# SISG
# 2015

# SISG Module 10:
# Quantitative Genetics

**20th Summer Institute in Statistical Genetics**

UNIVERSITY *of* WASHINGTON

(This page left intentionally blank.)

## SYLLABUS
## PRINCIPLES OF QUANTITATIVE GENETICS

INSTRUCTORS:

William Muir, Department of Animal Sciences, Purdue University
bmuir@purdue.edu

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

### LECTURE SCHEDULE

**Monday, 13 July**

| | | |
|---|---|---|
| 8:30 | 10:00 am | 1. Population Genetics Framework (Muir) |
| 10:00 | 10:30 am | Break |
| 10:30 | 12:00 | 2. Fisher's Variance Decomposition (Muir) |
| | | Background reading: LW Chapter 4 |
| 12:00 | 1:30 pm | Lunch |
| 1:30 | 3:00 pm | 3. Resemblance Between Relatives, Heritability (Muir) |
| | | 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 14 July**

| | | |
|---|---|---|
| 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 (Walsh) |
| | | Additional reading: WL Chapters |
| 12:00 | 1:30 pm | Lunch |
| 1:30 | 3:00 pm | 7. Mixed Models, BLUP Breeding Values, REML (Muir) |
| | | Background reading: LW Chapter 26 |
| | | Additional reading: WL Chapters 19, 20 |
| 3:00 | 3:30 pm | Break |
| 3:30 | 5:00 pm | 8. QTL/Association Mapping (Walsh) |
| | | Background reading: LW Chapters 15, 16 |
| Evening | | Open session (review, R, etc) |

**Wednesday, 15 July**

8:30 10:00 am 9. Tests for Molecular Signature of Selection (Walsh)
Background reading:
Additional reading:

10:00 10:30 am Break

10:30 12:00 10. More on Mixed Models, BLUP Breeding Values (Muir)
Additional reading: WL Chapters 8 - 10

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.

# Muir and Walsh
# Lecture 1
# Introduction to Quantitative Genetics

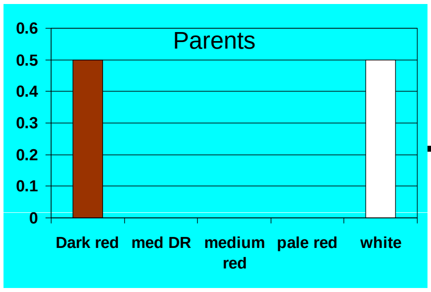
Population Genetics Foundation

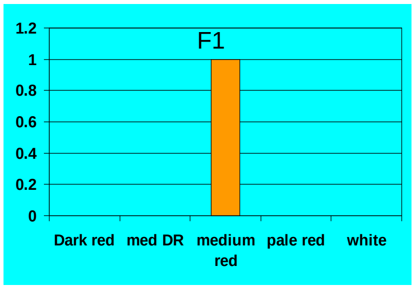
## Quantitative Genetics

- Quantitative Traits
  - Continuous variation
  - Varies by amount rather than kind
  - Height, weight, IQ
- What is the Basis for Quantitative Variation?

# Mendelian bases for Quantitative Genetics

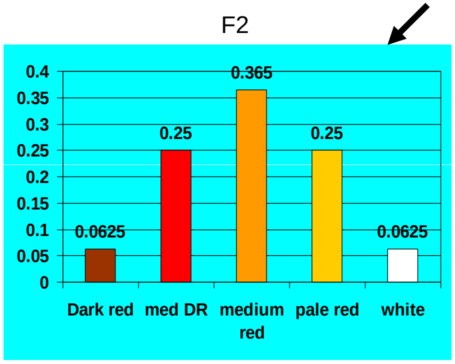
Early experiments by Nilsson-Ehle (1908)
Wheat color





F2

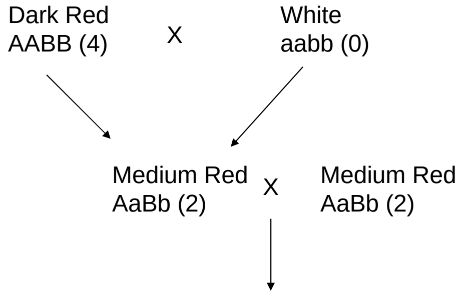
Blending?



Would Blending Explain this?

What Mode of Inheritance Would Explain This?

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



Dark Red
AABB (4)

X

White
aabb (0)

Medium Red
AaBb (2)

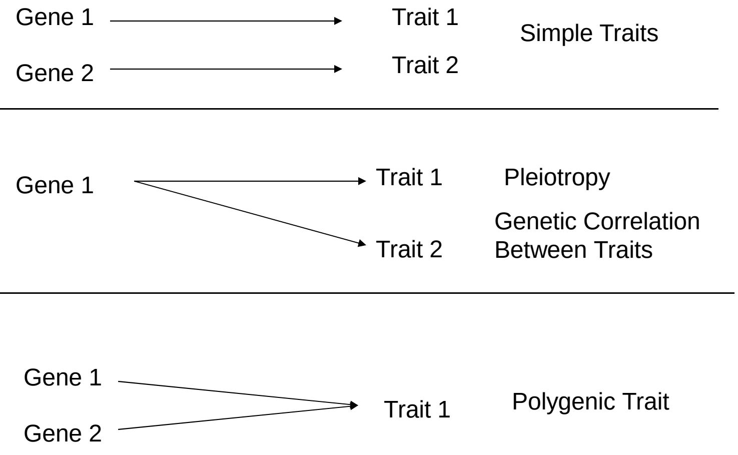
X

Medium Red
AaBb (2)

| Dark Red<br>AABB (4) | Medium<br>Dark Red<br>AABb (3)<br>AaBB (3) | Medium Red<br>AaBb (2)<br>AAbb (2)<br>aaBB (2) | Pale Red<br>Aabb (1)<br>aaBb (1) | White<br>aabb (0) |
|---|---|---|---|---|
| 1/16 | 4/16 | 6/16 | 4/16 | 1/16 |

## Gene Effects

Usual Mendelian Concept

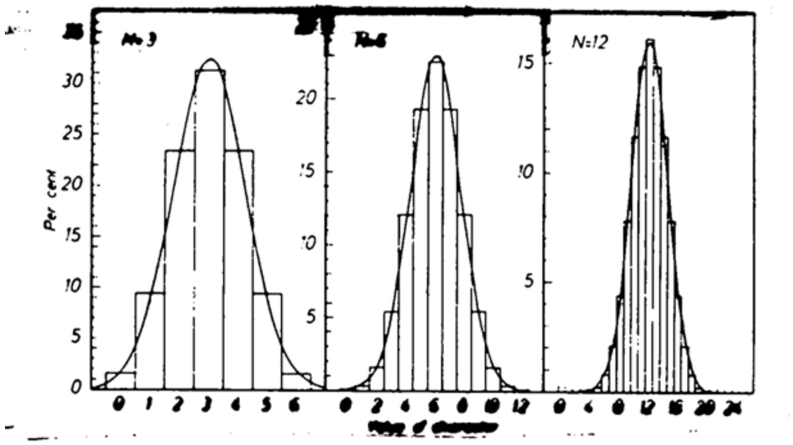


Gene 1 → Trait 1
Gene 2 → Trait 2
Simple Traits

Gene 1 → Trait 1, Trait 2
Pleiotropy
Genetic Correlation
Between Traits

Gene 1, Gene 2 → Trait 1
Polygenic Trait

## What happens to the distribution as the number of loci increases?

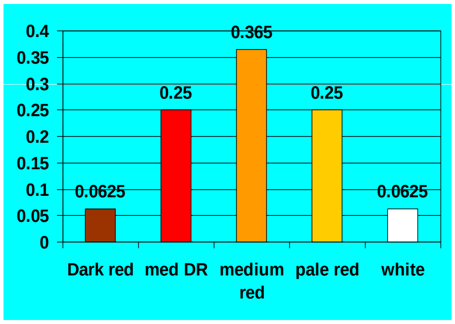


A continuous distribution emerges

## Stability of Distribution

From Previous Example with Wheat

F2



What is the expected Distribution of the F3? F4?

1. Will it stay the same?
2. Will we see a reduction in variability?
3. Will we see and increase in variability?

Need to know concepts of probability

## Important Concepts of Probability Compound Events

Two Events are Independent if
Knowledge of one event tells us nothing about the probability of occurrence of the other event

---

In General

Pr(A and B)=Pr(A|B)xPr(B)

With Independence

Pr(A|B)=Pr(A)

thus

Pr(A and B)=Pr(A)xPr(B)

## What is the Consequence of Random Mating on Genotypic Frequencies

- Assume a Perfect World
  - – No Forces Changing Allele frequency
    - No: mutation, migration, selection, genetic drift
  - – Equal Allele Frequencies in the Sexes
  - – Autosomal Inheritance
  - – Random Mating

# GENERATION 0

- Allow The Genotypic Frequencies To Be Any Arbitrary Values

| | | <u>genotypic frequency</u> |
|---|---|---|
| P( AA ) | = | X |
| P( Aa ) | = | Y |
| P( aa ) | = | Z |

such that X + Y + Z = 1

# Allele Frequencies

$$\mathrm{P(\ A\ ) = P(AA) + \tfrac{1}{2}\,P(Aa)}$$

$$\mathrm{P(\ A\ ) = X + {}^{1}\!/_{2}\,Y}$$

$$= p$$

$$\mathrm{P(\ a\ ) = P(aa) + \tfrac{1}{2}\,P(Aa)}$$

$$\mathrm{P(\ a\ ) = Z + {}^{1}\!/_{2}\,Y}$$

$$= q$$

$p + q = 1$

# Frequency of Mating

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

Random Mating=independence

# Expected genotypic frequencies that result from matings (Gen 1).

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

Conditional Probabilities given genotypes of parents

# GENERATION 1

$$
\begin{aligned}
\mathrm{P}(\mathrm{AA}_{\mathrm{offsping}}) &= 1(\mathrm{X}^2) + \tfrac{1}{2}(2\mathrm{XY}) + \tfrac{1}{4}(\mathrm{Y}^2) \\
&= \mathrm{X}^2 + \mathrm{XY} + \tfrac{1}{4}\mathrm{Y}^2 \\
&= \left(\mathrm{X} + \tfrac{1}{2}Y\right)^2 \\
&= \left[\mathrm{P}(\mathrm{A})_{\mathrm{parents}}\right]^2
\end{aligned}
$$

Because

$$\mathrm{P}(\mathrm{A}) = \mathrm{X} + \tfrac{1}{2}\mathrm{Y} = p$$

Therefore for generation 1

$$\mathrm{P}(\mathrm{AA}_{\mathrm{offspring}}) = p^2$$

$$
\begin{aligned}
\mathrm{P}(\mathrm{Aa}_{\mathrm{offsping}}) &= \tfrac{1}{2}(2\mathrm{XY}) + 1(2\mathrm{XZ}) + \tfrac{1}{2}(\mathrm{Y}^2) + \tfrac{1}{2}(2\mathrm{YZ}) \\
&= \mathrm{XY} + 2\mathrm{XZ} + \tfrac{1}{2}\mathrm{Y}^2 + \mathrm{YZ} \\
&= 2\left(\mathrm{X} + \tfrac{1}{2}\mathrm{Y}\right)\left(\mathrm{Z} + \tfrac{1}{2}\mathrm{Y}\right) \\
&= 2pq
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{P}(\mathrm{aa}_{\mathrm{offsping}}) &= \tfrac{1}{4}(\mathrm{Y}^2) + \tfrac{1}{2}(2\mathrm{YZ}) + 1(\mathrm{Z}^2) \\
&= \tfrac{1}{4}\mathrm{Y}^2 + \mathrm{YZ} + \mathrm{Z}^2 \\
&= \left(\mathrm{Z} + \tfrac{1}{2}\mathrm{Y}\right)^2 \\
&= q^2
\end{aligned}
$$

# Generation 2
# Frequency of Matings

| | | male genotype | | |
|---|---|---|---|---|
| female genotype | | AA | Aa | aa |
| | frequency | ( $p^2$ ) | ( $2pq$ ) | ( $q^2$ ) |
| AA | ( $p^2$ ) | $p^4$ | $2p^3q$ | $p^2q^2$ |
| Aa | ( $2pq$ ) | $2p^3q$ | $4p^2q^2$ | $2pq^3$ |
| aa | ( $q^2$ ) | $p^2q^2$ | $2pq^3$ | $q^4$ |

# Expected genotypic frequencies that result from matings (Gen 1).

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

# Overall Genotypic Frequencies

$$\begin{aligned}
\mathrm{P}(\mathrm{AA}_{\mathrm{offsping}}) &= 1(p^4) + \tfrac{1}{2}(4p^3q) + \tfrac{1}{4}(4p^2q^2) \\
&= p^4 + 2p^3q + p^2q^2 \\
&= p^2\left(p^2 + 2pq + q^2\right) \\
&= p^2(p+q)^2 = p^2(1) \\
&= p^2
\end{aligned}$$

$$\mathrm{P}(\mathrm{Aa}_{\mathrm{offsping}}) = 2pq$$

$$\mathrm{P}(\mathrm{aa}_{\mathrm{offsping}}) = q^2$$

# Summary of genotypic frequencies by Generation

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

# Hardy-Weinberg Equilibrium
# or the Squared Law

If a population starts with any arbitrary distribution of genotypes, provided they are equally frequent in the two sexes, the proportions of genotypes (AA, Aa, aa), with initial allele frequencies p and q, will be in the proportion

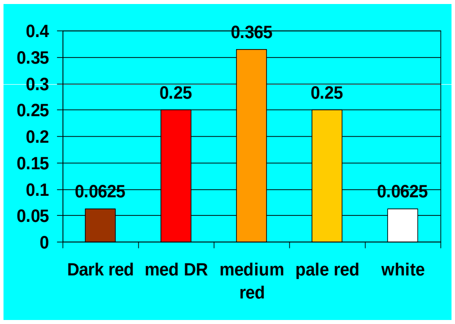
$$\left(p_{\mathrm{A}} + q_{\mathrm{a}}\right)^2 = p^2_{AA} + (2pq)_{Aa} + q^2_{aa}$$

after one generation of random mating and will remain in that distribution **until acted upon by other forces**

# Stability of Distribution

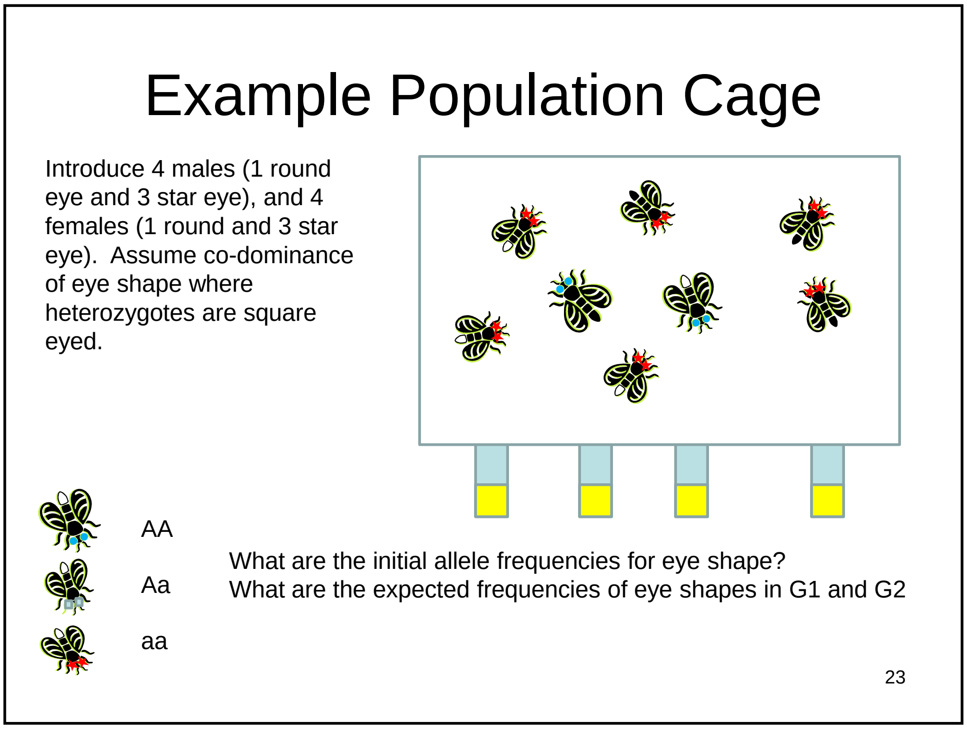
From Previous Example with Wheat

F2



What is the expected Distribution of the F3? F4?

1. **Will stay the same (assumptions?)**
2. Will see a reduction in variability (how, when)?
3. Will see and increase in variability(how, when)?



# Example Population Cage

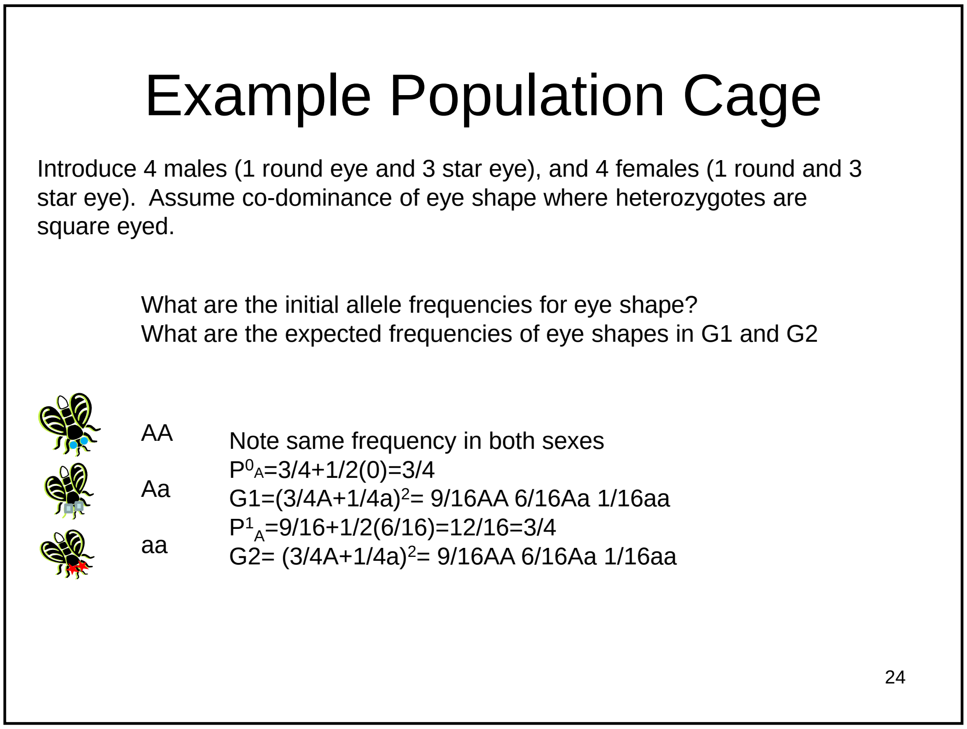
Introduce 4 males (1 round eye and 3 star eye), and 4 females (1 round and 3 star eye). Assume co-dominance of eye shape where heterozygotes are square eyed.

AA

Aa

aa

What are the initial allele frequencies for eye shape?
What are the expected frequencies of eye shapes in G1 and G2



# Example Population Cage

Introduce 4 males (1 round eye and 3 star eye), and 4 females (1 round and 3 star eye). Assume co-dominance of eye shape where heterozygotes are square eyed.

What are the initial allele frequencies for eye shape?
What are the expected frequencies of eye shapes in G1 and G2

AA

Aa

aa

Note same frequency in both sexes
$P^0{}_A$=3/4+1/2(0)=3/4
G1=(3/4A+1/4a)$^2$= 9/16AA 6/16Aa 1/16aa
$P^1{}_A$=9/16+1/2(6/16)=12/16=3/4
G2= (3/4A+1/4a)$^2$= 9/16AA 6/16Aa 1/16aa

What if Organisms Do not Mate But Release Gametes or Pollen to Search Out Each Other?

Do the Same Properties Hold?

# Fertilization

Given a female gamete carrying the ‘a’ allele, what is the probability it will be fertilized by a gamete (pollen) carrying the ‘A’ allele

If Independence
$P(A \mid a)=P(A)$

If not Independence
$P(A \mid a)=$ 0 to 1

Incompatibility

## What is the Consequence of Random Union of Gametes

- Assume a Perfect World
  - – No Forces Changing Allele Frequency
  - – Equal Allele Frequencies in the Sexes
  - – Autosomal Inheritance

## GENERATION 0

lets allow the genotypic frequencies to be any arbitrary value and the allelee frequencies to be the appropriate function of those values.

| | | genotypic frequency |
|---|---|---|
| P( AA ) | = | X |
| P( Aa ) | = | Y |
| P( aa ) | = | Z |

such that X + Y + Z = 1

## Allele Frequency

$$P(\,A\,) = X + {}^{1}\!/_{2}\,Y$$

$$= p$$

$$P(\,a\,) = Z + {}^{1}\!/_{2}\,Y$$

$$= q$$

$p + q = 1$

# With Independence

| | | male gamete/frequency | |
|---|---|---|---|
| | | A<br>( $p$ ) | a<br>( $q$ ) |
| female gamete/<br>frequency | A<br>( $p$ ) | AA<br>( $p^2$ ) | Aa<br>( $pq$ ) |
| | a<br>( $q$ ) | Aa<br>( $pq$ ) | aa<br>( $q^2$ ) |

Random Union of Gametes Produces the Same Outcome as Random Mating

# What Happens If The Allele Frequencies Are Not Equal Between The Sexes?

| Generation 0<br>Gametic Frequencies | | Male Gamete Frequency<br>A<br>$p_m^0$ | <br>a<br>$q_m^0$ |
|---|---|---|---|
| Female Gamete frequency | A $p_f^0$ | $p_m^0 p_f^0$<br>AA | $p_f^0 q_m^0$<br>Aa |
| | a $q_f^0$ | $p_m^0 q_f^0$<br>Aa | $q_m^0 q_f^0$<br>aa |

Genotypic Frequencies Generation 1

Each genotype is equally divided between sexes

## Gametic Frequencies Produced by Adults in First Generation

**Frequency of Homozygous Class + ½ frequency of heterozygous class**

$$p_m^1 = p_f^1 = p_m^0 p_f^0 + \frac{1}{2}\left(p_m^0 q_f^0 + p_f^0 q_m^0\right)$$

Because an equal number of both sexes are produced from each mating

**Note, because gametic frequencies are now equal in sexes**

$$p_m^1 = p_f^1 = p^1$$

# Generation 2

| Generation 1 Gametic Frequencies | | Male Gamete Frequency<br>A<br>$p^1$ | a<br>$q^1$ |
|---|---|---|---|
| Female Gamete frequency | A $p^1$ | $p^1 p^1 = \left(p^1\right)^2$<br>AA | $p^1 q^1$<br>Aa |
| | a $q^1$ | $p^1 q^1$<br>Aa | $q^1 q^1 = \left(q^1\right)^2$<br>aa |

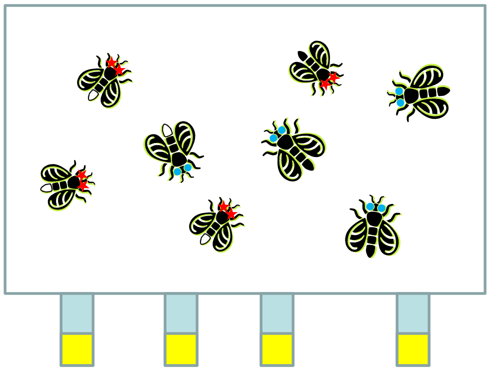
Genotypic Frequencies in Generation 2

Population in HW by 2nd generation

The one generation delay before reaching HW can be very informative

# Example

Introduce 4 males (1 round eye and 3 star eye), and 4 females (3 round and 1 star eye). Assume co-dominance of eye shape where heterozygotes are square eyed.



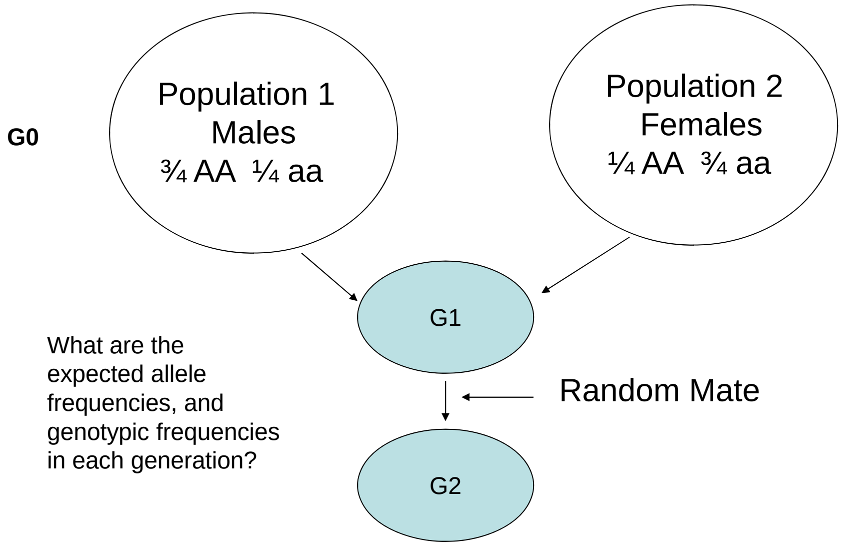
AA

Aa

aa

What are the initial allele frequencies for eye shape?
What are the expected frequencies of eye shapes in G1 and G2

# Example Cross Between Populations

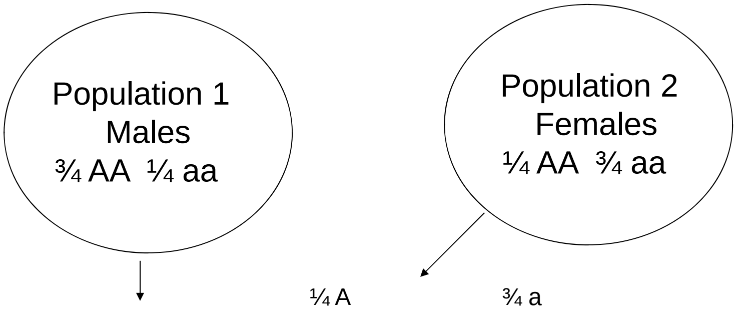

G0

Population 1
Males
¾ AA ¼ aa

Population 2
Females
¼ AA ¾ aa

What are the expected allele frequencies, and genotypic frequencies in each generation?

