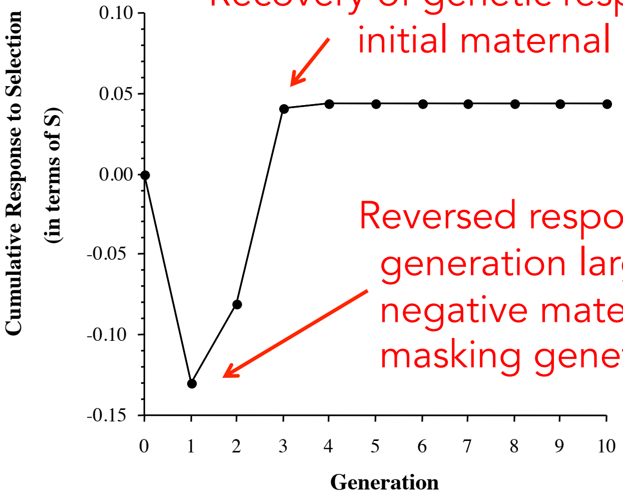
With maternal effects, a covariance between BV and maternal effect arises, with $\sigma_{A,M} = m\,\sigma^2_A \,/\, (\, 2 - m\,)$

The response thus becomes

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

## Response to a single generation of selection

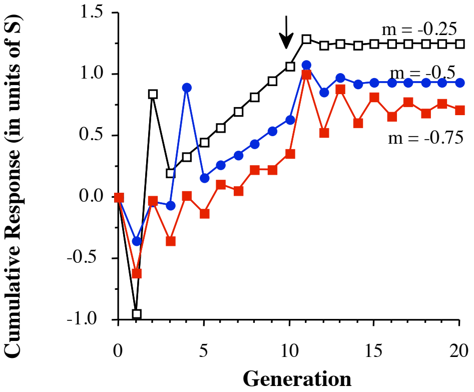
$h^2 = 0.11$, $m = -0.13$ (litter size in mice)



Recovery of genetic response after initial maternal correlation decays

Reversed response in 1st generation largely due to negative maternal correlation masking genetic gain

## Selection occurs for 10 generations and then stops



$h^2 = 0.35$

# Additional material

Unlikely to be covered in class

# Selection on Threshold Traits

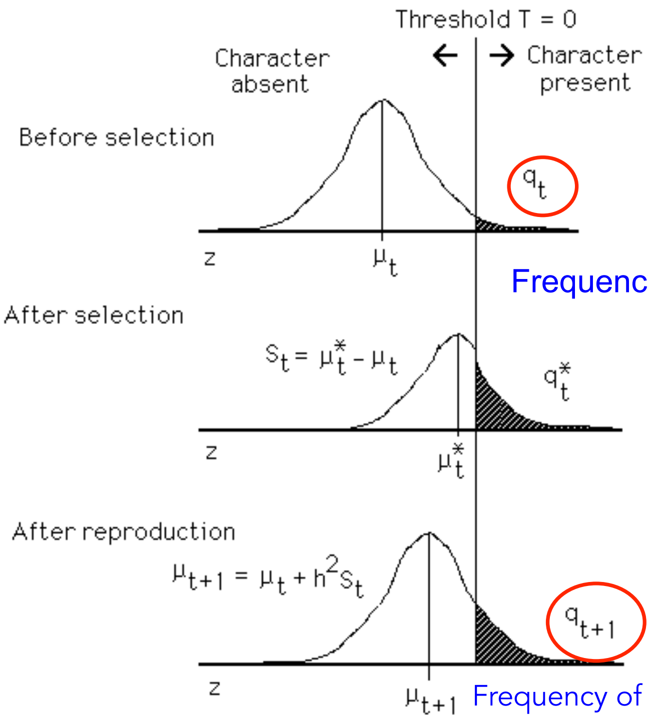
Response on a binary trait is a special case of response on a continuous trait

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) = Prob(state one | z). Details in LW Chapter 14.

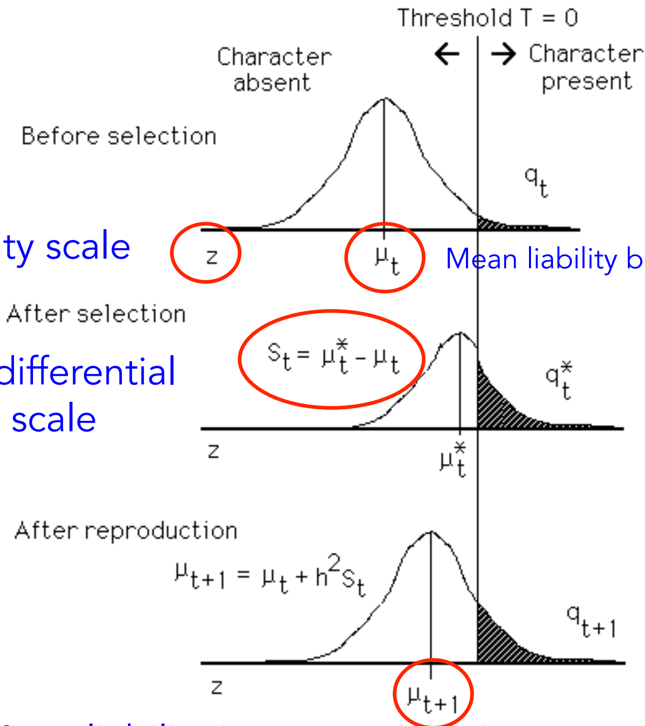


Observe: trait values are either 0,1. Pop mean = q (frequency of the 1 trait)

Frequency of trait

Want to map from q onto the underlying liability scale z, where breeder's equation $R_z = h^2 S_z$ holds

Frequency of character state on in next generation

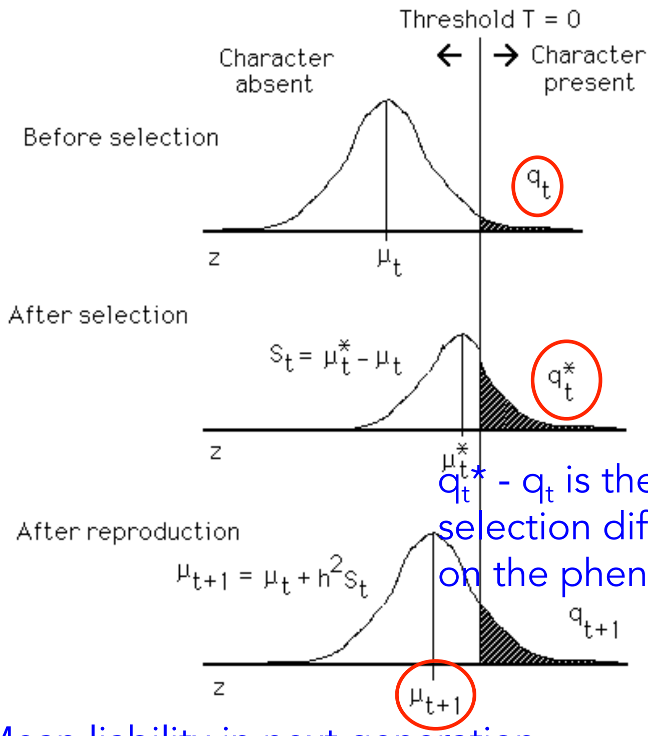


Liability scale

Mean liability before selection

Selection differential on liability scale

Mean liability in next generation



$q_t^* - q_t$ is the selection differential on the phenotypic scale

Mean liability in next generation

## 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, $z_{[1-q]}$ = `qnorm(1-q)`, with qnorm(0.95) returning 1.644854

## 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

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

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

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

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

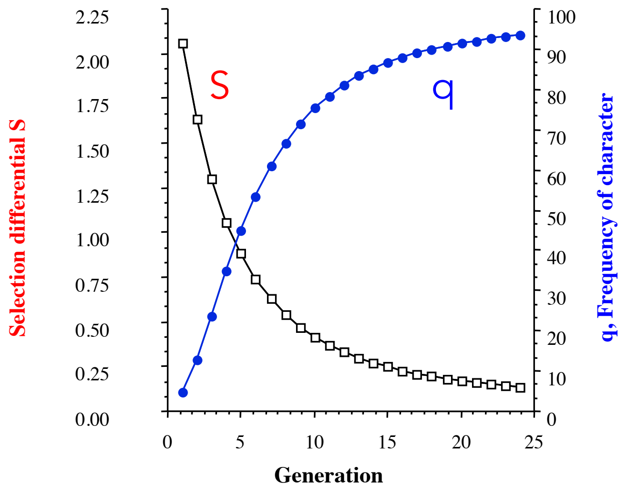
When z is normally distributed, this reduces to

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

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 iterative to obtain selection response

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



# 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:

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

2 parents

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

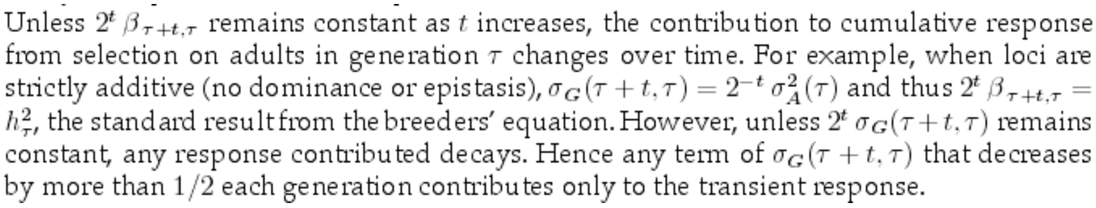
4 grandparents
Selection diff $S_1$
$\beta_{3,1}$ is the regression of relative in generation 3 on their gen 1 relatives

# Ancestral Regressions

## More generally,

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

The general expression cov($z_T$,$z_t$), where we keep track of the actual generation, as oppose to 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 $2^t\,\beta_{\tau+t,\tau} = h^2_\tau$, 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**, freq(AB gamete) = freq(A)freq(B)

With positive linkage disequilibrium, f(AB) > f(A)f(B), so that AB gametes are more frequent

With negative linkage disequilibrium, f(AB) < f(A)f(B), so that AB gametes are less frequent

# Additive variance with LD:

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

Genic variance: value of Var(A)
in the absence of disequilibrium
function of allele frequencies

$$\sigma^2\left(\sum_{k=1}^{n}\left(a_1^{(k)}+a_2^{(k)}\right)\right)=2\sum_{k=1}^{n}\sigma^2\left(a^{(k)}\right)+4\sum_{k<j}^{n}\sigma\left(a^{(j)},a^{(k)}\right)$$

$$=2\sum_{k=1}^{n}C_{kk}+4\sum_{k<j}^{n}C_{jk}$$

$$\sigma_A^2=\sigma_a^2+d$$

Additive variance

Disequilibrium contribution. Requires covariances
between allelic effects at different loci

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

Selection-induced changes in d change $\sigma^2_A$, $\sigma^2_z$, $h^2$

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

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

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,

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

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

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

Change in variance from selection

Change in d = change from recombination plus change from selection

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

Recombination Selection

In terms of change in d,

$$\begin{aligned}\Delta d(t) = \Delta\sigma_{z(t)}^2 &= \Delta\sigma_A^2(t) \\ &= -\frac{d(t)}{2} + \frac{h^4(t)}{2}\,\delta\left(\sigma_{z(t)}^2\right)\end{aligned}$$

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,

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

# 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

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

Var(A) = 0.6*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

# Specific models of selection-induced changes in variances

Proportional reduction model: constant fraction k of variance removed

$$\sigma^2_{z*} = (1-\kappa)\,\sigma^2_z$$

$$\delta\left(\sigma^2_z\right) = \sigma^2_{z*} - \sigma^2_z = -\kappa\,\sigma^2_z$$

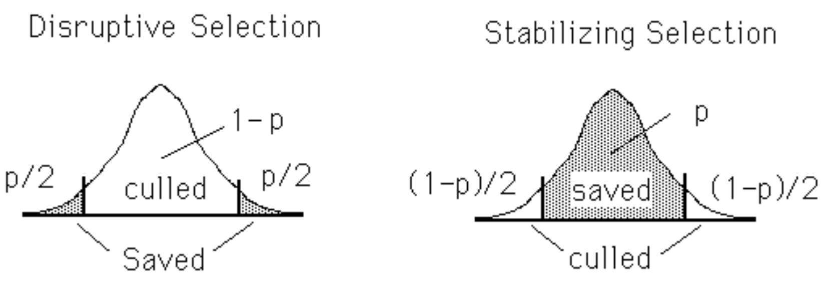
Bulmer equation simplifies to

$$d(t+1) = \frac{d(t)}{2} - \frac{\kappa}{2}\,h^2(t)\,\sigma^2_A(t)$$

$$= \frac{d(t)}{2} - \frac{\kappa}{2}\,\frac{[\,\sigma^2_a + d(t)\,]^2}{\sigma^2_z + d(t)}$$

Closed-form solution to equilibrium $h^2$

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



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

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

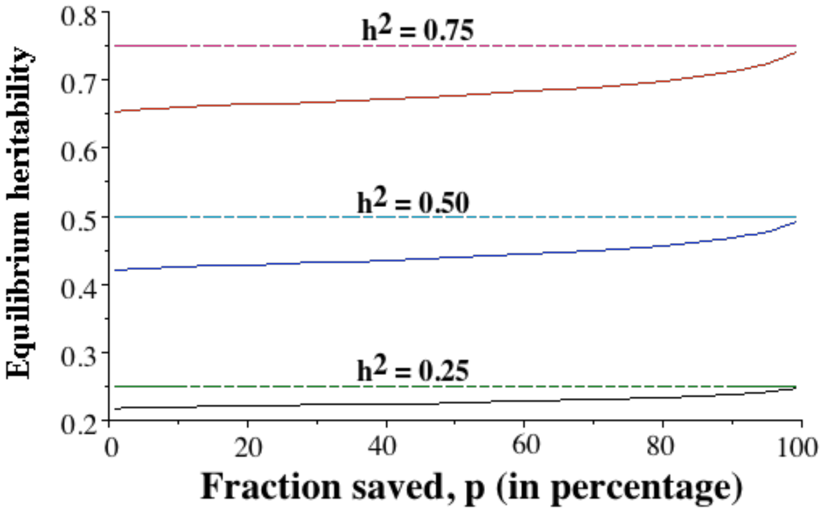
**Stabilizing Truncation Selection:** Middle fraction $p$ of the distribution saved

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

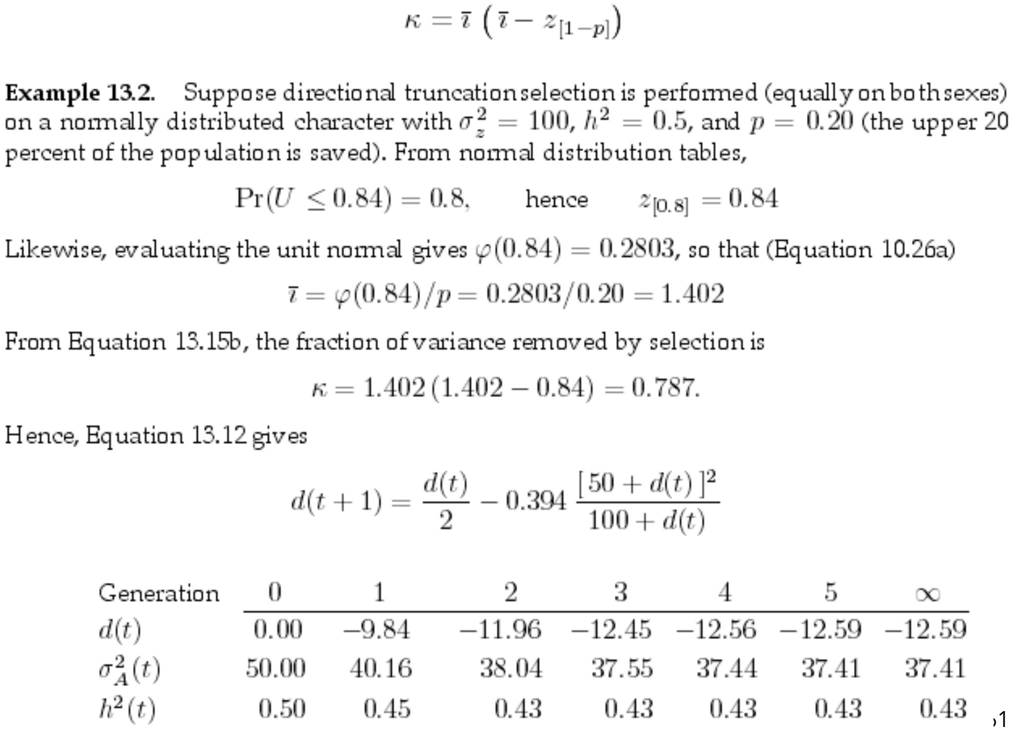
**DisruptiveTruncation Selection:** Uppermost and lowermost $p/2$ saved

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

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



# Directional truncation selection



$$\kappa = \bar\imath\,\left(\bar\imath - z_{[1-p]}\right)$$

**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,

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

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

$$\bar\imath = \varphi(0.84)/p = 0.2803/0.20 = 1.402$$

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

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

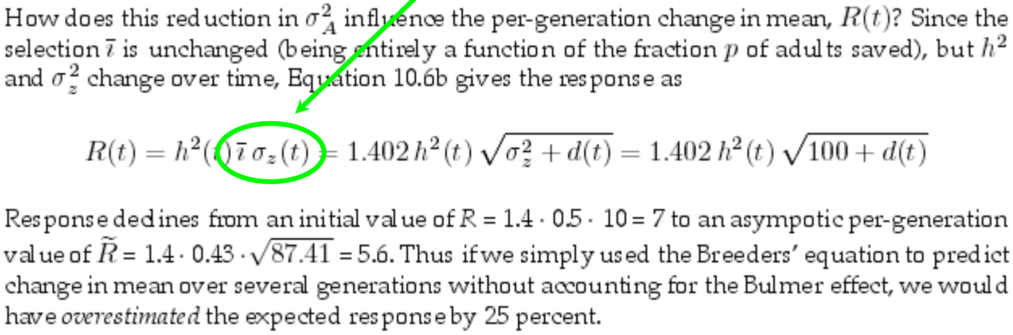
Hence, Equation 13.12 gives

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

| 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 |

## 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 $\bar\imath$ 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

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

Response declines from an initial value of $R$ = 1.4 · 0.5 · 10 = 7 to an asympotic per-generation value of $\widetilde{R}$ = 1.4 · 0.43 · $\sqrt{87.41}$ = 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

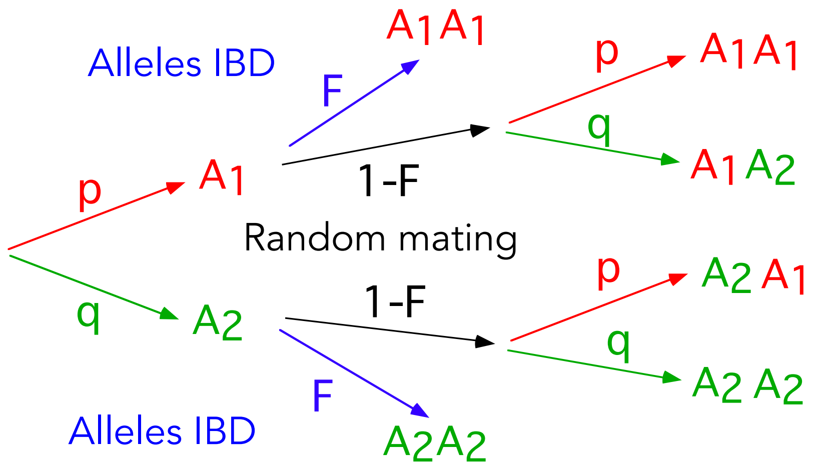
Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

# 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

# Genotype frequencies under inbreeding

- The inbreeding coefficient, F
- F = Prob(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



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

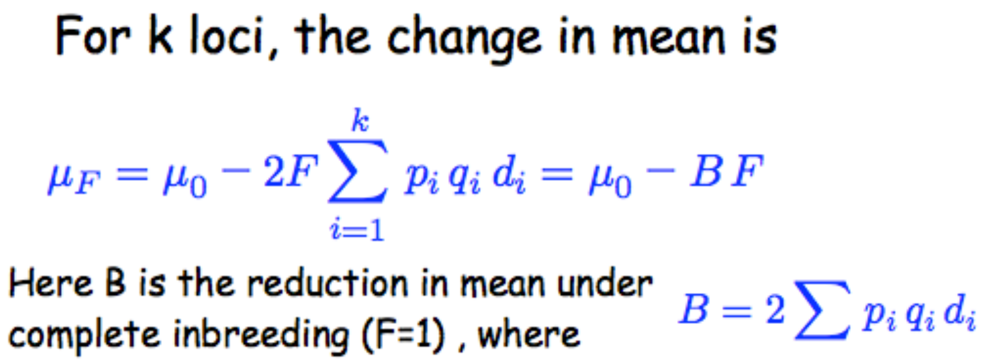
# Changes in the mean under inbreeding

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

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

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

$$\boxed{\mu_F = \mu_0 - 2Fpqd}$$



For k loci, the change in mean is

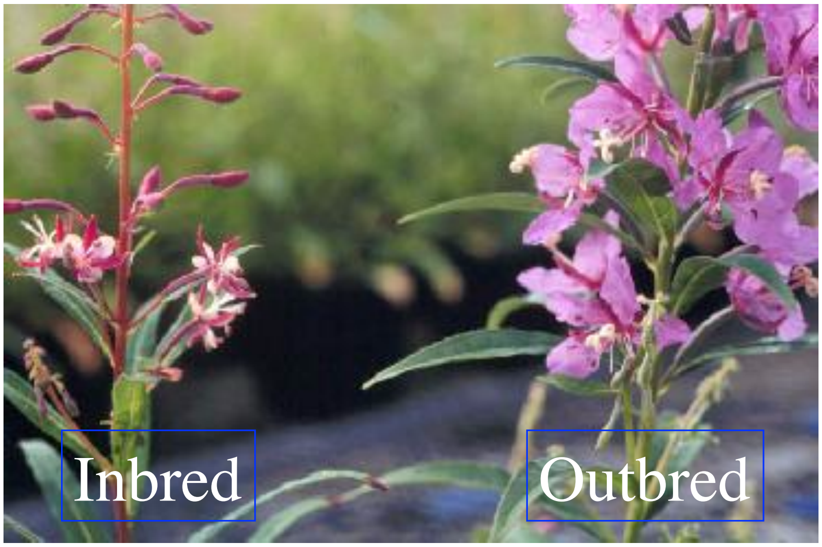
$$\mu_F = \mu_0 - 2F \sum_{i=1}^{k} p_i\, q_i\, d_i = \mu_0 - B\,F$$

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

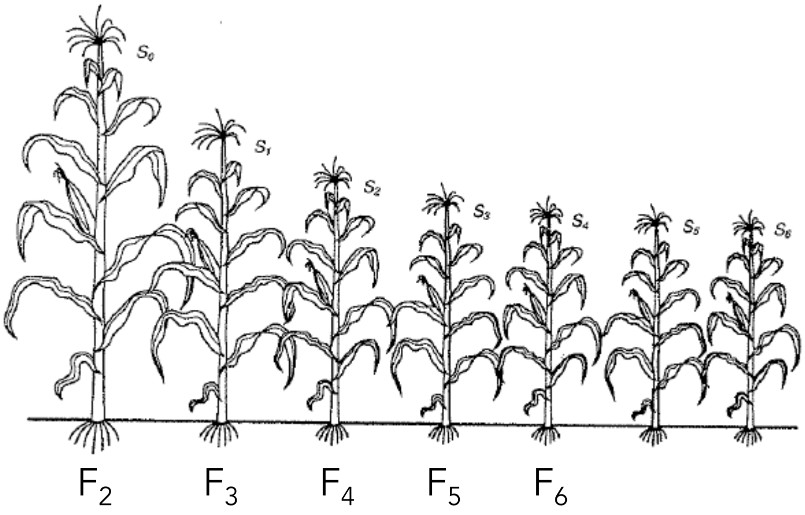
$$B = 2 \sum p_i\, q_i\, d_i$$

- There will be a change of mean value if dominance is present (d not 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

# Inbreeding Depression and Fitness traits



# Inbreeding depression



Example for maize height

# Fitness traits and inbreeding depression

- Often seen that inbreeding depression is strongest on fitness-relative 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

# 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.

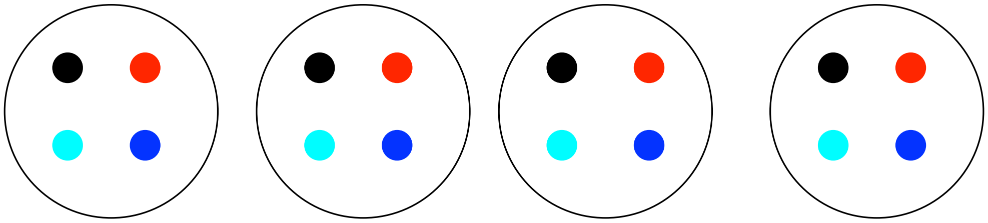
# Inbred depression in largely selfing lineages

- 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

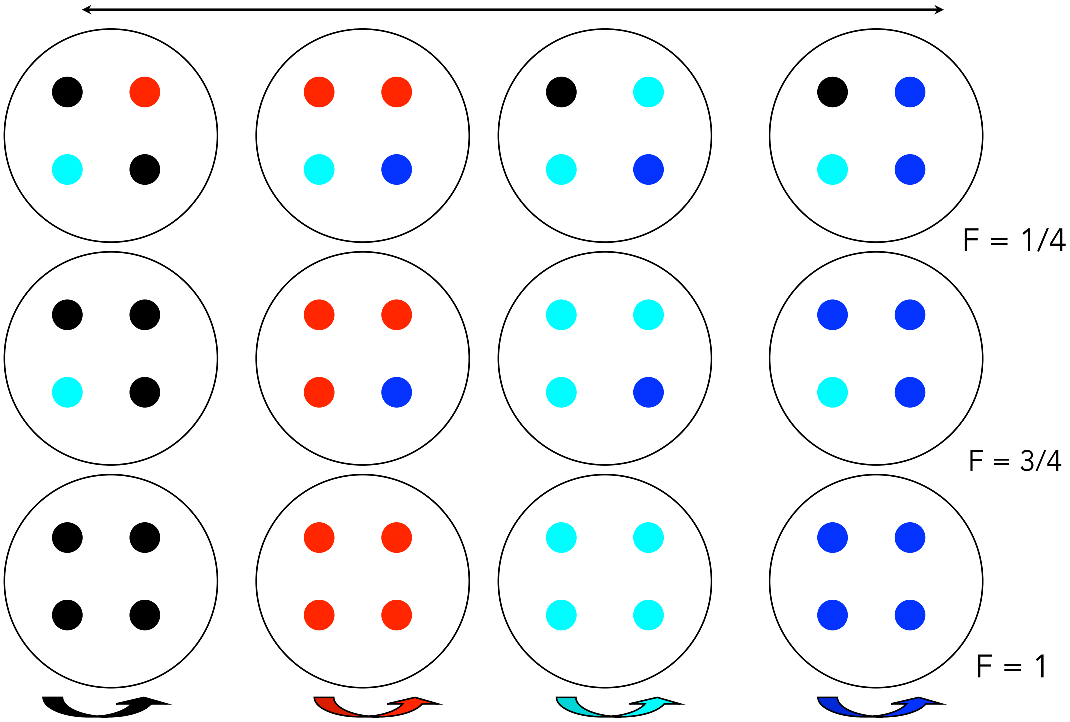
# Variance Changes Under Inbreeding

Inbreeding reduces variation within each population

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



F = 0



Between-group variance increases with F

F = 1/4

F = 3/4

F = 1

Within-group variance decreases with F

# 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 | $(1\text{-}F)\ V_A$ | 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 3).