# Generation 1


Population 1
Males
¾ AA ¼ aa

Population 2
Females
¼ AA ¾ aa

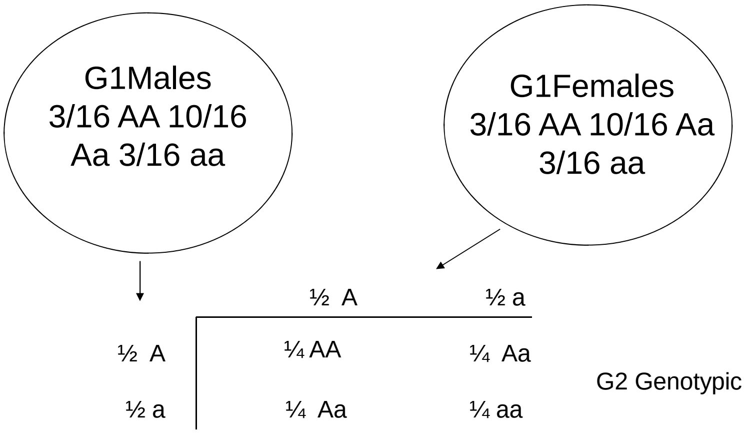
| | ¼ A | ¾ a |
|---|---|---|
| ¾ A | 3/16 AA | 9/16 Aa |
| ¼ a | 1/16 Aa | 3/16 aa |

G1 Genotypic Freq

$p_1=3/16 + ½(9/16 + 1/16)=8/16= ½$

G1 Allele Freq

# Generation 2



G1Males
3/16 AA 10/16 Aa 3/16 aa

G1Females
3/16 AA 10/16 Aa 3/16 aa

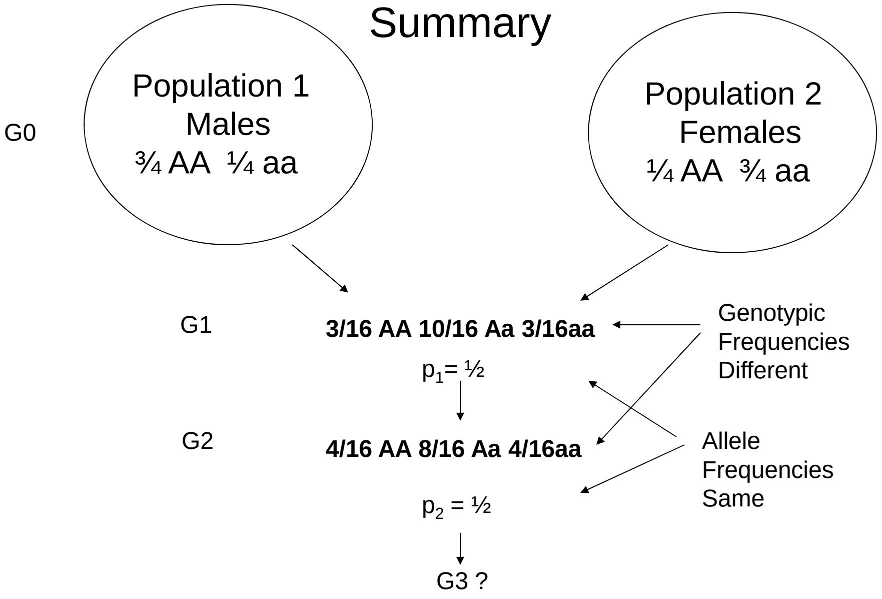
| | ½ A | ½ a |
|---|---|---|
| ½ A | ¼ AA | ¼ Aa |
| ½ a | ¼ Aa | ¼ aa |

G2 Genotypic Freq

$p_2$= ¼ + ½( ¼ + ¼ )= ½

G2 Allele Freq

# Summary



Population 1
Males
¾ AA ¼ aa

Population 2
Females
¼ AA ¾ aa

G0

G1 **3/16 AA 10/16 Aa 3/16aa**

$p_1$= ½

G2 **4/16 AA 8/16 Aa 4/16aa**

$p_2$ = ½

G3 ?

Genotypic
Frequencies
Different

Allele
Frequencies
Same

# Important Example

Sample
Genotype
Distribution

$$( p_{A} + q_{a} )^2 = p^2_{AA} + ( 2pq )_{Aa} + q^2_{aa} =$$

HWE
Expected
Distribution

**3/16 AA**
**10/16 Aa**
**3/16 aa**

Too Many
heterozygotes
in Sample

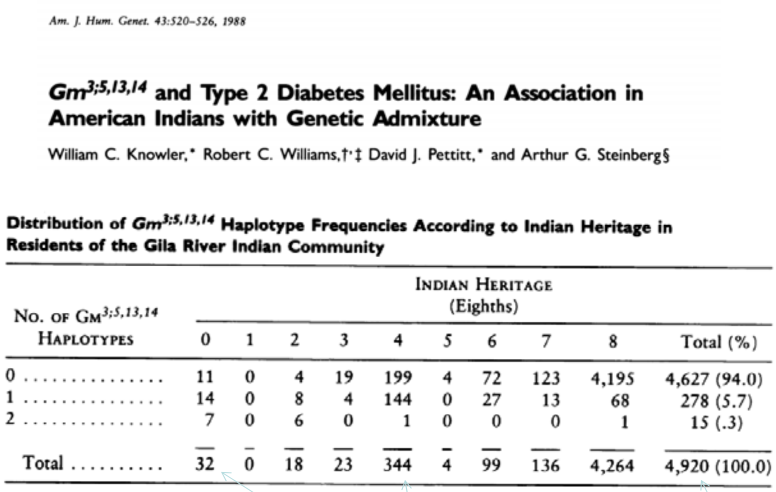
**4/16 AA**
**8/16 Aa**
**4/16aa**

This is also an example of a genomic pattern of recent crossing between populations with immigrants of one sex

Question 1: will this pattern affect all loci? (yes and no, why?)
This is said to be a pattern of demography, why?

Question 2:If the immigration only occurs once, will the pattern of demography disappear with time? If so, how many generations will it take?

Question 3: If an equal number of males and females were among the migrants, and mating was totally at random, regardless of origin, will there be a pattern of excess of heterozygotes?



*Am. J. Hum. Genet. 43:520–526, 1988*

**$Gm^{3;5,13,14}$ and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture**

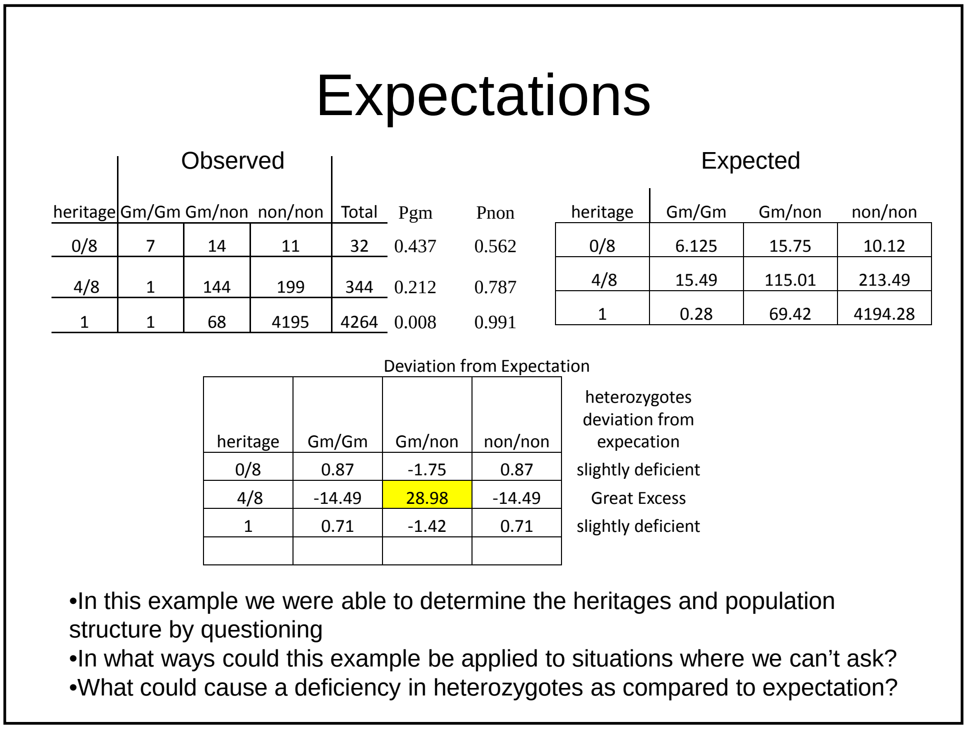
William C. Knowler,* Robert C. Williams,†‡ David J. Pettitt,* and Arthur G. Steinberg§

**Distribution of $Gm^{3;5,13,14}$ Haplotype Frequencies According to Indian Heritage in Residents of the Gila River Indian Community**

| No. of $Gm^{3;5,13,14}$ Haplotypes | Indian Heritage (Eighths) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total (%) |
| 0 | 11 | 0 | 4 | 19 | 199 | 4 | 72 | 123 | 4,195 | 4,627 (94.0) |
| 1 | 14 | 0 | 8 | 4 | 144 | 0 | 27 | 13 | 68 | 278 (5.7) |
| 2 | 7 | 0 | 6 | 0 | 1 | 0 | 0 | 0 | 1 | 15 (.3) |
| Total | 32 | 0 | 18 | 23 | 344 | 4 | 99 | 136 | 4,264 | 4,920 (100.0) |

Consider just No Indian Heritage (0), 50% (4) and pure American Indian (8)

Would you expect each of these sub-populations to conform to H-W distributions? Why?



# Expectations

Observed

| heritage | Gm/Gm | Gm/non | non/non | Total | Pgm | Pnon |
|---|---|---|---|---|---|---|
| 0/8 | 7 | 14 | 11 | 32 | 0.437 | 0.562 |
| 4/8 | 1 | 144 | 199 | 344 | 0.212 | 0.787 |
| 1 | 1 | 68 | 4195 | 4264 | 0.008 | 0.991 |

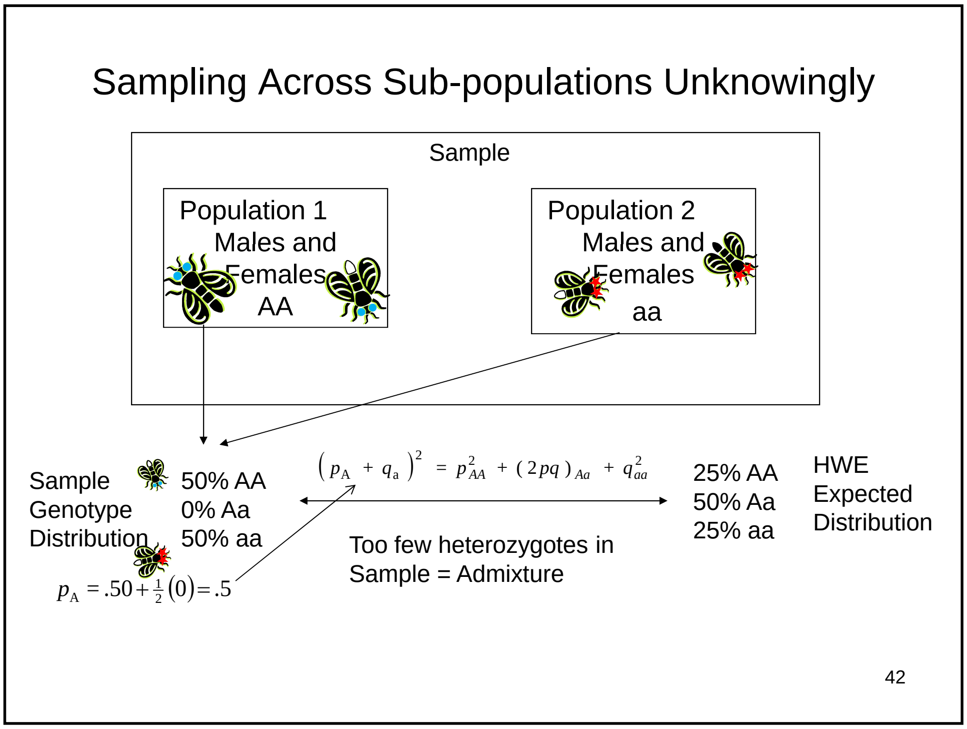
Expected

| heritage | Gm/Gm | Gm/non | non/non |
|---|---|---|---|
| 0/8 | 6.125 | 15.75 | 10.12 |
| 4/8 | 15.49 | 115.01 | 213.49 |
| 1 | 0.28 | 69.42 | 4194.28 |

Deviation from Expectation

| heritage | Gm/Gm | Gm/non | non/non | heterozygotes deviation from expecation |
|---|---|---|---|---|
| 0/8 | 0.87 | -1.75 | 0.87 | slightly deficient |
| 4/8 | -14.49 | 28.98 | -14.49 | Great Excess |
| 1 | 0.71 | -1.42 | 0.71 | slightly deficient |
| | | | | |

- •In this example we were able to determine the heritages and population structure by questioning
- •In what ways could this example be applied to situations where we can't ask?
- •What could cause a deficiency in heterozygotes as compared to expectation?



# Sampling Across Sub-populations Unknowingly

Sample
Population 1 Males and Females AA
Population 2 Males and Females aa

Sample Genotype Distribution

50% AA
0% Aa
50% aa

$p_A = .50 + \frac{1}{2}(0) = .5$

$(p_A + q_a)^2 = p^2_{AA} + (2pq)_{Aa} + q^2_{aa}$

Too few heterozygotes in Sample = Admixture

25% AA
50% Aa
25% aa

HWE Expected Distribution

# Admixture Pima Indians

Combined pure American Indian (8) and Non (0)

| | Gm/Gm | Gm/non | non/non | Total | Pgm | Pnon |
|---|---|---|---|---|---|---|
| total | 8 | 82 | 4206 | 4296 | 0.0114 | 0.9885 |

Expected

| | Gm/Gm | Gm/non | non/non |
|---|---|---|---|
| total | 0.55 | 96.88 | 4198.55 |

Deviation from Expectation

| | Gm/Gm | Gm/non | non/non | heterozygotes deviation from expecation |
|---|---|---|---|---|
| total | 7.44 | -14.88 | 7.44 | Too Few |

## Sampling Across Populations = Admixture

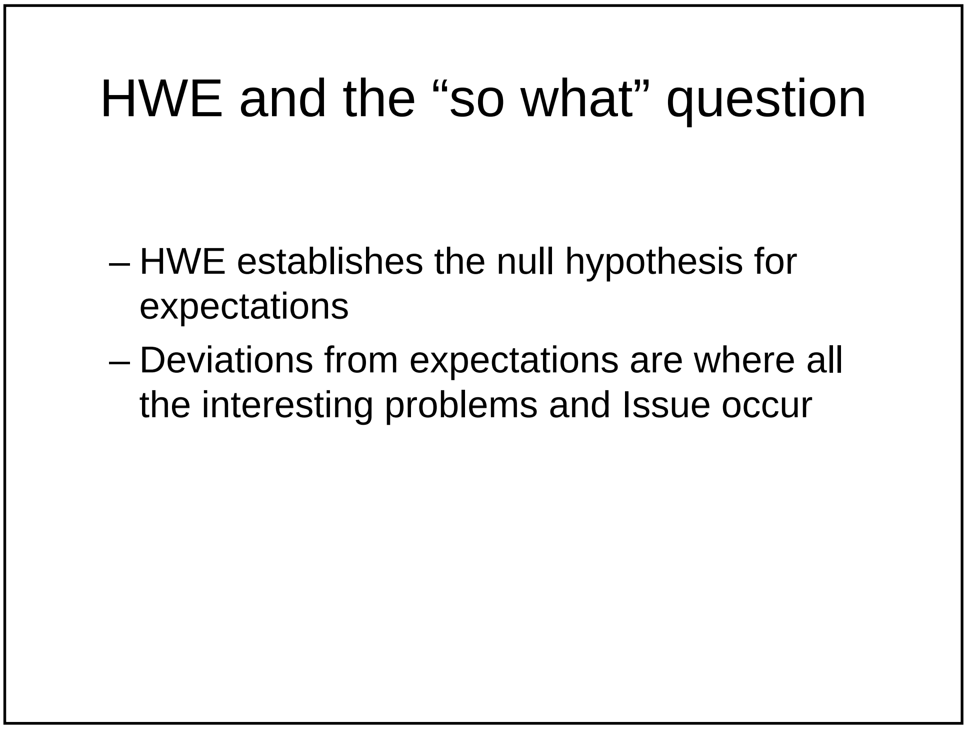
Too few heterozygotes in sample as compared to HWE
= Signal of Admixture

Admixture results from sampling across populations

Question 1: will this pattern affect all loci? (yes and no, why?)

This is also said to be a pattern of demography, why

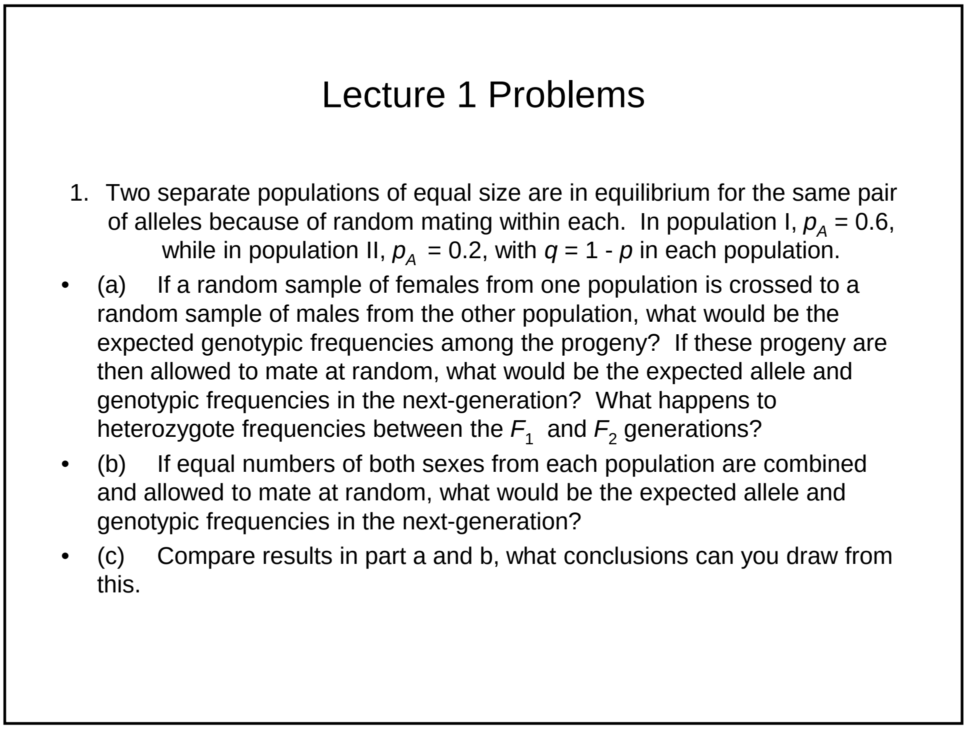
Question 2:If the races remain distinct (pure) will the signature of demography disappear with time? If so, how many generations?



# HWE and the “so what” question

– HWE establishes the null hypothesis for expectations

– Deviations from expectations are where all the interesting problems and Issue occur



## Lecture 1 Problems

1. Two separate populations of equal size are in equilibrium for the same pair of alleles because of random mating within each. In population I, $p_A$ = 0.6, while in population II, $p_A$ = 0.2, with $q = 1 - p$ in each population.

- (a) If a random sample of females from one population is crossed to a random sample of males from the other population, what would be the expected genotypic frequencies among the progeny? If these progeny are then allowed to mate at random, what would be the expected allele and genotypic frequencies in the next-generation? What happens to heterozygote frequencies between the $F_1$ and $F_2$ generations?
- (b) If equal numbers of both sexes from each population are combined and allowed to mate at random, what would be the expected allele and genotypic frequencies in the next-generation?
- (c) Compare results in part a and b, what conclusions can you draw from this.

# Muir Lecture 2

Quantitative Traits
Fisher Decomposition
Covariance Between Relatives

## Quantitative Traits

- Phenotype ($Y$)
  - Continuous (Weight)
  - semi-continuous scale (Egg number)
  - Some Discrete Traits (Disease Resistance)
    - Underlying distribution assumed continuous
- Polygenic ($G$)
- Environmentally Influenced ($E$)

$$Y_i = G_i + E_i$$

## Variances

- Genetic sources of variation
  - Partially underlie trait variation
  - Inferred from statistical sources of variation
- Statistical Sources of variation
  - Variation among and within identifiable groups (families)

## Partitioning sources of variation general case

Classic ANOVA model for two Treatments, ***A*** and ***B*,** each with 2 levels

Treatment A

| Treatment B | $A_1$ | $A_2$ |
|---|---|---|
| $B_1$ | | |
| $B_2$ | | |

interaction effect

$$Y_{ij} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{(ij)}$$

Main Effects

Random error

# Applied to genetics

Treatment A results from alleles from the Father
Treatment B results from alleles from the Mother

Female Parent Allele

| Male parent Allele | $\alpha_1^f$ | $\alpha_2^f$ |
|---|---|---|
| $\alpha_1^m$ | $Y_{11}$ | $Y_{12}$ |
| $\alpha_2^m$ | $Y_{21}$ | $Y_{22}$ |

Intra-locus interaction (dominance)

Model $$Y_{ij} = \mu + \alpha_i^f + \alpha_j^m + \delta_{ij} + \varepsilon_{(ij)}$$

If we assume allele effects from mother are the same as from the father (no imprinting) $$Y_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij} + \varepsilon_{(ij)}$$

# Allelic effects and intra locus interactions

Fit model

$$Y_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij}$$

$$E(\varepsilon) = 0$$

Female Parent Allele

| Male parent Allele | | $p_1$<br>$\alpha_1$ | $p_2$<br>$\alpha_2$ |
|---|---|---|---|
| $p_1$ | $\alpha_1$ | $Y_{11}$ | $Y_{12}$ |
| $p_2$ | $\alpha_2$ | $Y_{21}$ | $Y_{22}$ |

$$\alpha_1 = \sum_{j=1}^{2} p_j Y_{1j} - \mu$$

$$\alpha_2 = \sum_{j=1}^{2} p_j Y_{2j} - \mu$$

$$\alpha_1 = \sum_{i=1}^{2} p_i Y_{i1} - \mu \qquad \alpha_2 = \sum_{i=1}^{2} p_i Y_{i2} - \mu \qquad \mu = \sum_{i}^{2}\sum_{j}^{2} p_i p_j Y_{ij}$$

$$\delta_{ij} = Y_{ij} - \mu - \alpha_i - \alpha_j$$

Questions:
What is the expected frequency of $Y_{11}$?
What assumption is being made to find expected frequency?
Will these expectations be valid if there is admixture or recent crossing?

# Genetic Variances

In General

$$Mean = \overline{Y} = \sum f_i Y_i$$

$$Variance = \sigma_Y^2 = \sum f_i Y_i^2 - \left(\sum f_i Y_i\right)^2$$

So

$$\text{Additive Variance} = \sigma_A^2 = \sum f_i \alpha_i^2 - \left(\sum f_i \alpha_i\right)^2$$

$$\sigma_A^2 = 2\sum_i p_i \alpha_i^2$$

$$\text{Dominance Variance} = \sigma_D^2 = \sum_i \sum_j p_i p_j \delta_{ij}^2 - \left(\sum f_{ij} \delta_{ij}\right)^2$$

$$\sigma_D^2 = \sum_i \sum_j p_i p_j \delta_{ij}^2$$

# Additive Variation

Additive variation=variation due to effects of single alleles

Alleles are passed on in haploid state, thus only the effect of a single allele is inherited from one parent

Breeding for superior traits or natural selection is based on what can be passed on in the haploid state, i.e. single allele effects

Additive variation=useable variation

# Non-additive Variation

- Dominance Variation
  - Due to **intra**-locus interaction
  - Requires both alleles at a locus to express
  - Cannot be passed on by one parent
  - Not useable for selective breeding
- Epistatic Variation
  - Due to **inter**-locus interactions
  - Requires interaction of 2,3, or 4 alleles at two loci
  - Not useable for selective breeding
    - Yes 2 alleles at different loci can be inherited in the haploid state but recombination in following generation(s) will break up

# Example
## Genetic Effects, known genotypes

In a randomly mating population

Female Parent Allele

Male parent Allele

| | | $p_1 = \frac{3}{4}$<br>$\alpha_1$ | $p_2 = \frac{1}{4}$<br>$\alpha_2$ |
|---|---|---|---|
| $\frac{3}{4}$ | $\alpha_1$ | 12 | 10 |
| $\frac{1}{4}$ | $\alpha_2$ | 10 | 4 |

$\alpha_1 = .75$  $\alpha_2 = -2.25$

Fit

$$Y_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij}$$

$$E(\varepsilon) = 0$$

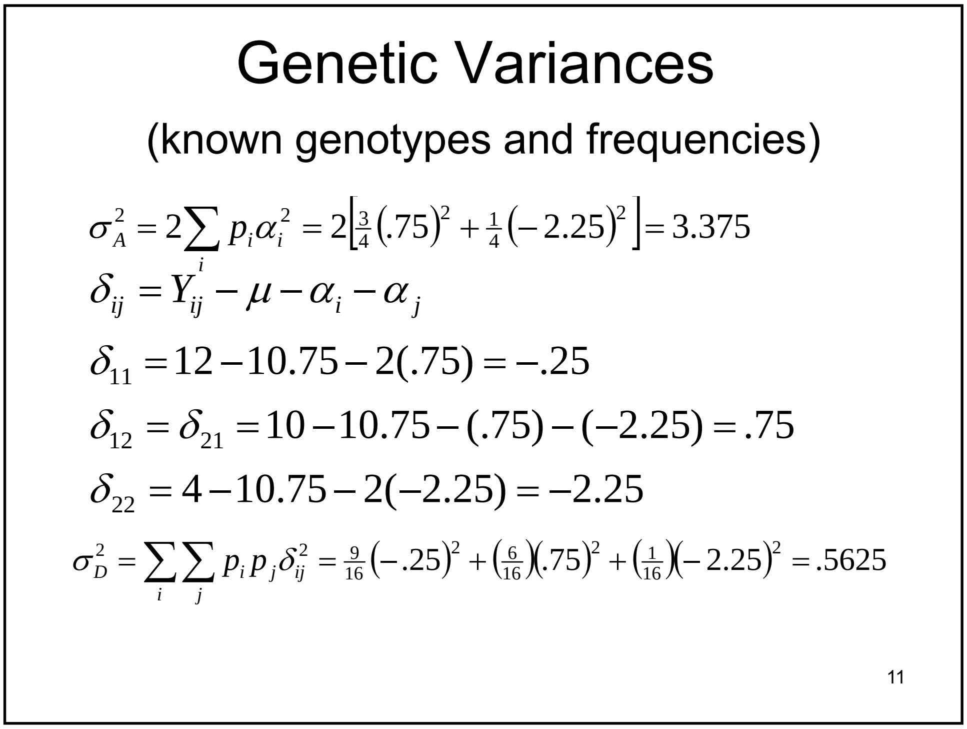
$$\alpha_1 = \tfrac{3}{4}(12) + \tfrac{1}{4}(10) - 10.75 = .75$$

$$\alpha_2 = \tfrac{3}{4}(10) + \tfrac{1}{4}(4) - 10.75 = -2.25$$

$$\mu = \tfrac{9}{16}(12) + \tfrac{3}{16}(10) + \tfrac{3}{16}(10) + \tfrac{1}{16}(4) = 10.75$$

Note: $\sum_i p_i\alpha_i = p_1\alpha_1 + p_2\alpha_2 = \tfrac{3}{4}(.75) + \tfrac{1}{4}(-2.25) = 0$

By construct!



# Genetic Variances

(known genotypes and frequencies)

$$\sigma_A^2 = 2\sum_i p_i \alpha_i^2 = 2\left[\tfrac{3}{4}(.75)^2 + \tfrac{1}{4}(-2.25)^2\right] = 3.375$$

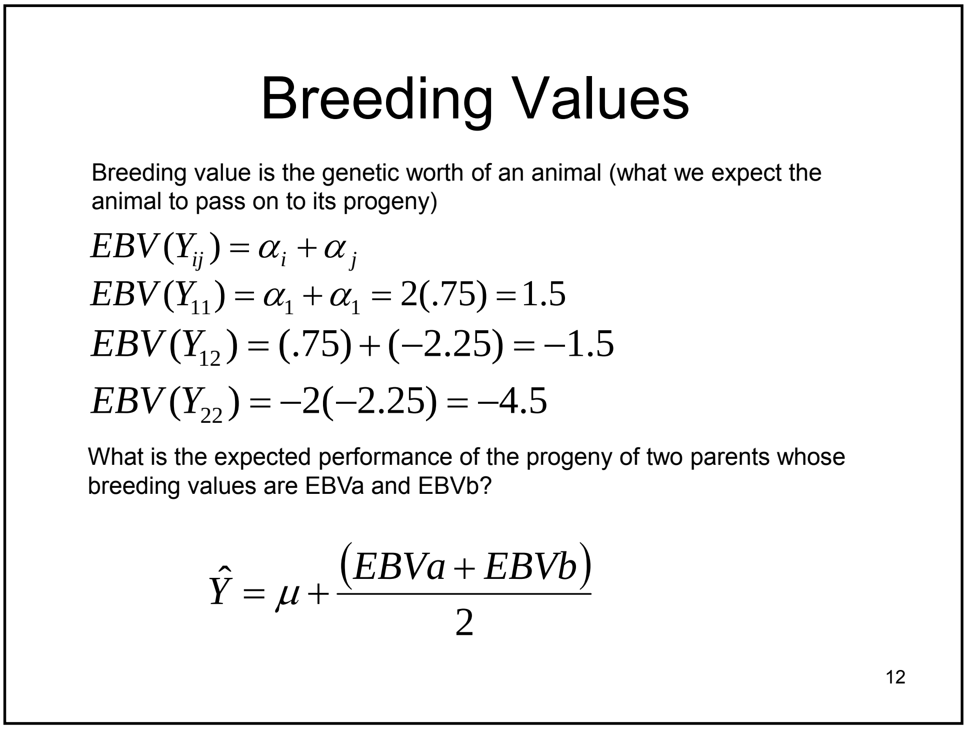
$$\delta_{ij} = Y_{ij} - \mu - \alpha_i - \alpha_j$$

$$\delta_{11} = 12 - 10.75 - 2(.75) = -.25$$

$$\delta_{12} = \delta_{21} = 10 - 10.75 - (.75) - (-2.25) = .75$$

$$\delta_{22} = 4 - 10.75 - 2(-2.25) = -2.25$$

$$\sigma_D^2 = \sum_i \sum_j p_i p_j \delta_{ij}^2 = \tfrac{9}{16}(-.25)^2 + \left(\tfrac{6}{16}\right)(.75)^2 + \left(\tfrac{1}{16}\right)(-2.25)^2 = .5625$$



# Breeding Values

Breeding value is the genetic worth of an animal (what we expect the animal to pass on to its progeny)

$$EBV(Y_{ij}) = \alpha_i + \alpha_j$$

$$EBV(Y_{11}) = \alpha_1 + \alpha_1 = 2(.75) = 1.5$$

$$EBV(Y_{12}) = (.75) + (-2.25) = -1.5$$

$$EBV(Y_{22}) = -2(-2.25) = -4.5$$

What is the expected performance of the progeny of two parents whose breeding values are EBVa and EBVb?

$$\hat{Y} = \mu + \frac{(EBVa + EBVb)}{2}$$

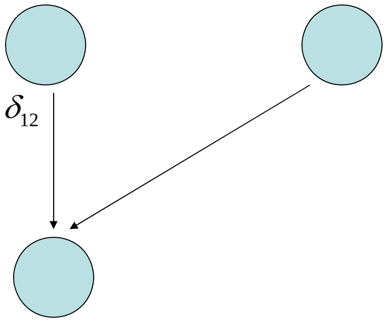
## Covariance Between Relatives

- Needed to associate genetic source of variance with statistical source of variation
  - Separate genetic sources of variation into additive vs. non-additive
  - Separate genetic sources of variation from environmental
- The entire underpinning to quantitative genetics

## Genetic Covariance

- Relatives are more likely to share alleles than non-relatives
  - Sharing Alleles = Identical by Descent (IBD)
    - Identically same allele can be traced to an ancestor
    - Statistical Concept
  - Not to be confused with Alike in State (AIS)
    - If two alleles are AIS they maybe IBD
    - If two alleles are not AIS then they cannot be IBD

# Single Parent-offspring Covariance



P1

P2

$A_1A_2$

$A_3A_4$

$G_P = \alpha_1 + \alpha_2 + \delta_{12}$

100% of the time parent and offspring will share one and only one allele)

½ $A_1A_x$ $\quad G_O = \alpha_1 + \alpha_x + \delta_{1x}$

½ $A_2A_x$ $\quad G_O = \alpha_2 + \alpha_x + \delta_{2x}$

## Single Parent-offspring Covariance

1-IBD

$$Cov(G_P, G_O) = \tfrac{1}{2} Cov(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_1 + \alpha_x + \delta_{1x})$$
$$+ \tfrac{1}{2} Cov(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_2 + \alpha_x + \delta_{2x})$$

1-IBD

$$Cov(G_P, G_O) = \tfrac{1}{2} \begin{pmatrix} Cov(\alpha_1, \alpha_1) + Cov(\alpha_1, \alpha_x) + Cov(\alpha_1, \delta_{1x}) + \\ Cov(\alpha_2, \alpha_1) + Cov(\alpha_2, \alpha_x) + Cov(\alpha_2, \delta_{1x}) + \\ Cov(\delta_{12}, \alpha_1) + Cov(\delta_{12}, \alpha_x) + Cov(\delta_{12}, \delta_{1x}) \end{pmatrix}$$

$$+ \tfrac{1}{2} \begin{pmatrix} Cov(\alpha_1, \alpha_2) + Cov(\alpha_1, \alpha_x) + Cov(\alpha_1, \delta_{2x}) + \\ Cov(\alpha_2, \alpha_2) + Cov(\alpha_2, \alpha_x) + Cov(\alpha_2, \delta_{2x}) + \\ Cov(\delta_{12}, \alpha_2) + Cov(\delta_{12}, \alpha_x) + Cov(\delta_{12}, \delta_{2x}) \end{pmatrix}$$

## In General: Covariance between effects

$$Cov(\alpha_i, \alpha_j) = \begin{cases} 0 & \text{if i ≠ j, i.e. not IBD} \\ \frac{1}{2}\sigma_A^2 & \text{if i = j, i.e. IBD} \end{cases} \qquad \sigma_A^2 = 2\sum_i p_i \alpha_i^2$$

Additivity is a function of single alleles

$$Cov(\alpha_i, \delta_{ij}) = 0$$

By construct (residuals are found as deviations from main effects)

$$Cov(\delta_{ij}, \delta_{km}) = \begin{cases} 0 & \text{if i j≠ km, i.e. both not IBD} \\ \sigma_D^2 & \text{if i j= km, i.e. both IBD} \end{cases} \qquad \sigma_D^2 = \sum_i \sum_j p_i p_j \delta_{ij}^2$$

Dominance is a function of two alleles at the same locus, it is estimated here as the failure of both alleles at that locus to be additive.

Question: What is epistasis a function of?

## Single Parent-offspring Covariance

1-IBD

$$Cov(G_P, G_O) = \tfrac{1}{2}Cov(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_1 + \alpha_x + \delta_{1x})$$
$$+ \tfrac{1}{2}Cov(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_2 + \alpha_x + \delta_{2x})$$

1-IBD

$$Cov(G_P, G_O) = \tfrac{1}{2}\left(\tfrac{1}{2}\sigma_A^2 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0\right)$$
$$+ \tfrac{1}{2}\left(0 + 0 + 0 + \tfrac{1}{2}\sigma_A^2 + 0 + 0 + 0 + 0 + 0\right)$$

$$Cov(G_P, G_O) = \tfrac{1}{2}\sigma_A^2$$

## Summary Single Parent-Offspring

| Number of IBD | Probability of Sharing | Contribution to variances |
|---|---|---|
| 0 IBD alleles | 0 | 0 |
| 1 IBD alleles | 1 | $\frac{1}{2}\sigma_A^2$ |
| 2 IBD alleles | 0 | $\sigma_A^2+\sigma_D^2$ |

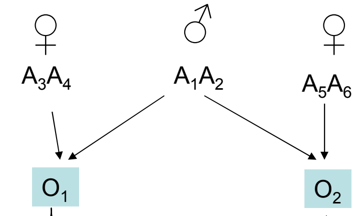
$$Cov(G_P, G_O)$$

Total covariance=Sum Probability x contribution=

$$\sigma_{P.O} = (1)\tfrac{1}{2}\sigma_A^2 + (0)(\sigma_A^2+\sigma_D^2)$$

$$\sigma_{P.O} = \tfrac{1}{2}\sigma_A^2$$

## Collateral: Half sibs



♀ $A_3A_4$ ♂ $A_1A_2$ ♀ $A_5A_6$

$O_1$ $O_2$

|  | $A_1$ | $A_2$ |
|---|---|---|
| $A_3$ | ¼ | ¼ |
| $A_4$ | ¼ | ¼ |

|  | $A_1$ | $A_2$ |
|---|---|---|
| $A_5$ | ¼ | ¼ |
| $A_6$ | ¼ | ¼ |

$$Cov(G_{O_1}, G_{O_2})$$

## Collateral: Number of IBD Alleles

Sib $O_1$ possible genotypes

Sib $O_2$ possible genotypes

|  | $A_1A_3$<br>1/4 | $A_1A_4$<br>1/4 | $A_2A_3$<br>1/4 | $A_2A_4$<br>1/4 |
|---|---|---|---|---|
| $A_1A_5$ ¼ | 1 | 1 | 0 | 0 |
| $A_1A_6$ ¼ | 1 | 1 | 0 | 0 |
| $A_2A_5$ ¼ | 0 | 0 | 1 | 1 |
| $A_2A_6$ ¼ | 0 | 0 | 1 | 1 |

## Collateral: Half sibs

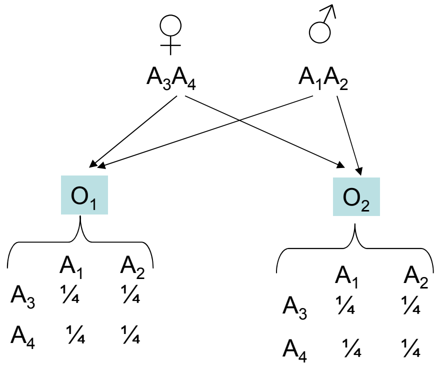
| Number of IBD | Probability of Sharing | Contribution to variances |
|---|---|---|
| 0 IBD alleles | 8/16 | 0 |
| 1 IBD alleles | 8/16 | $\frac{1}{2}\sigma_A^2$ |
| 2 IBD alleles | 0 | $\sigma_A^2 + \sigma_D^2$ |

$$Cov(G_{O_1}, G_{O_2})$$

Total covariance=Sum Probability x contribution=

$$\sigma_{HS} = (\tfrac{8}{16})\tfrac{1}{2}\sigma_A^2 = \tfrac{1}{4}\sigma_A^2$$

## Collateral: Full sibs



♀ $A_3A_4$

♂ $A_1A_2$

$O_1$

|  | $A_1$ | $A_2$ |
|---|---|---|
| $A_3$ | ¼ | ¼ |
| $A_4$ | ¼ | ¼ |

$O_2$

|  | $A_1$ | $A_2$ |
|---|---|---|
| $A_3$ | ¼ | ¼ |
| $A_4$ | ¼ | ¼ |

$$Cov(G_{O_1}, G_{O_2}) = \sigma_{FS}$$

## Collateral: Number of IBD Alleles

Sib $O_1$ possible genotypes

Sib $O_2$ possible genotypes

|  | $A_1A_3$<br>1/4 | $A_1A_4$<br>1/4 | $A_2A_3$<br>1/4 | $A_2A_4$<br>1/4 |
|---|---|---|---|---|
| $A_1A_3$ ¼ | 2 | 1 | 1 | 0 |
| $A_1A_4$ ¼ | 1 | 2 | 0 | 1 |
| $A_2A_3$ ¼ | 1 | 0 | 2 | 1 |
| $A_2A_4$ ¼ | 0 | 1 | 1 | 2 |

## Collateral: Full sibs

| Number of IBD | Probability of Sharing | Contribution to variances |
|---|---|---|
| 0 IBD alleles | 4/16 | 0 |
| 1 IBD alleles | 8/16 | $\frac{1}{2}\sigma_A^2$ |
| 2 IBD alleles | 4/16 | $\sigma_A^2+\sigma_D^2$ |

$Cov(G_{O_1},G_{O_2})$ Total covariance=Sum Probability x contribution=

$$\sigma_{FS}=\left(\tfrac{8}{16}\right)\left(\tfrac{1}{2}\sigma_A^2\right)+\left(\tfrac{4}{16}\right)\left(\sigma_A^2+\sigma_D^2\right)=\tfrac{1}{2}\sigma_A^2+\tfrac{1}{4}\sigma_D^2$$

Question: genetically, are you more similar to your mother or full sister ?

## Covariance Between Relatives in General

$Cov(G_x,G_y)$

| Number of IBD | Probability of Sharing | Contribution to variances |
|---|---|---|
| 1 IBD alleles | $\Pr(1)_{xy}$ | $\frac{1}{2}\sigma_A^2$ |
| 2 IBD alleles | $\Pr(2)_{xy}$ | $\sigma_A^2+\sigma_D^2$ |

Let $a_{xy}=\frac{1}{2}\Pr(1)_{xy}+\Pr(2)_{xy}$ $u_{xy}=\Pr(2)_{xy}$

Additive relationship

Dominance Relationship

Total covariance=Sum contributions=

$$Cov(G_x,G_y)=a_{xy}\sigma_A^2+u_{xy}\sigma_D^2$$

With higher order effects (epistasis)

$$Cov(G_x,G_y)=a_{xy}\sigma_A^2+u_{xy}\sigma_D^2+a_{xy}^2\sigma_{AA}^2+a_{xy}u_{xy}\sigma_{AD}^2+u_{xy}^2\sigma_{DD}^2+\ldots$$

# Problem 1

- Falconer (1981) reported a partially dominant gene in the mouse called *pg* “pygmy.” At six weeks of age, it produces the following average weight phenotypes in grams:
- + / + : 14, + / *pg* : 12, *pg* / *pg* : 6
- 
- (a) What is the additive effect of each “+” substitution? What would be the expected mean, additive and dominance variance in a population with $p_+ = 0.8$, $q_{pg} = 0.2$ under random mating? ?
- 
- (b) If $p_+ = q_{pg}$ , What average effect of an allele substitution? what would be the mean additive and dominance variance?

Problem 1a Answer

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | + | **pg** | | | | |
| | | **0.8** | **0.2** | | Alpha(i) | | |
| + | **0.8** | **14** | **12** | **13.6** | 0.56 | 0.25088 | |
| **pg** | **0.2** | **12** | **6** | **10.8** | -2.24 | 1.00352 | |
| | | **13.6** | **10.8** | **13.04** | 0 | 1.2544 | |
| | Alpha(i) | 0.56 | -2.24 | 0 | | | |
| | var | 0.25088 | 1.00352 | 1.2544 | | **2.5088** | sig2(a) |

Dominance

| | | |
|---|---|---|
| | 0.8 | 0.2 |
| 0.8 | -0.16 | 0.64 |
| 0.2 | 0.64 | -2.56 |

**0.4096** sig2(d)

Problem 1b Answer

Lesson: Additive genetics variance is dependent on allele frequencies

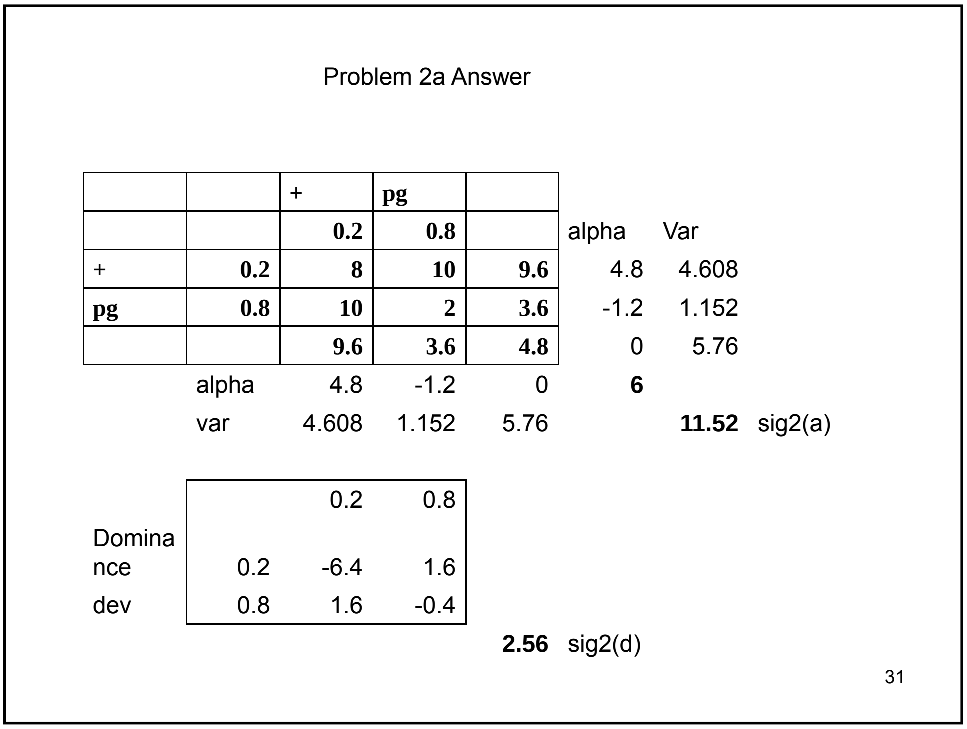
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | + | **pg** | | | | |
| | | **0.5** | **0.5** | | alpha | pqa$^2$ | |
| + | **0.5** | **14** | **12** | **13** | 2 | 2 | |
| **pg** | **0.5** | **12** | **6** | **9** | -2 | 2 | |
| | | **13** | **9** | **11** | 0 | 4 | |
| | alpha | 2 | -2 | 0 | **4** | | |
| | pa$^2$ | 2 | 2 | 4 | | **8** | sig2(a) |

| | | | |
|---|---|---|---|
| | | 0.5 | 0.5 |
| Dominance | 0.5 | -1 | 1 |
| dev | 0.5 | 1 | -1 |

**1** sig2(d)

# Problem 2

- Consider the following phenotypes:
- $A_1A_1$ = 8 $A_1A_2$ = 10 $A_2A_2$ = 2
- (a) If $p$ = 0.2, $q$ = 0.8 , What is the effect of an allele substitution? what would be the mean additive and dominance variance?
- (b) If $p$ = 0.8, $q$ = 0.2 What is the effect of an allele substitution? What would be the expected mean, additive and dominance variance?
- (c) Considering these results, what are the limitations of working backward and drawing conclusions about gene action from calculations of variance components?