# 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.

# 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 $F_1$,

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

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 $10^6$ | 13 x $10^6$ ha |
| Sorghum | 48 | 40 | 19 | 13 x $10^6$ | 9 x $10^6$ ha |
| Sunflower | 60 | 50 | 30 | 7 x $10^6$ | 6 x $10^6$ ha |
| Rice | 12 | 30 | 4 | 15 x $10^6$ | 6 x $10^6$ ha |

# 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

# 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 detassle 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

# 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**

$$F_2 = F_1 - \boxed{\frac{F_1 - \overline{P}}{n}}$$

**H/n**

$$H_{F_2} = H_{F_1}\left(1 - \frac{1}{n}\right)$$

**Only 1/n of heterosis lost vs. 1/2**

# 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

# Types of crosses

- The $F_1$ from a cross of lines A x 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 F1 offspring of B x C.
  - Denoted A x (B x C)
- A double (or four-way) cross is (A x B) x (C x D), the offspring from crossing an A x B $F_1$ with a C x D $F_1$.

# 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:
  - mean(A x {B x C}) = [mean(A x B) + mean(A x C)]/2
- The mean of a double (or four-way) cross is the average of all the single crosses,
  - mean({A x B} x {C x D}) = [mean(AxC) + mean(AxD) + mean(BxC) + mean(BxD)]/4

# 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 heteroic effects is

$$\frac{z_{AB} + z_{BA}}{2} - \frac{z_{AA} + z_{BB}}{2} = h^{I}_{AB}$$

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

$$z_{C\,AB} = \frac{2g^{I}_{C} + g^{I}_{A} + g^{I}_{B}}{4} + \frac{h^{I}_{CA} + h^{I}_{CB}}{2} + \frac{g^{M}_{A} + g^{M}_{B}}{2} + h^{M}_{AB} + g^{M}_{B}{}^{0} + \frac{r^{I}_{ab}}{2}$$

New individual heterosis of C x AB cross

Maternal genetic heteroic effect

"Recombinational loss" --- decay of the $F_1$ heterosis in the $F_2$

One estimate (confounded) of maternal heterosis

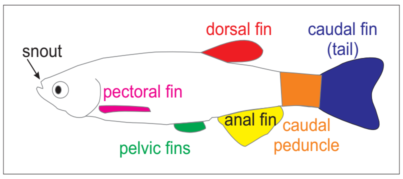
$$z_{C\,AB} = \frac{z_{CA} + z_{CB}}{2} = h^{M}_{AB} + \frac{r^{I}_{ab}}{2}$$

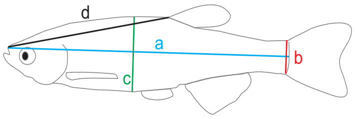
# Lecture 6: Correlated Characters

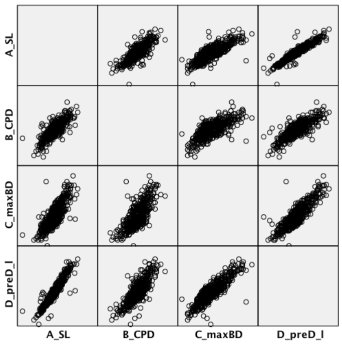
Background Reading: L&W chapter 21
Additional Reading: W&L Chapter 34
(Correlated response sections)

Steve Chenoweth lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

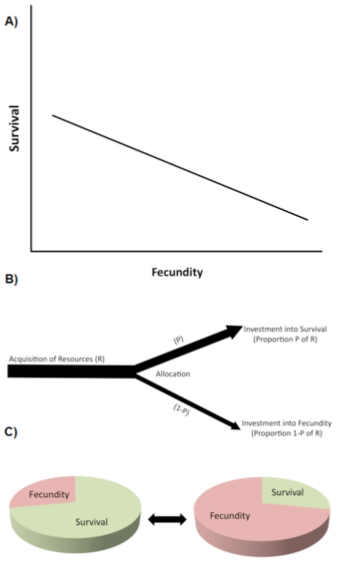
## Many quantitative traits are correlated





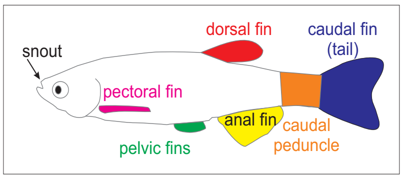


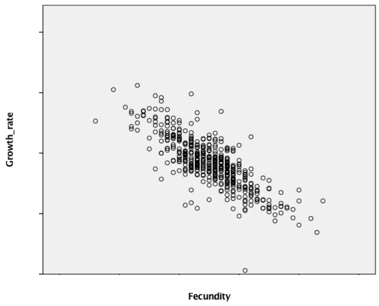
## Life History Evolution

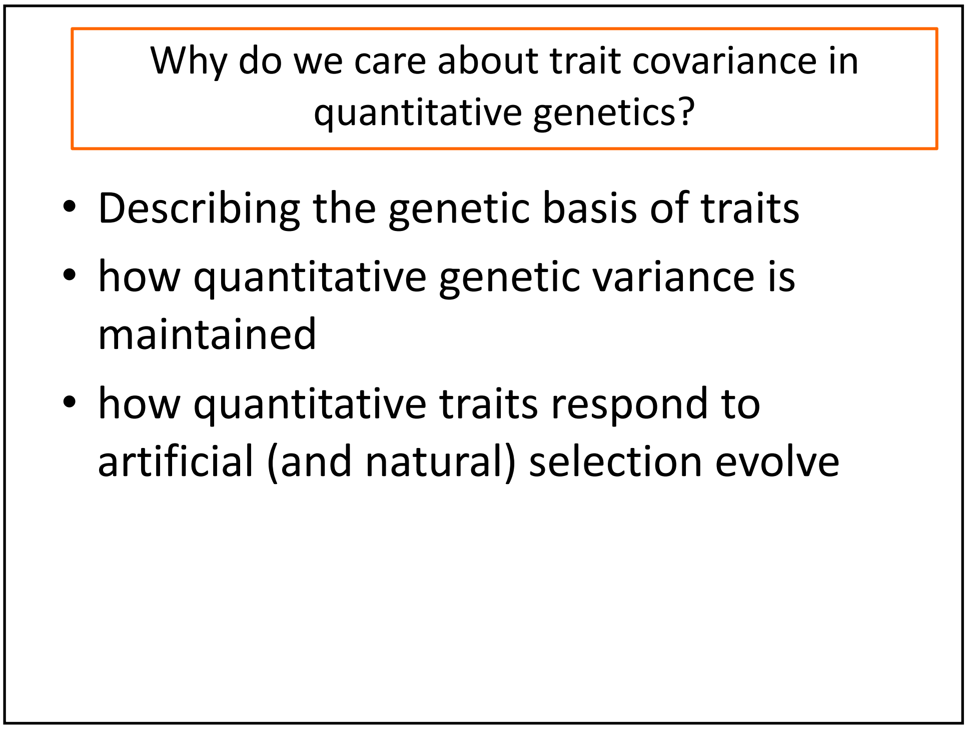


Fabian, D. & Flatt, T. (2012) Life History Evolution. Nature Education Knowledge 3(10):24

## Many quantitative traits are correlated

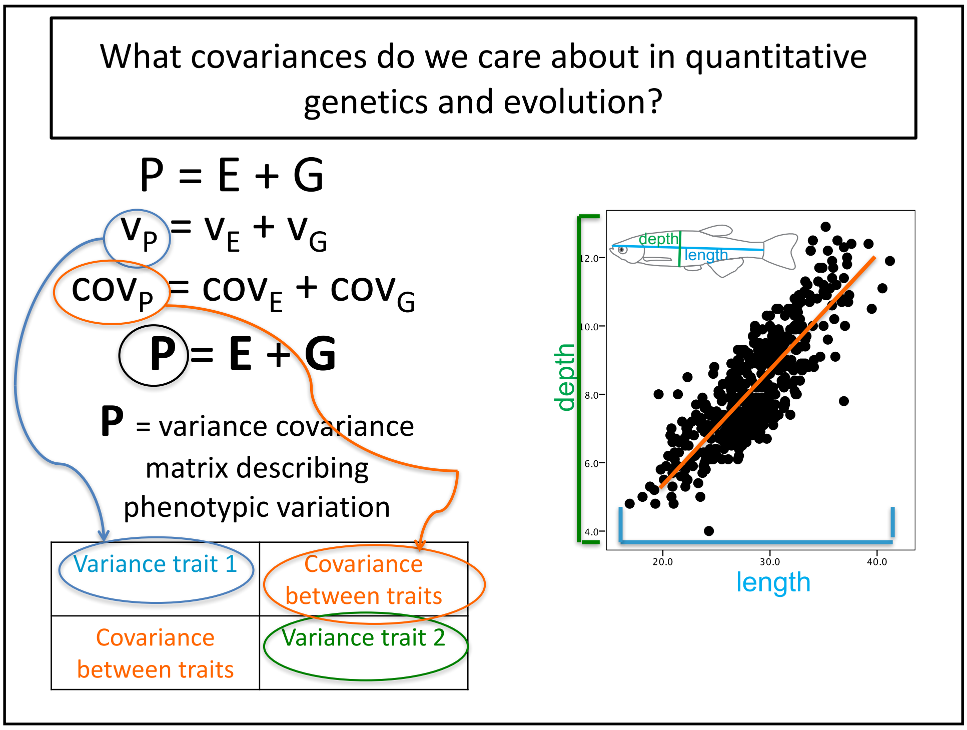






## Why do we care about trait covariance in quantitative genetics?

- Describing the genetic basis of traits
- how quantitative genetic variance is maintained
- how quantitative traits respond to artificial (and natural) selection evolve



## What covariances do we care about in quantitative genetics and evolution?

$$P = E + G$$

$$V_P = V_E + V_G$$

$$COV_P = COV_E + COV_G$$

$$\mathbf{P} = \mathbf{E} + \mathbf{G}$$

**P** = variance covariance matrix describing phenotypic variation

| | |
|---|---|
| Variance trait 1 | Covariance between traits |
| Covariance between traits | Variance trait 2 |

**P** – the phenotypic variance-covariance matrix

- Estimated directly from the observed phenotype recorded for each individual
- Underlies the estimation of *partial regression coefficients* of selection:

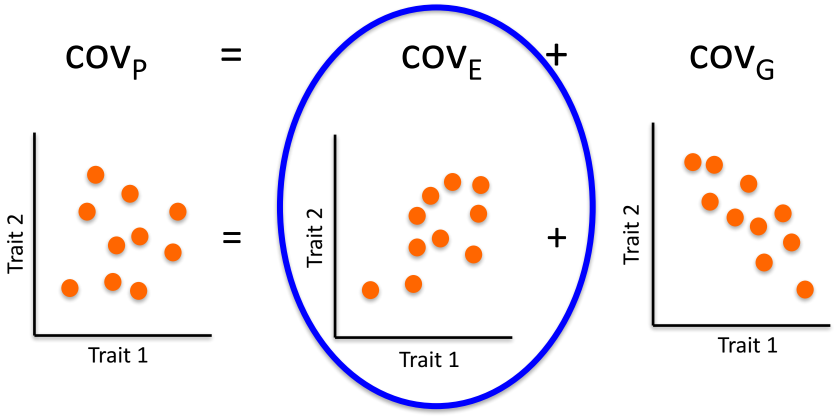
$$\beta = \mathbf{P}^{-1}\mathrm{s}$$

s = selection differential,
= mean of selected individuals – population mean

More on this in the next lecture…

What covariance matrices do we care about in quantitative genetics and evolution?

**P = E + G**



$\mathrm{COV}_P = \mathrm{COV}_E + \mathrm{COV}_G$

# Environmental effects

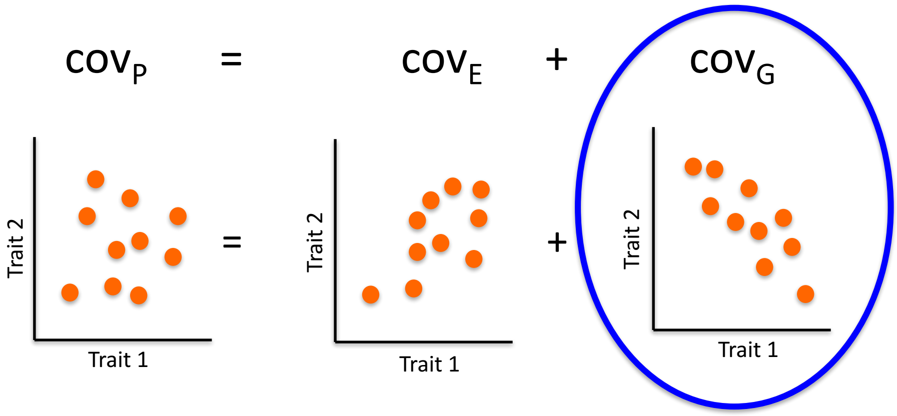
- Variation among individuals in their environmental experience can generate phenotypic differences, and affect multiple traits
  - E.g., nutrition environment
- Typically difficult to *know* and *measure* the environmental variation

  **E = P – G**
- Partition out environmental causes to focus on genetic

What covariance matrices do we care about in quantitative genetics and evolution?

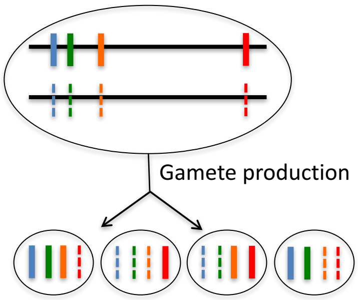
**P = E + G**

$$\mathrm{cov}_P = \mathrm{cov}_E + \mathrm{cov}_G$$



# Cause of genetic covariance

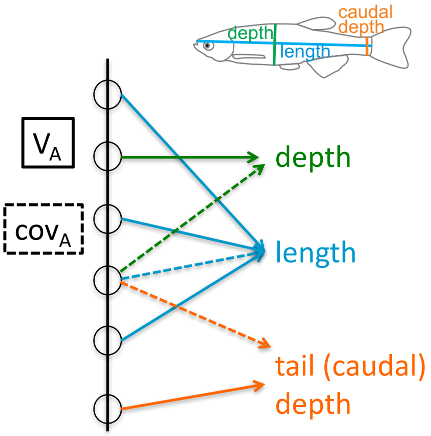
1. **Linkage** -alleles at different loci found together in same genotype more often than expected by chance
   - Physical
   - Selection
   - Non random mating

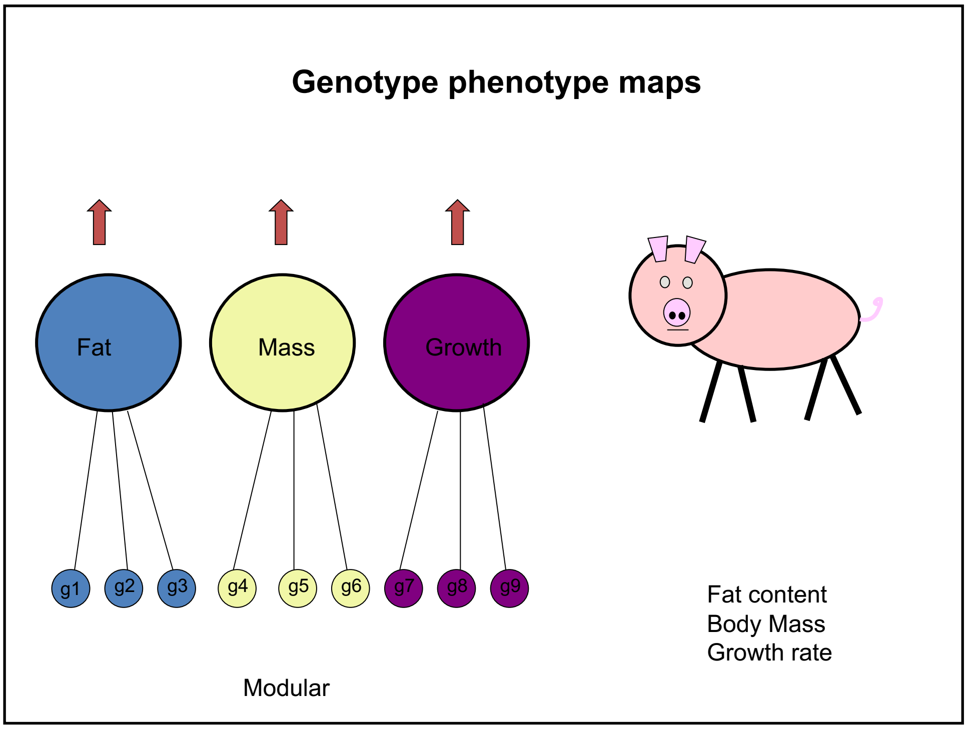


# Cause of genetic covariance

2. **Pleiotropy** -same genes affecting both traits

- Pleiotropy is considered primary cause of genetic covariance
  - more persistent than linkage, which can readily be broken down by recombination

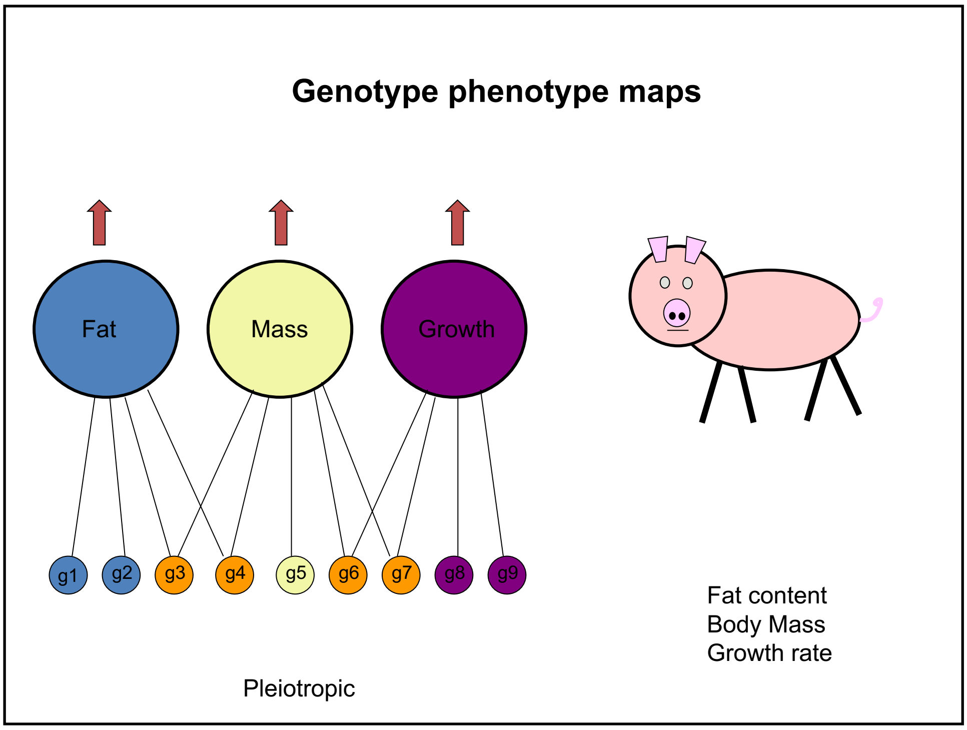




## Genotype phenotype maps

Fat content
Body Mass
Growth rate

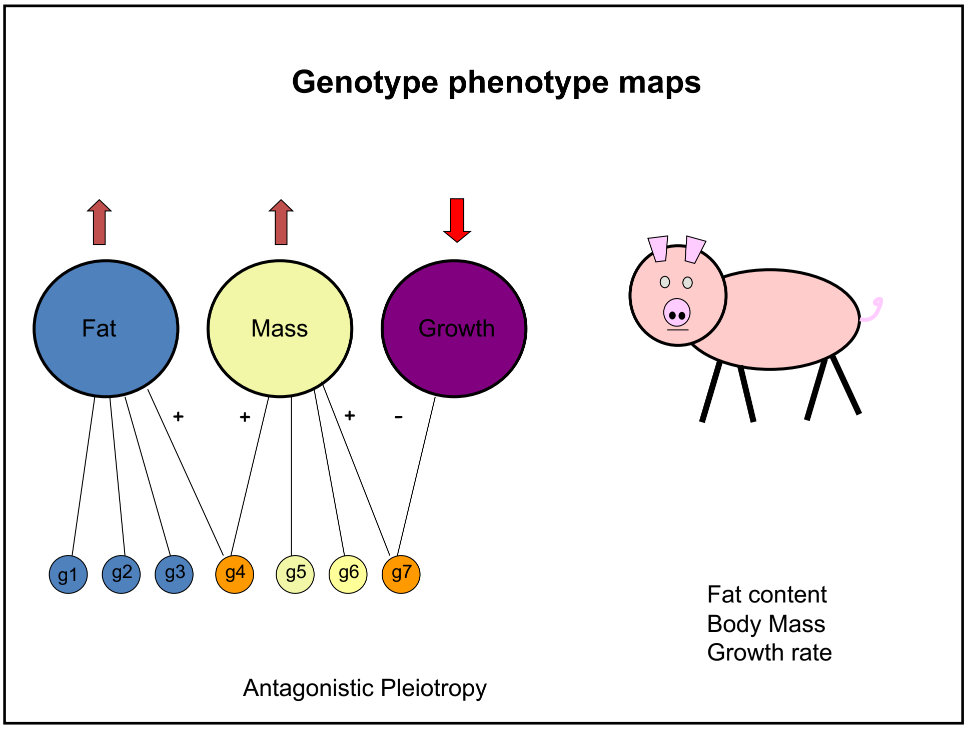
Modular



## Genotype phenotype maps

Fat content
Body Mass
Growth rate

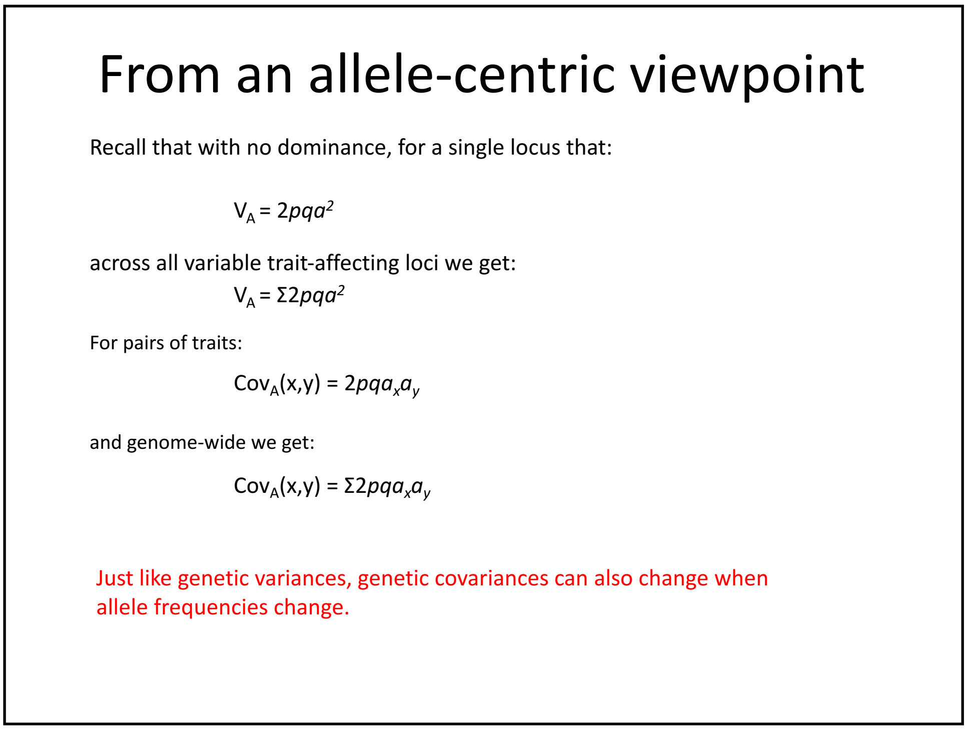
Pleiotropic



## Genotype phenotype maps

Fat content
Body Mass
Growth rate

Antagonistic Pleiotropy



# From an allele-centric viewpoint

Recall that with no dominance, for a single locus that:

$$V_A = 2pqa^2$$

across all variable trait-affecting loci we get:

$$V_A = \Sigma 2pqa^2$$

For pairs of traits:

$$\mathrm{Cov}_A(x,y) = 2pqa_xa_y$$

and genome-wide we get:

$$\mathrm{Cov}_A(x,y) = \Sigma 2pqa_xa_y$$

Just like genetic variances, genetic covariances can also change when allele frequencies change.

# Covariance or Correlation?

- Covariances are on a scale of trait products, like a variance is on a scale of trait values squared.
  - Hard to think about and compare

- Difficult to compare them directly so we often think about them in terms of correlations:
  - Correlations more intuitive and easier to compare

$$r_A = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x) \times \operatorname{var}(y)}}$$

## Interpreting Genetic Correlations

- Genetic correlations are bound between –1 and 1

- The **sign indicates only the *net* directionality** of pleiotropic effects. Whether standing variation affects trait 1 and trait 2 in similar ways.