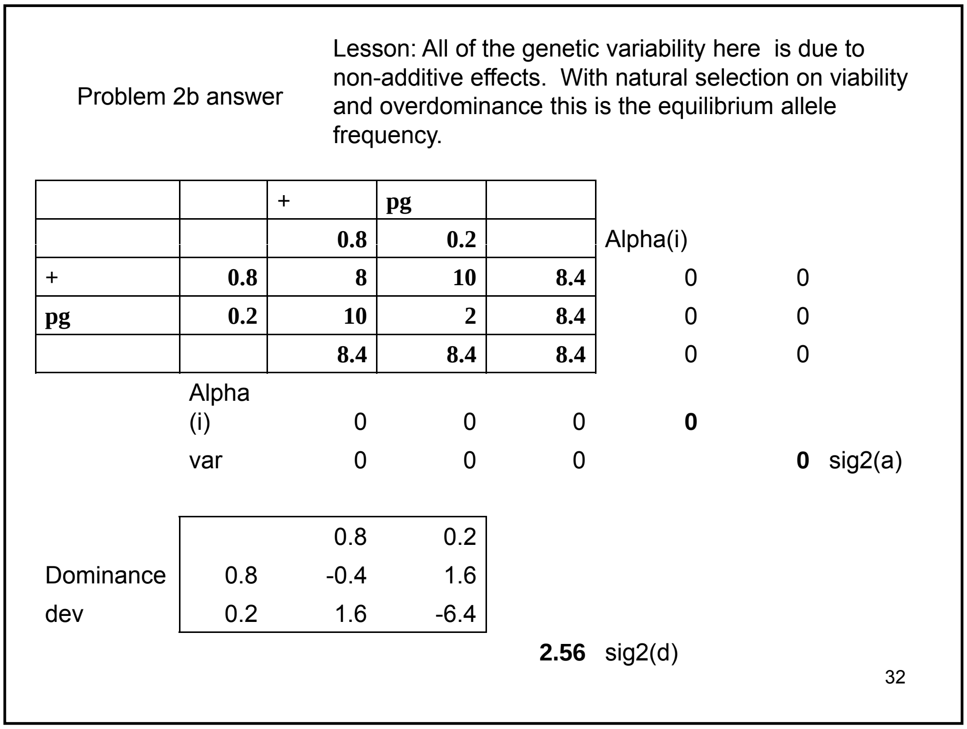


Problem 2a Answer

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | + | **pg** | | | | |
| | | **0.2** | **0.8** | | alpha | Var | |
| + | **0.2** | **8** | **10** | **9.6** | 4.8 | 4.608 | |
| **pg** | **0.8** | **10** | **2** | **3.6** | -1.2 | 1.152 | |
| | | **9.6** | **3.6** | **4.8** | 0 | 5.76 | |
| | alpha | 4.8 | -1.2 | 0 | **6** | | |
| | var | 4.608 | 1.152 | 5.76 | | **11.52** | sig2(a) |

| | | | |
|---|---|---|---|
| | | 0.2 | 0.8 |
| Dominance | 0.2 | -6.4 | 1.6 |
| dev | 0.8 | 1.6 | -0.4 |

**2.56** sig2(d)



Problem 2b answer

Lesson: All of the genetic variability here is due to non-additive effects. With natural selection on viability and overdominance this is the equilibrium allele frequency.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | + | **pg** | | | | |
| | | **0.8** | **0.2** | | Alpha(i) | | |
| + | **0.8** | **8** | **10** | **8.4** | 0 | 0 | |
| **pg** | **0.2** | **10** | **2** | **8.4** | 0 | 0 | |
| | | **8.4** | **8.4** | **8.4** | 0 | 0 | |
| | Alpha (i) | 0 | 0 | 0 | **0** | | |
| | var | 0 | 0 | 0 | | **0** | sig2(a) |

| | | | |
|---|---|---|---|
| | | 0.8 | 0.2 |
| Dominance | 0.8 | -0.4 | 1.6 |
| dev | 0.2 | 1.6 | -6.4 |

**2.56** sig2(d)

# Muir Lecture 3

Computation of Additive Relationships

$a_{xy}$

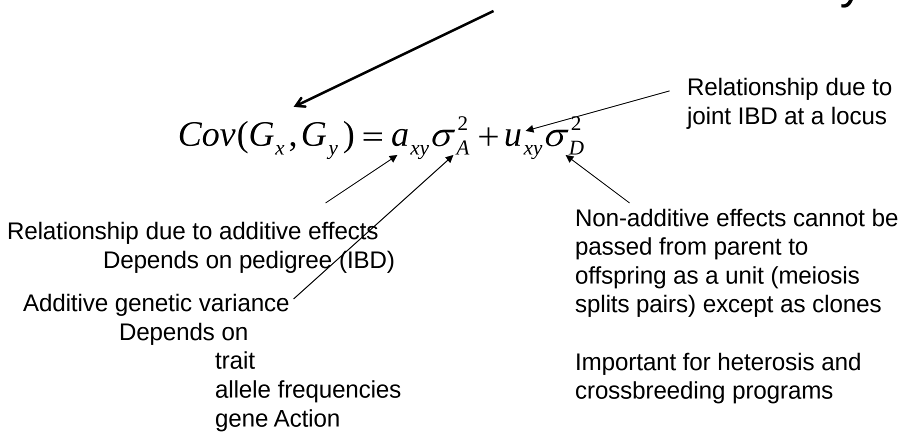
Genetic covariance between relatives

$$Cov(G_x, G_y) = a_{xy}\sigma_A^2 + u_{xy}\sigma_D^2$$

Heritability

$h^2$ and $H^2$

# Covariance Between Relatives x and y



$$Cov(G_x, G_y) = a_{xy}\sigma_A^2 + u_{xy}\sigma_D^2$$

Relationship due to joint IBD at a locus

Relationship due to additive effects
Depends on pedigree (IBD)

Additive genetic variance
Depends on
trait
allele frequencies
gene Action

Non-additive effects cannot be passed from parent to offspring as a unit (meiosis splits pairs) except as clones

Important for heterosis and crossbreeding programs

# Recursive Method to Compute Additive Relationships $a_{xy}$

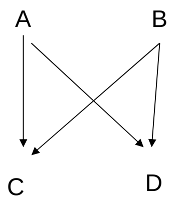
- Setup pedigree table from oldest to youngest (parents must occur before offspring, genes flow in one direction)
- For parents unknown, assume they are unrelated to each other and non-inbred
- Compute following from oldest to youngest

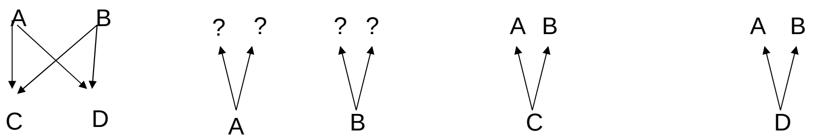
$$a_{ij} = a_{ji} = \frac{\left(a_{i,j_s} + a_{i,j_d}\right)}{2}$$

$$a_{ii} = 1 + \frac{a_{i_s,i_d}}{2}$$

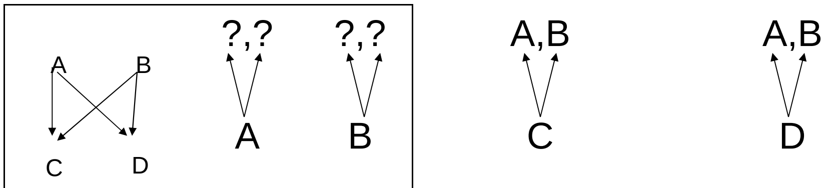
# Example 1

In the following pedigree find the additive relationship between all individuals



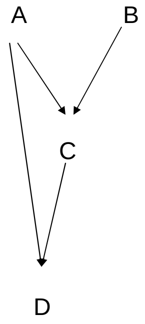


|   | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 0 | ½ ($a_{AA}$+$a_{AB}$)<br>½ (1+0)= ½ | ½ ($a_{AA}$+$a_{AB}$)<br>½ (1+0)= ½ |
| B |   | 1 | ½($a_{BA}$+$a_{BB}$)<br>½ (0+1)= ½ | ½ ($a_{BA}$+$a_{BB}$)<br>½ (0+1)= ½ |
| C |   |   | 1+ ½ $a_{AB}$<br>1+0=1 | ½ ($a_{CA}$+$a_{CB}$)<br>½ (½ +½)= ½ |
| D |   |   |   | 1+ ½ $a_{AB}$<br>1+0=1 |

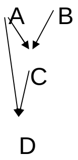


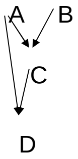
|   | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 0 | ½ | ½ |
| B | 0 | 1 | ½ | ½ |
| C | ½ | ½ | 1 | ½ |
| D | ½ | ½ | ½ | 1 |

# Example 2



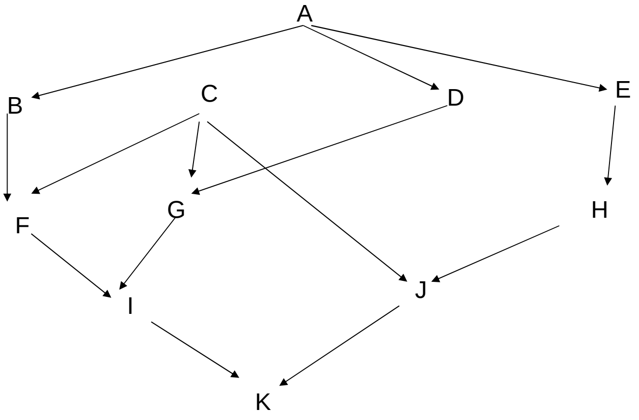
Find all relationships note that D is the result of mating relatives

|  | ?,?<br>A | ?,?<br>B | A,B<br>C | A,C<br>D |
|---|---|---|---|---|
| A | 1 | 0 | ½ $(a_{AA}+a_{AB})$ | ½ $(a_{AA}+a_{AC})$ |
| B | sym | 1 | ½ $(a_{BB}+a_{BA})$ | ½ $(a_{BA}+a_{BC})$ |
| C | sym | | 1+ ½ $a_{AB}$ | ½ $(a_{CC}+a_{AC})$ |
| D | | | sym | 1+ ½ $a_{AC}$ |

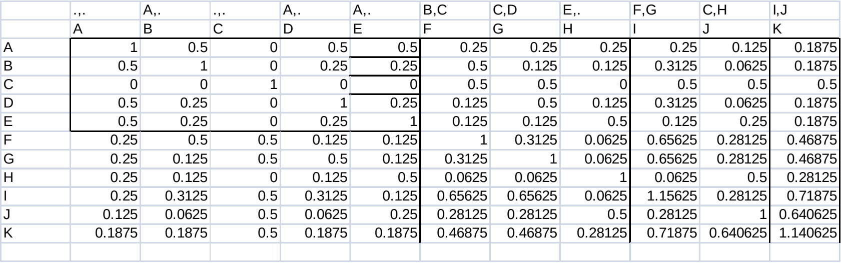
|  | ?,?<br>A | ?,?<br>B | A,B<br>C | A,C<br>D |
|---|---|---|---|---|
| A | 1 | 0 | ½ (1+0)= ½ | ½ (1+ ½)=3/4 |
| B | 0 | 1 | ½ (0+1)= ½ | ½ (0+ ½)=1/4 |
| C | sym | | [1+ ½ (0)] | ½ (1+ ½)=3/4 |
| D | | | sym | [1 + ½(½)] =5/4 |

# Example

Given the following pedigree, find all additive relationship and inbreeding coefficients. Assume animals not given are unrelated and not inbred. Pedigree From Wright (1922) given in Lynch and Walsh p 139.



# A Matrix by Excel



| | .,. | A,. | .,. | A,. | A,. | B,C | C,D | E,. | F,G | C,H | I,J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K |
| A | 1 | 0.5 | 0 | 0.5 | 0.5 | 0.25 | 0.25 | 0.25 | 0.25 | 0.125 | 0.1875 |
| B | 0.5 | 1 | 0 | 0.25 | 0.25 | 0.5 | 0.125 | 0.125 | 0.3125 | 0.0625 | 0.1875 |
| C | 0 | 0 | 1 | 0 | 0 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 0.5 |
| D | 0.5 | 0.25 | 0 | 1 | 0.25 | 0.125 | 0.5 | 0.125 | 0.3125 | 0.0625 | 0.1875 |
| E | 0.5 | 0.25 | 0 | 0.25 | 1 | 0.125 | 0.125 | 0.5 | 0.125 | 0.25 | 0.1875 |
| F | 0.25 | 0.5 | 0.5 | 0.125 | 0.125 | 1 | 0.3125 | 0.0625 | 0.65625 | 0.28125 | 0.46875 |
| G | 0.25 | 0.125 | 0.5 | 0.5 | 0.125 | 0.3125 | 1 | 0.0625 | 0.65625 | 0.28125 | 0.46875 |
| H | 0.25 | 0.125 | 0 | 0.125 | 0.5 | 0.0625 | 0.0625 | 1 | 0.0625 | 0.5 | 0.28125 |
| I | 0.25 | 0.3125 | 0.5 | 0.3125 | 0.125 | 0.65625 | 0.65625 | 0.0625 | 1.15625 | 0.28125 | 0.71875 |
| J | 0.125 | 0.0625 | 0.5 | 0.0625 | 0.25 | 0.28125 | 0.28125 | 0.5 | 0.28125 | 1 | 0.640625 |
| K | 0.1875 | 0.1875 | 0.5 | 0.1875 | 0.1875 | 0.46875 | 0.46875 | 0.28125 | 0.71875 | 0.640625 | 1.140625 |

amatrix LECTURE EXAMPLE by Excel.xlsx

1. $a_{xy}$ is a covariance between individuals due to shared IBD alleles, not a probability or a correlation ($0<a_{xy}<2$).
2. $a_{xy}$=2 x coefficient of relationship (Malecot, 1948)

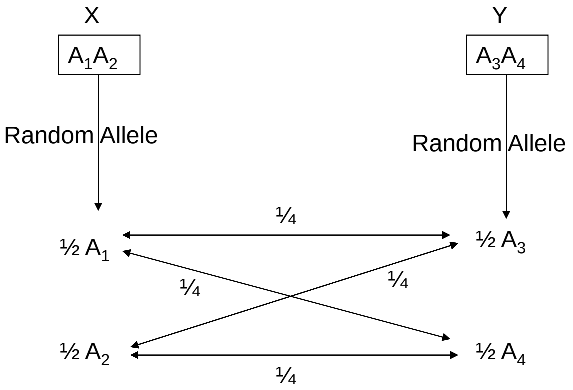
# Coefficient of Relationship

## (Malecot, 1948)

p(randomly chosen allele at a locus in individual x is IBD with a randomly chosen allele at that locus in individual y)

## Probability IBD non-related individuals

coefficient of relationship =p(randomly chosen allele at a locus in individual x is IBD with a randomly chosen allele at that locus in individual y)



X

Y

$A_1A_2$

$A_3A_4$

Random Allele

Random Allele

½ $A_1$

½ $A_3$

½ $A_2$

½ $A_4$

¼

¼

¼

¼

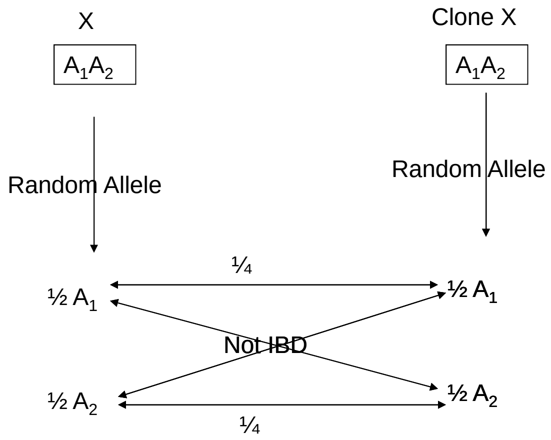
P(IBD between unrelated individuals)=
P(IBD)= 1/4P($A_1$=$A_3$)+ 1/4P($A_1$=$A_4$)+ 1/4P($A_2$=$A_3$)+ 1/4P($A_2$=$A_4$)=0
$a_{xy}$=2P(IBD)=0
this is why off diagonal element are 0 for non-related individuals

## What is the coefficient of relationship of an individual with itself?

- This is the same as asking what is the coefficient of relationship between the individual and its clone
  - First consider a non-inbred clone
  - From this find the additive relationship ($a_{xy}$) between non-inbred clones

## Probability IBD **between** non-inbred Clones

Non inbred means that the P(IBD) **within** and individual is 0 or $P(A_1=A_2)=0$



X

Clone X

$A_1A_2$

$A_1A_2$

Random Allele

Random Allele

¼

½ $A_1$

½ $A_1$

Not IBD

½ $A_2$

½ $A_2$

¼

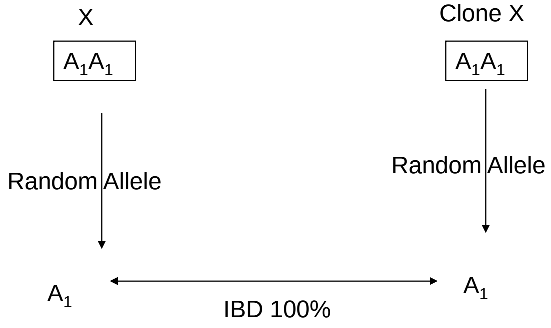
P(IBD between Non-Inbred Clones)=1/2
$a_{xy}$=2P(IBD)=1
This is why the diagonal elements in A are 1 for non-inbred individuals

## Next: Consider the other extreme: What is the coefficient of relationship between completely inbred clones?

Convert this probability to the additive relationship ($a_{xy}$) between perfectly inbred clones

$a_{xy}$=2P(IBD)

## Completely inbred means both two alleles at a locus are IBD



X

$A_1A_1$

Clone X

$A_1A_1$

Random Allele

Random Allele

$A_1$

$A_1$

IBD 100%

P(IBD Perfectly Inbred Clones)=1
$a_{xy}$=2P(IBD)=2 which is the maximum value for a diagonal element in the A matrix
Inbreeding is why some of the diagonal element are > 1 and is a method to estimate the inbreeding coefficient $F_x=a_{xx}-1$. In this example $F_x=2-1=1$

# Additive Relationship Matrix (A)

| | .,. | A,. | .,. | A,. | A,. | B,C | C,D | E,. | F,G | C,H | I,J |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K |
| A | 1 | 0.5 | 0 | 0.5 | 0.5 | 0.25 | 0.25 | 0.25 | 0.25 | 0.125 | 0.1875 |
| B | 0.5 | 1 | 0 | 0.25 | 0.25 | 0.5 | 0.125 | 0.125 | 0.3125 | 0.0625 | 0.1875 |
| C | 0 | 0 | 1 | 0 | 0 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 0.5 |
| D | 0.5 | 0.25 | 0 | 1 | 0.25 | 0.125 | 0.5 | 0.125 | 0.3125 | 0.0625 | 0.1875 |
| E | 0.5 | 0.25 | 0 | 0.25 | 1 | 0.125 | 0.125 | 0.5 | 0.125 | 0.25 | 0.1875 |
| F | 0.25 | 0.5 | 0.5 | 0.125 | 0.125 | 1 | 0.3125 | 0.0625 | 0.65625 | 0.28125 | 0.46875 |
| G | 0.25 | 0.125 | 0.5 | 0.5 | 0.125 | 0.3125 | 1 | 0.0625 | 0.65625 | 0.28125 | 0.46875 |
| H | 0.25 | 0.125 | 0 | 0.125 | 0.5 | 0.0625 | 0.0625 | 1 | 0.0625 | 0.5 | 0.28125 |
| I | 0.25 | 0.3125 | 0.5 | 0.3125 | 0.125 | 0.65625 | 0.65625 | 0.0625 | 1.15625 | 0.28125 | 0.71875 |
| J | 0.125 | 0.0625 | 0.5 | 0.0625 | 0.25 | 0.28125 | 0.28125 | 0.5 | 0.28125 | 1 | 0.640625 |
| K | 0.1875 | 0.1875 | 0.5 | 0.1875 | 0.1875 | 0.46875 | 0.46875 | 0.28125 | 0.71875 | 0.640625 | 1.140625 |

1. What is the inbreeding coefficient for individual K?
2. $F_k$=1.14-1=.14
3. What is the additive relationship between individuals J and K
4. $a_{jk}$=.64
5. What is the additive genetic covariance between individuals J and K for trait T1? for trait T2?

$$Cov(G_x, G_y) = a_{xy}\sigma_A^2$$

$$Cov_{T1}(G_x, G_y) = a_{xy}\sigma_{A_{T1}}^2 \ = .64\sigma_{A_{T1}}^2$$

$$Cov_{T2}(G_x, G_y) = a_{xy}\sigma_{A_{T2}}^2 \ = .64\sigma_{A_{T2}}^2$$

## Covariance Between Relatives Common Environmental Causes $E_c$

Common Farm

Common Maternal

Common Cage or Pen

# Environmental Effects

Common Environmental Effects Ec
Can contribute to resemblance of relatives
Can be corrected for using mixed models allowing for correlated residuals

## Summary Resemblance Among Relatives for a Given Trait

| Relative Pair | *Cov* |
|---|---|
| Any Set of Relatives | $Cov(G_x, G_y) = a_{xy}\sigma_A^2 + \sigma_{E_c(xy)}$ |
| Parent-Offspring | $\frac{1}{2}\sigma_A^2 + \sigma_{E_c(P.O)}$ |
| Half-Sib | $\frac{1}{4}\sigma_A^2 + \sigma_{E_c(HS)}$ |
| Full-Sib | $\frac{1}{2}\sigma_A^2 + \sigma_{E_c(FS)}$ |

## Heritability

Broad Sense: Proportion of the phenotypic variation due to genetic causes

$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2}$$

Useful to determine to what extent genetics vs environment impact a trait

Narrow Sense: Proportion of the phenotypic variation due to additive genetic effects

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

Useful to determine to what extent directional selection can improve a trait

# Examples of Heritabilities

| Organism | Trait | $h^2$ |
|---|---|---|
| Humans | | |
| | Height | 0.85 |
| | Serum IG | 0.45 |
| Pigs | | |
| | Back-fat thickness | 0.70 |
| | Daily weight-gain | 0.30 |
| | Litter size | 0.05 |
| Fruit flies | | |
| | Abdominal bristles | 0.50 |
| | Body size | 0.40 |
| | Ovary size | 0.30 |
| | Egg production | 0.20 |

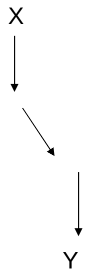
# Estimation of Narrow Sense Heritability

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2} = \frac{\sigma_A^2}{\sigma_G^2 + \sigma_\varepsilon^2}$$

- Three Approaches
  - Regression
  - Analysis of Variance
  - Maximum Likelihood (ML), REML, Gibbs Estimation of Variance Components
- All based on resemblance between relatives

# Regression

## Ancestor-Descendent Pairs



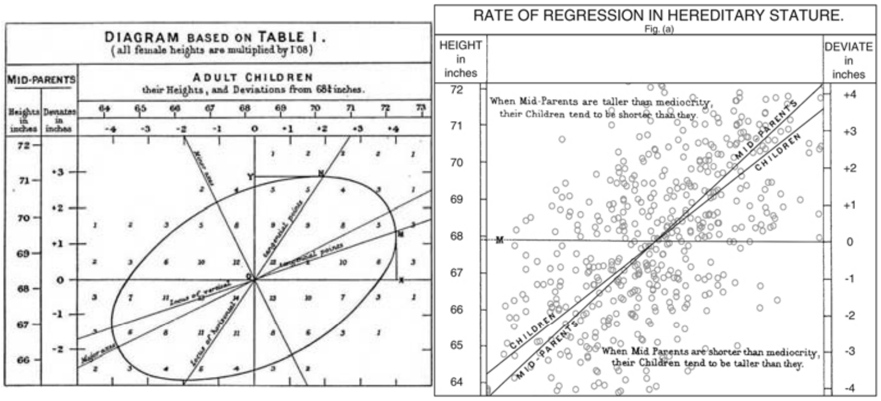
$$Cov(G_x, G_y) = a_{xy}\sigma^2_A$$

Parent-offspring
Grand-parent-grand son
Great Uncle-Nephew
Etc

# Regression

- Francis Galton a half-cousin to C. Darwin established the principle of what he termed "regression to mediocrity.“
  - studied the inheritance of height in humans
  - noticed that extremely tall fathers tended to have sons shorter than themselves, and extremely short fathers tended to have sons taller than themselves.
  - The offspring seemed to regress to the median, or "mediocrity.”

# regression to mediocrity



*European Journal of Human Genetics* (2009) **17,** 1070–107

The regression coefficient later become known as the heritability

# Ancestor-Descendent Pairs

- First Case
  - All pairs have same additive relationship
    - Parent-Offspring
      - $a_{xy}$=1/2
    - Grandparent-Grandchildren
      - $a_{xy}$=1/4
  - Assume Additive Genetic Variance Only
  - No Environmental Covariance

| Pair(i) | Ancestor (X) | Descendent (Y) |
|---|---|---|
| 1 | $X_1$ | $Y_1$ |
| 2 | $X_2$ | $Y_2$ |
| . | | |
| i | $X_i$ | $Y_i$ |
| n | $X_n$ | $Y_n$ |

Expected covariance

$$Cov(G_x, G_y) = a_{xy}\sigma_A^2$$

Estimated Covariance

$$Co\hat{v}(G_x, G_y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

Set expected covariance= estimated and solve for additive variance component

$$Cov(G_x, G_y) = Co\hat{v}(G_x, G_y)$$

$$a_{xy}\sigma_A^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

$$\hat{\sigma}_A^2 = \left(\frac{1}{a_{xy}}\right)\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

Heritability is the ratio of additive to phenotypic variance

$$\hat{h}^2 = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_X^2}$$

Phenotypic Variance= $\hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$

$$\hat{h}^2 = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_X^2} = \left(\frac{1}{a_{xy}}\right)\frac{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}}{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \left(\frac{1}{a_{xy}}\right)\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

## Regression on One Parent: Example Butterfat (kg)

| Parent<br>Dam<br>(X) | Offspring<br>Daughter<br>(Y) |
|---|---|
| 150 | 132 |
| 102 | 122 |
| 129 | 104 |
| 127 | 103 |
| 149 | 112 |
| 133 | 130 |
| 164 | 140 |
| 150 | 148 |
| 124 | 120 |
| 141 | 168 |

$$\sum X_i = 1{,}369 \qquad \sum Y_i = 1{,}279$$

$$\sum X_i^2 = 190{,}217 \qquad \sum X_i Y_i = 176{,}447$$

$$\hat{\sigma}_X^2 = \frac{190{,}217 - \frac{(1{,}369)^2}{10}}{10-1} = 311.2$$

$$Co\hat{v}(G_x, G_y) = \frac{176{,}447 - \frac{(1{,}369)(1{,}279)}{10}}{10-1} = 150.2$$

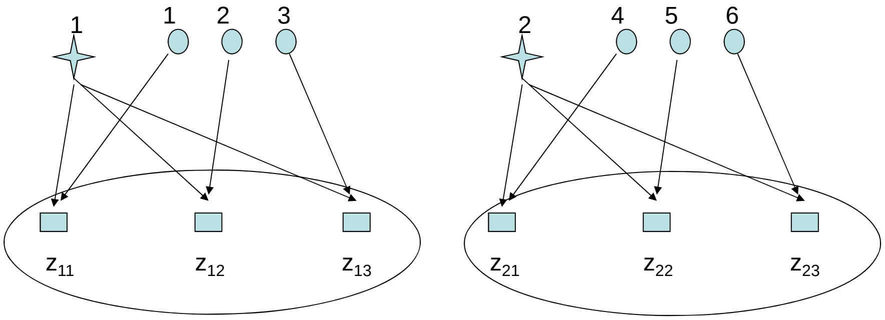
$$a_{xy} = \tfrac{1}{2}$$

$$\therefore$$

$$\sigma_A^2 = \left(\frac{1}{\frac{1}{2}}\right)(150.2) = 300.4$$

$$h^2 = \frac{300.4}{311.2} = .96$$

## Co-lateral Data (Sibs, Cousins, etc.)



Phenotypic Data Only on Sibs

$a_{xy}$=1/4 among half sibs, $a_{xy}$=0 otherwise

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

# Quantifying an association due to group ownership

- Groups
  - Environmental
    - Fields
    - Pens
    - Location
  - Genetic
    - Family
    - Lineage

•Note color variation in field
•Those in same plot share same drainage and nutrients
•Those in same plot have **correlated** yields due to these factors
•How do you measure that correlation?
•How do you control for that variation

# The extent to which observations are correlated due to group ownership is the **intra-class correlation**

$$r_{Ic} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Becomes large when between group differences are large

Becomes small when within group differences are small

| Factor | df | MS | E(MS) |
|---|---|---|---|
| Among Groups | b-1 | $MS_b = SS_b/(b-1)$ | $\sigma_w^2 + n\sigma_b^2$ |
| Within group | b(n-1) | $MS_w = SS_w/b(n-1)$ | $\sigma_w^2$ |

# Groups as Related Individuals (Families)

$$r_{Ic} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

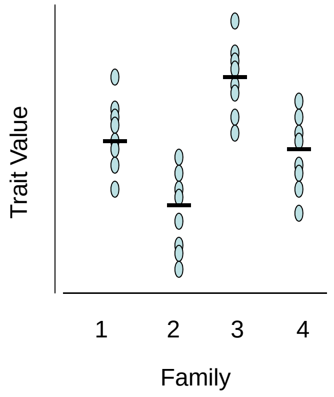
Becomes large when between family differences are large

Becomes small when within family differences are small

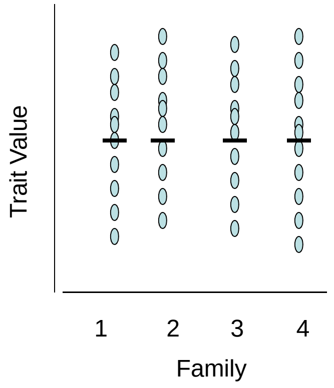
How does the correlation relate to heritability?