- The **magnitude** indicates how much genetic variation is shared between traits.

### Interpreting Genetic Correlations

- $r_A > 0$ ?

  Genetic variance in both traits controlled by some genes which are the same. These variable loci cause an increase in trait 1 and an increase in trait 2.

- $r_A = 1$ ?

  Perfectly shared control, essentially the "same trait" from a genetic perspective

- $r_A < 0$ ?

  Partially shared genetic basis, BUT genes which increase trait 1 lead to a decrease in trait 2

- $r_A = 0$ ?

  No genes which exhibit genetic variance affect trait 1 and trait 2 together.

  *BUT remember it's the net effect cross loci, there may indeed be pleiotropy but opposing effects cancel each other out,*

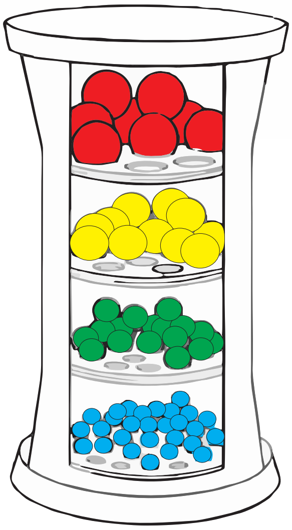
## How do we estimate genetic correlations?

- **Method 1**: Artificial selection experiments
  - Correlated responses to direct selection
  - *Genetic covariance among traits affects the evolution of trait means*

- **Method 2:** Using the same statistical machinery as to estimate $V_A$, we can estimate $\text{cov}_A$
  - Phenotypic data on >1 trait + pedigree
  - Breeding values:
    - $V_A$ = variance in breeding values
    - $\text{cov}_A$ = covariance in breeding values of two traits

# Method 1: Selection *of* vs. *for*

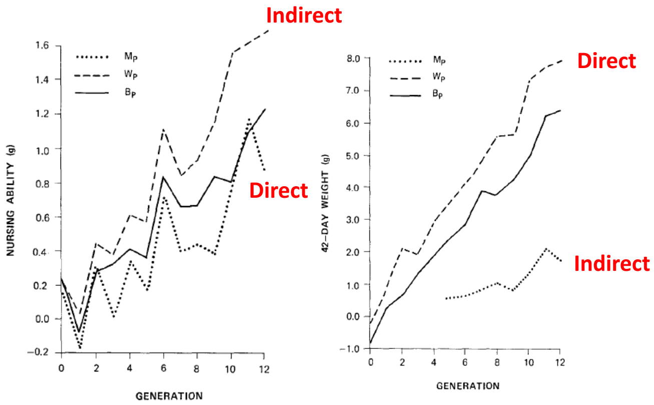
Selection ***for*** small balls results in selection ***of*** small, <u>blue</u> balls



There is a **correlated response** in colour to selection for size

Sober, E. 1984. The Nature of Selection.

## Nagai et al 1978 selected for nursing ability and body weight in mice



....... Nursing ability
------ Weight (42 days)
Solid: index of both

*Nagai et al. 1978, Genetics, 88,761-780.*

$$r^2_{\mathrm{A(nursing,weight)}} = \frac{CR_n}{R_n} \; \frac{CR_w}{R_w}$$

Where:

R = response to selection<br>
*CR* = correlated response to selection<br>
$CR_Y = i_X\, h_X\, h_Y\, rA\, \sigma_{PY}$

*Where*<br>
$i_X$ = selection intensity<br>
$h^2$ = realised heritability (R = $Sh^2$ thus, $h^2 = R/S$)<br>
$\sigma_{PY}$ = phenotypic variance in Trait Y

Empirically CR can be estimated from the selection response in the indirectly selected trait

See Also Falconer & Mackay 1996, Chapter 19

| Trait selected | Nursing | Weight |
|---|---|---|
| Response Nursing | **0.080** | *0.134* |
| Response weight | *0.197* | **0.680** |

$$r_{\mathrm{A(nursing,weight)}} = \left( \frac{CR_n}{R_n} \; \frac{CR_w}{R_w} \right)^{1/2}$$

= [(0.134/0.080) x (0.197/0.680)] ½

= 0.485 ½

= 0.70

*Nagai et al. 1978, Genetics, 88,761-780.*

EVOLUTION OF FLORAL DISPLAY IN *EICHHORNIA PANICULATA* (PONTEDERIACEAE): DIRECT AND CORRELATED RESPONSES TO SELECTION ON FLOWER SIZE AND NUMBER

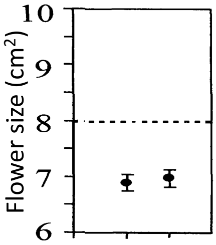
*Evolution,* 54(5), 2000, pp. 1533–1545

ANNE C. WORLEY[1] AND SPENCER C. H. BARRETT[2]
*Department of Botany, University of Toronto, 25 Willcocks Street, Toronto ON, M5S-3B2, Canada*
[2]*E-mail: barrett@botany.utoronto.ca*

Water hyacinth

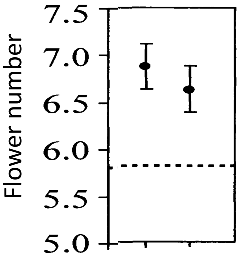
http://labs.eeb.utoronto.ca/barrett/Pho
l

Hypothesis: a plant has finite resources that can be allocated to either **larger** or **more** flowers to attract pollinators & increase reproductive success. This will result in a negative correlation between flower number and size.



Apply selection to decrease flower size for 2 generations

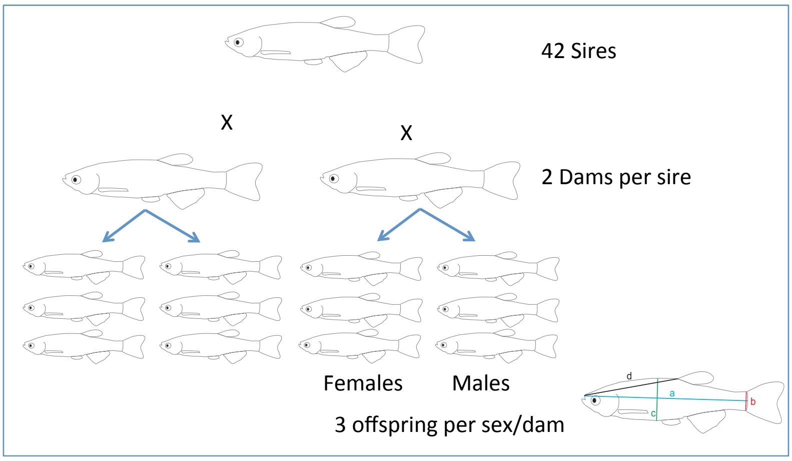
- Estimated $h^2$ from relatives = 0.48
- realised $h^2$ from selection response= 0.45



Flowers also had:
more nectar
more pollen
bigger ovules

Selection lines also had more flowers.
Realised genetic correlation: r = -0.6

# Method 2: Estimation in a breeding design



Paternal full-sib half-sib breeding design for the analysis of zebrafish body size and shape

# Method 2: Estimation in a breeding design

**Observational model**

To estimate genetic variance we use the following random effects general linear model to describe our breeding design, which is actually a nested ANOVA:

$$z_{ijk} = \mu + s_i + d_{ij} + e_{ijk}$$

where:

$z_{ijk}$ is the trait value of the *k*th offspring of the *j*th dam which was mated to the *i*th sire,
$\mu$ is the population mean,
$s_i$ is the effect due to the *i*th sire,
$d_{ij}$ is the effect due to the the *j*th dam mated to the *i*th sire, and
$e_{ijk}$ is the unexplained residual.

*Sire*, *dam within sire* are all RANDOM EFFECTS in this model.

# Linking observational components to causal genetic components of (co)variance

For a single trait the total phenotypic variance is simply given by:

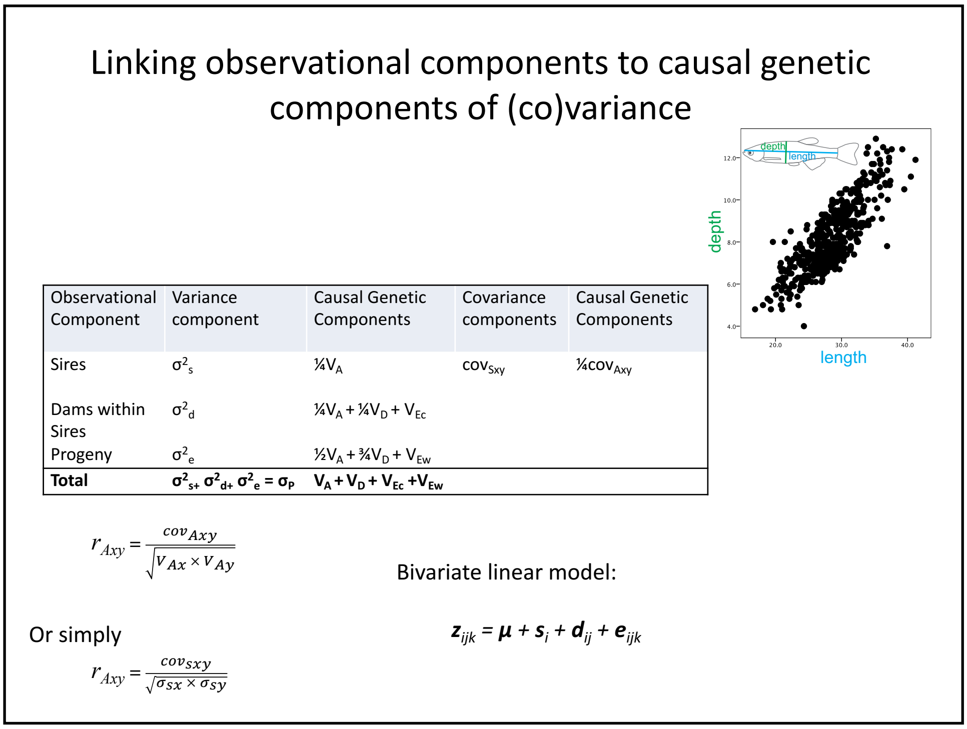
$$\sigma^2_z \ = \sigma^2_s + \sigma^2_d + \sigma^2_e$$

As Bruce showed you, we now know what's inside these variance components in terms of genetic effects:

| Observational Component | Variance component | Causal Genetic Components |
|---|---|---|
| Sires | $\sigma^2_s$ | $¼V_A$ |
| Dams within Sires | $\sigma^2_d$ | $¼V_A + ¼V_D + V_{Ec}$ |
| Progeny | $\sigma^2_e$ | $½V_A + ¾V_D + V_{Ew}$ |
| **Total** | $\boldsymbol{\sigma^2_{s+}\, \sigma^2_{d+}\, \sigma^2_e = \sigma_P}$ | $\mathbf{V_A + V_D + V_{Ec} + V_{Ew}}$ |

$$h^2 = V_A/V_P = \frac{4\ \sigma^2_s}{\sigma^2_{s+}\ \sigma^2_{d+}\ \sigma^2_e}$$

We can now extend this to deal with covarianes



# Linking observational components to causal genetic components of (co)variance

depth
length
12.0
10.0
8.0
6.0
4.0
20.0
30.0
40.0
depth
length

| Observational Component | Variance component | Causal Genetic Components | Covariance components | Causal Genetic Components |
|---|---|---|---|---|
| Sires | $\sigma^2{}_s$ | $¼V_A$ | $cov_{Sxy}$ | $¼cov_{Axy}$ |
| Dams within Sires | $\sigma^2{}_d$ | $¼V_A + ¼V_D + V_{Ec}$ | | |
| Progeny | $\sigma^2{}_e$ | $½V_A + ¾V_D + V_{Ew}$ | | |
| **Total** | $\boldsymbol{\sigma^2{}_{s+}\,\sigma^2{}_{d+}\,\sigma^2{}_e = \sigma_P}$ | $\boldsymbol{V_A + V_D + V_{Ec} + V_{Ew}}$ | | |

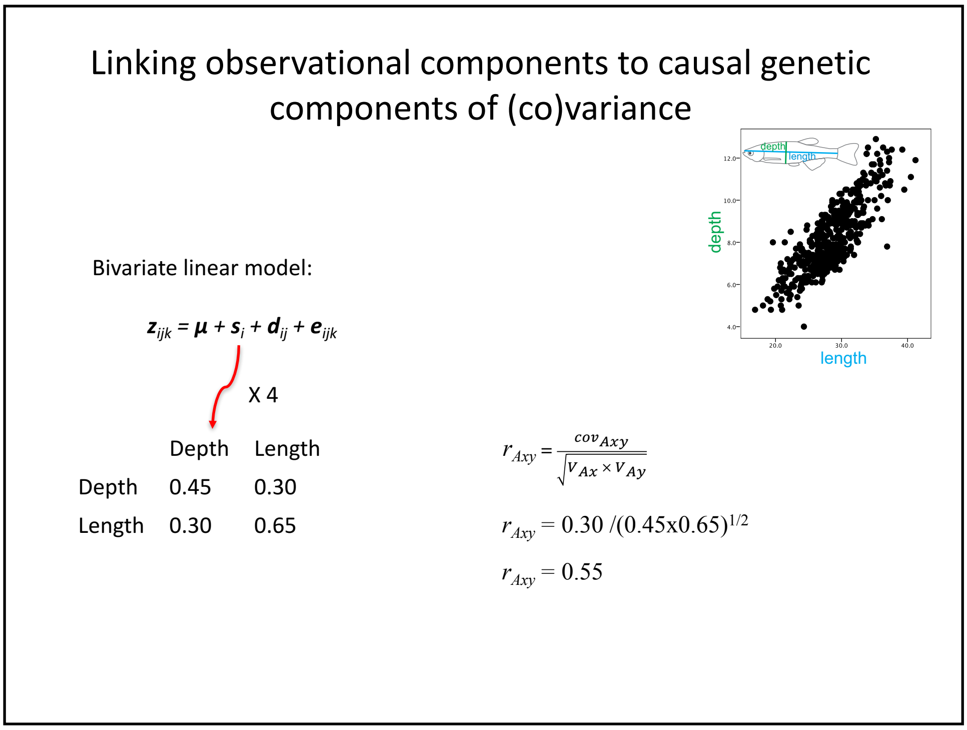
$$r_{Axy} = \frac{cov_{Axy}}{\sqrt{V_{Ax} \times V_{Ay}}}$$

Bivariate linear model:

Or simply

$$z_{ijk} = \mu + s_i + d_{ij} + e_{ijk}$$

$$r_{Axy} = \frac{cov_{Sxy}}{\sqrt{\sigma_{Sx} \times \sigma_{Sy}}}$$



# Linking observational components to causal genetic components of (co)variance

depth
length
12.0
10.0
8.0
6.0
4.0
20.0
30.0
40.0
depth
length

Bivariate linear model:

$$z_{ijk} = \mu + s_i + d_{ij} + e_{ijk}$$

X 4

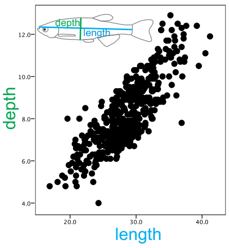
| | Depth | Length |
|---|---|---|
| Depth | 0.45 | 0.30 |
| Length | 0.30 | 0.65 |

$$r_{Axy} = \frac{cov_{Axy}}{\sqrt{V_{Ax} \times V_{Ay}}}$$

$$r_{Axy} = 0.30\ /(0.45\text{x}0.65)^{1/2}$$

$$r_{Axy} = 0.55$$

## Linking observational components to causal genetic components of (co)variance



Bivariate linear model:

$$z_{ijk} = \mu + s_i + d_{ij} + e_{ijk}$$

X 4

$$\begin{array}{c|cc} & \text{Depth} & \text{Length} \\ \hline \text{Depth} & 0.45 & 0.30 \\ \text{Length} & 0.30 & 0.65 \end{array}$$

The genetic variance-covariance matrix, **G**

$$r_{Axy} = \frac{cov_{Axy}}{\sqrt{V_{Ax} \times V_{Ay}}}$$

$$r_{Axy} = 0.30 /(0.45 \text{ x } 0.65)^{1/2}$$

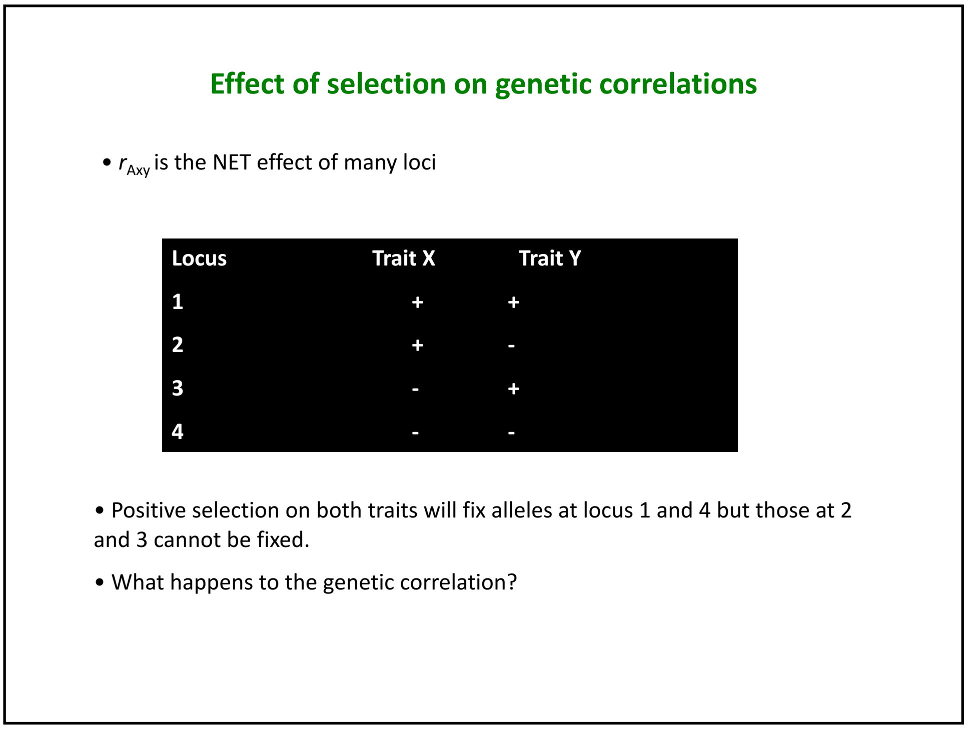
$$r_{Axy} = 0.55$$

# FAQs

“My estimate of the genetic correlation is greater than 1! Isn’t supposed to be bounded between -1 and 1?“

“Okay then so why can’t I just correlate sire means, it would be a product moment correlation and won’t go out of bounds…”

“I just about killed myself (my student) breeding thousands of animals but my genetic correlation has a huge standard error”

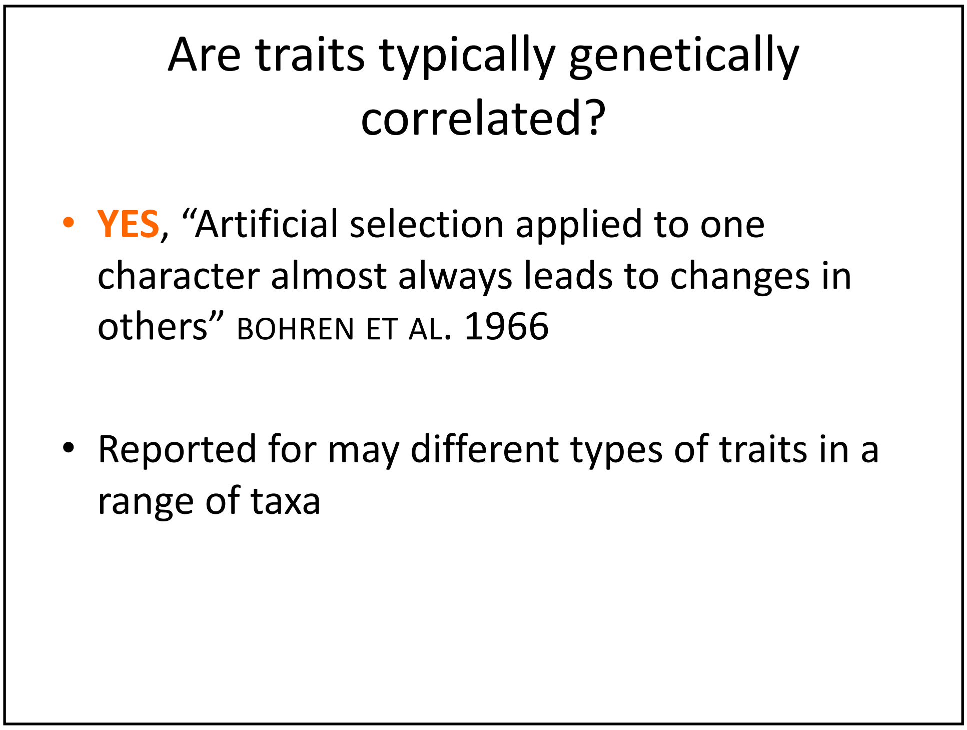


## Effect of selection on genetic correlations

- $r_{Axy}$ is the NET effect of many loci

| Locus | Trait X | Trait Y |
|---|---|---|
| 1 | + | + |
| 2 | + | - |
| 3 | - | + |
| 4 | - | - |

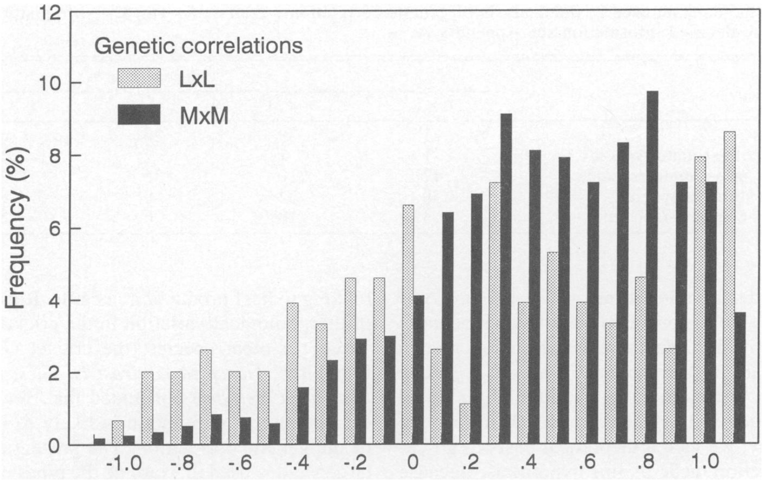
- Positive selection on both traits will fix alleles at locus 1 and 4 but those at 2 and 3 cannot be fixed.
- What happens to the genetic correlation?



## Are traits typically genetically correlated?

- **YES**, "Artificial selection applied to one character almost always leads to changes in others" BOHREN ET AL. 1966
- Reported for may different types of traits in a range of taxa

L = life history M= morphology



Roff, D.A. 1996. Evolution 50: 1392-1403

## Extending genetic correlations

- Thus far our focus has been on traits expressed **by the same individual** at a **specific point in time.**
- Genetic correlations routinely measured between:
  - Growth stages
  - Sexes
  - Environments

## What is the genetic correlation among environments?

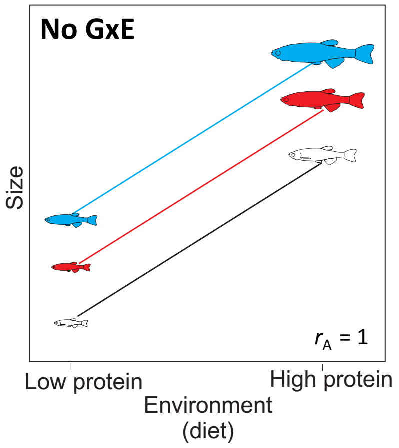
- Falconer (1952) had the idea to treat the same trait, measured in two different environments, as two different traits, and estimate the genetic correlation between these two "traits"

- If there is no GxE, $r_G = 1.0$
  - Alleles (genotypes) have the same effect on the trait (relative to the population mean) in each environment – parallel reaction norms
  - Selection in one environment will cause the trait to change value in other environments too

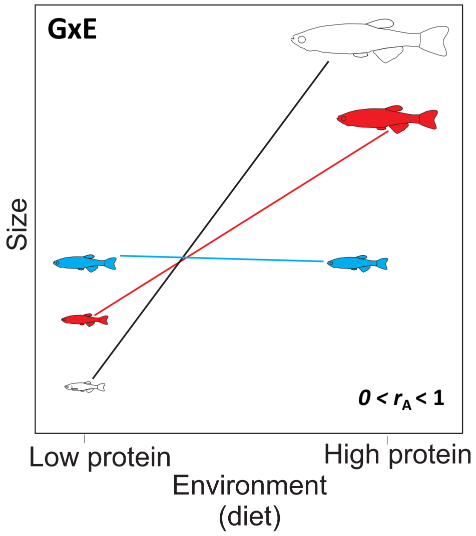
## Cross – Environment genetic correlations

Cross environment **G**

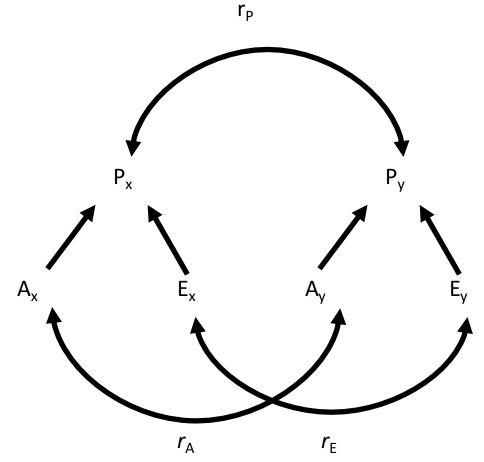
| | **Body size in High** | **Body size in Environment B** |
|---|---|---|
| Body size in Environment A | Variance in size within Env A | Covariance between size in Env A & B |
| Body size in Environment B | Covariance between size in Env A & B | Variance in size within Env B |



Selection for bigger fish on high protein diet will cause correlated evolution of larger size when the population was fed on a low protein diet



## Summarising relationships between traits



$r_A$ = correlation of breeding values for traits x and y.

Due to pleiotropy and linkage disequilibrium.

$r_E$ = correlation of environmental deviations for traits x and y.

Due to exposure of two traits to the same environment.

Contains nonadditive genetic effects.

# Lecture 7: An Introduction to Evolutionary Quantitative Genetics

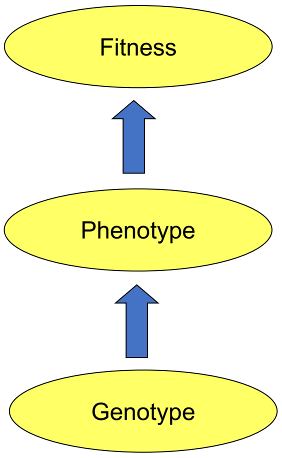
Background Reading: W&L v1. chapter 28, 29,
W&L v2: Chapter 34
Additional Reading: W&L v1. Chapter 27

Steve Chenoweth lecture notes
Introduction to Quantitative
Genetics
SISG, Brisbane
6 – 7 Feb 2017

## Outline

1. Measuring natural selection on multiple traits
2. Predicting multi-trait responses to selection
3. Genetic constraints: when natural selection ≠ adaptation
4. What processes maintain genetic variance in complex traits?

## Two Maps in Evolutionary Quantitative Genetics



**Quantitative Genetic Tools**

1. **Fitness – Phenotype Map**

- selection gradients /surfaces/differentials

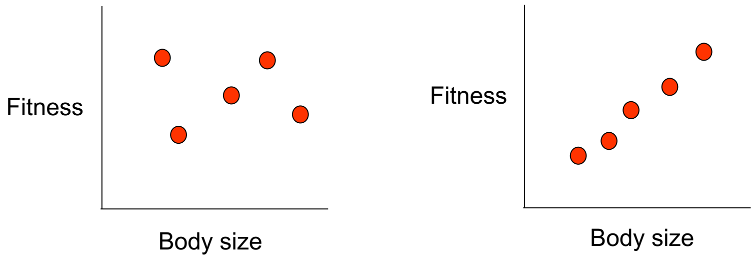
**2. Genotype-Phenotype Map**

**Indirect**
- Genetic variance, heritability
- heritability/genetic correlations (**G**-matrix)

**Directly**
- QTL mapping
- Genome Wide Association studies

## 1. Measuring Natural Selection on Quantitative Traits

## Covariance between trait and fitness



## What is fitness?

**Absolute fitness**

*Number of descendants an individual leaves at the start of the next generation*

***Note:*** no info on the rate of change under selection

**Relative fitness:** *of a specific phenotype/ genotype is its fitness relative to the weighted average fitness of all other phenotypes/ genotypes within the population*

### Example (phenotypic): body size and fitness in cane toads

| male | size | #mates | fecundity<br># eggs/mating | absolute<br>fitness | relative<br>fitness |
|---|---|---|---|---|---|
| 1 | 145 | 1 | 25820 | 25820 | 1.164027843 |
| 2 | 128 | 1 | 22670 | 22670 | 1.022018249 |
| 3 | 148 | 0 | 0 | 0 | 0 |
| 4 | 138 | 2 | 7230 | 14460 | 0.651891658 |
| 5 | 141 | 3 | 15986 | 47958 | 2.16206225 |

**Absolute fitness:**

Male 4 = 2 x 7,230 = 14460

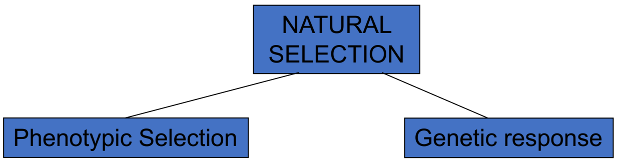
**Relative fitness**

*absolute fitness(male 4) / mean absolute fitness*

= 14,460 / [ (25,820 + 22,670 + 0 + 14,460 + 47,958)/5]
= 14460 / 22182
= 0.65

adapted from Walsh, 2007

## The quantitative genetic view of selection


NATURAL SELECTION

Phenotypic Selection

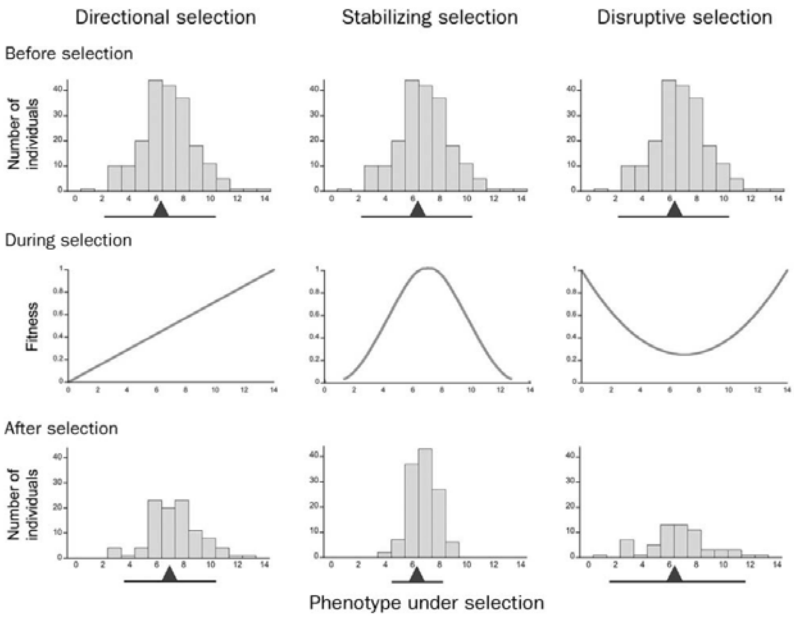
Genetic response

Selection gradient

genetic variation

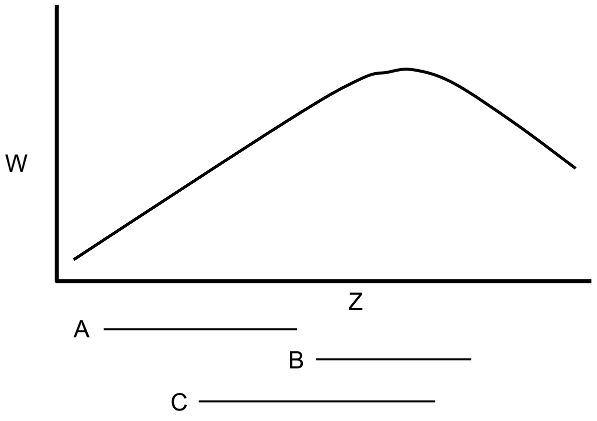
- **Phenotypic Selection:** Consistent difference in fitness among phenotypes, acting within a single generation.
- **Response to selection:** Change in population mean phenotype from one generation to the next.
- Thus selection acts on **phenotypes** but its effect on evolution (change in allele frequencies) depends on the *mapping of phenotype to genotype.*

The three forms of phenotypic selection



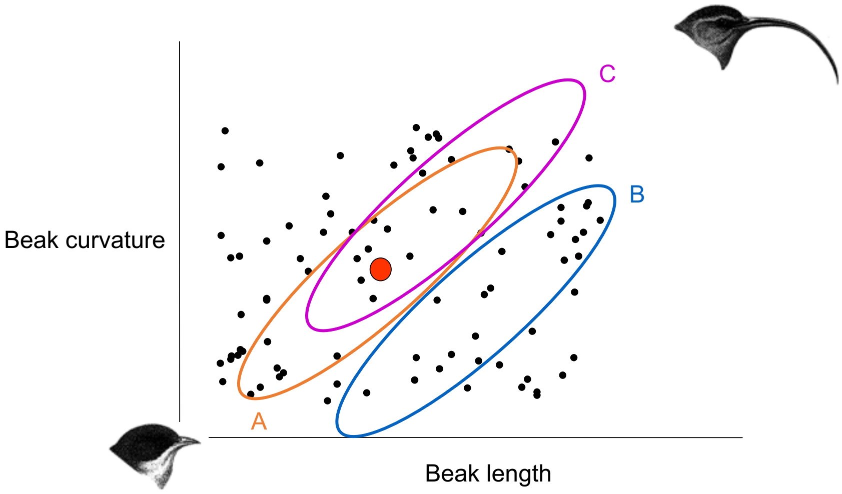
Kingsolver and Pfennig, 2007

## Combinations of forms may exist



Endler, 1986

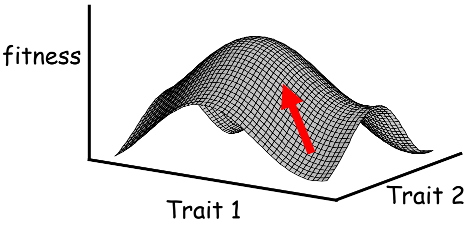
## Correlational selection



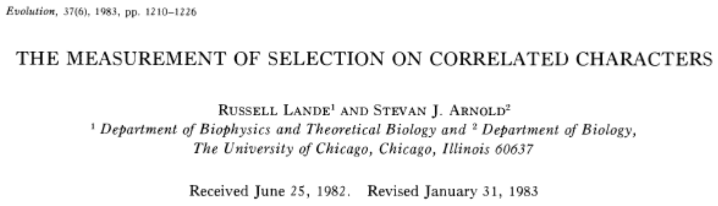
Can lead to the evolution of highly correlated traits

Endler, 1986

## Correlational selection



How can we compare natural selection across traits, species and populations?



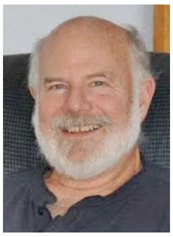
*Evolution*, 37(6), 1983, pp. 1210–1226

THE MEASUREMENT OF SELECTION ON CORRELATED CHARACTERS

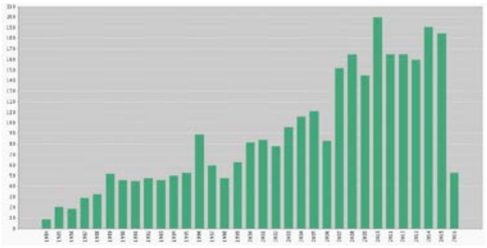
RUSSELL LANDE[1] AND STEVAN J. ARNOLD[2]
[1] *Department of Biophysics and Theoretical Biology and* [2] *Department of Biology,*
*The University of Chicago, Chicago, Illinois 60637*

Received June 25, 1982. Revised January 31, 1983

Russ Lande



Steve Arnold



- Landmark paper in evolutionary genetics (2930 citations)
- Uses multiple linear regression to estimate “selection gradients”
- Easy to collect data and compare selection
- Use to predict evolution

Fitness is a surface: *w = f(z) + error*

What is the form of *f(z) ?* Linear, flat, bumpy

Schluter 2000

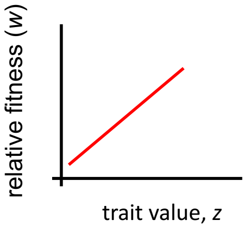
## Directional selection gradients

**Univariate: single trait**

$w = \alpha + \beta z + e$

*Simple linear regression*

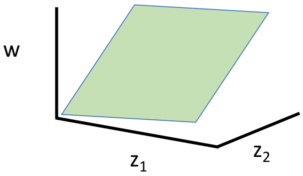
$\beta = cov(z,w) / var(z)$



**Multivariate: multiple traits**

$w = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_n z_n + e$

*Multiple linear regression*



Selection is represented as a **vector** of partial regression coefficients

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_n \end{pmatrix}$$

Lande and Arnold 1983 Evolution36:1210-1226

### How the Horned Lizard Got Its Horns

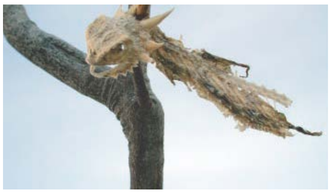
Kevin V. Young,[1] Edmund D. Brodie Jr.,[1] Edmund D. Brodie III[2*]

Flat tailed horned lizard
*Phrynosoma mcalli*

+

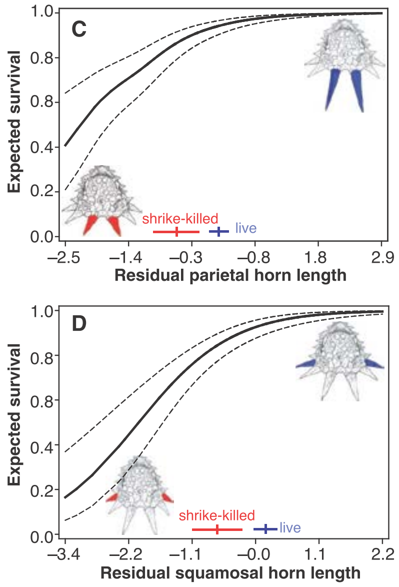
Loggerhead shrike
*Lanius ludovicianus*

=



www.sciencemag.org SCIENCE VOL 304 2 APRIL 2004 65

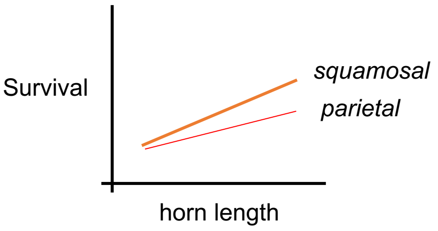
# Visualisations



# Directional selection gradients

**Interpretation**

- $w = \alpha + \beta z + e$



Survival

*squamosal*

*parietal*

horn length

*Survival = 0.0945 x squamosal horn length + intercept*

*P = 0.007*

*Survival = 0.0549 x parietal horn length + intercept*

*P = 0.055*

- An *increase* in one phenotypic standard deviation in squamosal horn length *increases* survival by 9%

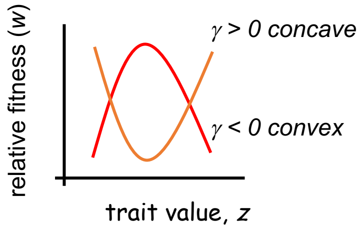
  $\beta$ = cov(z,w) / var(z)

## Quadratic and correlational selection gradients

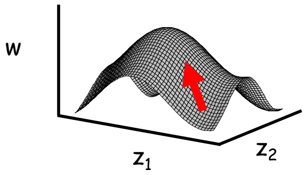
**Univariate: single trait**

$w = \alpha + \beta z + \gamma/2\ z^2$

*Quadratic regression*



**Multivariate: multiple traits**



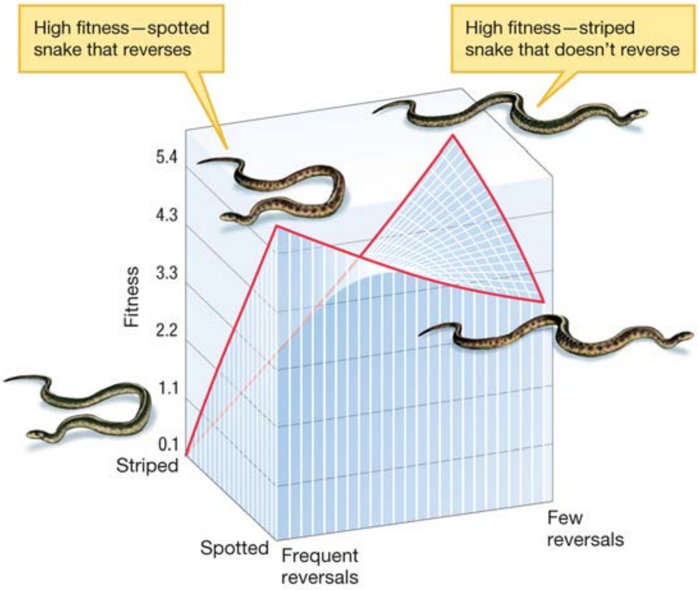
$w = \alpha + \beta_1 z_1 + \beta_2 z_2 + \gamma_1/2\ z_1 + \gamma_2/2z_2^2 + \gamma_{12} z_1 z_2$

Nonlinear selection is represented as a **MATRIX** of partial regression coefficients

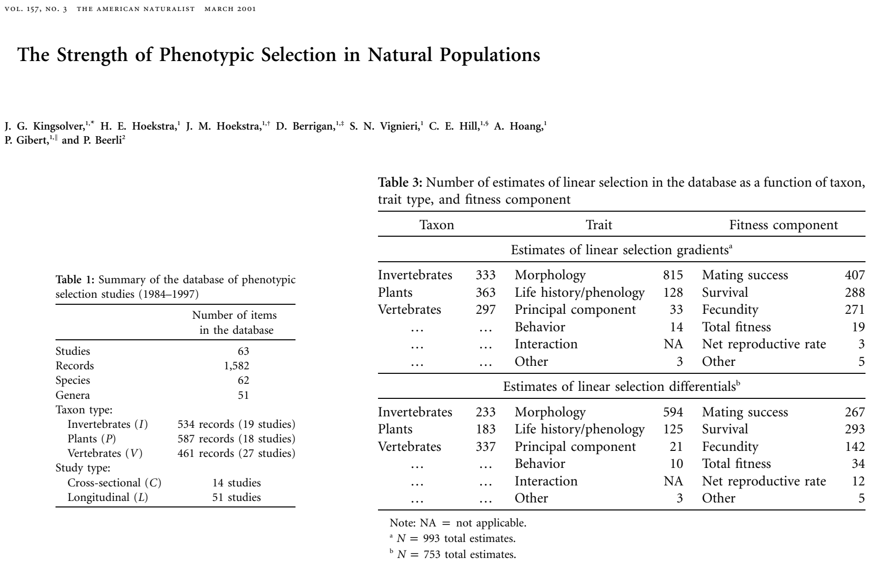
$$\gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}$$

## Correlational selection: Garter Snakes

*Evolution*, 46(5), 1992, pp. 1284–1298



## How strong is selection in nature?


vol. 157, no. 3 the american naturalist march 2001

### The Strength of Phenotypic Selection in Natural Populations

J. G. Kingsolver,[1,*] H. E. Hoekstra,[1] J. M. Hoekstra,[1,†] D. Berrigan,[1,‡] S. N. Vignieri,[1] C. E. Hill,[1,§] A. Hoang,[1]
P. Gibert,[1,‖] and P. Beerli[2]

**Table 1:** Summary of the database of phenotypic selection studies (1984–1997)

| | Number of items in the database |
|---|---|
| Studies | 63 |
| Records | 1,582 |
| Species | 62 |
| Genera | 51 |
| Taxon type: | |
| Invertebrates (*I*) | 534 records (19 studies) |
| Plants (*P*) | 587 records (18 studies) |
| Vertebrates (*V*) | 461 records (27 studies) |
| Study type: | |
| Cross-sectional (*C*) | 14 studies |
| Longitudinal (*L*) | 51 studies |

**Table 3:** Number of estimates of linear selection in the database as a function of taxon, trait type, and fitness component

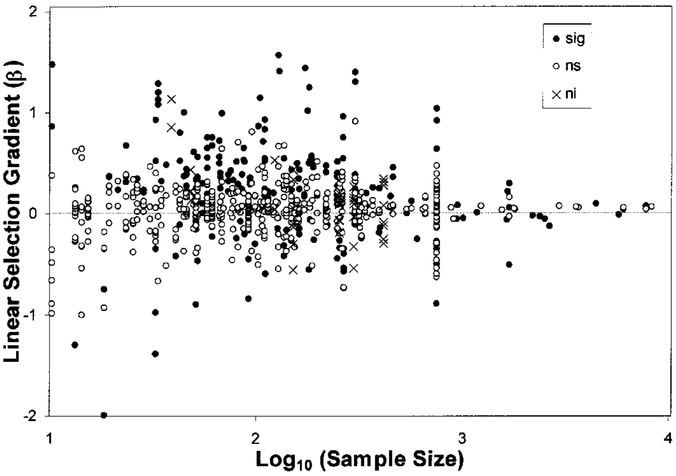
| Taxon | | Trait | | Fitness component | |
|---|---|---|---|---|---|
| Estimates of linear selection gradients[a] | | | | | |
| Invertebrates | 333 | Morphology | 815 | Mating success | 407 |
| Plants | 363 | Life history/phenology | 128 | Survival | 288 |
| Vertebrates | 297 | Principal component | 33 | Fecundity | 271 |
| … | … | Behavior | 14 | Total fitness | 19 |
| … | … | Interaction | NA | Net reproductive rate | 3 |
| … | … | Other | 3 | Other | 5 |
| Estimates of linear selection differentials[b] | | | | | |
| Invertebrates | 233 | Morphology | 594 | Mating success | 267 |
| Plants | 183 | Life history/phenology | 125 | Survival | 293 |
| Vertebrates | 337 | Principal component | 21 | Fecundity | 142 |
| … | … | Behavior | 10 | Total fitness | 34 |
| … | … | Interaction | NA | Net reproductive rate | 12 |
| … | … | Other | 3 | Other | 5 |

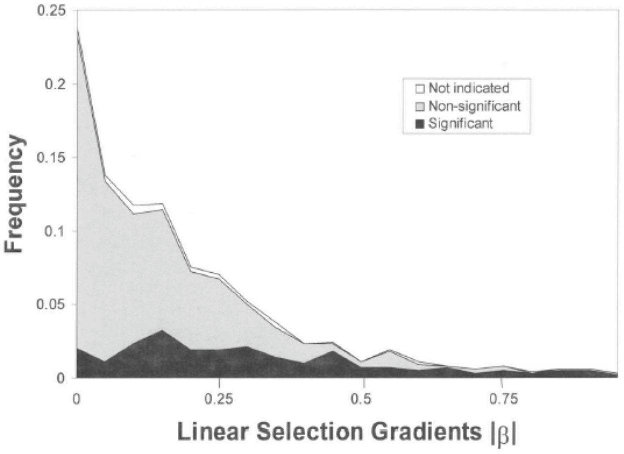
Note: NA = not applicable.

[a] $N = 993$ total estimates.

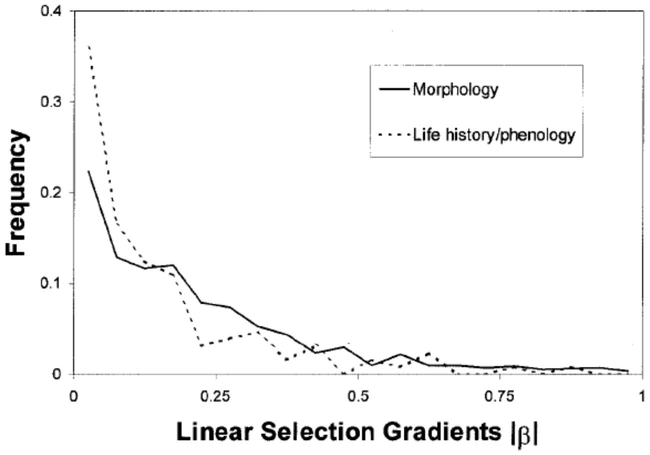
[b] $N = 753$ total estimates.

Kingsolver, J. G.,H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, P. Beerli. 2001. *The strength of phenotypic selection in natural populations.* The American Naturalist 157:245-261.


Linear Selection Gradient (β)
2
1
0
-1
-2
1
2
3
4
Log$_{10}$ (Sample Size)
● sig
○ ns
× ni


Frequency
0.25
0.2
0.15
0.1
0.05
0
0
0.25
0.5
0.75
Linear Selection Gradients |β|
□ Not indicated
▨ Non-significant
■ Significant

## Morphology > Life history



## Sexual selection is surprisingly strong in nature

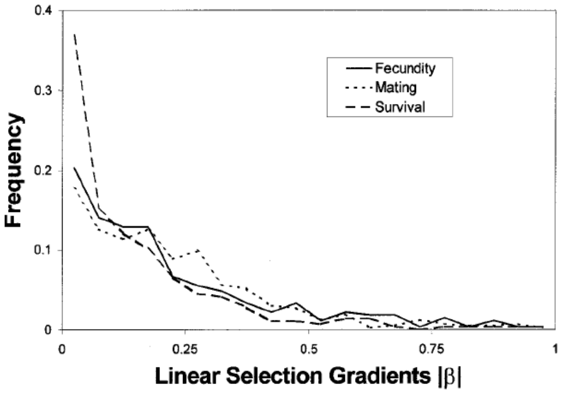


Figure 5: Frequency distribution of the absolute values of the linear selection gradient estimates ($|\beta|$) binned at 0.05 value intervals, for selection via three different components of fitness: fecundity (*solid line*, $N = 271$), mating success (*short dashes*, $N = 407$), and survival (*long dashes*, $N = 288$).

## Disruptive selection appears as common as stabilising selection

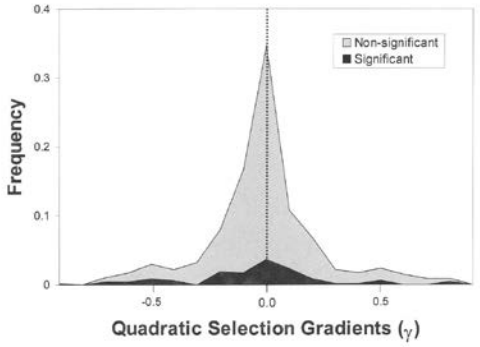


**Figure 8:** Frequency distribution (in %) of the quadratic selection gradient estimates ($\gamma$) binned at 0.10 value intervals ($N$ = 465 estimates). The distributions are stacked according to the statistical significance (at the $P$ = .05 level) of each individual estimates: black indicates significantly different from 0; grey indicates not significant.