Concept : If a trait is heritable then individuals within a family should be more similar (concordance) than individuals between families.
Below are two traits from the **same families**, which trait has the higher heritability?

Trait 1 (e.g. height)



Trait 2 (e.g. %fat)



Var(B)=2.5 Var(W)=.5

Vp=Var(B)+Var(W)=3

$r_{Ic} = \frac{2.5}{2.5+.5} = .83$

Var(B)=0 Var(W)=3

Vp=Var(B)+Var(W)=3

$r_{Ic} = \frac{0}{0+3} = 0$

Trait 2 is lowly heritable

- The phenotypic covariance among members of the same group equals the variance between groups

Note the i subscript is the same

$$
\begin{aligned}
&Cov(within\_\ family) = \sigma(z_{ij}, z_{ik}) \\
&= \sigma\left[\left(\mu + b_i + w_{ij}\right), \left(\mu + b_i + w_{ik}\right)\right] \\
&= \sigma\left(b_i, b_i\right) + \sigma\left(b_i, w_{ik}\right) + \sigma\left(w_{ij}, b_i\right) + \sigma\left(w_{ij}, w_{ik}\right) \\
&= \sigma_b^2
\end{aligned}
$$

# The Among Family Variance Component

Variance due to Among Family differences= Covariance within a Family

$$\sigma_b^2 = \sigma_{wf}$$

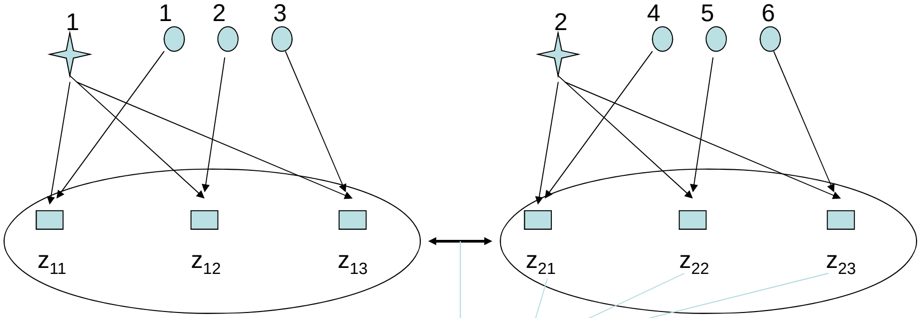
$$\sigma_{wf} = a_{xy}\sigma_A^2$$

$$a_{xy} = \text{genetic relationship among individuals within a family}$$

If there is no covariance with a group, then the individuals in that group are not correlated. Note that the within group covariance can be zero for 2 reasons: 1) the members are not related, or 2) the trait is not influenced by alleles

$r_g = 0$ $\sigma_A^2 = 0$

# Half Sib



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

$$\sigma_z^2 = \sigma_b^2 + \sigma_{w(HS)}^2$$

variation among families (circles) is due to differences between half sib families=covariance within half-sib families

within group variation is not just a measure of environmental variation, but also includes genetic variation not accounted for by the half-sib covariance

$$\sigma_b^2 = \sigma_{HS} = \tfrac{1}{4}\sigma_A^2$$

$$\sigma_{w(HS)}^2 = \tfrac{3}{4}\sigma_A^2 + \sigma_\varepsilon^2$$

# ANOVA Computational Formulas

| Factor | df | MS | E(MS) |
|---|---|---|---|
| Among Families | s-1 | $MS_b = SS_b/(s-1)$ | $\sigma_{w(HS)}^2 + d\sigma_b^2$ |
| Within Families | s(d-1) | $MS_w = SS_w/d(s-1)$ | $\sigma_{w(HS)}^2$ |

# Method of Moments

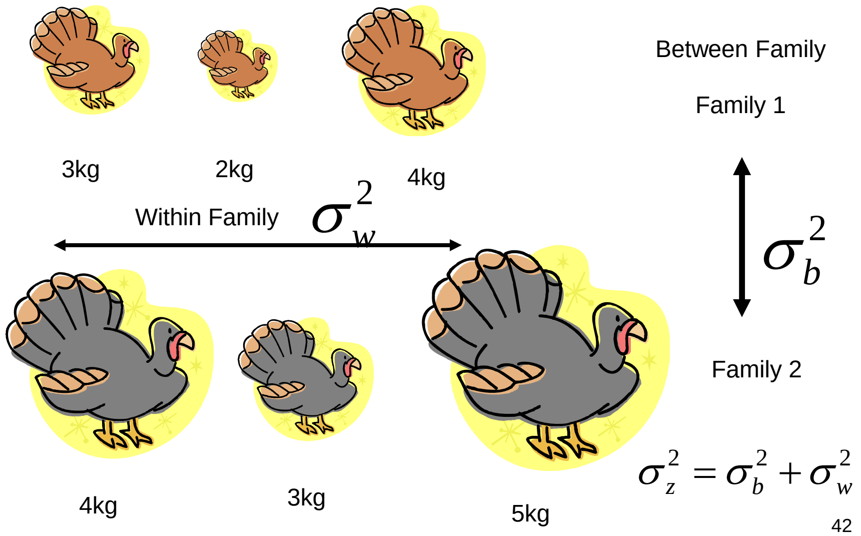
Set Expected mean squares equal to estimated mean squares and solve

$$\hat{\sigma}_b^2 = \frac{MS_b - MS_w}{d} \qquad \hat{\sigma}_{w(HS)}^2 = MS_w$$

$$\sigma_b^2 = \sigma_{HS} = \tfrac{1}{4}\sigma_A^2 \qquad \sigma_{w(HS)}^2 = \tfrac{3}{4}\sigma_A^2 + \sigma_\varepsilon^2$$

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \cong \frac{4\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_{w(HS)}^2}$$

# Example (Half-Sib Families)



Between Family

Family 1

3kg 2kg 4kg

Within Family $\sigma_w^2$

$\sigma_b^2$

Family 2

4kg 3kg 5kg

$\sigma_z^2 = \sigma_b^2 + \sigma_w^2$

# Turkey Example

| Source | df | ss | ms | E(ms) |
|---|---|---|---|---|
| Among Family | 1 | 1.5 | 1.5 | $\sigma^2_{w(HS)} + 3\sigma^2_b$ |
| Within Family | 4 | 4 | 1 | $\sigma^2_{w(HS)}$ |

$$\hat{\sigma}^2_b = (1.5-1)/3 = .167 \qquad \hat{\sigma}^2_{w(HS)} = 1$$

$$h^2 \cong \frac{4(.167)}{1.167} = .566$$

s.e.=formulas given in notes

Typically very large

# Bias in Estimates

- The resemblance between relatives can be impacted by non-additive effects (dominance and epistasis)
  - Non-additive effects can occur when relatives share more than one IBD allele, i.e. dominance is due to shared IBD alleles within a locus and epistasis is due to shared IBD alleles between loci.
  - The parent-offspring resemblance is least biased for estimating narrow sense heritability ($h^2$), but maybe inflated by AxA epistasis.
  - The resemblance between clones is most biased, because it is inflated by all non-additive effects, but is the best estimator of broad sense heritability ($H^2$)

# Discussion

If for a given trait the broad sense and narrow sense heritabilites are as follows, in each which would be more effective at improving the trait, a breeding program, improving management, neither or both?

- $H^2$=.9, $h^2$=.1
- $H^2$=.1, $h^2$=.1
- $H^2$=.9, $h^2$=.9

# Answer

- $H^2$=.9, $h^2$=.1
  - **Neither**: The high broad sense heritability indicates that there is little environmental effects so management does not influence. Low narrow sense indicates that selective breeding will not be very useful. (cloning will work, but will not improve the trait, only reproduce what is there. Development of inbred lines will produce similar effect as cloning)
- $H^2$=.1, $h^2$=.1
  - **Management:** low broad sense indicates primarily impacted by environment effects, genes are not important. Hence breeding will not work (note: most reproductive traits fall in this category)
- $H^2$=.9, $h^2$=.9
  - **Breeding**: High narrow sense indicates that selective breeding will be very effective and environment does not have much effect on the trait (note: human height falls in this category)

# Problem set 3

- 1. Continuing from the previous problem set, Falconer (1981) reported a partially dominant gene in the mouse called *pg* “pygmy.” At six weeks of age, they produce the following average weight phenotypes in grams (the actual weight of the heterozygote was 12, but it was reduced to 10 for this example):
- + / + : 14, + / *pg* : 10, *pg* / *pg* : 6
- If the population of mice is randomly mating with *p*+ = 0.8, $q^{pg}$ = 0.2
- Assuming no Environmental Effects, what are the narrow and broad sense heritabilities for this trait?
- If the environmental variance is 2, what is the narrow and broad sense heritability?

# Answer 1

A. Assuming no Environmental Effects, what are the narrow and broad sense heritabilities for this trait?

$$\sigma_A^2 = 2.51$$

$$\sigma_D^2 = .4096$$

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2 = 2.51 + .409 + 0 = 2.919$$

$$h^2 = \frac{2.51}{2.919} = .86$$

$$H^2 = \frac{2.919}{2.919} = 1$$

B. Assuming Environmental Variance=2, what are the narrow and broad sense heritabilities for this trait?

$$\sigma_A^2 = 2.51$$

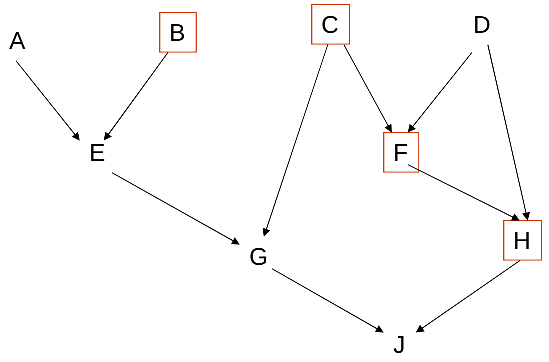
$$\sigma_D^2 = .4096$$

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2 = 2.51 + .409 + 2 = 4.919$$

$$h^2 = \frac{2.51}{4.919} = .51$$

$$H^2 = \frac{2.919}{4.919} = .59$$

## 2. Find the additive relationship matrix for the following pedigree



## Answer

```
A={1     0      0     0     0.5  0    0.25   0      0.125,
  0     1      0     0     0.5  0    0.25   0      0.125,
  0     0      1     0     0    0.5  0.5    0.25   0.375,
  0     0      0     1     0    0.5  0      0.75   0.375,
  0.5   0.5    0     0     1    0    0.5    0      0.25,
  0     0      0.5   0.5   0    1    0.25   0.75   0.5,
  0.25  0.25   0.5   0     0.5  0.25 1      0.125  0.5625,
  0     0      0.25  0.75  0    0.75 0.125  1.25   0.6875,
  0.125 0.125  0.375 0.375 0.25 0.5  0.5625 0.6875 1.0625};
```

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

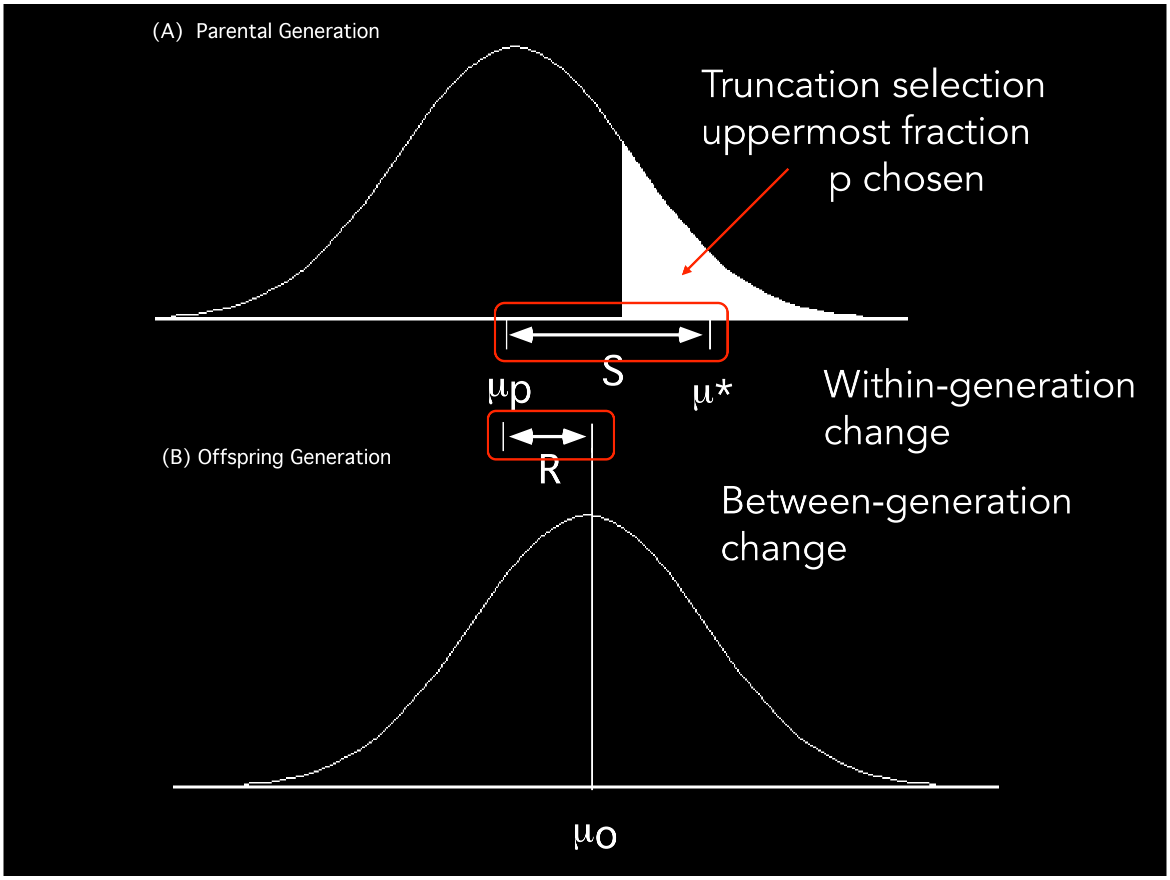
Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 13 – 15 July 2015

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



(A) Parental Generation

Truncation selection uppermost fraction p chosen

Within-generation change

(B) Offspring Generation

Between-generation change

# 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

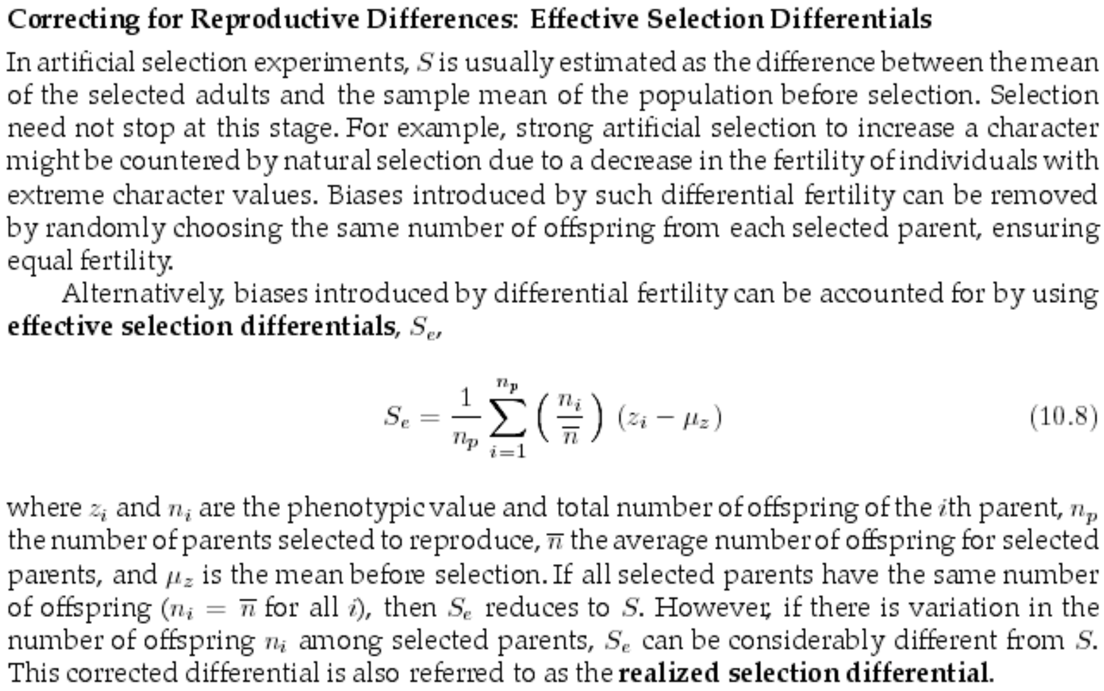
- Recall that $S = (S_f + S_m)/2$
- An issue that arises in plant breeding is pollen control --- is the pollen from plants that have also been selected?
- Not the case for traits (i.e., yield) scored after pollination. In this case, $S_m = 0$, so response only half that with pollen control
- Tradeoff: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

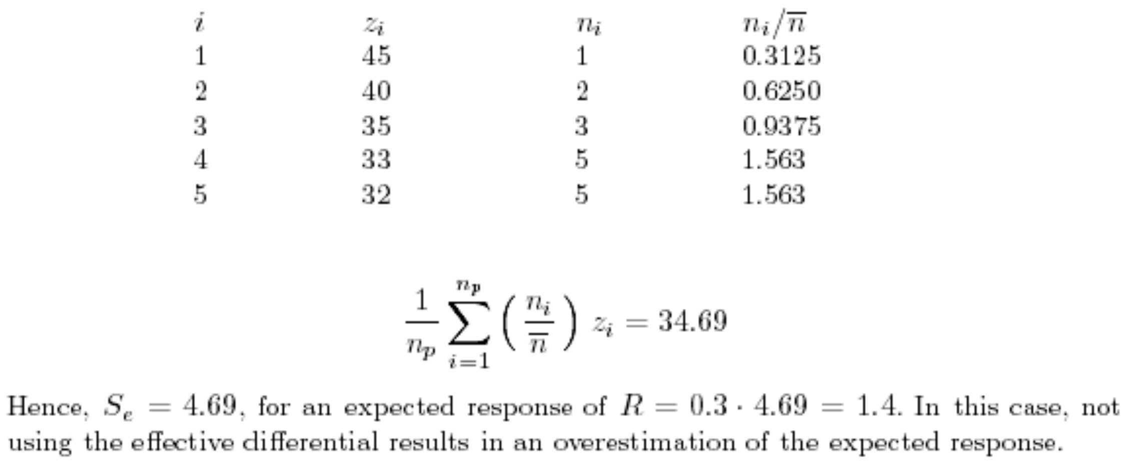
In artificial selection experiments, $S$ is usually estimated as the difference between the mean of the selected adults and the sample mean of the population before selection. Selection need not stop at this stage. For example, strong artificial selection to increase a character might be countered by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

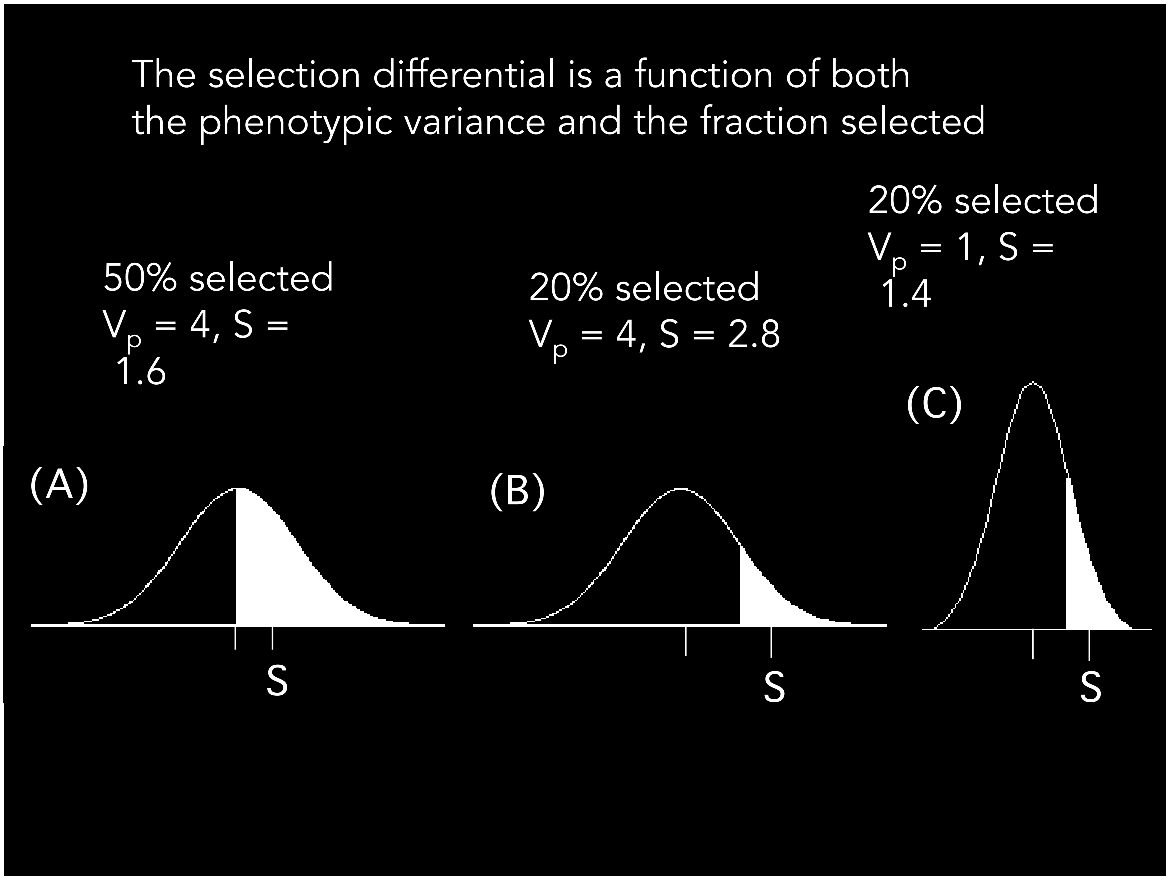
$$\frac{1}{n_p}\sum_{i=1}^{n_p}\left(\frac{n_i}{\overline{n}}\right)\, z_i = 34.69$$

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

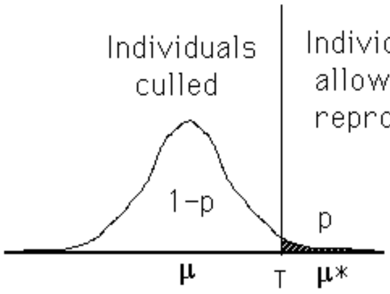
As the previous example shows, populations with the same selection differential (S) may experience very different amounts of selection

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

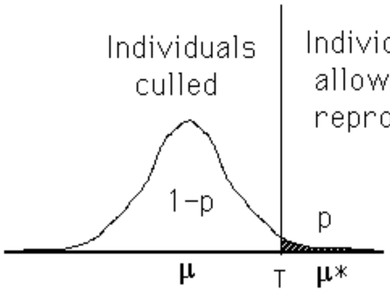
$$i = \frac{S}{\sqrt{V_P}} = \frac{S}{\sigma_p}$$