**Kingsolver, J. G. and D. W. Pfennig. 2007. Patterns and power of phenotypic selection in nature. Bioscience 57:561-571.**

*Articles*

# Patterns and Power of Phenotypic Selection in Nature

JOEL G. KINGSOLVER AND DAVID W. PFENNIG

*Phenotypic selection occurs when individuals with certain characteristics produce more surviving offspring than individuals with other characteristics. Although selection is regarded as the chief engine of evolutionary change, scientists have only recently begun to measure its action in the wild. These studies raise numerous questions: How strong is selection, and do different types of traits experience different patterns of selection? Is selection on traits that affect mating success as strong as selection on traits that affect survival? Does selection tend to favor larger body size, and, if so, what are its consequences? We explore these questions and discuss the pitfalls and future prospects of measuring selection in natural populations.*

*Keywords: adaptive landscape, Cope's rule, natural selection, rapid evolution, sexual selection*

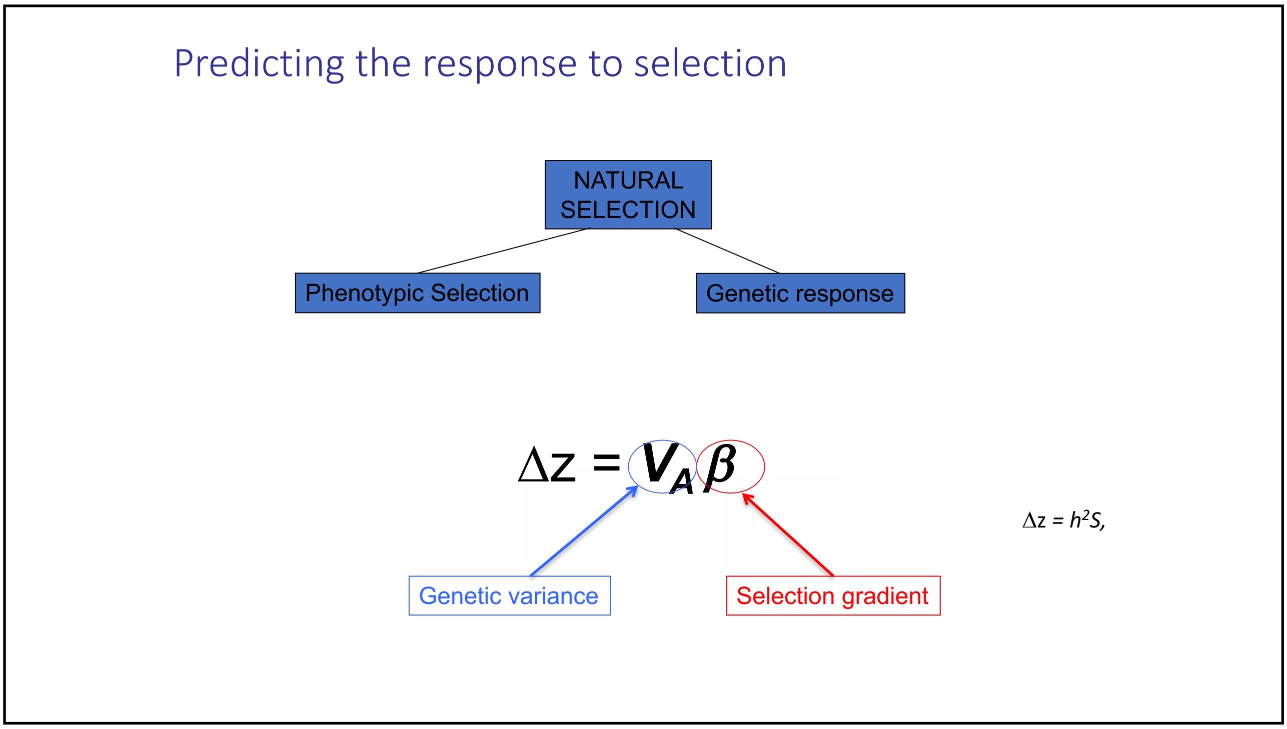
**Phenotypic selection occurs when individuals with** different characteristics (i.e., different phenotypes) differ in their survival, fecundity, or mating success. The idea of phenotypic selection traces back to Darwin and Wallace (1858), and selection is widely accepted as the primary cause of adaptive evolution within natural populations. Yet Darwin never attempted to measure selection in nature, and in the century following the publication of *On the Origin of Species* (Darwin 1859), selection was generally regarded as too weak to be observed directly in natural populations. Several now-classic demonstrations of selection in the wild were published between 1950 and 1975, most notably the case of industrial melanism in peppered moths (Kettlewell 1973). As late as the 1970s, however, industrial melanism remained the primary example of selection in action.

The view that selection is too weak to be measured in the wild has changed dramatically. In the past 25 years, selection has been detected and quantified in hundreds of populations in nature (Endler 1986, Kingsolver et al. 2001, Hereford et al. 2004). Indeed, there are literally thousands of estimates of phenotypic selection in natural populations (Endler 1986, Kingsolver et al. 2001). These data demonstrate that selection occurs routinely in nature and that researchers can measure its action. We are therefore in a position to ask more general questions about phenotypic selection: How strong is selection? Does selection always tend to increase (or decrease) trait values, or are other patterns possible? Do different types of traits experience different patterns or levels of selection? Is selection on traits that affect survival stronger than on those that affect only mating success? In this article, we explore these and other questions about the patterns and power of phenotypic selection in nature.

### What is selection, and how does it work?

Selection is the nonrandom differential survival or reproduction of phenotypically different individuals. Thus, selection requires variation, whereby individuals differ in some of their characteristics, and differential reproduction, whereby some individuals have more surviving offspring than others because of their distinctive characteristics. Those individuals that do have more surviving offspring are said to have higher fitness (note that fitness is a relative, not an absolute, measure). When the characteristics under selection show heredity (i.e., when parents pass on some of their characteristics to their offspring), selection will lead to evolutionary change in these characteristics. Indeed, when populations exhibit variation, heredity, and differential reproduction for a trait, evolution by natural selection will occur. Because these three conditions are met for many traits in many populations, evolution by natural selection is widespread.

*Joel G. Kingsolver (e-mail: jgking@bio.unc.edu) and David W. Pfennig (e-mail: dpfennig@email.unc.edu) work in the Department of Biology at the University of North Carolina, Chapel Hill, NC 27599.*



## Predicting the response to selection

NATURAL SELECTION

Phenotypic Selection

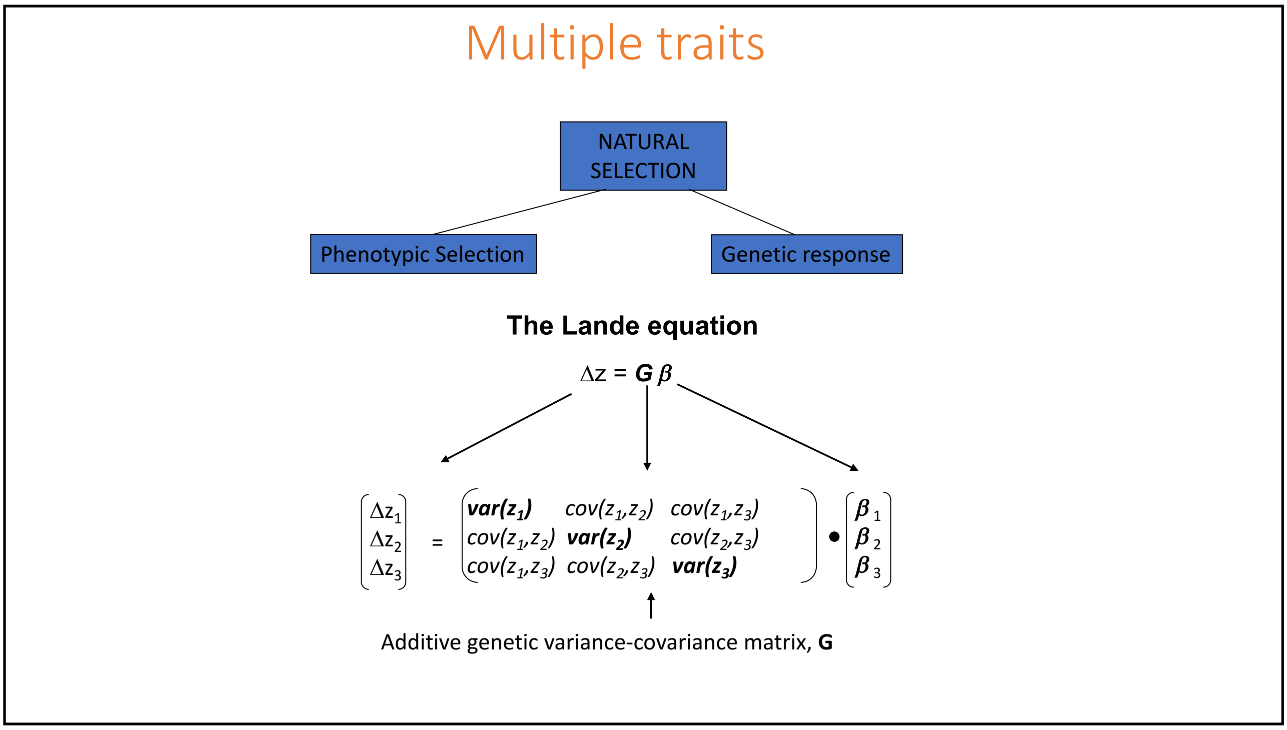
Genetic response

$$\Delta z = V_A \beta$$

Genetic variance

Selection gradient

$\Delta z = h^2 S,$



## Multiple traits

NATURAL SELECTION

Phenotypic Selection

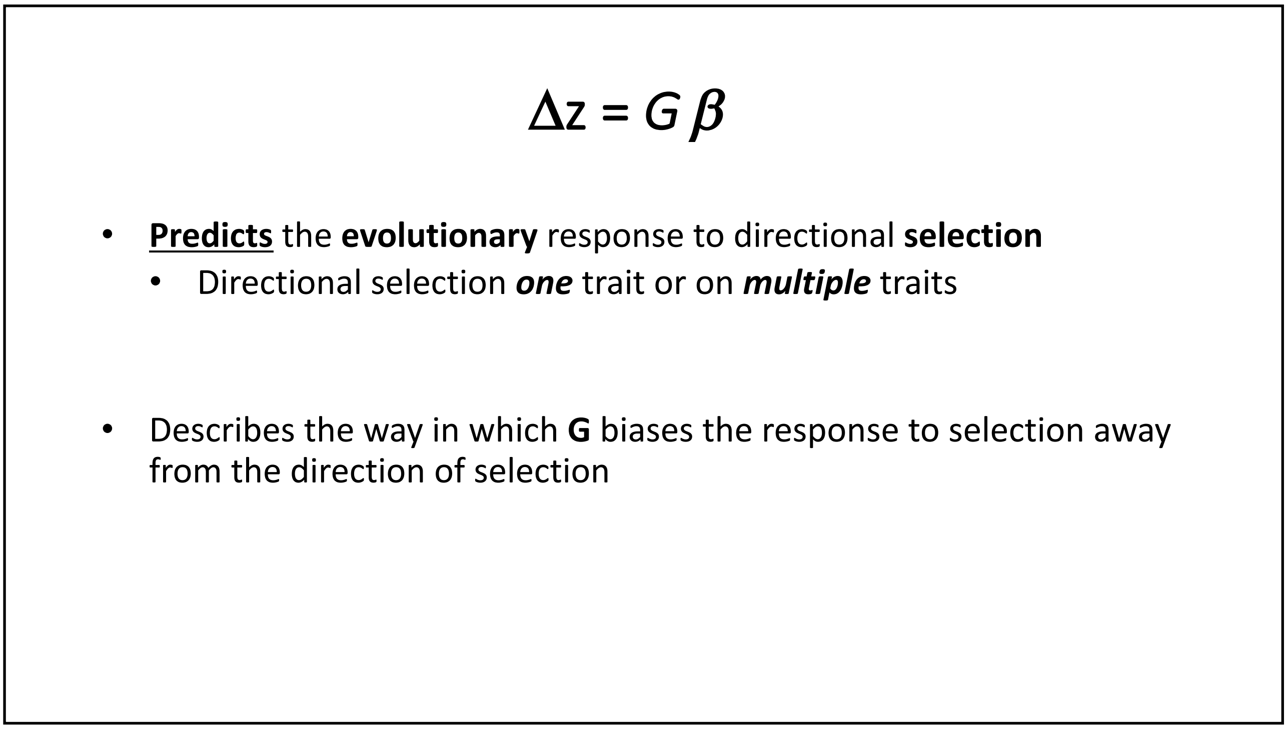
Genetic response

**The Lande equation**

$$\Delta z = \mathbf{G}\,\beta$$

$$\begin{pmatrix} \Delta z_1 \\ \Delta z_2 \\ \Delta z_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{var(z_1)} & cov(z_1,z_2) & cov(z_1,z_3) \\ cov(z_1,z_2) & \boldsymbol{var(z_2)} & cov(z_2,z_3) \\ cov(z_1,z_3) & cov(z_2,z_3) & \boldsymbol{var(z_3)} \end{pmatrix} \bullet \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

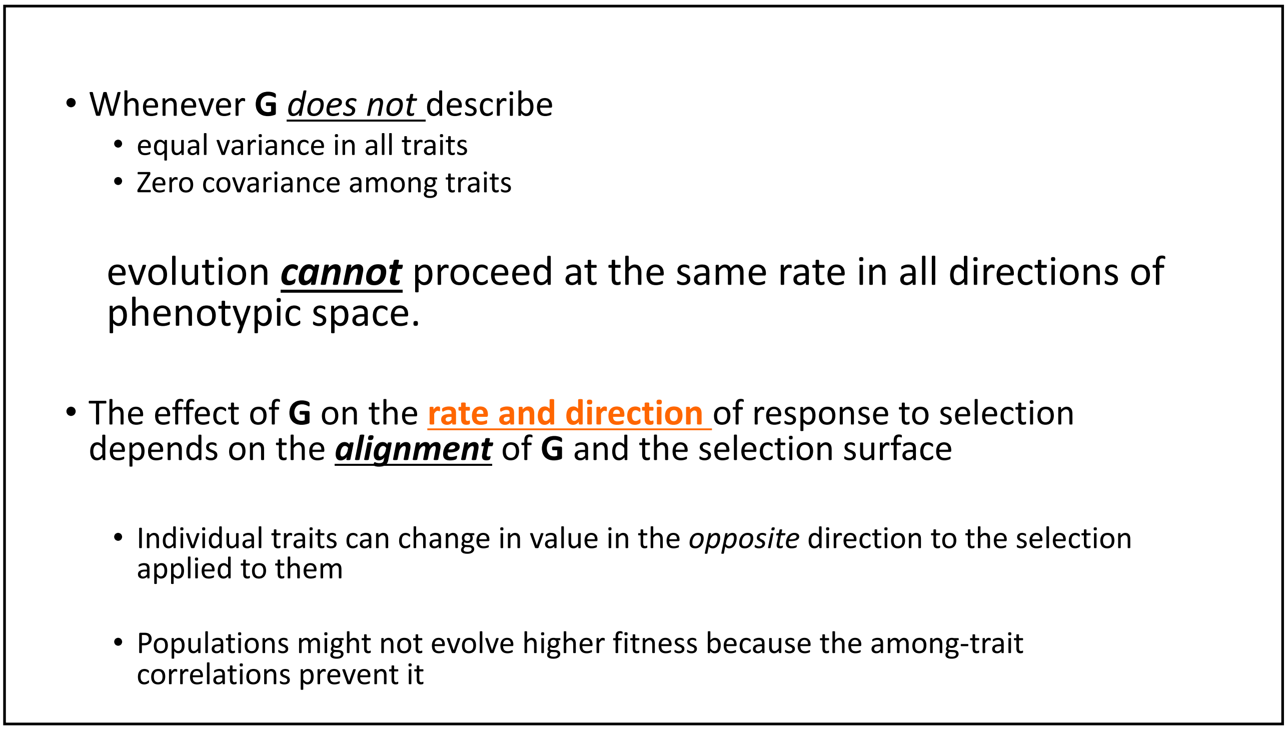
Additive genetic variance-covariance matrix, **G**



$$\Delta z = G\,\beta$$

- **<u>Predicts</u>** the **evolutionary** response to directional **selection**
  - Directional selection ***one*** trait or on ***multiple*** traits

- Describes the way in which **G** biases the response to selection away from the direction of selection



- Whenever **G** *<u>does not</u>* describe
  - equal variance in all traits
  - Zero covariance among traits

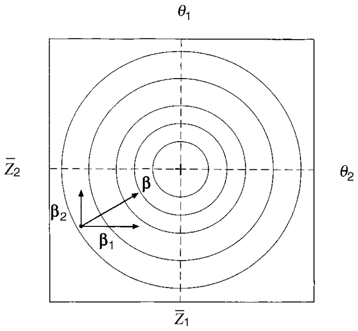
evolution ***<u>cannot</u>*** proceed at the same rate in all directions of phenotypic space.

- The effect of **G** on the **<u>rate and direction</u>** of response to selection depends on the ***<u>alignment</u>*** of **G** and the selection surface
  - Individual traits can change in value in the *opposite* direction to the selection applied to them
  - Populations might not evolve higher fitness because the among-trait correlations prevent it

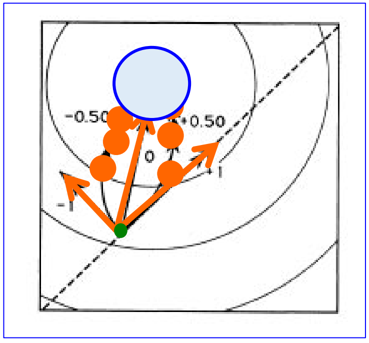
# The adaptive landscape

- A heuristic for thinking about how populations evolve
- Developed by Simpson in 1944, and used by Lande (1979)

Arnold et al. 2001



Walsh & Lynch Fig. 32.2. http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html#2B



- Circles = fitness isoclines (everything along the line has equal fitness
- Two traits, with equal variance

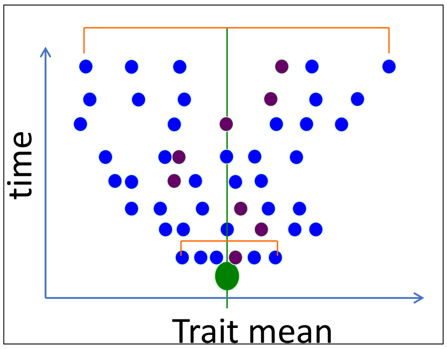
- **No covariance between traits**, evolution proceeds directly uphill for maximum increase in fitness (i.e., along β)
- **Complete covariance (+1 or -1),** only one trait increases in fitness, and the population never climbs the peak
- **Moderate covariance** (+0.5 or -0.5) the population takes a curved path, and approaches the peak much more slowly (each arrow head = a generation of change)

# Genetic Constraints

- The genetic variance shared among traits (their covariance) can markedly affect the RATE and DIRECTION of total phenotypic evolution, and the response of individual traits
- If selection favours a trait combination with little genetic variance, the rate of evolution will be slow
  - Populations might become extinct before gaining sufficient absolute fitness

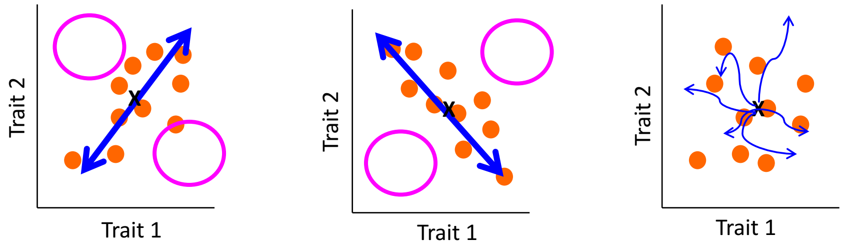
# Random Genetic Drift

- Occurs in finite populations – proportional to $N_e$
- Causes population <u>mean phenotypes</u> to diverge
  - Variation among populations at time $t$ proportional to **G** at time 0



## Random Genetic Drift

- Greatest divergence in *direction* of most genetic variance



## What maintains (quantitative) genetic variation within populations?

How does genetic variance evolve within populations?

Maintenance of genetic variance remains a major unresolved question in Quantitative Genetics

## Conflicting observations

1. **selection on traits is common, and fairly strong:**

"phenotypic selection in many natural populations is strong enough to cause substantial evolutionary changes in tens to hundreds of generations, which is a very short timescale in evolutionary terms" PG 565 KINGSOLVER AND PFENNIG 2007

2. **Traits are heritable:**

"If one's sole interest in performing a quantitative-genetic analysis is to demonstrate that the character of interest is heritable, there is probably little point in expending the effort. The outcome is virtually certain. Almost every character in almost every species that has been studied intensely exhibits nonzero heritability." PG 174 LYNCH AND WALSH 1998

## Practical Importance

- AGRICULTURE: How genetic variance is maintained will affect how we can apply artificial selection, and what the responses will be

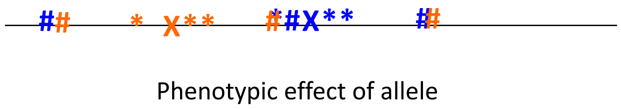
- BIOMEDICAL: The nature of genetic variation will affect how we can go about identifying causal genetic variants of human diseases

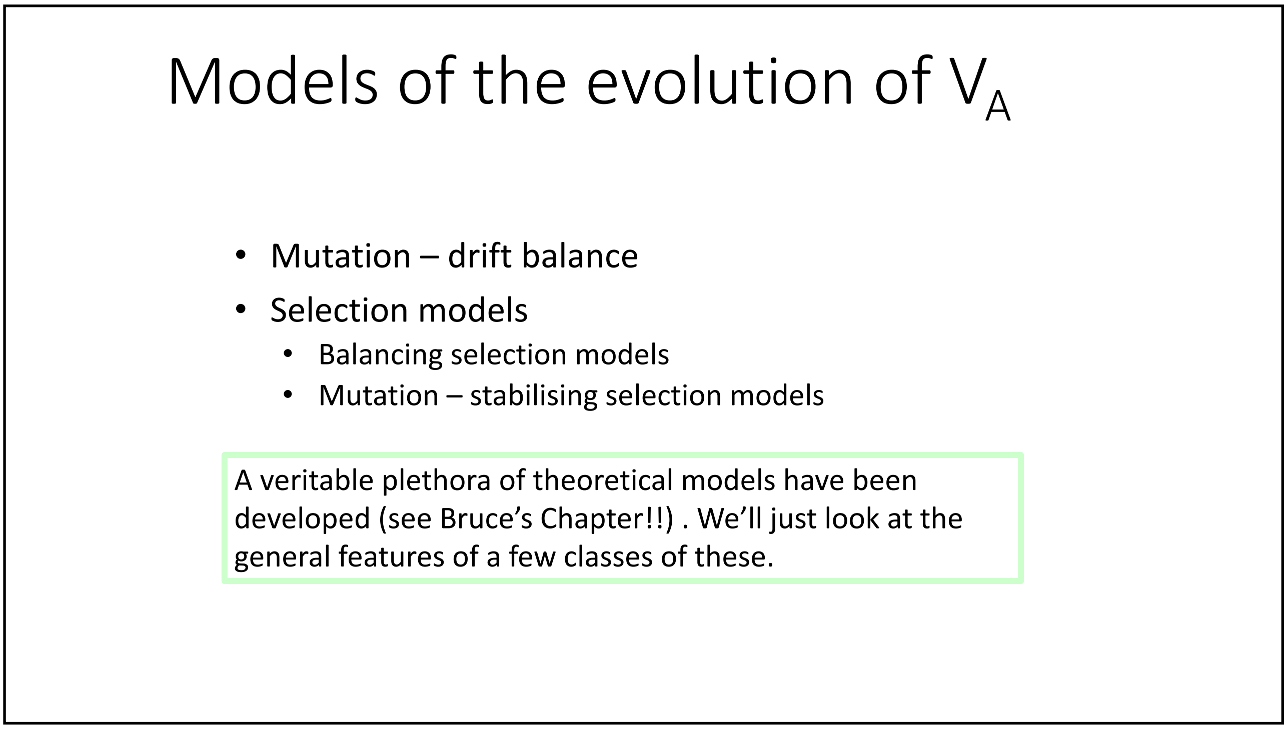
## Predicating the effects of evolutionary process on **G**

We cannot exactly predict the evolution of **G** because it depends on *unknown* details of the genetic architecture underpinning **G**

- Frequencies of alleles
- Number of loci
- Effects of alleles on phenotypic trait of interest AND on fitness

- Do many loci with many alleles of small effect contribute to a trait?
  - Each allele would be under weak selection, and change little in frequency, resulting in the maintenance of high levels of variance
- Do mutations change the effect of an allele relative to the effect before mutation, or are all allelic effects possible?

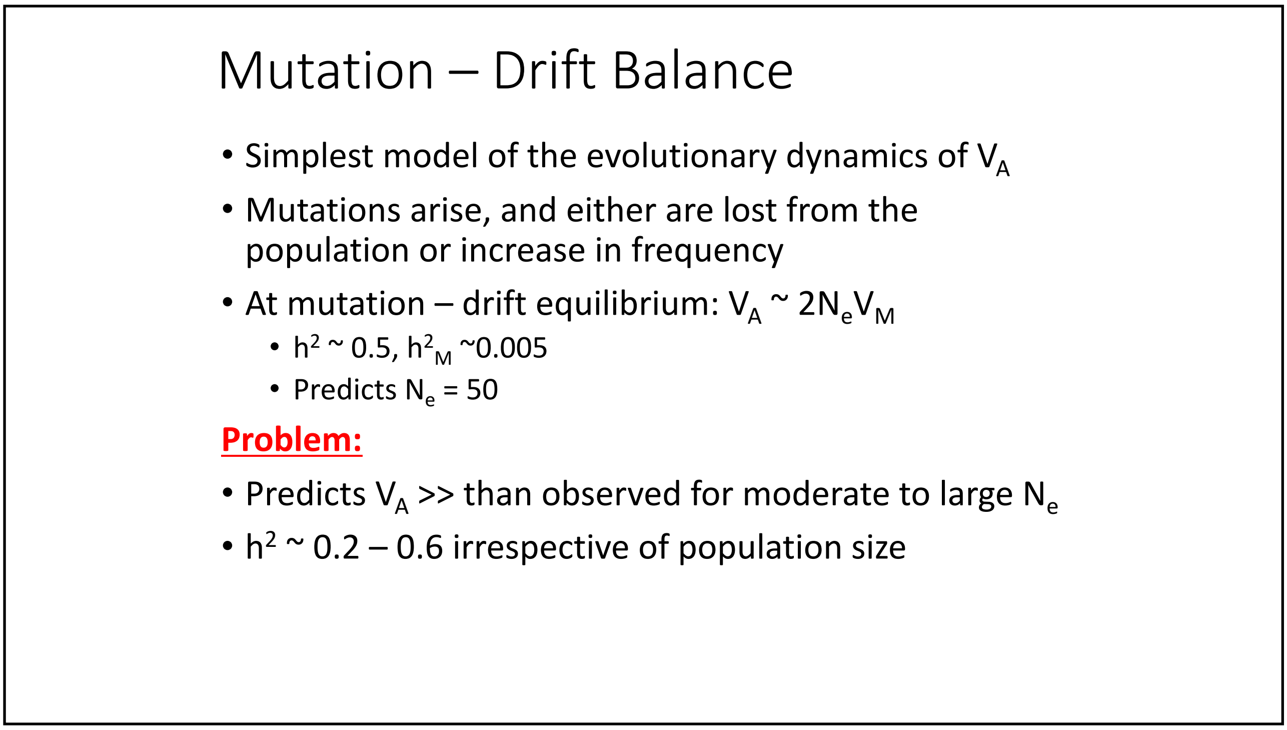




# Models of the evolution of $V_A$

- Mutation – drift balance
- Selection models
  - Balancing selection models
  - Mutation – stabilising selection models

A veritable plethora of theoretical models have been developed (see Bruce's Chapter!!) . We'll just look at the general features of a few classes of these.



# Mutation – Drift Balance

- Simplest model of the evolutionary dynamics of $V_A$
- Mutations arise, and either are lost from the population or increase in frequency
- At mutation – drift equilibrium: $V_A \sim 2N_eV_M$
  - $h^2 \sim 0.5$, $h^2_M$ ~0.005
  - Predicts $N_e = 50$

**Problem:**

- Predicts $V_A$ >> than observed for moderate to large $N_e$
- $h^2 \sim 0.2 – 0.6$ irrespective of population size

## Mutation – Drift Balance

- Drift cannot be the whole story, but:
- Alleles are considered "effectively neutral" (fate determined by drift, not selection) when:

$s < 1/2Ne$

For Ne = 50, mutational fate determined by drift when $s < 0.01$

- Estimates of s from new mutations:
  - $s <$ ~0.01 for those affecting morphological traits
  - $s$ ~ 0.02 for those affecting fitness components
- Chance sampling of alleles under weak selection (mildly deleterious effects on fitness) likely **contributes** to the maintenance of standing genetic variance in finite populations

## Models with Selection

**Classical View (MSB) – T. H. Morgan & Hermann Muller**

- Wildtype allele has highest fitness in any particular environment
- Variation due to recurrent deleterious mutations

Fig. 1.30. Barton et al. 2007 "Evolution"

**Balancing View – Theodosius Dobzhansky**

- Balancing selection maintains variation
- Allows rapid adaptation to ever changing environment

Fig. 1.36. Barton et al. 2007 "Evolution"

## Models with Selection

1. Selection ***maintains*** variation – **balancing selection** models

- Rare alleles favoured
- Need to understand how alleles become fitter as they become rarer.

**a. Heterozygote advantage**

- Rare alleles will mostly be present in heterozygotes
  - $q^2$ vs $2pq$
- Several specific examples, e.g. Sickle cell anemia in the presence of malaria
- Can't be the *only* mechanism - haploid taxa have abundant genetic variance
  - Can't have heterozygotes with only one copy of the gene

b. Frequency-dependent selection

- Rare alleles are directly favoured
  - Inbreeding avoidance mating incompatibilities
  - Batesian mimicry predator avoidance
  - Intra-specific competition avoidance

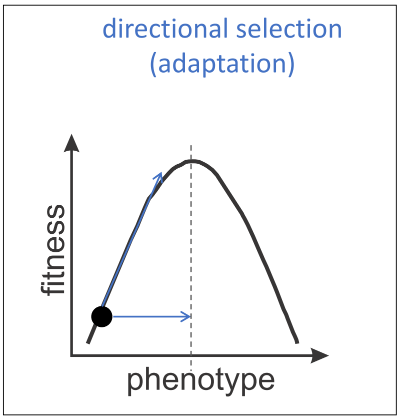
c. Fluctuating selection

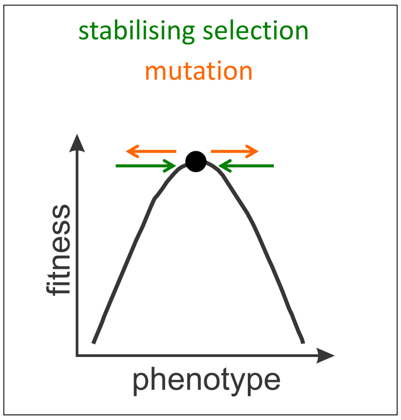
  - spatial or temporal variation in the alleles with the highest fitness can maintain polymorphism of the population
    - Requires some fairly restrictive simplifying assumptions,
  - Includes fitness differences of alleles in females versus males

## 2. Mutation-Selection Balance

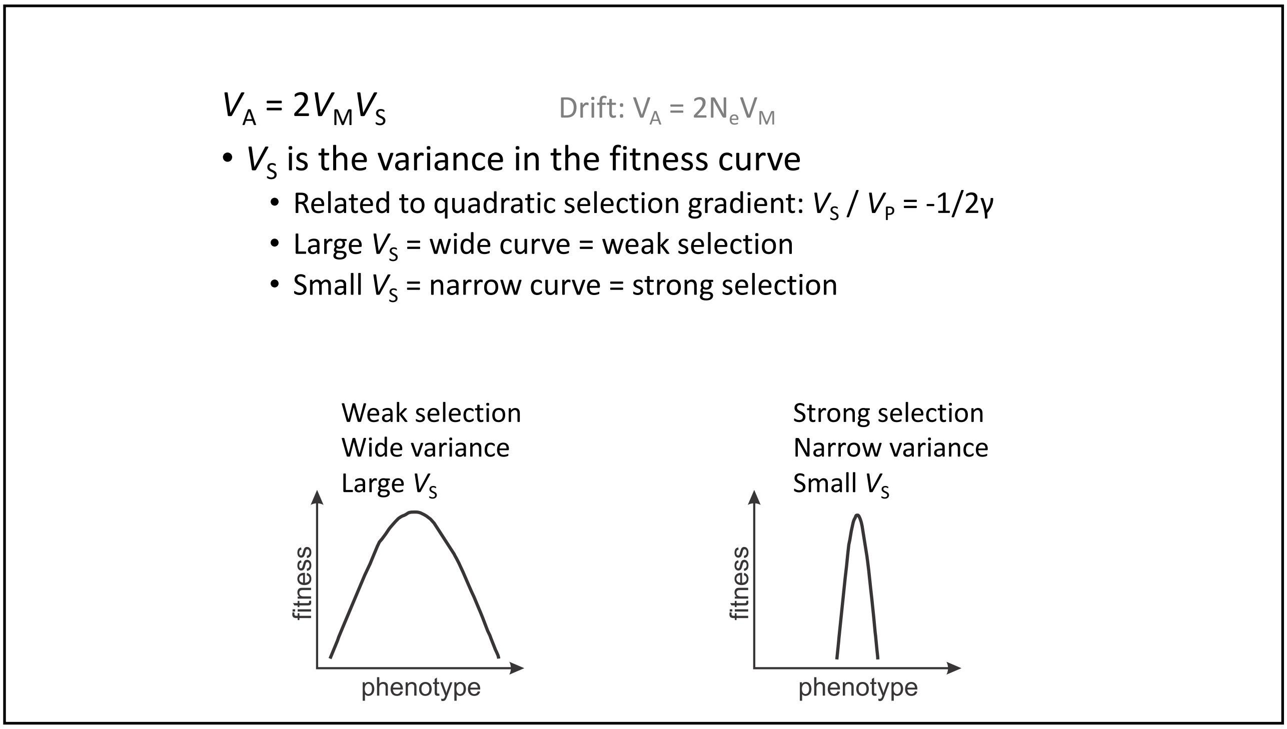
Considered the most generally applicable quantitative genetic model of the maintenance of additive genetic variance

- Rate at which genetic variation is removed by selection is exactly matched by the rate at which it is introduced by mutation
- *Assumes* population is under **stabilising selection**
- Few studies report significant stabilising selection in contemporary populations (*Kingsolver* et al.)
- Phenotypes (morphology) stay the same over long periods of time


directional selection
(adaptation)


stabilising selection
mutation

- Stabilising selection = **variance reducing selection**
- Assume that the average trait value in the population has the greatest fitness, and therefore all mutations reduce fitness, and are eliminated

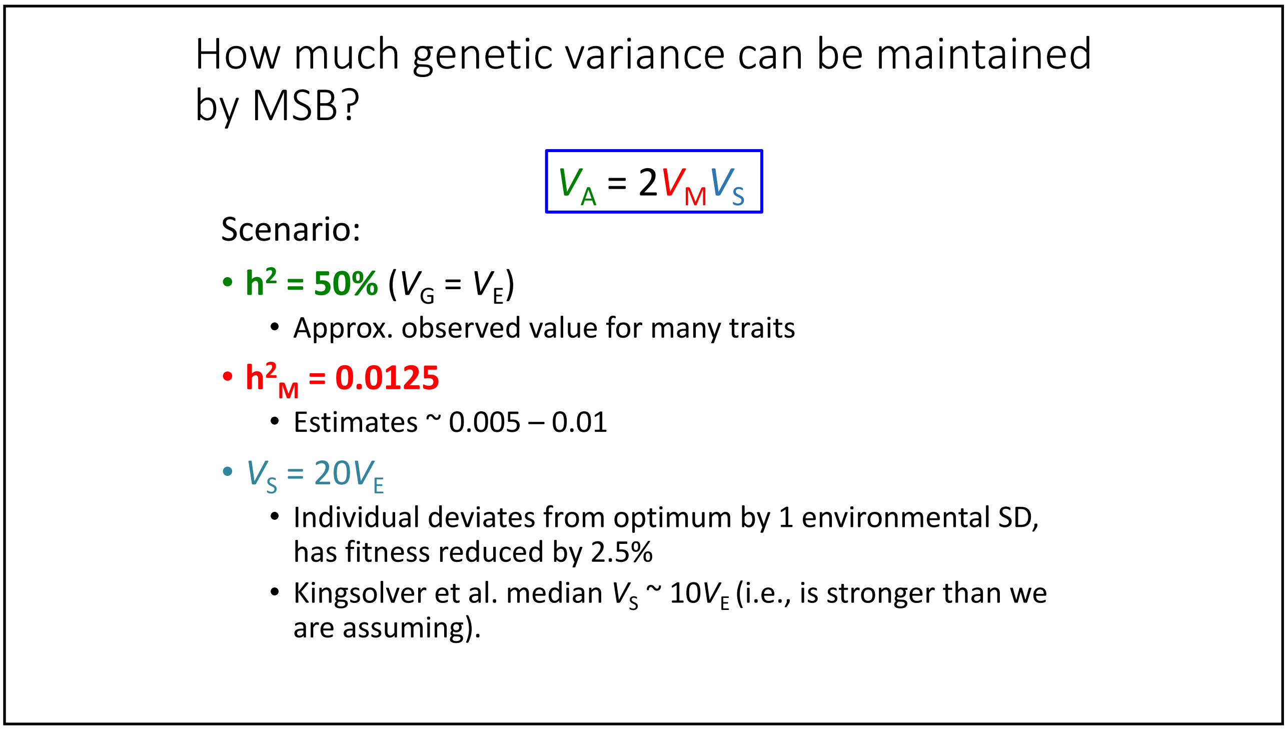


$V_A = 2V_MV_S$ Drift: $V_A = 2N_eV_M$

- $V_S$ is the variance in the fitness curve
  - Related to quadratic selection gradient: $V_S / V_P = -1/2\gamma$
  - Large $V_S$ = wide curve = weak selection
  - Small $V_S$ = narrow curve = strong selection

Weak selection
Wide variance
Large $V_S$

Strong selection
Narrow variance
Small $V_S$



## How much genetic variance can be maintained by MSB?

$$\boxed{V_A = 2V_MV_S}$$

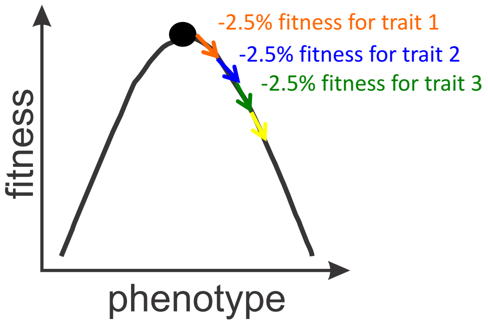
Scenario:

- **$h^2$ = 50%** ($V_G = V_E$)
  - Approx. observed value for many traits
- **$h^2_M$ = 0.0125**
  - Estimates ~ 0.005 – 0.01
- $V_S = 20V_E$
  - Individual deviates from optimum by 1 environmental SD, has fitness reduced by 2.5%
  - Kingsolver et al. median $V_S$ ~ $10V_E$ (i.e., is stronger than we are assuming).

# MSB

Can work quite well (predict genetic variance of plausible magnitude) when we think about a single trait BUT

- Many traits individually under (weak) stabilising selection, fitness implausibly low


-2.5% fitness for trait 1
-2.5% fitness for trait 2
-2.5% fitness for trait 3
fitness
phenotype

# MSB

- Mutation rate to allow ~50% $h^2$ is fairly high
  - Implies many loci affect each trait (per locus mutation rates are much lower)
  - If each gene/mutation affected each trait independently, there aren't enough genes
- **Pleiotropy *must* be pervasive**
  - The same allele (mutation) affects multiple traits, and fitness

## If pleiotropy is pervasive…

Imagine:

- Each allele affects your trait of interest, and also decreases fitness [assuming stabilising selection]

$V_G = V_M/s$ average selection against alleles

Estimates of *s* from new mutations:

*s* < ~0.01 to ~ 0.02

Estimates of $V_M$:

$V_M$ ~0.005 – 0.01

MSB works if average selection is a little weaker than we think it might be, or if mutational variance is a little greater than we think it might be.

For $h^2$ = 50%, av *s* must be ~ 0.001 to 0.01

## Maintenance of Quantitative Genetic Variance

- No theoretical model predicts observed levels of $V_A$ for realistic values of other parameters ($V_M$, $V_S$, $N_e$), with realistic simplifying assumptions
- Some models of MSB seem plausible, but we really don't know enough about the mutation rate or fitness effects of new mutations
- Is evidence that BS maintains variance in at least some traits

## Balancing selection or mutation?

Unknown whether the high levels of $V_A$ in populations is due to:

- Balancing selection maintaining variation
- Mutation – selection balance

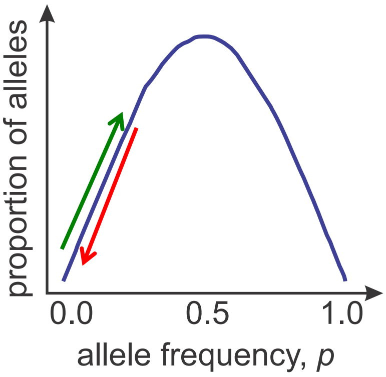
Key predictions that allow us to distinguish between them

- Allele frequencies
- Allelic effects on fitness

## Key predictions distinguishing between models with selection

**<u>Balancing selection</u>**: alleles at intermediate frequencies

- Rare alleles have positive effects on fitness
- Selection increases their frequency
- When they are no longer rare, they cease to be under positive selection

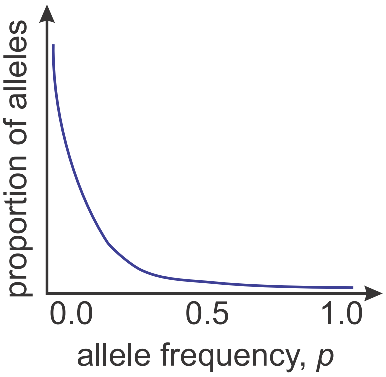


# Key predictions distinguishing between models with selection

- **Mutation-selection balance**: most alleles will be at low frequency
  - Most alleles = new mutations
    - By definition, new mutations are rare
  - Most alleles = low fitness
    - Selection is keeping them rare, eliminating them

# Key predictions distinguishing between models with selection

Under MSB model, standing genetic variance must be made up of many low frequency alleles, and few high frequency alleles

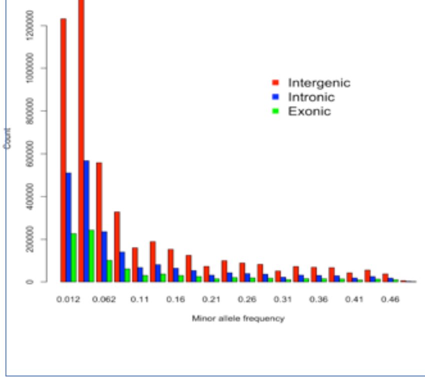


# What do we know about allele frequencies?

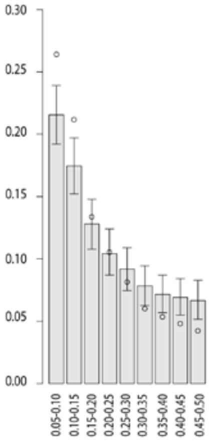
***molecular genetic*** data generally support many rare allele distribution, consistent with MSB

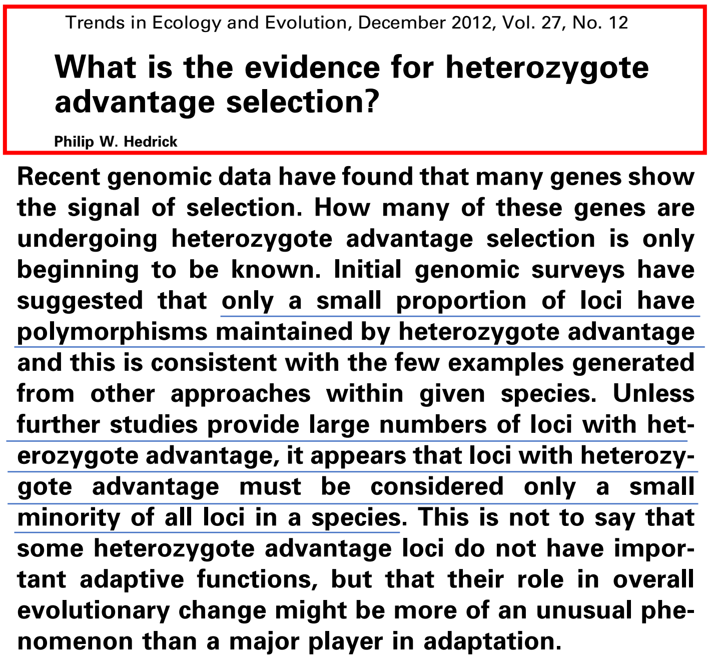
"minor" allele = 2nd most frequent allele; has to be <0.5

Allele frequency distribution in Steve's population of *D. serrata*



Allele frequency distribution of loci affecting gene expression variation in mustard relative – Josephs et al. 2016





Trends in Ecology and Evolution, December 2012, Vol. 27, No. 12

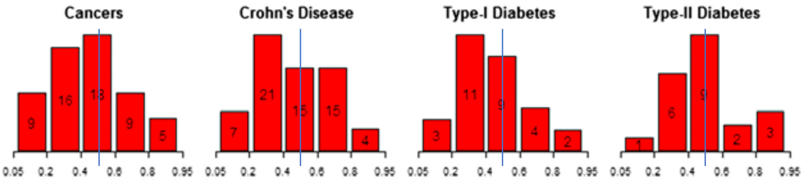
**What is the evidence for heterozygote advantage selection?**

Philip W. Hedrick

**Recent genomic data have found that many genes show the signal of selection. How many of these genes are undergoing heterozygote advantage selection is only beginning to be known. Initial genomic surveys have suggested that only a small proportion of loci have polymorphisms maintained by heterozygote advantage and this is consistent with the few examples generated from other approaches within given species. Unless further studies provide large numbers of loci with heterozygote advantage, it appears that loci with heterozygote advantage must be considered only a small minority of all loci in a species. This is not to say that some heterozygote advantage loci do not have important adaptive functions, but that their role in overall evolutionary change might be more of an unusual phenomenon than a major player in adaptation.**

# What's the fitness effects of these alleles?

From Park et al. 2011 PNAS. 108:18026–18031

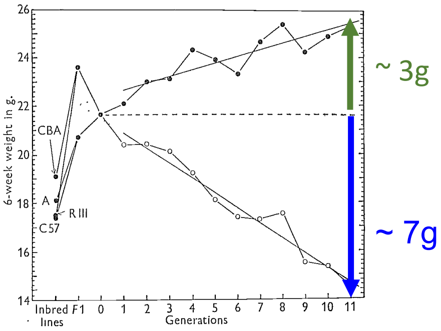


Frequency of “risk” alleles – alleles that increase the likelihood that you will get the disease

- Distributions are left skewed – more alleles that increase risk occur at low frequencies

## Genetic variation is less in the direction of high fitness than in the direction of low fitness

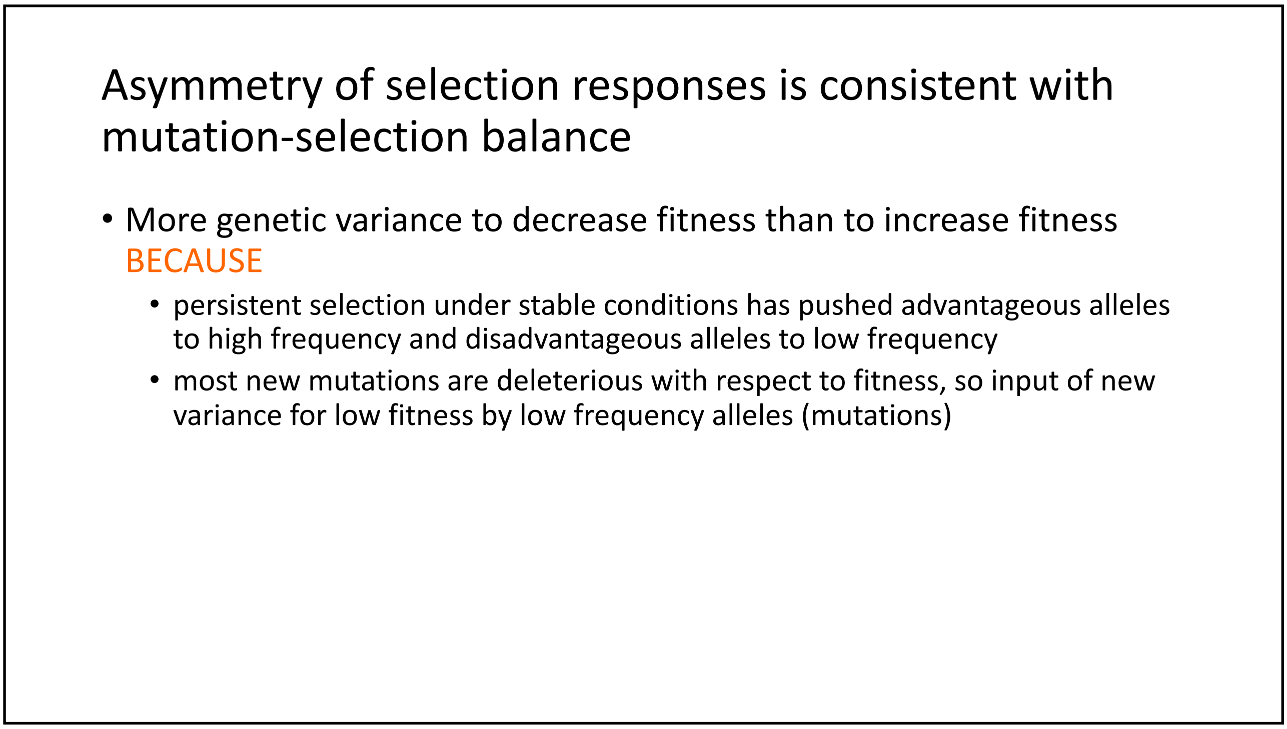
Artificial selection for decreased fitness causes more evolution than selection for increased fitness



Falconer, D. S. 1953. Selection for large and small size in mice. Journal of Genetics 51:470-501.

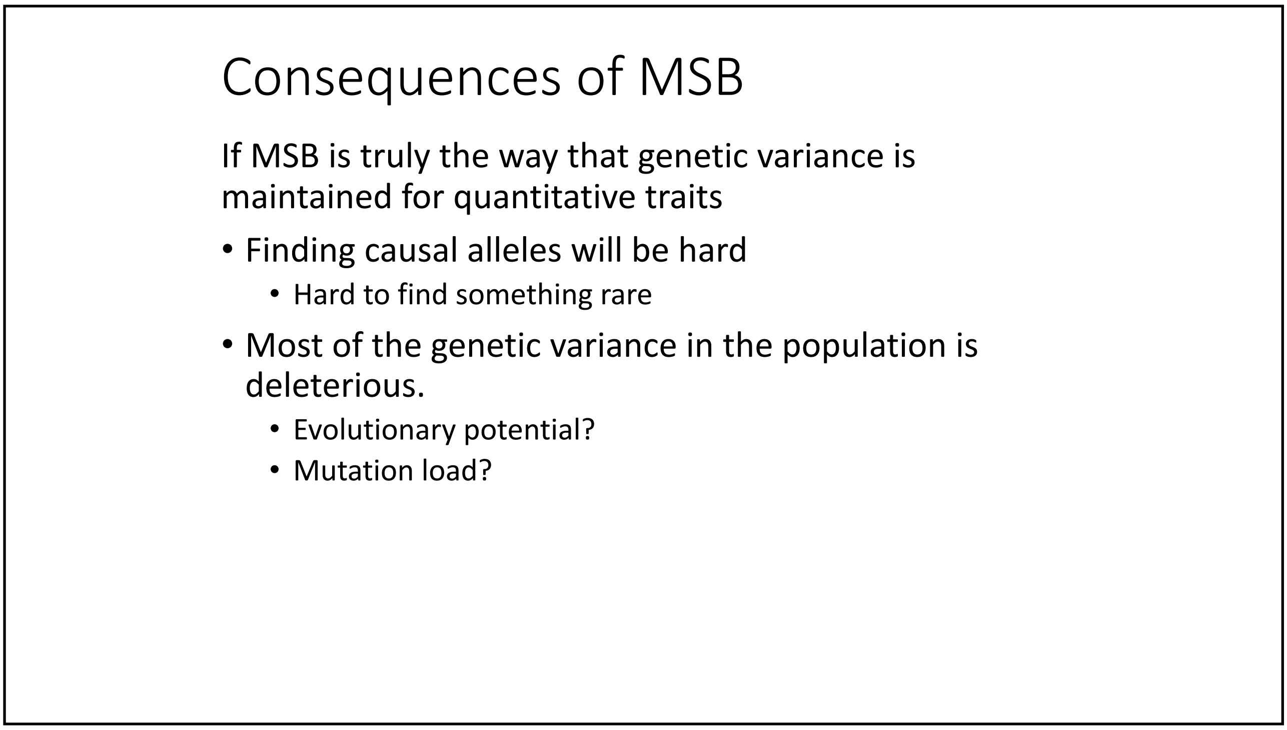
Frankham (1990)

- 30 bi-directional artificial selection experiments on fitness (components)
- **80%** report greater response for decreased fitness



## Asymmetry of selection responses is consistent with mutation-selection balance

- More genetic variance to decrease fitness than to increase fitness BECAUSE
  - persistent selection under stable conditions has pushed advantageous alleles to high frequency and disadvantageous alleles to low frequency
  - most new mutations are deleterious with respect to fitness, so input of new variance for low fitness by low frequency alleles (mutations)



## Consequences of MSB

If MSB is truly the way that genetic variance is maintained for quantitative traits

- Finding causal alleles will be hard
  - Hard to find something rare
- Most of the genetic variance in the population is deleterious.
  - Evolutionary potential?
  - Mutation load?