# 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



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

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

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, $h = r_{PA}$

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

Response = Intensity * Accuracy * spread in Va

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

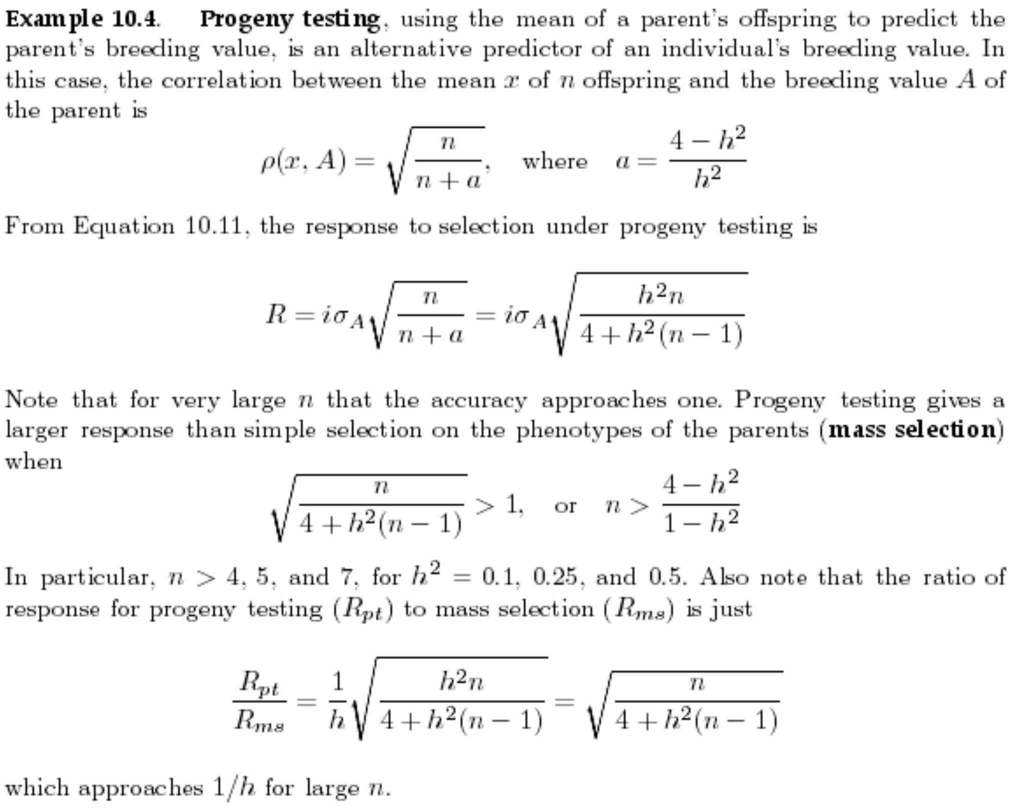
# Accuracy of selection

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 |

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

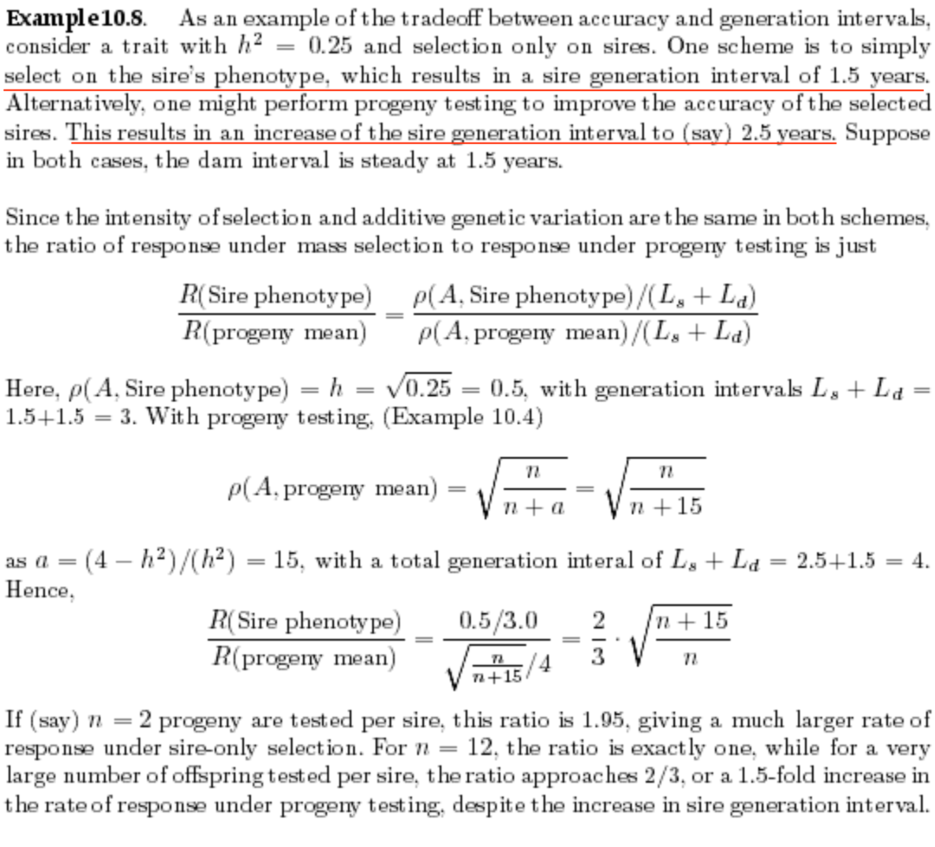
$$L_d = \frac{2 \cdot 400 + 3 \cdot 600 + 4 \cdot 100 + 5 \cdot 40}{400 + 600 + 100 + 40} = 2.81$$

# 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)

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

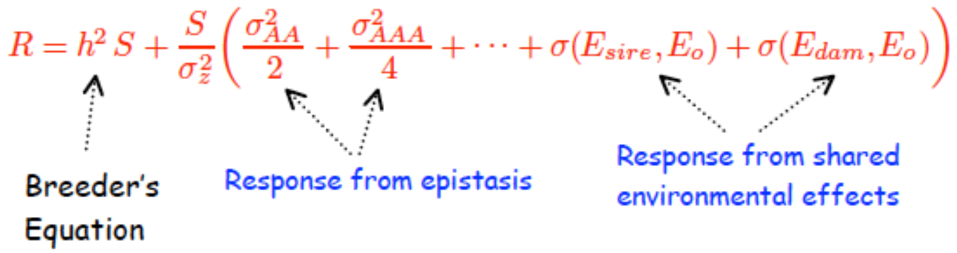
as $a = (4 - h^2)/(h^2) = 15$, with a total generation interal of $L_s + L_d = 2.5{+}1.5 = 4$. Hence,

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

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
τ generations of no selection (but recombination)

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

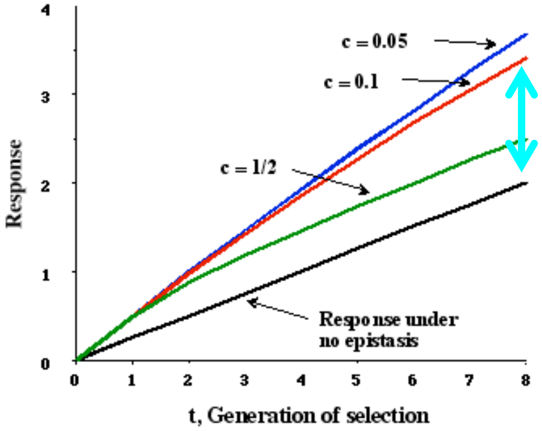
$R_{AA}$ has a limiting value given by

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

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

- Autotraploids 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

$$\sigma(z_p, z_o) = \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{6},$$

Single-generation response

$$R = S\left(h^2 + \frac{\sigma_D^2}{3\sigma_z^2}\right)$$

Response to t generations of selection with constant selection differential S

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

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

Response remaining after t generations of selection followed by 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,

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

**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

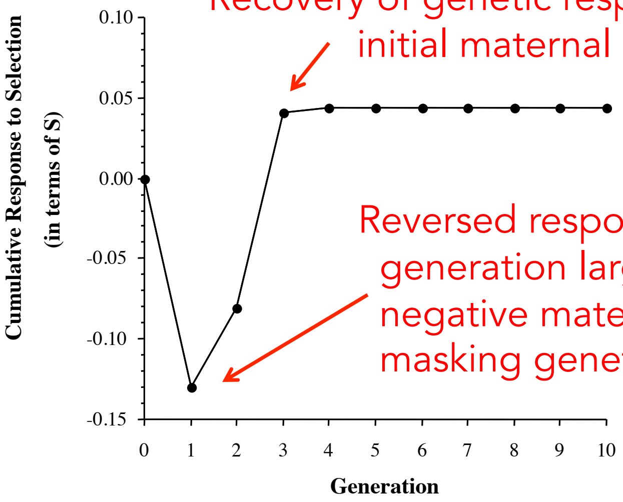
With maternal effects, a covariance between BV and maternal effect arises, with $\sigma_{A,M} = m\,\sigma^2_A\,/\,(2 - m)$

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

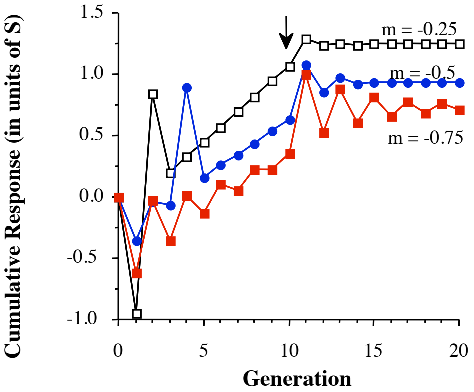
$h^2 = 0.11$, $m = -0.13$ (litter size in mice)



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

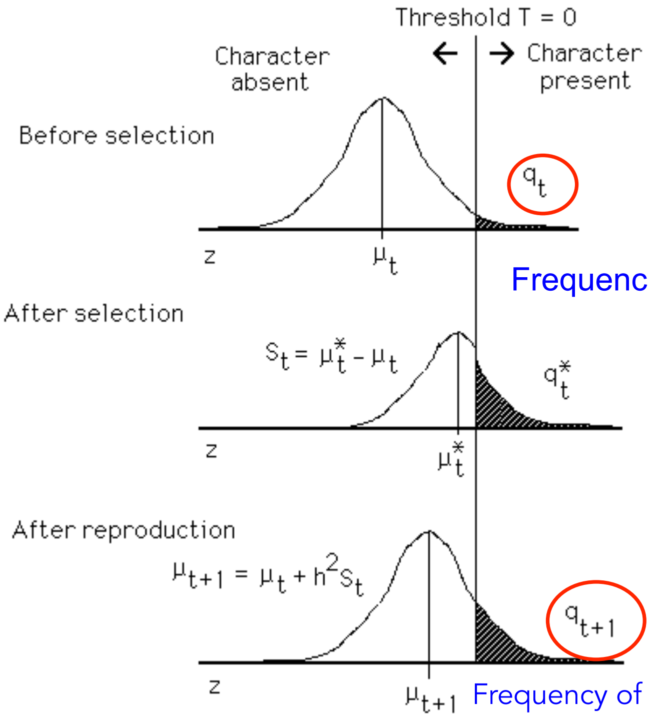
Response on a binary trait is a special case of response on a continuous trait

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.

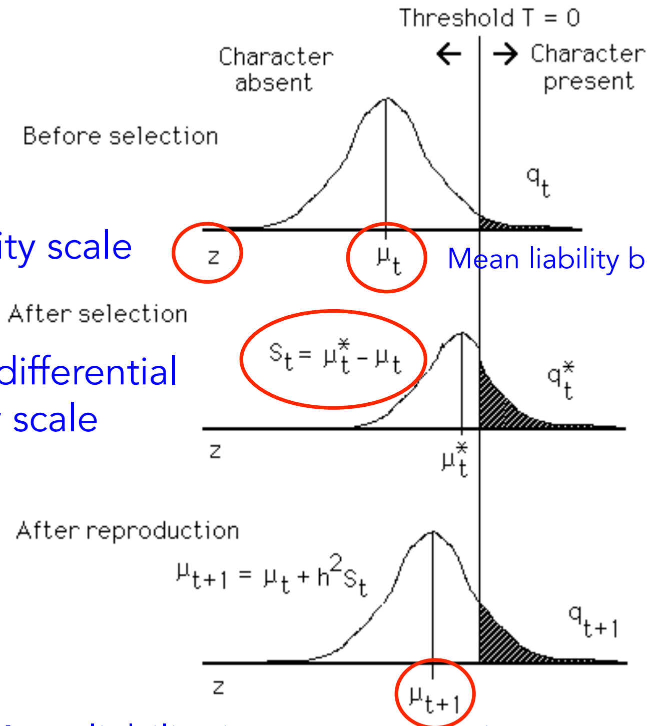


Frequency of trait

Frequency of character state on in next generation

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

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

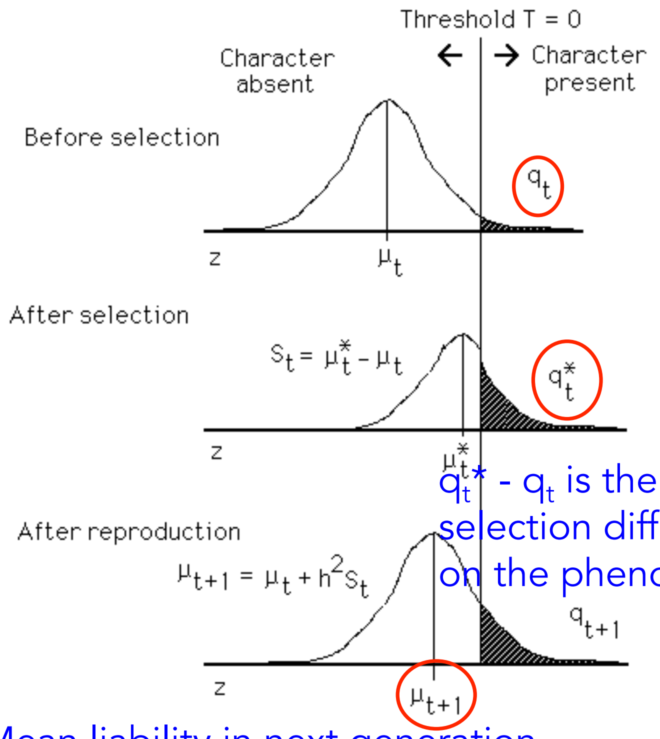


Liability scale

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 $z \geq 0$

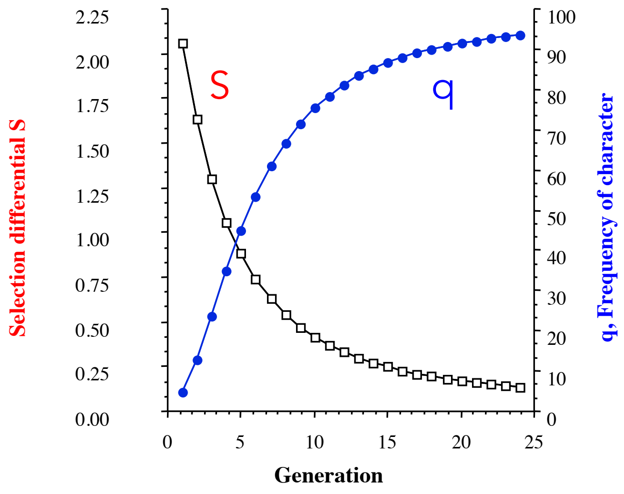
When z is normally distributed, this reduces to

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

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:

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

8 great-grand parents
$S_0$ is there selection differential
$\beta_{3,0}$ is the regression coefficient for an offspring at time 3 on a great-grandparent From time 0

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

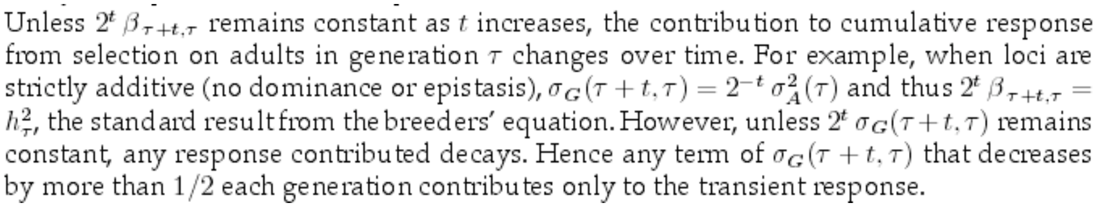
2 parents

# 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 $2^t\,\beta_{\tau+t,\tau} = h_\tau^2$, the standard result from the breeders' equation. However, unless $2^t\,\sigma_G(\tau+t,\tau)$ remains constant, any response contributed decays. Hence any term of $\sigma_G(\tau+t,\tau)$ that decreases by more than $1/2$ each generation contributes only to the transient response.

## Changes in the Variance under Selection

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

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

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

Under linkage equilibrium, 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

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

Recombination Selection

In terms of change in d,

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

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

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

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

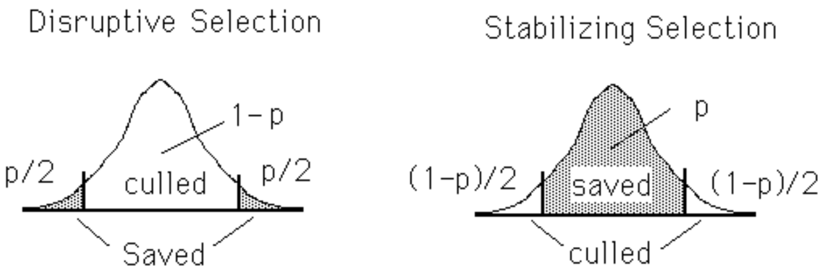
Bulmer equation simplifies to

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

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

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

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

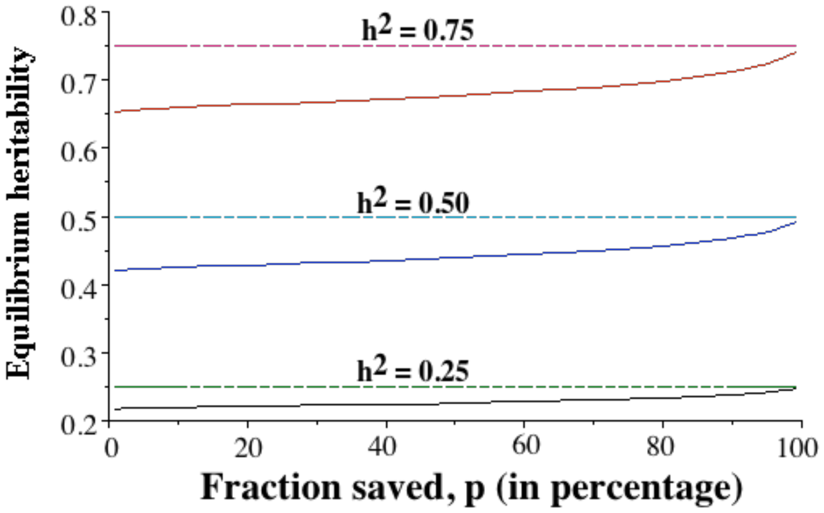
**Stabilizing Truncation Selection:** Middle fraction $p$ of the distribution saved

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

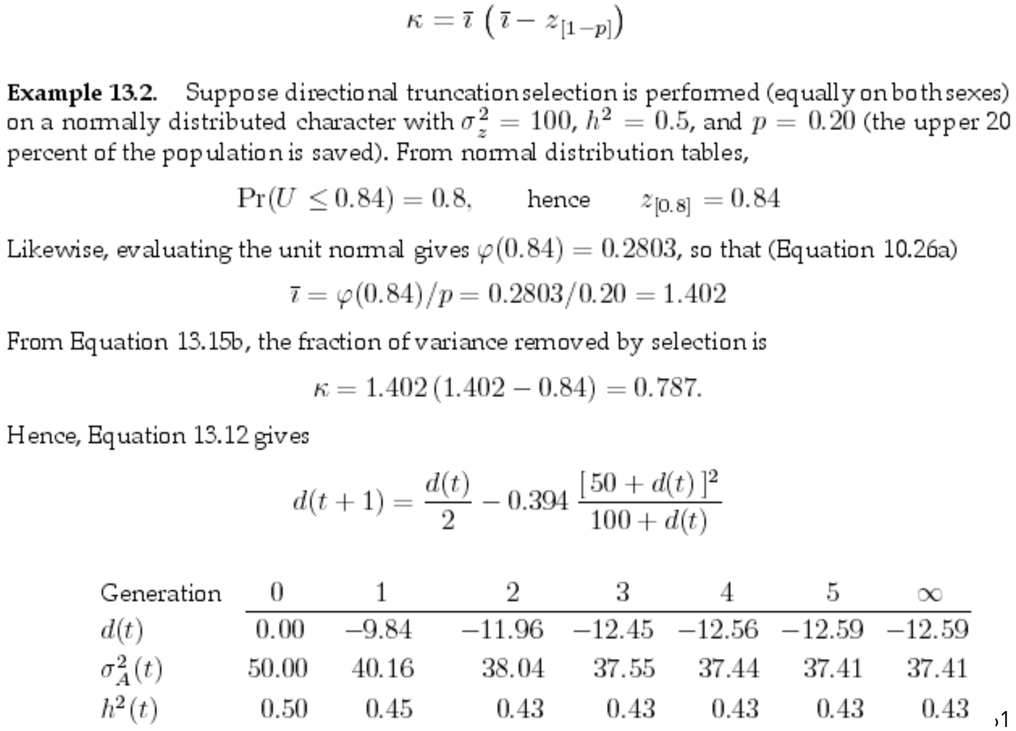
**DisruptiveTruncation Selection:** Uppermost and lowermost $p/2$ saved

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

## 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,

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

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

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

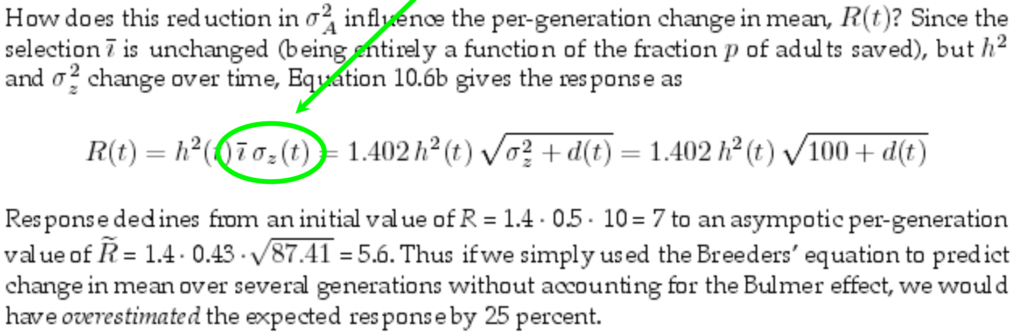
Hence, Equation 13.12 gives

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

| 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

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

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

# Lecture 5 Inbreeding and Crossbreeding

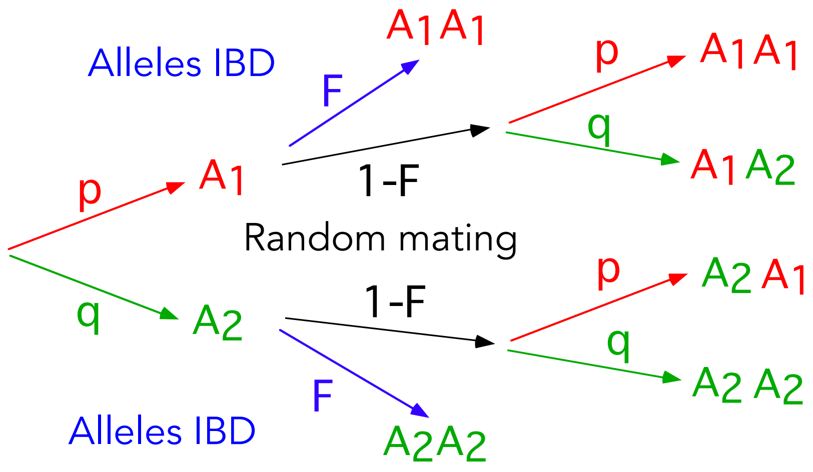
Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 13 – 15 July 2015

# 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 | $(1-F)p^2$ | $p^2 + Fpq$ |
| $A_2A_1$ | 0 | (1-F)2pq | (1-F)2pq |
| $A_2A_2$ | Fq | $(1-F)q^2$ | $q^2 + Fpq$ |

# Changes in the mean under inbreeding

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

freq($A_1$) = p, 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

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

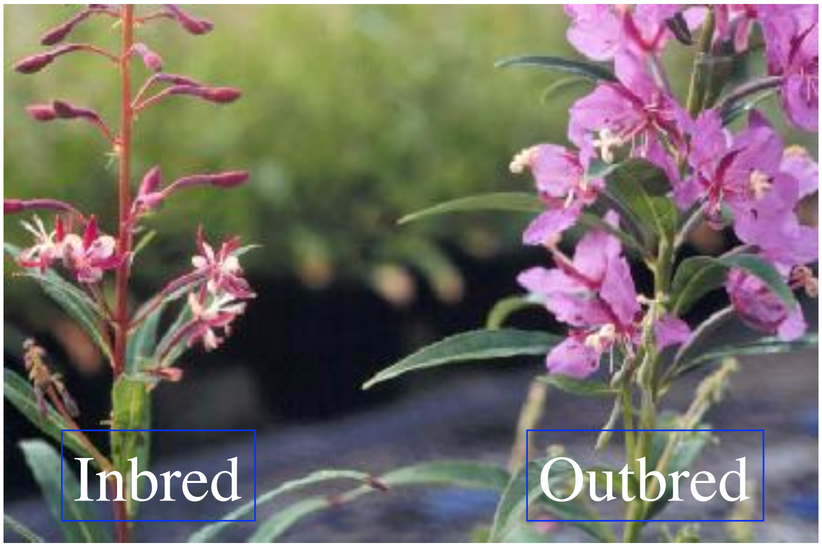
**For k loci, the change in mean is**

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

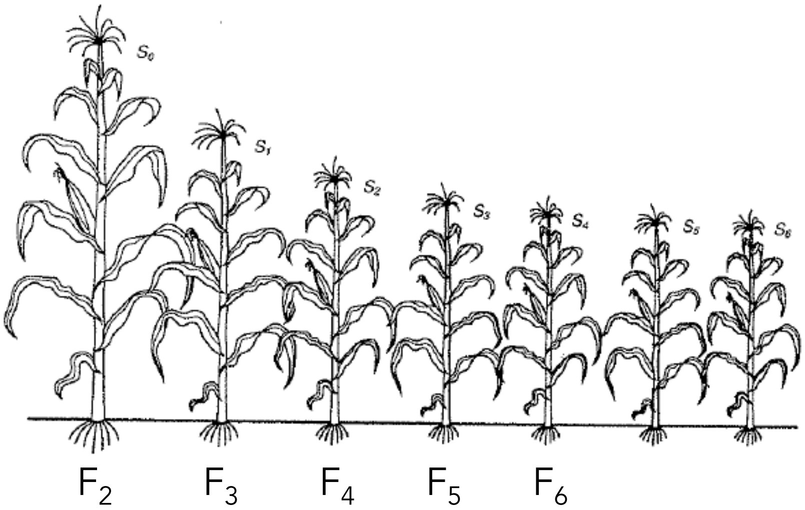
**Here B is the reduction in mean under complete inbreeding (F=1) , where** $B = 2 \sum p_i\, q_i\, d_i$

- 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.