# Lecture 8
# QTL and Association mapping

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
6 – 7 Feb 2017

# Part I
# QTL mapping and the use of inbred line crosses

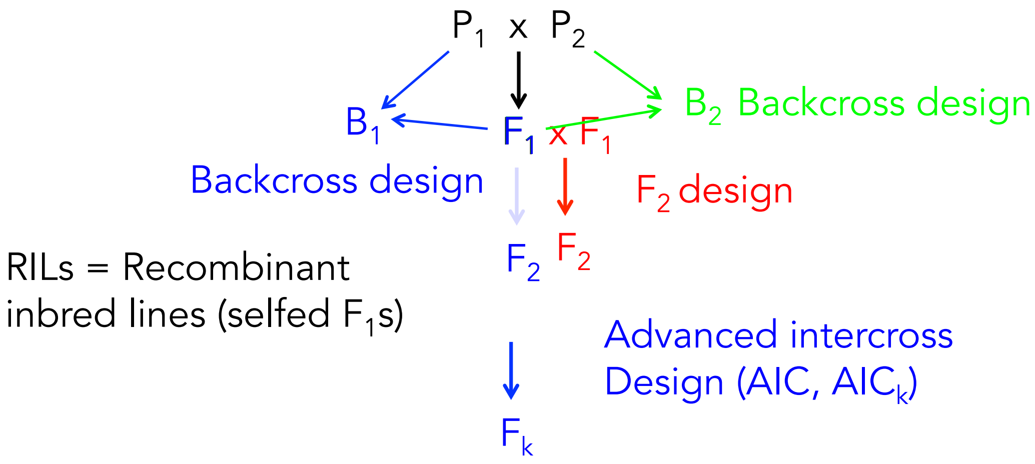
- QTL mapping tries to detect small (20-40 cM) chromosome segments influencing trait variation
  - Relatively crude level of resolution
- QTL mapping performed either using inbred line crosses or sets of known relatives
  - Uses the simple fact of an excess of parental gametes

Key idea: Looking for marker-trait associations in collections of relatives

If (say) the mean trait value for marker genotype MM is statistically different from that for genotype mm, then the M/m marker is linked to a QTL

One can use a random collection of such markers spanning a genome (a genomic scan) to search for QTLs

# Experimental Design: Crosses



$P_1$ x $P_2$

$B_1$

Backcross design

$F_1$ x $F_1$

$B_2$ Backcross design

$F_2$ design

$F_2$ $F_2$

RILs = Recombinant inbred lines (selfed $F_1$s)

Advanced intercross Design (AIC, $AIC_k$)

$F_k$

# Experimental Designs: Marker Analysis

Single marker analysis

Flanking marker analysis (interval mapping)

Composite interval mapping

Interval mapping plus additional markers

Multipoint mapping

Uses all markers on a chromosome simultaneously

# Conditional Probabilities of QTL Genotypes

The basic building block for all QTL methods is $Pr(Q_k \mid M_j)$ --- the probability of QTL genotype $Q_k$ given the marker genotype is $M_j$.

$$Pr(Q_k \mid M_j) = \frac{Pr(Q_k M_j)}{Pr(M_j)}$$

Consider a QTL linked to a marker (recombination Fraction = c). Cross MMQQ x mmqq. In the F1, all gametes are MQ and mq

In the F2, freq(MQ) = freq(mq) = (1-c)/2,
freq(mQ) = freq(Mq) = c/2

Hence, Pr(MMQQ) = Pr(MQ)Pr(MQ) = $(1\text{-}c)^2/4$

Pr(MMQq) = 2Pr(MQ)Pr(Mq) = 2c(1-c) /4

Pr(MMqq) = Pr(Mq)Pr(Mq) = $c^2$ /4

Why the 2? MQ from father, Mq from mother, OR
MQ from mother, Mq from father

Since Pr(MM) = 1/4, the conditional probabilities become

Pr(QQ | MM) = Pr(MMQQ)/Pr(MM) = $(1\text{-}c)^2$

Pr(Qq | MM) = Pr(MMQq)/Pr(MM) = 2c(1-c)

Pr(qq | MM) = Pr(MMqq)/Pr(MM) = $c^2$

How do we use these?

# Expected Marker Means

The expected trait mean for marker genotype $M_j$ is just

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(\, Q_k \,|\, M_j \,)$$

For example, if QQ = 2a, Qq = a(1+k), qq = 0, then in the F2 of an MMQQ/mmqq cross,

$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c)$$

- If the trait mean is significantly different for the genotypes at a marker locus, it is linked to a QTL

- A small MM-mm difference could be (i) a tightly-linked QTL of small effect or (ii) loose linkage to a large QTL

# Linear Models for QTL Detection

The use of differences in the mean trait value for different marker genotypes to detect a QTL and estimate its effects is a use of linear models.

One-way ANOVA.

Value of trait in kth individual of marker genotype type i

$$z_{ik} = \mu + b_i + e_{ik}$$

Effect of marker genotype i on trait value

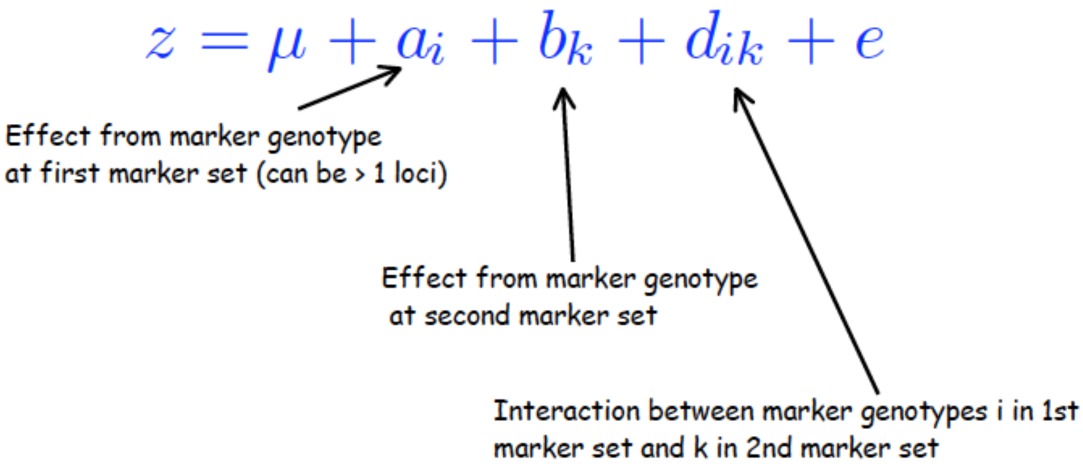
$$z_{ik} = \mu + b_i + e_{ik}$$

Detection: a QTL is linked to the marker if at least one of the $b_i$ is significantly different from zero

Estimation: (QTL effect and position): This requires relating the $b_i$ to the QTL effects and map position

# Detecting epistasis

One major advantage of linear models is their flexibility. To test for epistasis between two QTLs, use ANOVA with an interaction term



$$z = \mu + a_i + b_k + d_{ik} + e$$

Effect from marker genotype at first marker set (can be > 1 loci)

Effect from marker genotype at second marker set

Interaction between marker genotypes i in 1st marker set and k in 2nd marker set

# Detecting epistasis

$$z = \mu + a_i + b_k + d_{ik} + e$$

- At least one of the $a_i$ significantly different from 0 ---- QTL linked to first marker set

- At least one of the $b_k$ significantly different from 0 ---- QTL linked to second marker set

- At least one of the $d_{ik}$ significantly different from 0 ---- interactions between QTL in sets 1 and two

Problem: Huge number of potential interaction terms (order $m^2$, where m = number of markers)

# Maximum Likelihood Methods

ML methods use the entire distribution of the data, not just the marker genotype means.

More powerful that linear models, but not as flexible in extending solutions (new analysis required for each model)

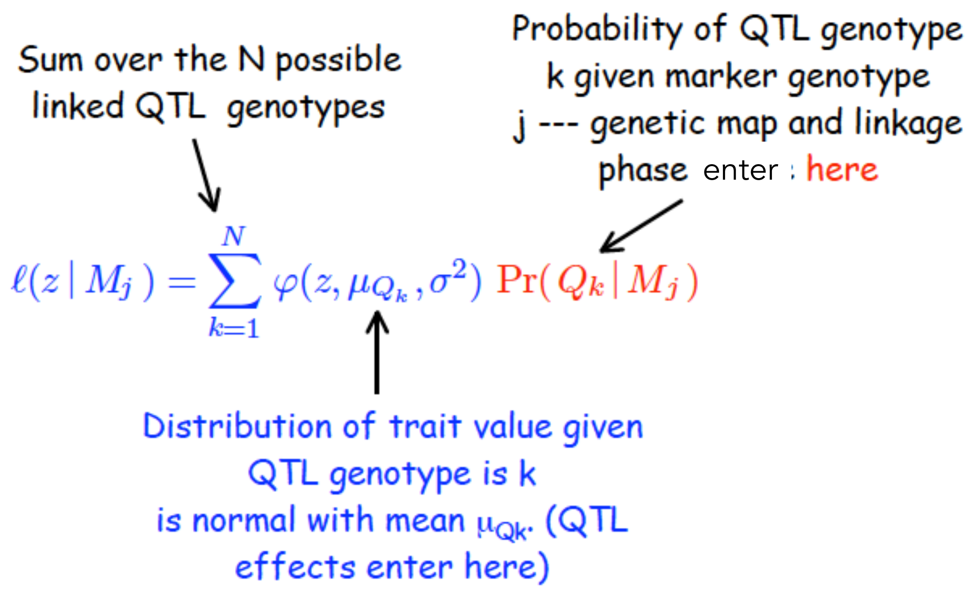
Basic likelihood function:

Trait value given marker genotype is type j

$$\ell(z \,|\, M_j\,) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2)\ \Pr(\,Q_k \,|\, M_j\,)$$

This is a mixture model

# Maximum Likelihood Methods



Sum over the N possible linked QTL genotypes

Probability of QTL genotype k given marker genotype j --- genetic map and linkage phase enter : here

$$\ell(z \,|\, M_j\,) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2)\ \Pr(\,Q_k \,|\, M_j\,)$$

Distribution of trait value given QTL genotype is k is normal with mean $\mu_{Qk}$. (QTL effects enter here)

ML methods combine both detection and estimation of QTL effects/position.

Test for a linked QTL given from by the Likelihood Ratio (or LR ) test

Maximum of the likelihood under a no-linked QTL model

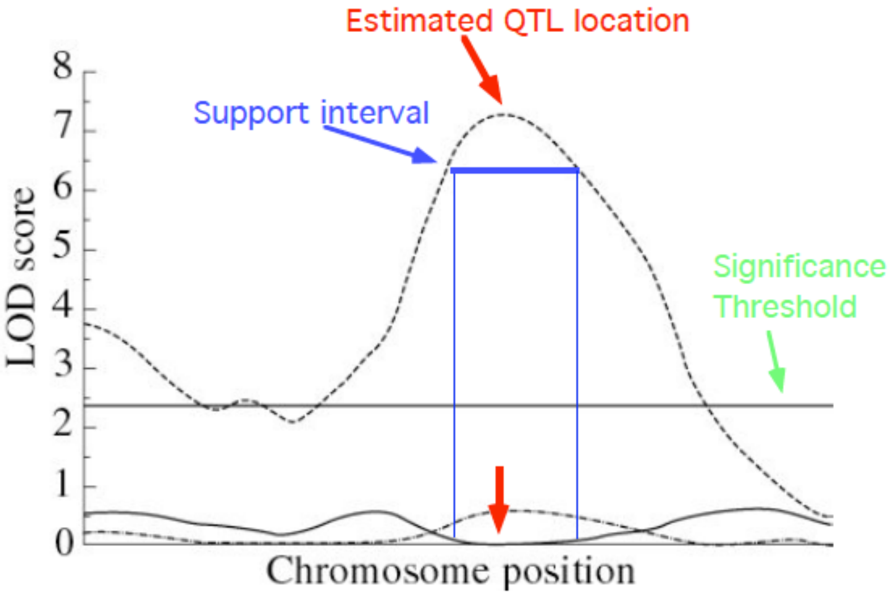
$$\mathrm{LR} = -2\ln\,\frac{\max\,\ell_r(\mathbf{z})}{\max\,\ell(\mathbf{z})}$$

Maximum of the full likelihood

The LR score is often plotted by trying different locations for the QTL (i.e., values of c) and computing a LOD score for each

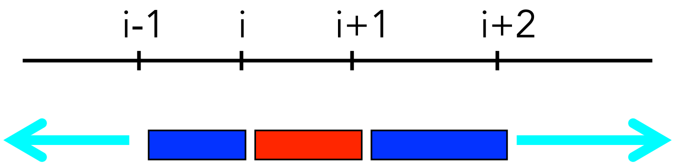
$$\mathrm{LOD}(c) = -\log_{10}\left[\frac{\max\,\ell_r(\mathbf{z})}{\max\,\ell(\mathbf{z},c)}\right] = \frac{\mathrm{LR}(c)}{2\,\ln 10} \simeq \frac{\mathrm{LR}(c)}{4.61}$$

## A typical QTL map from a likelihood analysis



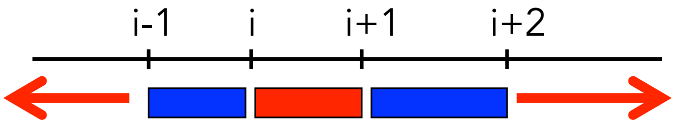
# Interval Mapping with Marker Cofactors

Consider interval mapping using the markers i and i+1. QTLs linked to these markers, but outside this interval, can contribute (falsely) to estimation of QTL position and effect



Interval being mapped

Now suppose we also add the two markers flanking the interval (i-1 and i+2)



Inclusion of markers i-1 and i+2 fully account for any linked QTLs to the left of i-1 and the right of i+2

Interval mapping + marker cofactors is called Composite Interval Mapping (CIM)

CIM works by adding an additional term to the linear model,

$$\sum_{k \neq i, i+1} b_k \, x_{kj}$$

CIM also (potentially) includes unlinked markers to account for QTL on other chromosomes.

# Power and Precision

While modest sample sizes are sufficient to detect a QTL of modest effect (power), large sample sizes are required to map it with any precision

With 200-300 $F_2$, a QTL accounting for 5% of total variation can be mapped to a 40cM interval

Over 10,000 $F_2$ individuals are required to map this QTL to  a 1cM interval

# Power and Repeatability:  The Beavis Effect

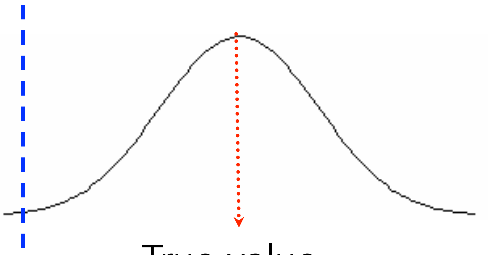
QTLs with low power of detection tend to have their effects ***overestimated***, often very dramatically

As power of detection increases, the overestimation of detected QTLs becomes far less serious

This is often called the Beavis Effect, after Bill Beavis who first noticed this in simulation studies. This phenomena is also called the winner's curse in statistics (and GWAS)

# Beavis Effect

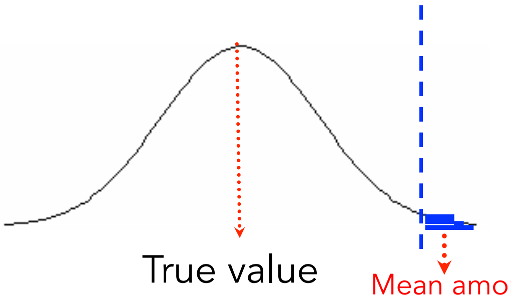
Also called the “winner’s curse” in the GWAS literature



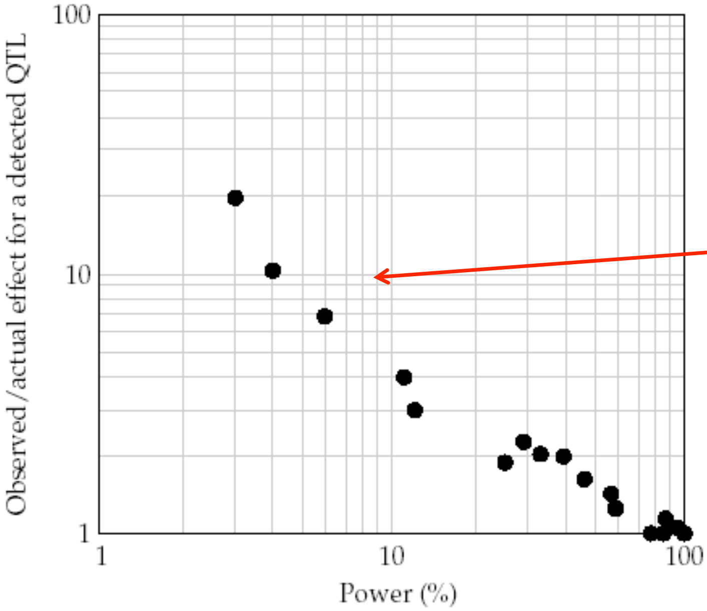
Distribution of
the realized value of an
effect in a sample

High power setting: Most realizations are to the right of the significance threshold. Hence, the average value given the estimate is declared significant (above the threshold) is very close to the true value.

In low power settings, most realizations are below the significance threshold, hence most of the time the effect is scored as being nonsignificant

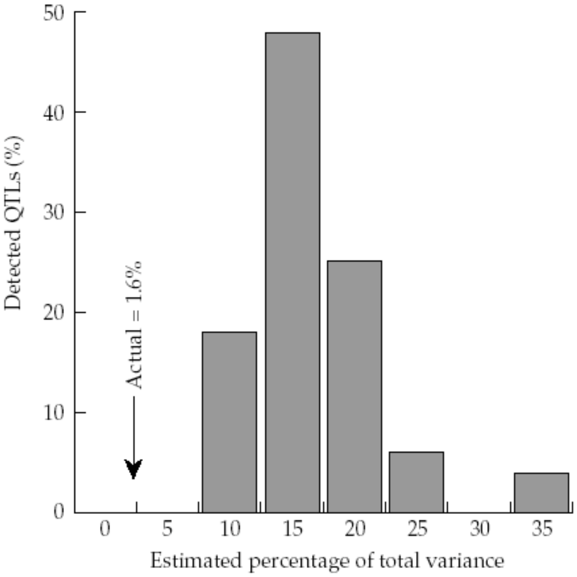


However, the mean of those declared significant is much larger than the true mean



Inflation at lower power

Inflation can be significant, esp. with low power



Beavis simulation: actual effect size is 1.6% of variation. Estimated effects (at significant markers) much higher

# Model selection

- With (say) 300 markers, we have (potentially) 300 single-marker terms and 300*299/2 = 44,850 epistatic terms
  - Hence, a model with up to p= 45,150 possible parameters
  - $2^p$ possible submodels = $10^{13,600}$ ouch!
- The issue of Model selection becomes very important.
- How do we find the best model?
  - Stepwise regression approaches
    - Forward selection (add terms one at a time)
    - Backwards selection (delete terms one at a time)
  - Try all models, assess best fit
  - Mixed-model (random effect) approaches

# Model Selection

Model Selection: Use some criteria to choose among a number of candidate models. Weight goodness-of-fit (L, value of the likelihood at the MLEs) vs. number of estimated parameters (k)

AIC = Akaike's information criterion
AIC = 2k - 2 Ln(L)

BIC = Bayesian information criterion (Schwarz criterion)
BIC = k*ln(n)/n - 2 Ln(L)/n
BIC penalizes free parameters more strongly than AIC

For both AIC & BIC, smaller value is better

# Model averaging

Model averaging: Generate a composite model by weighting (averaging) the various models, using AIC, BIC, or other

Idea: Perhaps no "best" model, but several models all extremely close. Better to report this "distribution" rather than the best one

One approach is to average the coefficients on the "best-fitting" models using some scheme to return a composite model

# Shrinkage estimators

Shrinkage estimates: Rather than adding interaction terms one at a time, a shrinkage method starts with all interactions included, and then shrinks most back to zero.

Under a Bayesian analysis, any effect is *random*. One can assume the effect for (say) interaction *ij* is drawn from a normal with mean zero and variance $\sigma^2_{ij}$

Further, the interaction-specific variances are themselves random variables drawn from a hyperparameter distribution, such as an inverse chi-square.

One then estimates the hyperparameters and uses these to predict the variances, with effects with small variances shrinking back to zero, and effects with large variances remaining in the model.

# What is a "QTL"

- A detected "QTL" in a mapping experiment is a region of a chromosome detected by linkage.
- Usually large (typically 10-40 cM)
- When further examined, most "large" QTLs turn out to be a linked collection of locations with increasingly smaller effects
- The more one localizes, the more subregions that are found, and the smaller the effect in each subregion
- This is called fractionation

# Limitations of QTL mapping

- Poor resolution (~20 cM or greater in most designs with sample sizes in low to mid 100's)
  - Detected "QTLs" are thus large chromosomal regions
- Fine mapping requires either
  - Further crosses (recombinations) involving regions of interest (i.e., RILs, NILs)
  - Enormous sample sizes
    - If marker-QTL distance is 0.5cM, require sample sizes in excess of 3400 to have a 95% chance of 10 (or more) recombination events in sample
    - 10 recombination events allows one to separate effects that differ by ~ 0.6 SD

# Limitations of QTL mapping (cont)

- "Major" QTLs typically fractionate
  - QTLs of large effect (accounting for > 10% of the variance) are routinely discovered.
  - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The Beavis effect:
  - When power for detection is low, marker-trait associations declared to be statistically significant significantly overestimate their true effects.
  - This effect can be very large (order of magnitude) when power is low.

# II: QTL mapping in Outbred Populations and Association Mapping

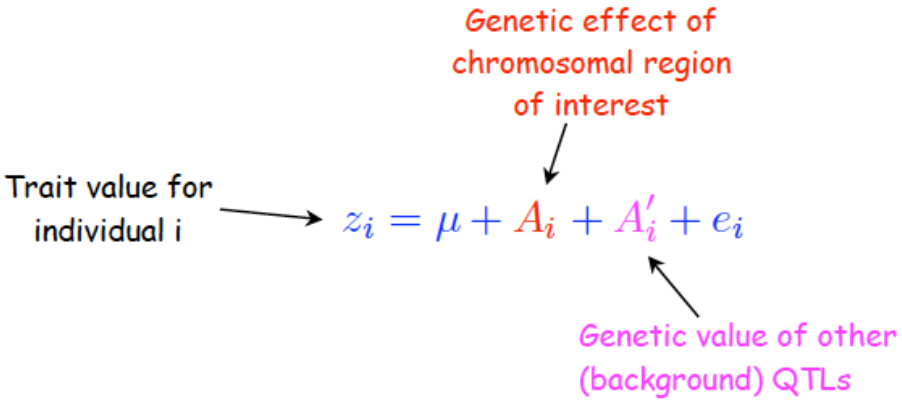
- Association mapping uses a set of very dense markers in a set of (largely) unrelated individuals
- Requires population level LD
- Allows for very fine mapping (1-20 kB)

# QTL mapping in outbred populations

- Much lower power than line-cross QTL mapping
- Each parent must be separately analyzed
- We focus on an approach for general pedigrees, as this leads us into association mapping

# General Pedigree Methods

Random effects (hence, variance component) method for detecting QTLs in general pedigrees



Genetic effect of chromosomal region of interest

Trait value for individual i

$$z_i = \mu + A_i + A_i' + e_i$$

Genetic value of other (background) QTLs

The model is rerun for each marker

$$z_i = \mu + A_i + A_i' + e_i$$

The covariance between individuals i and j is thus

Variance explained by the region of interest

Resemblance between relatives correction

$$\sigma(z_i, z_j) = R_{ij}\, \sigma^2_A + 2\Theta_{ij}\, \sigma^2_{A'}$$

Fraction of chromosomal region shared IBD between individuals i and j.

Variance explained by the background polygenes

Assume z is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R}\, \sigma^2_A + \mathbf{A}\, \sigma^2_{A'} + \mathbf{I}\, \sigma^2_e$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker data

Estimated from the pedigree

The resulting likelihood function is

$$\ell(\mathbf{z} \,|\, \mu, \sigma^2_A, \sigma^2_{A'}, \sigma^2_e) = \frac{1}{\sqrt{(2\pi)^n \,|\mathbf{V}|}} \exp\left[ -\frac{1}{2} (\mathbf{z} - \mu)^T\, \mathbf{V}^{-1}\, (\mathbf{z} - \mu) \right]$$

A significant $\sigma_A^2$ indicates a linked QTL.

# Association & LD mapping

Mapping major genes (LD mapping) vs. trying to Map QTLs (Association mapping)

Idea: Collect random sample of individuals, contrast trait means over marker genotypes

If a dense enough marker map, likely population level linkage disequilibrium (LD) between closely-linked genes

# LD: Linkage disequilibrium

D(AB) = freq(AB) - freq(A)*freq(B).
LD = 0 if A and B are independent. If LD not zero, correlation between A and B in the population

If a marker and QTL are linked, then the marker and QTL alleles are in LD in close relatives, generating a marker-trait association.

The decay of D: $D(t) = (1-c)^t D(0)$
here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

# Dense SNP Association Mapping

Mapping genes using known sets of relatives can be problematic because of the cost and difficulty in obtaining enough relatives to have sufficient power.

By contrast, it is straightforward to gather large sets of unrelated individuals, for example a large number of cases (individuals with a particular trait/disease) and controls (those without it).

With the very dense set of SNP markers (dense = very tightly linked), it is possible to scan for markers in LD in a random mating population with QTLs, simply because c is so small that LD has not yet decayed

These ideas lead to consideration of a strategy of

.