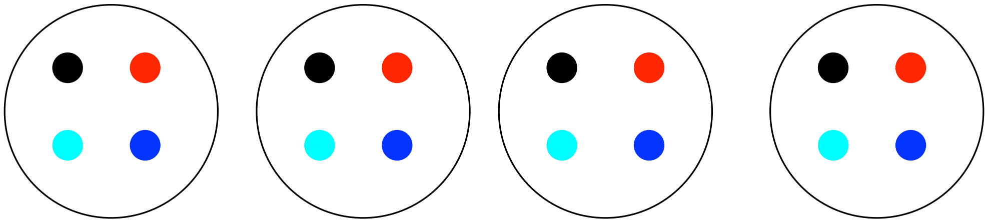
# Inbred depression in largely selfing lineages

- 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 and inbreeding depression

# Variance Changes Under Inbreeding

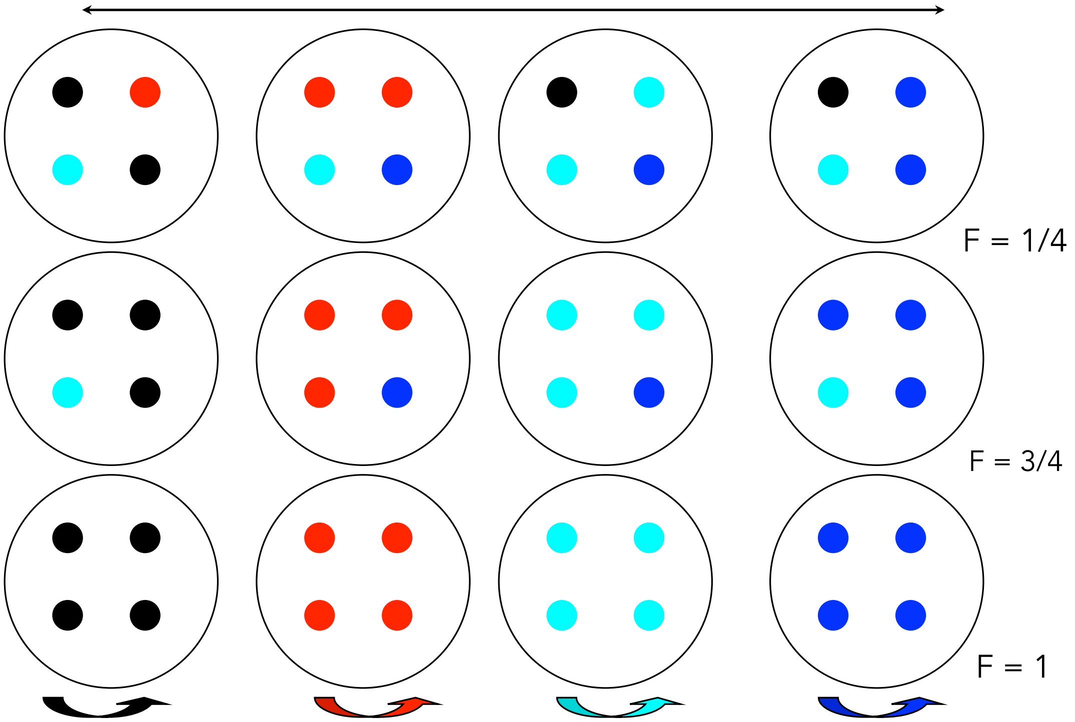
Inbreeding reduces variation within each population

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

$$z_{AB} = z + \frac{g^{I}_{A} + g^{I}_{B}}{2} + g^{M}_{B} + g^{M\,0}_{B} + h^{I}_{AB}$$

Maternal and grandmaternal effects from the B mothers

$$z_{AB} = z + \frac{g^{I}_{A} + g^{I}_{B}}{2} + g^{M}_{B} + g^{M\,0}_{B} + h^{I}_{AB}$$

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

$$z_{BA} = z + \frac{g^{I}_{A} + g^{I}_{B}}{2} + g^{M}_{A} + g^{M\,0}_{A} + h^{I}_{AB}$$

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

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

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

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

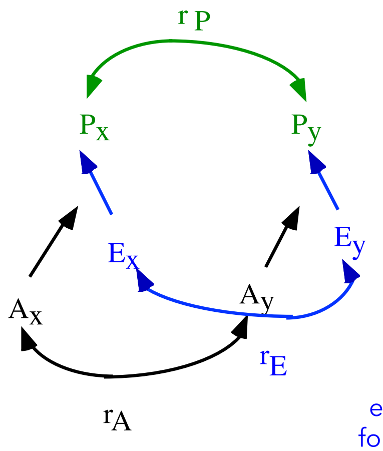
# Lecture 6:
# Selection on
# Multiple Traits

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 13 – 15 July 2015

## Genetic vs. Phenotypic correlations

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

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



Correlations between the breeding values of x and y within the individual can generate a phenotypic correlation

Likewise, the environmental values for the two traits within the individual could also be correlated

# Genetic & Environmental Correlations

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

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

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

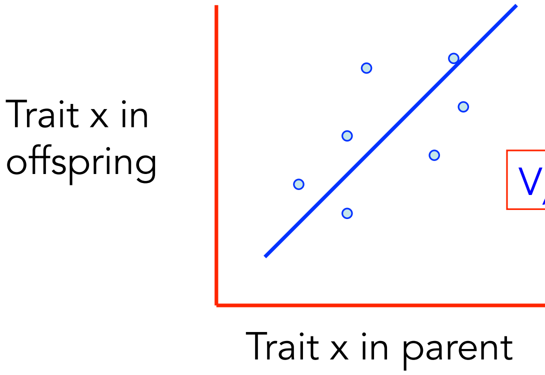
If heritability values are high for both traits, then the correlation in breeding values dominates the phenotypic corrrelation

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

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

# Estimating Genetic Correlations

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

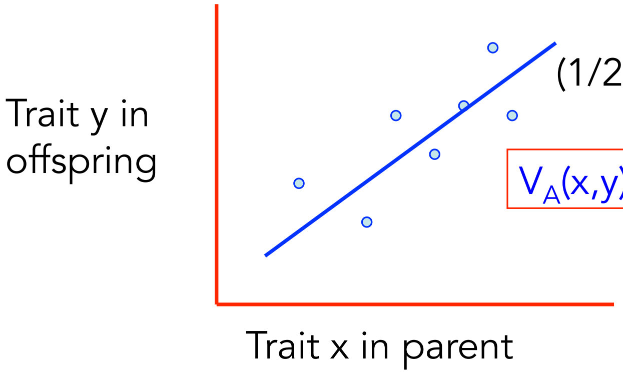


Slope = $(1/2)\ V_A(x)/V_P(x)$

$\boxed{V_A(x) = 2 * \text{slope} * V_P(x)}$

# Estimating Genetic Correlations

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



Slope = $(1/2)\ V_A(x,y)/V_P(x)$

$$\boxed{V_A(x,y) = 2 * \text{slope} * V_P(x)}$$

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

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

giving

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

Put another way,

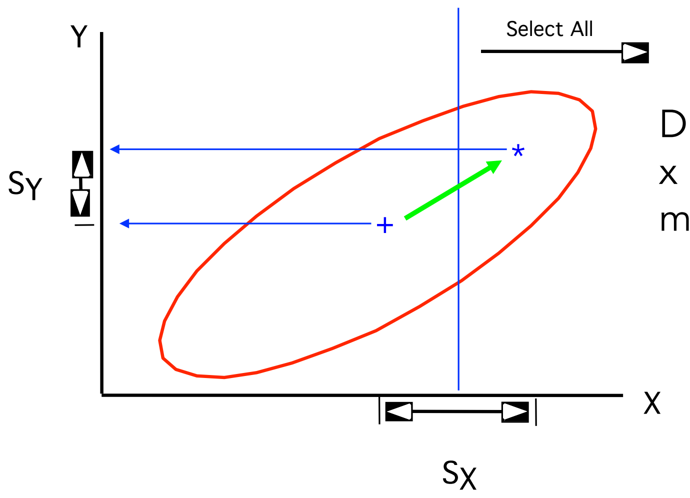
$$\mathrm{Cov}(x_O,y_P) = \mathrm{Cov}(y_O,x_P) = (1/2)\mathrm{Cov}(A_x,A_y)$$
$$\mathrm{Cov}(x_O,x_P) = (1/2)\ V_A\ (x) = (1/2)\mathrm{Cov}(A_x,\ A_x)$$
$$\mathrm{Cov}(y_O,y_P) = (1/2)\ V_A\ (y) = (1/2)\mathrm{Cov}(A_y,\ A_y)$$

Likewise, for half-sibs,

$$\mathrm{Cov}(x_{HS},y_{HS}) = (1/4)\ \mathrm{Cov}(A_x,A_y)$$
$$\mathrm{Cov}(x_{HS},x_{HS}) = (1/4)\ \mathrm{Cov}(A_x,A_x) = (1/4)\ V_A\ (x)$$
$$\mathrm{Cov}(y_{HS},y_{HS}) = (1/4)\ \mathrm{Cov}(A_y,A_y) = (1/4)\ V_A\ (y)$$

# Correlated Response to Selection

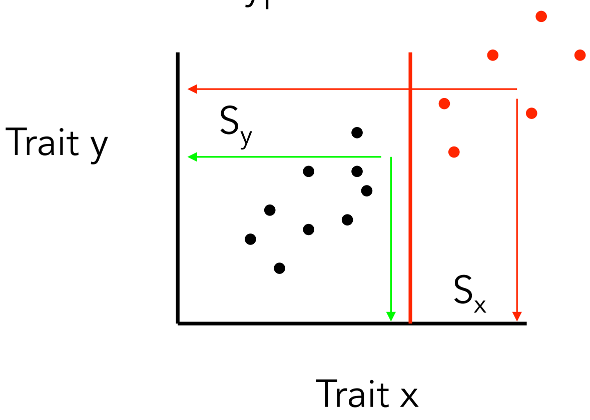
Direct selection of a character can cause a within-generation change in the mean of a phenotypically correlated character.



Direct selection on x also changes the mean of y

Phenotypic correlations induce within-generation changes

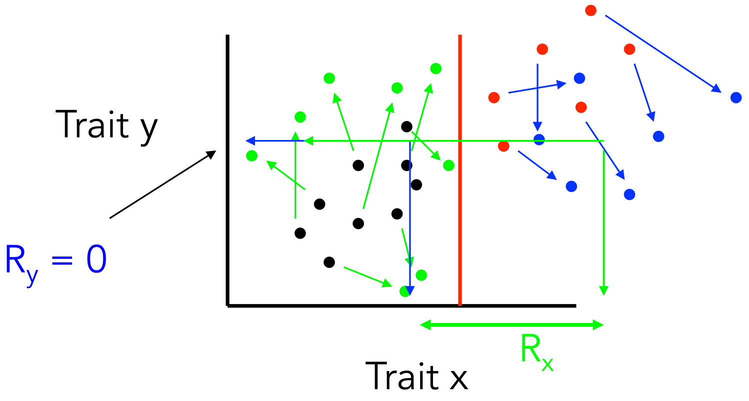
Phenotypic values



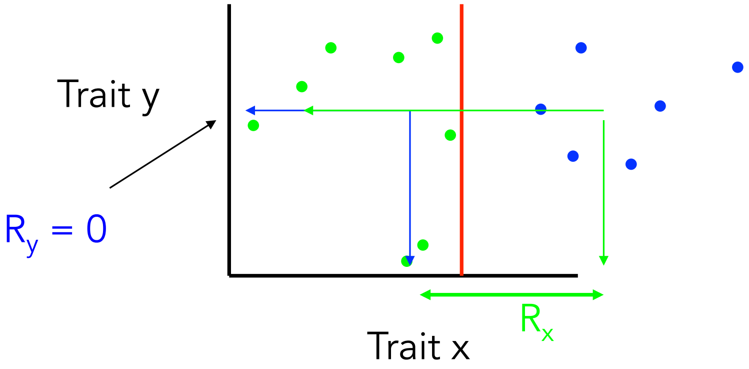
For there to be a between-generation change, the breeding values must be correlated. Such a change is called a correlated response to selection

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

Breeding values



Breeding values



# Predicting the correlated response

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

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

where

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

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

$$CR_y = b_{Ay|Ax} R_x$$

We can rewrite $CR_y = b_{Ay|Ax} R_x$ as follows

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

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

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

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

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

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

$$\boxed{CR_y = i_x\, h_x\, h_y\, r_A\, \sigma_P\,(y)}$$

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

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

# Direct vs. Indirect Response

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

$$\frac{CR_X}{R_X} = \frac{i_Y\, r_A\, \sigma_{AX}\, h_Y}{i_X\, h_X\, \sigma_{AX}} = \frac{i_Y\, r_A\, h_Y}{i_X\, h_X}$$

Hence, indirect selection gives a large response when

$$i_Y\, r_A\, h_Y > i_X\, h_X$$

- The selection intensity is much greater for y than x. This would be true if y were measurable in both sexes but x measurable in only one sex.

- Character y has a greater heritability than x, and the genetic correlation between x and y is high. This could occur if x is difficult to measure with precison but y is not.

# G x E

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

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

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

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

# Participatory breeding

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

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

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

$$i_Y \, r_A \, h_Y > i_X \, h_X$$

# Matrices

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

Dimensions given by rows x columns (r x c)

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

# Matrix Multiplication

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

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

In order to multiply two matrices, they must conform

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

# Matrix Multiplication

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

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

The identity matrix I serves the role of one in matrix multiplication: AI =A, IA = A

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

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

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

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

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

# The inverse serves the role of division in matrix multiplication

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

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

# The Multivariate Breeders' Equation

Suppose we are interested in the vector R of responses when selection occurs on n correlated traits

Let S be the vector of selection differentials.

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

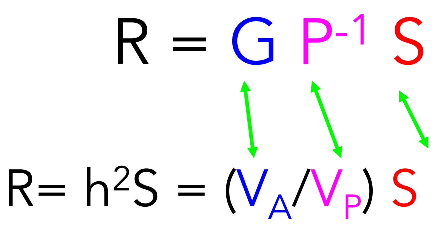
What is the multivariate version of this?

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

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

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

# The multivariate breeder's equation



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

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

Natural parallels
with univariate
breeder's equation

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

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

## Sources of within-generation change in the mean

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

Within-generation change in trait j

Change in mean from phenotypically correlated characters under direct selection

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

Change in mean from direct selection on trait j

## Within-generation change in the mean

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

## Response in the mean

Between-generation change (response) in trait j

Indirect response from genetically correlated characters under direct selection

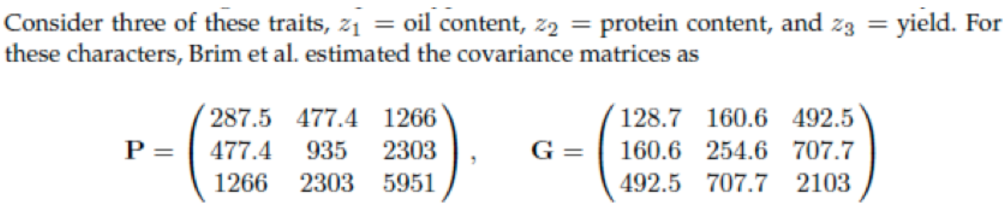
$$R_j = \sigma^2(A_j)\,\beta_j + \sum_{i \neq j} \sigma(A_j, A_i)\,\beta_i$$

Response from direct selection on trait j

Correlated response

Direct response

# Example in R



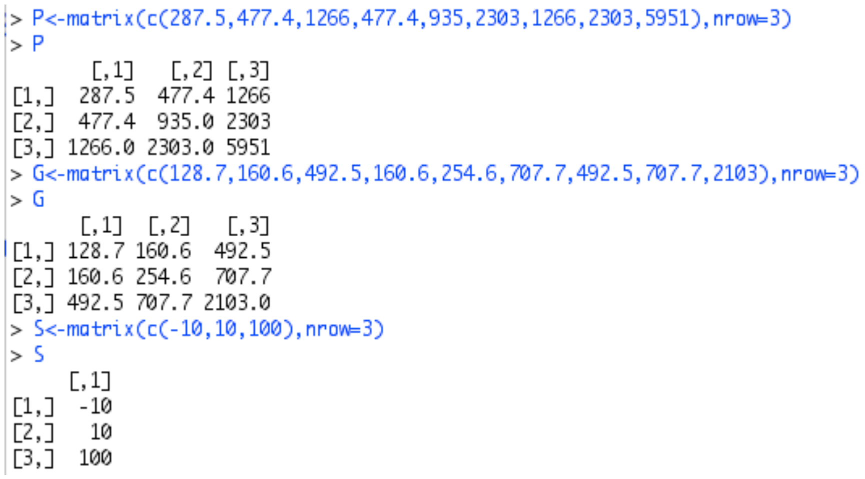
Consider three of these traits, $z_1$ = oil content, $z_2$ = protein content, and $z_3$ = yield. For these characters, Brim et al. estimated the covariance matrices as

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

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

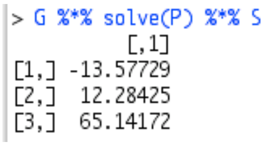
What is R? What is the nature of selection on each trait?

Enter G, P, and S



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

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

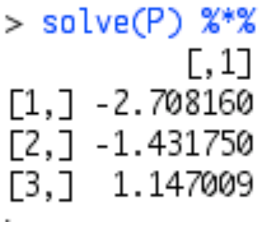


```
> G %*% solve(P) %*% S
          [,1]
[1,] -13.57729
[2,]  12.28425
[3,]  65.14172
```

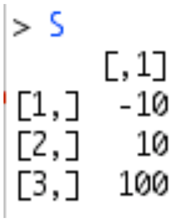
13.6 decrease in oil
12.3 increase in protein
65.1 increase in yield

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

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



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



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

β: targets of selection

S: observed within-generation change

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

## Quantifying Multivariate Constraints to Response

Is there genetic variation in the direction of selection?

Consider the following G and β:

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

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

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

What is the actual response?

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

# Constraints Imposed by Genetic Correlations

While β is the directional optimally favored by selection, the actual response is dragged off this direction, with R = G β.

**Example: Suppose**

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

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

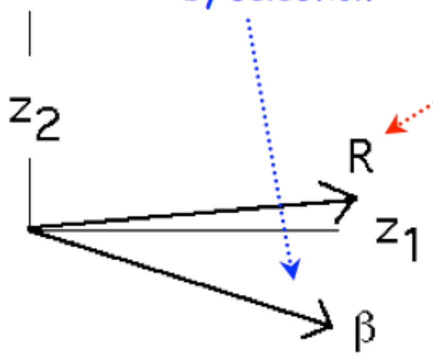
$$\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S} = \mathbf{P} = \begin{pmatrix} 20 & -10 \\ -10 & 40 \end{pmatrix}^{-1} \begin{pmatrix} 10 \\ -10 \end{pmatrix} = \begin{pmatrix} 0.43 \\ -0.14 \end{pmatrix}$$

What does the actual response look like?

$$\mathbf{R} = \mathbf{G}\boldsymbol{\beta} = \begin{pmatrix} 20 & 5 \\ 5 & 10 \end{pmatrix} \begin{pmatrix} 0.43 \\ -0.14 \end{pmatrix} = \begin{pmatrix} 7.86 \\ 0.71 \end{pmatrix}$$

Direction favored by selection

Direction of response



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

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

Hence, a vector has a length and a direction.

We can thus change a vector by both rotation and scaling

The length (or norm) of a vector x is denoted by ||x||

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

The (Euclidean) distance between two vectors x and y (of the same dimension) is

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

The angle θ between two vectors provides a measure for how they differ.

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

Vectors at right angles to each other, θ = 90° or 270° are said to be orthogonal.  If they have unit length as well, they are further said to be orthonormal.

## Matrices Describe Vector transformations

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

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

Thus A describes a *transformation* of the original coordinate system of x into a new coordinate system.

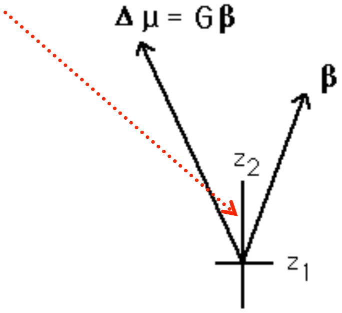
Example: Consider the following G and β:

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

The resulting angle between R and β is given by

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

For an angle of θ = 45 °



# Eigenvalues and Eigenvectors

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

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

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

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

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

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

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

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

Two nice features:

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

$\mathrm{trace}(A) = \Sigma_i \lambda_i$. The trace (sum of the diagonal elements) is is the sum of the eigenvalues

Note that det(A) = 0 if any only if at least one eigenvalue = 0

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

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

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

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

Example: Let's reconsider a previous G matrix

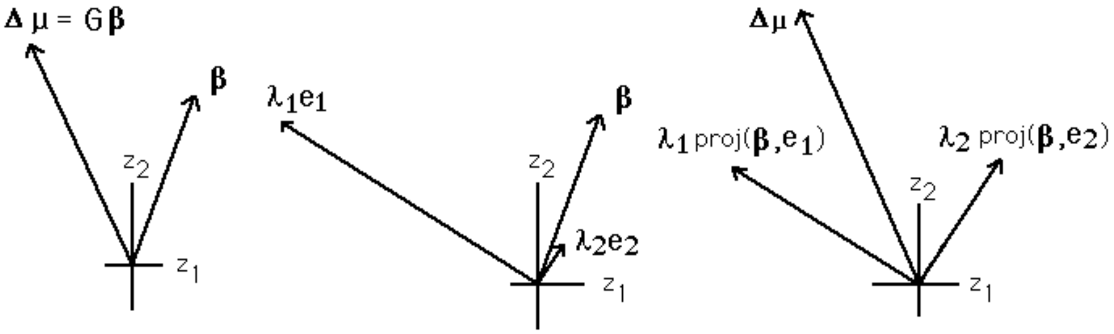
$$\begin{aligned}|\mathbf{G} - \lambda\mathbf{I}| &= \left|\begin{pmatrix} 4-\lambda & -2 \\ -2 & 2-\lambda \end{pmatrix}\right| \\ &= (4-\lambda)(2-\lambda) - (-2)^2 = \lambda^2 - 6\lambda + 4 = 0\end{aligned}$$

The solutions are

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

The corresponding eigenvectors become

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



Even though $\beta$ points in a direction very close of $e_2$, because most of the variation is accounted for by $e_1$, its projection is this dimension yields a much longer vector. The sum of these two projections yields the selection response R.