For example, using 30,000 equally spaced SNP in The 3000cM human genome places any QTL within 0.05cM of a SNP. Hence, for an association created t generations ago (for example, by a new mutant allele appearing at that QTL), the fraction of original LD still present is at least $(1-0.0005)^t$ ~ 1-exp(t*0.0005). Thus for mutations 100, 500, and 1000 generations old (2.5K, 12.5K, and 25 K years for humans), this fraction is 95.1%, 77.8%, 60.6%,

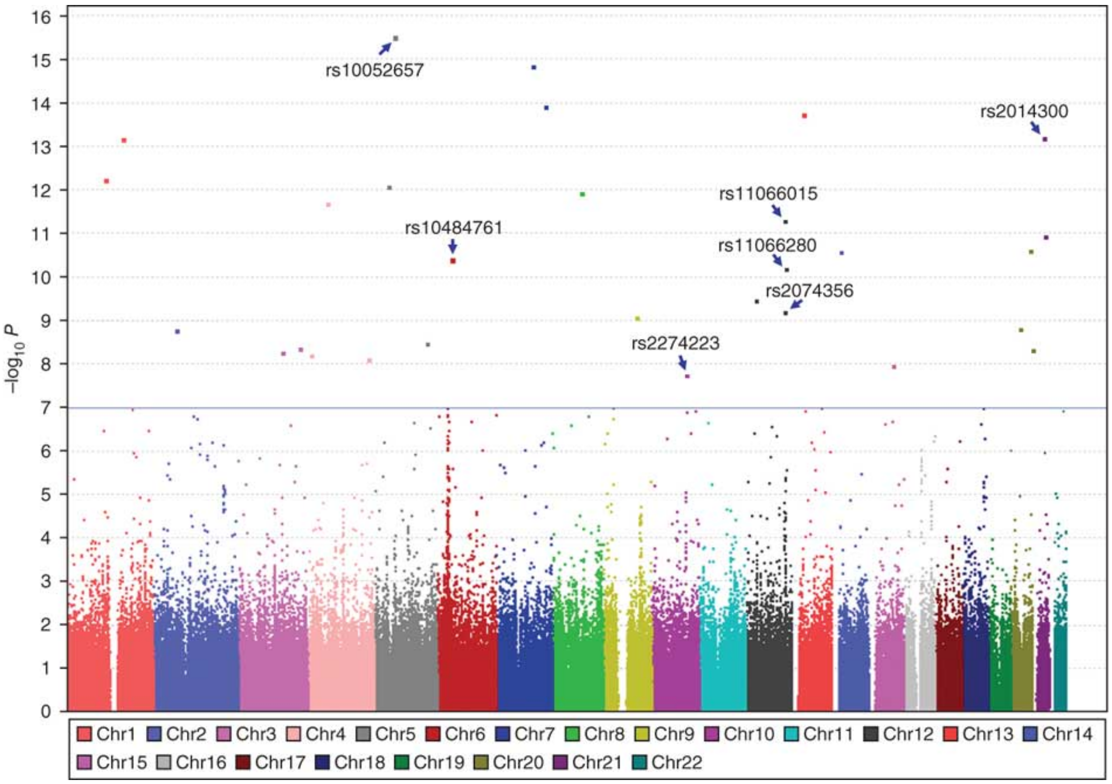
We thus have large samples and high disequilibrium, the recipe needed to detect linked QTLs of small effect

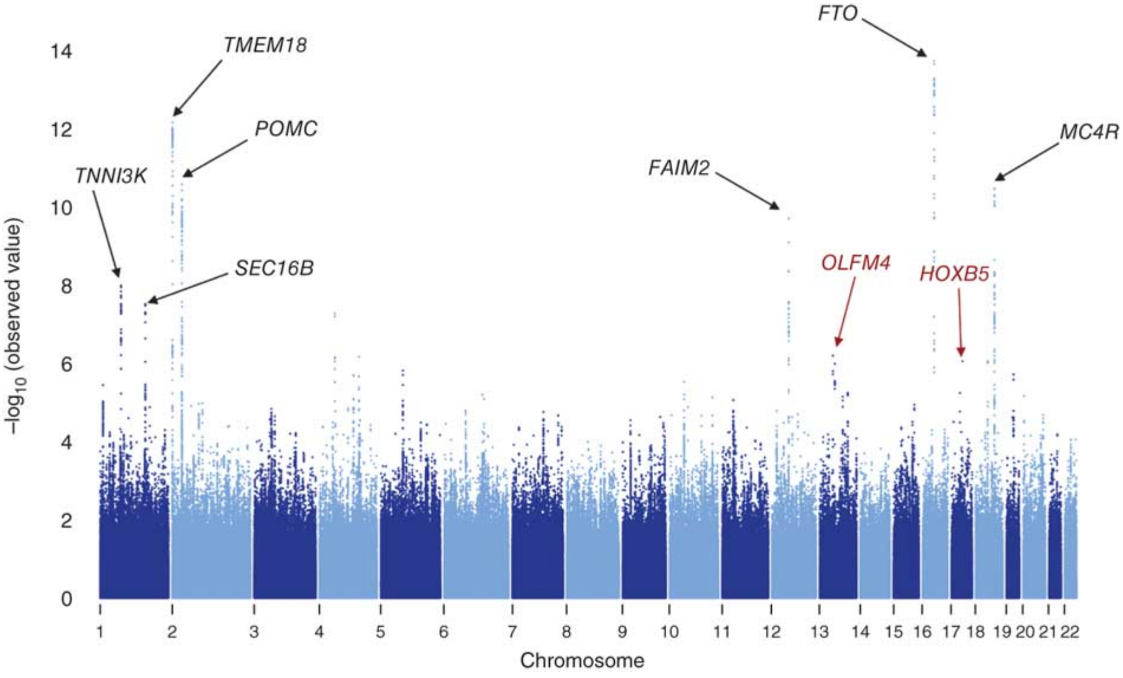
# Association mapping

- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a marker to generate marker-trait associations
  - LD decays very quickly unless very tight linkage
  - Hence, resolution on the scale of LD in the population(s) being studied ( 1 ~ 40 kB)
- Widely used since mid 1990's. Mainstay of human genetics, strong inroads in breeding, evolutionary genetics
- Power a function of the genetic variance of a QTL, not its mean effects

# Manhattan plots

- The results for a Genome-wide Association study (or GWAS) are typically displayed using a Manhattan plot.
  - At each SNP, -ln(p), the negative log of the p value for a significant marker-trait association is plotted. Values above a threshold indicate significant effects
  - Threshold set by Bonferroni-style multiple comparisons correction
  - With n markers, an overall false-positive rate of p requires each marker be tested using p/n.
  - With n = $10^6$ SNPs, p must exceed 0.01/$10^6$ or $10^{-8}$ to have a control of 1% of a false-positive


16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
$-\log_{10} P$
rs10052657
rs2014300
rs11066015
rs10484761
rs11066280
rs2074356
rs2274223
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22


FTO
TMEM18
POMC
MC4R
TNNI3K
FAIM2
SEC16B
OLFM4
HOXB5
14
12
10
8
6
4
2
0
$-\log_{10}$ (observed value)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

# Candidate Loci and the TDT

Often try to map genes by using case/control contrasts, also called association mapping.

The frequencies of marker alleles are measured in both a
case sample -- showing the trait (or extreme values)
control sample -- not showing the trait

The idea is that if the marker is in tight linkage, we might expect LD between it and the particular DNA site causing the trait variation.

Problem with case-control approach (and association mapping in general): Population Stratification can give false positives.

When population being sampled actually consists of several distinct subpopulations we have lumped together, marker alleles may provide information as to which group an individual belongs. If there are other risk factors in a group, this can create a false association btw marker and trait

Example. The Gm marker was thought (for biological reasons) to be an excellent candidate gene for diabetes in the high-risk population of Pima Indians in the American Southwest. Initially a very strong association was observed:

| $Gm^+$ | Total | % with diabetes |
|---|---|---|
| Present | 293 | 8% |
| Absent | 4,627 | 29% |

| $Gm^+$ | Total | % with diabetes |
|---|---|---|
| Present | 293 | 8% |
| Absent | 4,627 | 29% |

Problem: freq($Gm^+$) in Caucasians (lower-risk diabetes Population) is 67%, $Gm^+$ rare in full-blooded Pima

The association was re-examined in a population of Pima that were 7/8th (or more) full heritage:

| $Gm^+$ | Total | % with diabetes |
|---|---|---|
| Present | 17 | 59% |
| Absent | 1,764 | 60% |

# Linkage vs. Association

The distinction between linkage and association is subtle, yet critical

Marker allele M is associated with the trait if Cov(M,y) is not 0

While such associations can arise via linkage, they can also arise via population structure.

Thus, association DOES NOT imply linkage, and linkage is not sufficient for association

# Transmission-disequilibrium test (TDT)

The TDT accounts for population structure. It requires sets of relatives and compares the number of times a marker allele is transmitted (T) versus not-transmitted (NT) from a marker heterozygote parent to affected offspring.

Under the hypothesis of no linkage, these values should be equal, resulting in a chi-square test for lack of fit:

$$\chi^2_{td} = \frac{(T - NT)^2}{(T + NT)}$$

Scan for type I diabetes in Humans. Marker locus D2S152

| Allele | T | NT | $\chi^2$ | p |
|---|---|---|---|---|
| 228 | 81 | 45 | 10.29 | 0.001 |
| 230 | 59 | 73 | 1.48 | 0.223 |
| 240 | 36 | 24 | 2.30 | 0.121 |

$$\chi^2 = \frac{(81 - 45)^2}{(81 + 45)} = 10.29$$

# Accounting for population structure

- Three classes of approaches proposed
  - 1) Attempts to correct for common pop structure signal (genomic control, regression/ PC methods)
  - 2) Attempts to first assign individuals into subpopulations and then perform association mapping in each set (Structure)
  - 3) Mixed models that use all of the marker information (Tassle, EMMA, many others)
    - These can also account for cryptic relatedness in the data set, which also causes false-positives.

# Genomic Control

Devlin and Roeder (1999). Basic idea is that association tests (marker presence/absence vs. trait presence/absence) is typically done with a standard 2 x 2 $\chi^2$ test.

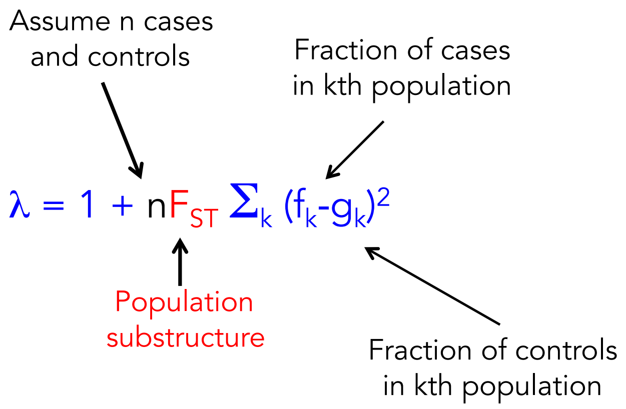
When population structure is present, the test statistic now follows a scaled $\chi^2$, so that if S is the test statistic, then $S/\lambda \sim \chi^2_1$ (so $S \sim \lambda\chi^2_1$)

The inflation factor $\lambda$ is given by

$$\lambda = 1 + nF_{ST} \Sigma_k (f_k - g_k)^2$$

Note that this departure from a $\chi^2$ increases with sample size n

# Genomic Control



Assume n cases and controls

Fraction of cases in kth population

$$\lambda = 1 + nF_{ST} \sum_k (f_k - g_k)^2$$

Population substructure

Fraction of controls in kth population

Genomic control attempts to estimate λ directly from our distribution of test statistics S

# Estimation of λ

The mean of a $\chi^2_1$ is one. Hence, since $S \sim \lambda\chi^2_1$ and we expect most test statistic values to be from the null (no linkage), one estimator of λ is simply the mean of S, the mean value of the test statistics.

The problem is that this is not a particular robust estimator, as a few extreme values of S (as would occur with linkage!) can inflate λ over its true value.

A more robust estimator is offered from the medium (50% value) of the test statistics, so that for m tests

$$\widehat{\lambda} = \frac{\text{medium}\,(S_1, \cdot \cdot ; S_m)}{0.456}$$

# Structured Association Mapping

Pritchard and Rosenberg (1999) proposed Structured Association Mapping, wherein one assumes k subpopulations (each in Hardy-Weinberg).

Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier

Once individuals assigned to groups, association mapping without any correction can occur in each group.

# Regression Approaches

A third approach to control for structure is simply to include a number of markers, outside of the SNP of interest, chosen because they are expected to vary over any subpopulations

How might you choose these in a sample? Try those markers (read STRs) that show the largest departure from Hardy-Weinberg, as this is expected in markers that vary the most over subpopulations.

Indicator (0 / 1) Variable
for SNP genotype k. Typically
k = 3, i.e.  AA, Aa aa

$$y = \mu + \sum_{k=1}^{n} \beta_k\, M_k + \sum_{j=1}^{m} \gamma_j\, b_j + e$$

Significant β indicates
marker-trait association

m unlinked markers that
vary across subpopulations.
$b_j$ = marker genotype indicator
variable

SNP marker
under consideration

Variations on this theme (eigenstrat) --- use all of the marker information to extract a set of significant PCs, which are then included in the model as cofactors

# Mixed-model approaches

- Mixed models use marker data to
  - Account for population structure
  - Account for cryptic relatedness
- Three general approaches:
  - Treat a single SNP as fixed
    - TASSLE, EMMA
  - Treat a single SNP as random
    - General pedigree method
  - Fit all of the SNPs at once
    - GBLUP

# Structure plus Kinship Methods

Association mapping in plants offer occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage).

$$Y = X\beta + Sa + Qv + Zu + e$$

Fixed effects in blue, random effects in red

This is a mixed-model approach. The program TASSEL runs this model.

# Q-K method

$$Y = X\beta + Sa + Qv + Zu + e$$

$\beta$ = vector of fixed effects

$a$ = SNP effects

$v$ = vector of subpopulation effects (STRUCTURE)
$Q_{ij}$ = Prob(individual i in group j). Determined from STRUCTURE output

$u$ = shared polygenic effects due to kinship.
Cov($u$) = var(A)*A, where the relationship matrix A estimated from marker data matrix K, also called a GRM – a genomic relationship matrix

# Which markers to include in K?

- Best approach is to leave out the marker being tested (and any in LD with it) when construction the genomic relationship matrix
  - LOCO approach – leave out one chromosome (which the tested marker is linked to)
- Best approach seems to be to use most of the markers
- Other mixed-model approaches along these lines

# GBLUP

- The Q-K method tests SNPs one at a time, treating them as fixed effects
- The general pedigree method (slides 35-36) also tests one marker at a time, treating them as random effects
- Genomic selection can be thought of as estimating all of the SNP effects at once and hence can also be used for GWAS

# BLUP, GBLUP, and GWAS

- Pedigree information gives EXPECTED value of shared sites (i.e., ½ for full-sibs)
  - A matrix in BLUP
  - The actual realization of the fraction of shared genes for a particular pair of relatives can be rather different, due to sampling variance in segregation of alleles
  - GRM, genomic relationship matrix (or K or marker matrix M)
  - Hence "identical" relatives can differ significantly in faction of shared regions
  - Dense marker information can account for this

# The general setting

- Suppose we have n measured individuals (the n x 1 vector **y** of trait values)
- The n x n relationship matrix **A** gives the relatedness among the sampled individuals, where the elements of **A** are obtained from the pedigree of measured individuals
- We may also have p (>> n) SNPs per individual, where the n x p marker information matrix **M** contains the marker data, where $M_{ij}$ = score for SNP j (i.e., 0 for 00, 1 for 10, 2 for 11) in individual i.

# Covariance structure of random effects

- A critical element specifying the mixed model is the covariance structure (matrix) of the vector **u** of random effects
- Standard form is that Cov(**u**) = variance component * matrix of known constants
  - This is the case for pedigree data, where **u** is typically the vector of breeding values, and the pedigree defines a relationship matrix **A**, with Cov(**u**) = Var(A) * **A,** the additive variance times the relationship matrix
  - With marker data, the covariance of random effects are functions of the marker information matrix **M**.
    - If **u** is the vector of p marker effects, then Cov(**u**) = Var(m) * $\mathbf{M}^T\mathbf{M}$, the marker variance times the covariance structure of the markers.

$$Y = X\beta + Zu + e$$

Pedigree-based BV estimation: (BLUP)
$u_{nx1}$ = vector of BVs, Cov(u) = Var(A) $\mathbf{A}_{nxn}$

Marker-based BV estimation: (GBLUP)
$u_{nx1}$ = vector of BVs, Cov(u) = Var(m) $\mathbf{M}^T\mathbf{M}$ (n x n)

GWAS: $u_{px1}$ = vector of marker effects,
Cov(u) = Var(m) $\mathbf{M}\mathbf{M}^T$ (p x p)

Genomic selection: predicted vector of breeding values from marker effects (genetic breeding values),
$GBV_{nx1} = \mathbf{M}_{nxp} u_{px1}$.
Note that Cov(GBV) = Var(m) $\mathbf{M}^T\mathbf{M}$ (n x n)

Many variations of these general ideas by adding additional assumptions on covariance structure.

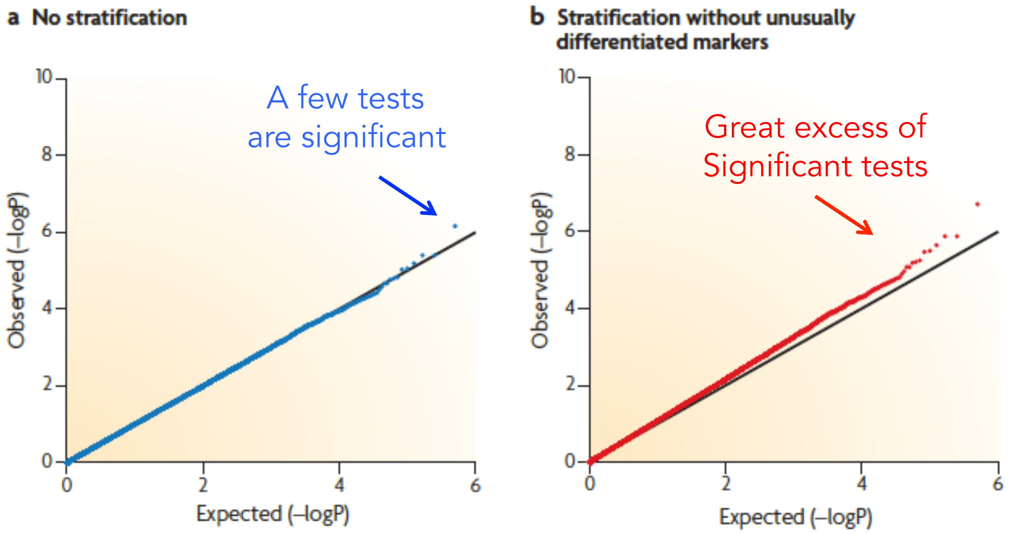
# GWAS Model diagnostics

## Genomic control λ as a diagnostic tool

- Presence of population structure will inflate the λ parameter
- A value above 1 is considered evidence of additional structure in the data
  - Could be population structure, cryptic relatedness, or both
  - A lambda value less that 1.05 is generally considered benign
- One issue is that if the true polygenic model holds (lots of sites of small effect), then a significant fraction will have inflated p values, and hence an inflated λ value.
- Hence, often one computes the λ following attempts to remove population structure. If the resulting value is below 1.05, suggestion that structure has been largely removed.

# P – P plots

- Another powerful diagnostic tool is the p-p plot.
- If all tests are drawn from the null, then the distribution of p values should be uniform.
  - There should be a slight excess of tests with very low p indicating true positives
- This gives a straight line of a log-log plot of observed (seen) and expected (uniform) p values with a slight rise near small values
  - If the fraction of true positives is high (i.e., many sites influence the trait), this also bends the p-p plot



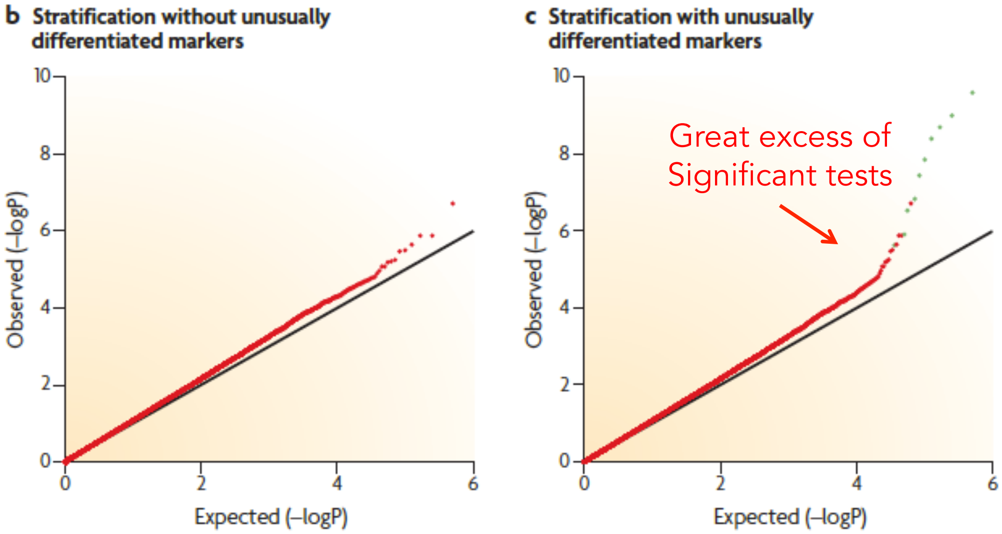
**a** No stratification

**b** Stratification without unusually differentiated markers

A few tests are significant

Great excess of Significant tests

Price et al. 2010 Nat Rev Gene 11: 459



Great excess of Significant tests

As with using λ, one should construct p-p following some approach to correct for structure & relatedness to see if they look unusual.

# Power of Association mapping

Q/q is the polymorphic site contributing to trait variation, M/m alleles (at a SNP) used as a marker

Let p be the frequency of M, and assume that Q only resides on the M background (complete disequilibrium)

| Haloptype | Frequency | effect |
|---|---|---|
| QM | rp | a |
| qM | (1-r)p | 0 |
| qm | 1-p | 0 |

| Haloptype | Frequency | effect |
|---|---|---|
| QM | rp | a |
| qM | (1-r)p | 0 |
| qm | 1-p | 0 |

Effect of m = 0

Effect of M = ar

Genetic variation associated with Q = $2(rp)(1-rp)a^2$
~ $2rpa^2$ when Q rare. Hence, little power if Q rare

Genetic variation associated with marker M is
$2p(1-p)(ar)^2 \sim 2pa^2r^2$

Ratio of marker/true effect variance is ~ r

Hence, if Q rare within the A class, even less power!

# Common variants

- Association mapping is only powerful for common variants
  - freq(Q) moderate
  - freq (r) of Q within M haplotypes modest to large
- Large effect alleles (a large) can leave small signals.
- The fraction of the actual variance accounted for by the markers is no greater than ~ ave(r), the average frequency of Q within a haplotype class
- Hence, don’t expect to capture all of Var(A) with markers, esp. when QTL alleles are rare but markers are common (e.g. common SNPs, $p > 0.05$)
- Low power to detect G x G, G x E interactions

"How wonderful that we have met with a paradox. Now we have some hope of making progress" -- Neils Bohr

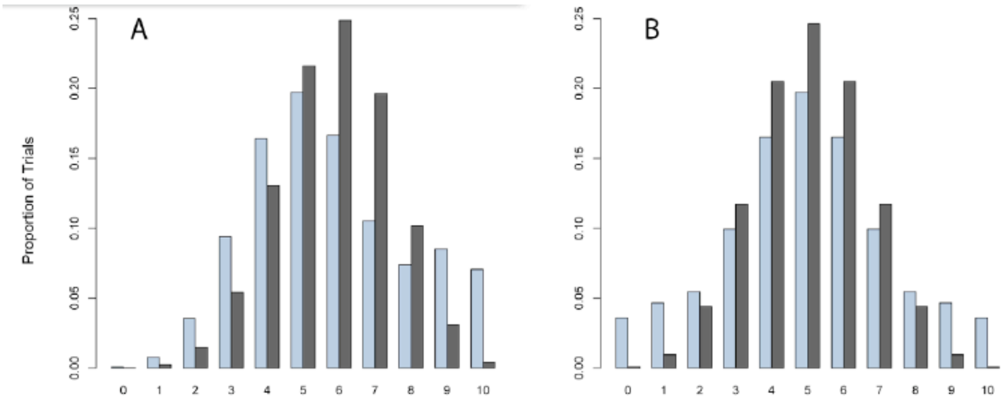
The case of the missing heritability

## The "missing heritability" pseudo-paradox

- A number of GWAS workers noted that the sum of their significant marker variances was much less (typically 10%) than the additive variance estimated from biometrical methods
- The "missing heritability" problem was birthed from this observation.
- Not a paradox at all
  - Low power means small effect (i.e. variance) sites are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests
  - Further, even if all markers are detected, only a fraction ~ r (the frequency of the causative site within a marker haplotype class) of the underlying variance is accounted for.

# Dealing with Rare Variants

- Many disease may be influenced by rare variants.
  - Problem: Each is rare and thus overall gives a weak signal, so testing each variant is out (huge multiple-testing problem)
  - However, whole-genome sequencing (or just sequencing through a target gene/region) is designed to pick up such variants
- Burden tests are one approach
  - Idea: When comparing case vs. controls, is there an overdispersion of mutations between the two categories?



Solid = random distribution over cases/controls
Blue = observed distribution

A: Variants only increase disease risk (excess at high values)

B: Variants can both increase (excess high values) and decrease risk (excess low values) --- inflation of the variance

# C($\alpha$) test

- Idea: Suppose a fraction $p_0$ of the sample are controls, $p_1$ = 1-$p_0$ are cases. Note these varies are fixed over all variants
- Let $n_i$ be the total number of copies of a rare variant i.
- Under binomial sampling, the expected number of variant i in the case group is ~ Bin($p_1$,$n_i$)
- Pool the observations of all such variants over a gene/region of interest and ask if the variance in the number in cases exceeds the binomial sampling variance $n_i p_1(1-p_1)$

# C($\alpha$) test (cont).

- Suppose m variants in a region, test statistic is of the form
- $\Sigma_i (y_i - n_i p_1)^2 - n_i p_1 (1-p_1)$
- $y_i$ = number of variant I in cases.
- This is observed variance minus binomial prediction
- This is scaled by a variance term to give a test statistic that is roughly normally distributed