## Realized Selection Gradients

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

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

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

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

Here

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

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

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

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

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

# Muir Lecture 7

Introduction to Mixed Models, BLUP Breeding Values and REML Estimates of Variance Components

## References

Searle, S.R. 1971 Linear Models, Wiley

Schaeffer, L.R., Linear Models and Computer Strategies in Animal Breeding

Schaeffer, LR http://www.aps.uoguelph.ca/~lrs/ABModels/NOTES/vcBAYES.pdf

Lynch and Walsh Chapter 26

Mrode, R.A. Linear Models for the Prediction of Animal Breeding Values

# MIXED MODEL

- Separates Independent variable into those that are
  - Fixed $\mathbf{Xb}$

    **X**=value of each fixed effect
    ***b***=linear regression coefficients
  - Random $\mathbf{Zu}$

    **Z**=incidence matrix of random effect, usually a 1 corresponding to each animal
    **u**=estimate of random effects (breeding value)

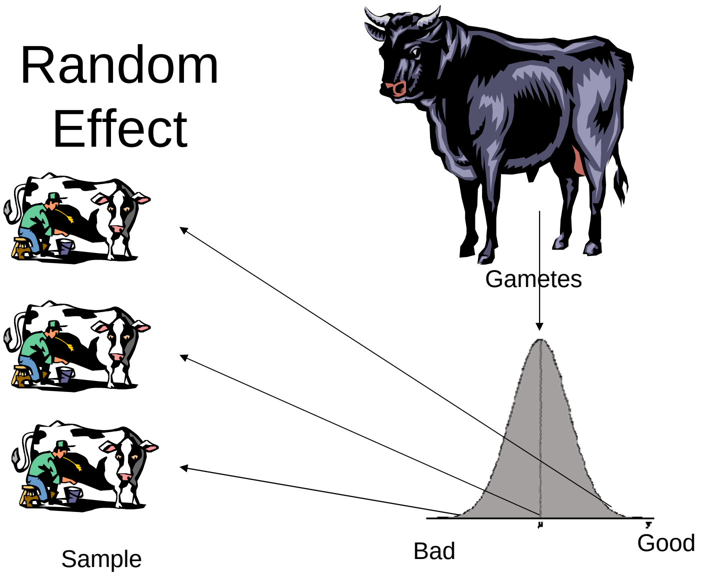
$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

More importantly model’s the variance structure

# Fixed and Random Effects

- Fixed Effect
  - – Inference Space only to those levels
  - – Herd, Year, Season, Parity, and Sex effect
- Random Effect
  - – Effect Sampled From A Distribution Of Effects
  - – Inference Space To The Population From Which The Random Effect Was Sampled

# Random Effect



Inference is to the genetic worth of the bull (breeding value)

## Variances In Mixed Models

$$\mathbf{Y} = \mathbf{Xb} + \underset{\substack{\uparrow \\ \text{EBV}}}{\mathbf{Zu}} + \mathbf{e}$$

$$V(\mathbf{b}) = \mathbf{0}$$

$$V(\mathbf{u}) = E(\mathbf{uu'}) = \mathbf{G}$$

$$V(\mathbf{e}) = E(\mathbf{ee'}) = \mathbf{R}$$

$$V(\mathbf{Y}) = V(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) = \mathbf{ZGZ'} + \mathbf{R}$$

Estimate the breeding values “**u”** and fixed effects simultaneously

The Maximum Likelihood Estimates of **b** and **u** give the mixed model equations (MME), These are also the Best Linear Unbiased Predictors (BLUP)

## Mixed Model Equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

Simplifications If $\mathbf{R} = \mathbf{I}\sigma_e^2$

Substitute $\mathbf{R}^{-1} = \left(\frac{1}{\sigma_e^2}\right)\mathbf{I}$ Then multiply both equation by $\sigma_e^2$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

# Simplifications

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\sigma_e^2\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

Assuming Additivity $\mathbf{G}=\mathbf{A}\sigma_a^2 \qquad \mathbf{G}^{-1}=\frac{1}{\sigma_a^2}\mathbf{A}^{-1}$

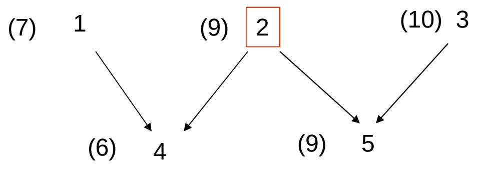
$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

Only Estimate of Ratio is Needed

Only inverse is needed

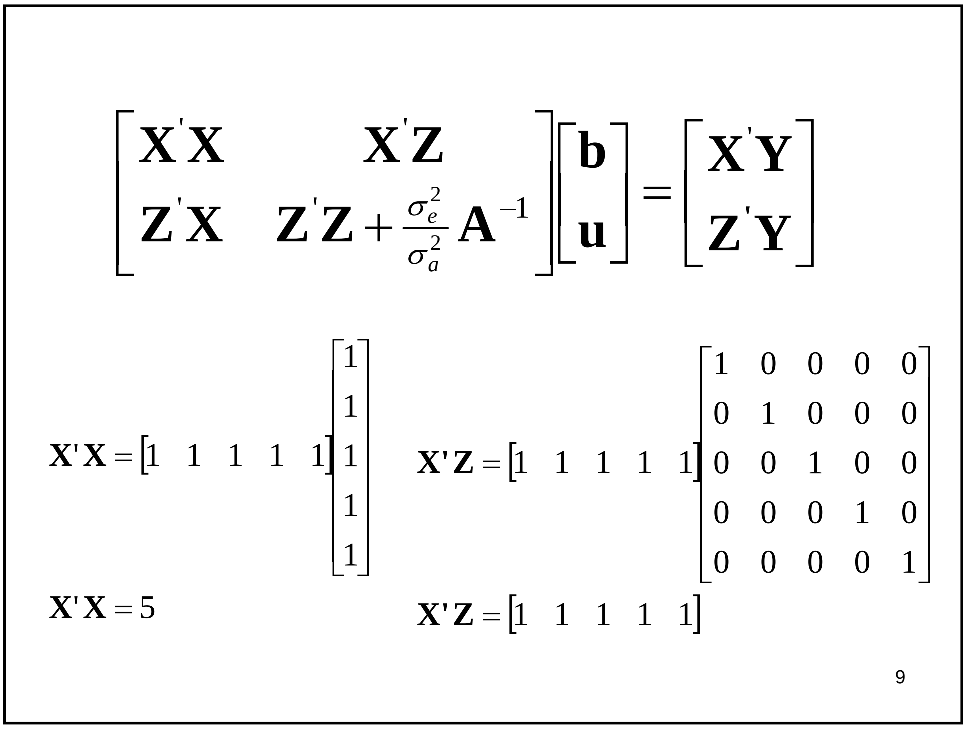
# Example 1

$$\mathbf{Y}=\begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$



$$b=[\mu]$$

$$\mathbf{X}=\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad Z=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{u}=\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} \quad \mathbf{e}=\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$



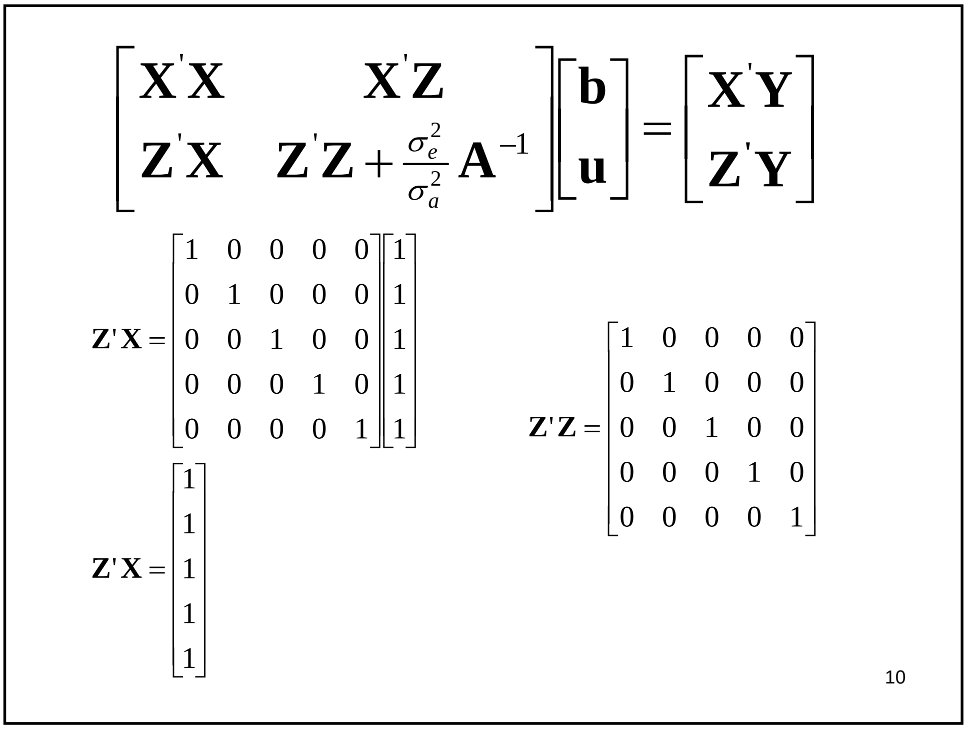
$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{u}\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{Y}\\ \mathbf{Z}'\mathbf{Y}\end{bmatrix}$$

$$\mathbf{X'X}=\begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$$

$$\mathbf{X'Z}=\begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}$$

$$\mathbf{X'X}=5$$

$$\mathbf{X'Z}=\begin{bmatrix}1 & 1 & 1 & 1 & 1\end{bmatrix}$$

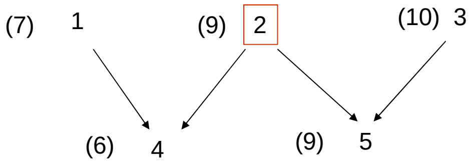


$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{u}\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{Y}\\ \mathbf{Z}'\mathbf{Y}\end{bmatrix}$$

$$\mathbf{Z'X}=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$$

$$\mathbf{Z'Z}=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}$$

$$\mathbf{Z'X}=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}$$



$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$$

Assume heritability=.5

$$\frac{\sigma_e^2}{\sigma_a^2} = 1$$

$$\mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} = \begin{bmatrix} \frac{5}{2} & \frac{1}{2} & 0 & -1 & 0 \\ \frac{1}{2} & 3 & \frac{1}{2} & -1 & -1 \\ 0 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & 0 & 3 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$

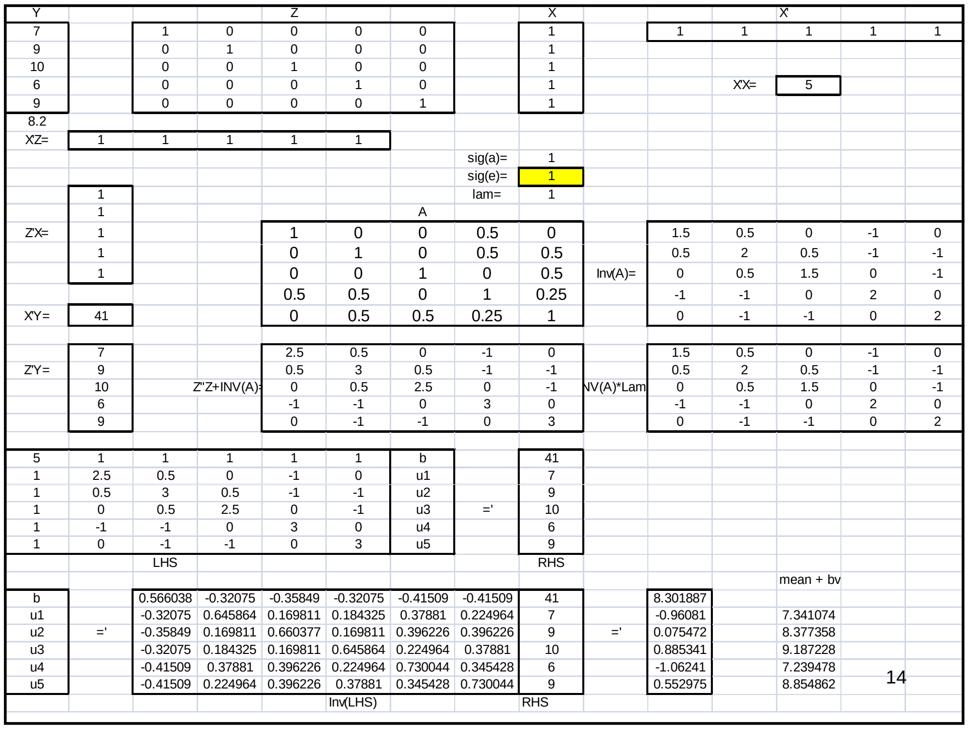
$$\mathbf{Z'Y} = \begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 41 \end{bmatrix}$$

# MME

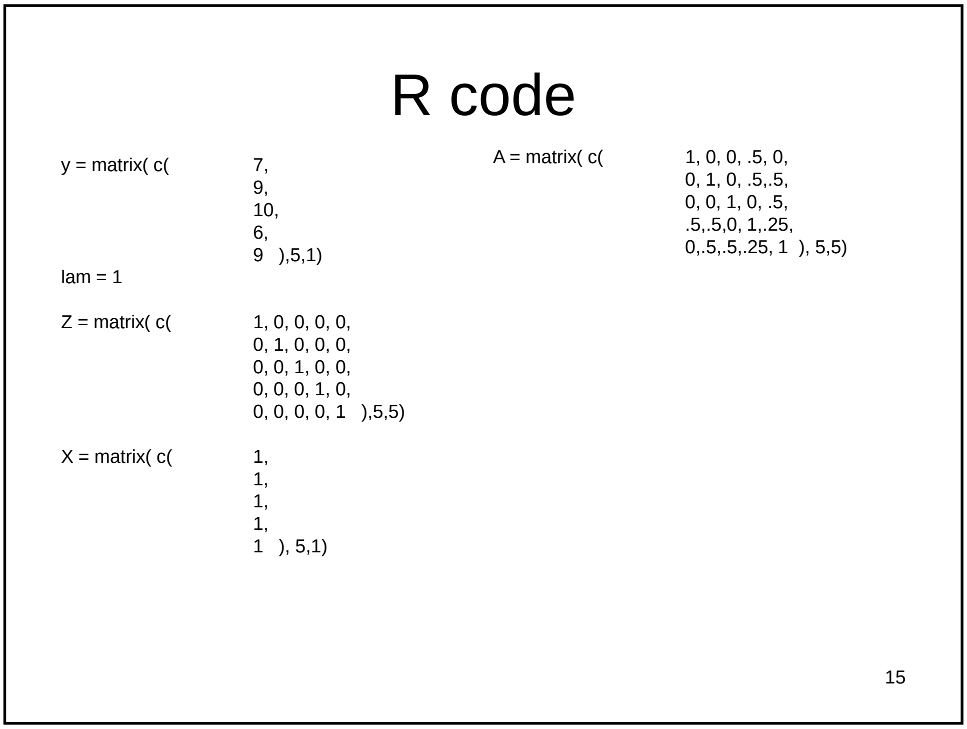
$$\begin{bmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & \frac{5}{2} & \frac{1}{2} & 0 & -1 & 0 \\ 1 & \frac{1}{2} & 3 & \frac{1}{2} & -1 & -1 \\ 1 & 0 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ 1 & -1 & -1 & 0 & 3 & 0 \\ 1 & 0 & -1 & -1 & 0 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} 41 \\ 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$

13



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | | | | Z | | | | X | | | | X' | | |
| 7 | | 1 | 0 | 0 | 0 | 0 | | 1 | | 1 | 1 | 1 | 1 | 1 |
| 9 | | 0 | 1 | 0 | 0 | 0 | | 1 | | | | | | |
| 10 | | 0 | 0 | 1 | 0 | 0 | | 1 | | | | | | |
| 6 | | 0 | 0 | 0 | 1 | 0 | | 1 | | | X'X= | 5 | | |
| 9 | | 0 | 0 | 0 | 0 | 1 | | 1 | | | | | | |
| 8.2 | | | | | | | | | | | | | | |
| X'Z= | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| | | | | | | | sig(a)= | 1 | | | | | | |
| | | | | | | | sig(e)= | 1 | | | | | | |
| | 1 | | | | | | lam= | 1 | | | | | | |
| | 1 | | | | | A | | | | | | | | |
| Z'X= | 1 | | | 1 | 0 | 0 | 0.5 | 0 | | 1.5 | 0.5 | 0 | -1 | 0 |
| | 1 | | | 0 | 1 | 0 | 0.5 | 0.5 | | 0.5 | 2 | 0.5 | -1 | -1 |
| | 1 | | | 0 | 0 | 1 | 0 | 0.5 | Inv(A)= | 0 | 0.5 | 1.5 | 0 | -1 |
| | | | | 0.5 | 0.5 | 0 | 1 | 0.25 | | -1 | -1 | 0 | 2 | 0 |
| X'Y= | 41 | | | 0 | 0.5 | 0.5 | 0.25 | 1 | | 0 | -1 | -1 | 0 | 2 |
| | | | | | | | | | | | | | | |
| | 7 | | | 2.5 | 0.5 | 0 | -1 | 0 | | 1.5 | 0.5 | 0 | -1 | 0 |
| Z'Y= | 9 | | | 0.5 | 3 | 0.5 | -1 | -1 | | 0.5 | 2 | 0.5 | -1 | -1 |
| | 10 | | Z"Z+INV(A)= | 0 | 0.5 | 2.5 | 0 | -1 | NV(A)*Lam | 0 | 0.5 | 1.5 | 0 | -1 |
| | 6 | | | -1 | -1 | 0 | 3 | 0 | | -1 | -1 | 0 | 2 | 0 |
| | 9 | | | 0 | -1 | -1 | 0 | 3 | | 0 | -1 | -1 | 0 | 2 |
| | | | | | | | | | | | | | | |
| 5 | 1 | 1 | 1 | 1 | 1 | b | | 41 | | | | | | |
| 1 | 2.5 | 0.5 | 0 | -1 | 0 | u1 | | 7 | | | | | | |
| 1 | 0.5 | 3 | 0.5 | -1 | -1 | u2 | | 9 | | | | | | |
| 1 | 0 | 0.5 | 2.5 | 0 | -1 | u3 | =' | 10 | | | | | | |
| 1 | -1 | -1 | 0 | 3 | 0 | u4 | | 6 | | | | | | |
| 1 | 0 | -1 | -1 | 0 | 3 | u5 | | 9 | | | | | | |
| | | LHS | | | | | | RHS | | | | | | |
| | | | | | | | | | | | | mean + bv | | |
| b | | 0.566038 | -0.32075 | -0.35849 | -0.32075 | -0.41509 | -0.41509 | 41 | | 8.301887 | | | | |
| u1 | | -0.32075 | 0.645864 | 0.169811 | 0.184325 | 0.37881 | 0.224964 | 7 | | -0.96081 | | 7.341074 | | |
| u2 | =' | -0.35849 | 0.169811 | 0.660377 | 0.169811 | 0.396226 | 0.396226 | 9 | =' | 0.075472 | | 8.377358 | | |
| u3 | | -0.32075 | 0.184325 | 0.169811 | 0.645864 | 0.224964 | 0.37881 | 10 | | 0.885341 | | 9.187228 | | |
| u4 | | -0.41509 | 0.37881 | 0.396226 | 0.224964 | 0.730044 | 0.345428 | 6 | | -1.06241 | | 7.239478 | | |
| u5 | | -0.41509 | 0.224964 | 0.396226 | 0.37881 | 0.345428 | 0.730044 | 9 | | 0.552975 | | 8.854862 | | |
| | | | | | Inv(LHS) | | | RHS | | | | | | |

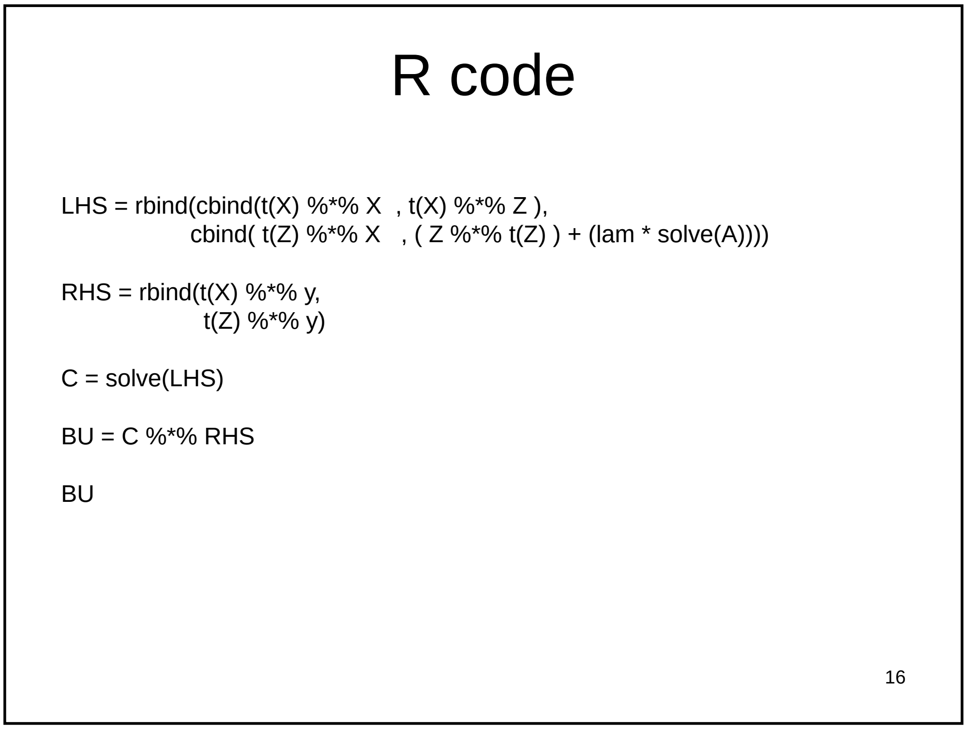
14



# R code

```
y = matrix( c(        7,
                      9,
                      10,
                      6,
                      9   ),5,1)
lam = 1

Z = matrix( c(        1, 0, 0, 0, 0,
                      0, 1, 0, 0, 0,
                      0, 0, 1, 0, 0,
                      0, 0, 0, 1, 0,
                      0, 0, 0, 0, 1   ),5,5)

X = matrix( c(        1,
                      1,
                      1,
                      1,
                      1   ), 5,1)
```

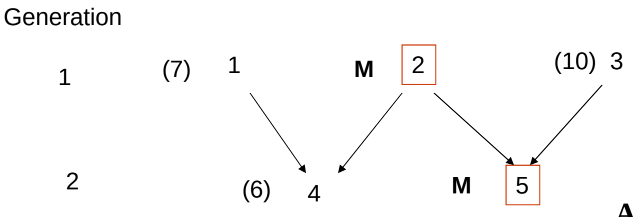
```
A = matrix( c(        1, 0, 0, .5, 0,
                      0, 1, 0, .5,.5,
                      0, 0, 1, 0, .5,
                      .5,.5,0, 1,.25,
                      0,.5,.5,.25, 1  ), 5,5)
```



# R code

```
LHS = rbind(cbind(t(X) %*% X  , t(X) %*% Z ),
            cbind( t(Z) %*% X   , ( Z %*% t(Z) ) + (lam * solve(A))))

RHS = rbind(t(X) %*% y,
             t(Z) %*% y)

C = solve(LHS)

BU = C %*% RHS

BU
```

# Missing Values (Sex Limited Traits)



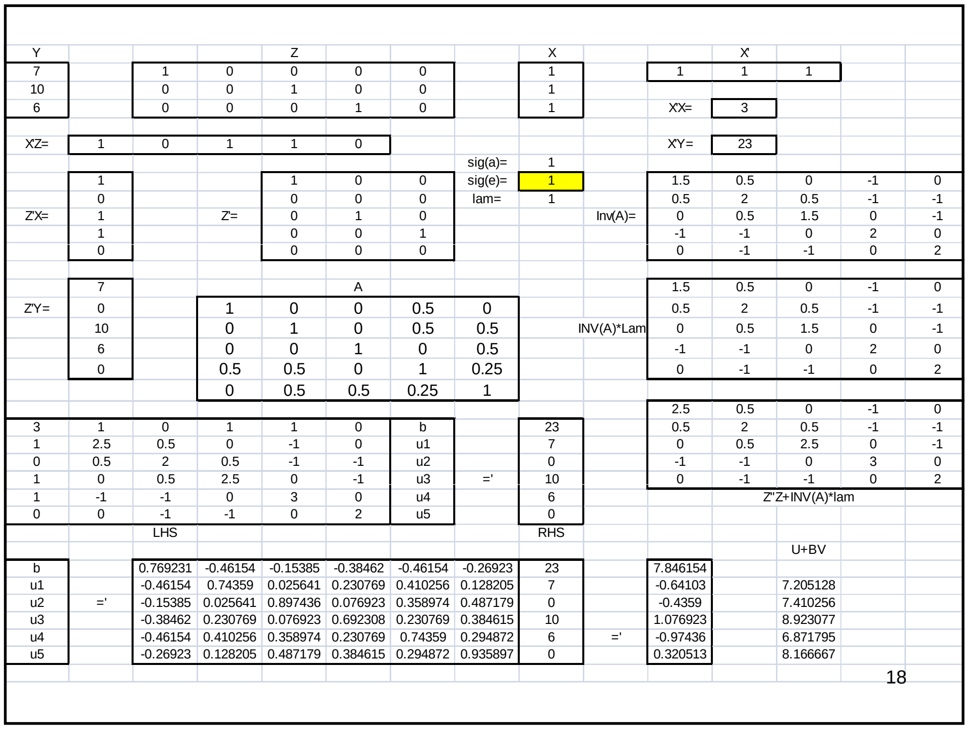
$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 7 \\ 10 \\ 6 \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_3 \\ e_5 \end{bmatrix} \qquad b = [\mu]$$

$$\mathbf{U} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Assume $h^2$=.5



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | | | | Z | | | | X | | | X' | | | |
| 7 | | 1 | 0 | 0 | 0 | 0 | | 1 | | 1 | 1 | 1 | | |
| 10 | | 0 | 0 | 1 | 0 | 0 | | 1 | | | | | | |
| 6 | | 0 | 0 | 0 | 1 | 0 | | 1 | | X'X= | 3 | | | |
| | | | | | | | | | | | | | | |
| X'Z= | 1 | 0 | 1 | 1 | 0 | | | | | X'Y= | 23 | | | |
| | | | | | | | sig(a)= | 1 | | | | | | |
| | 1 | | | 1 | 0 | 0 | sig(e)= | 🟨 1 | | 1.5 | 0.5 | 0 | -1 | 0 |
| | 0 | | | 0 | 0 | 0 | lam= | 1 | | 0.5 | 2 | 0.5 | -1 | -1 |
| Z'X= | 1 | | Z'= | 0 | 1 | 0 | | | Inv(A)= | 0 | 0.5 | 1.5 | 0 | -1 |
| | 1 | | | 0 | 0 | 1 | | | | -1 | -1 | 0 | 2 | 0 |
| | 0 | | | 0 | 0 | 0 | | | | 0 | -1 | -1 | 0 | 2 |
| | | | | | | | | | | | | | | |
| | 7 | | | | A | | | | | 1.5 | 0.5 | 0 | -1 | 0 |
| Z'Y= | 0 | | 1 | 0 | 0 | 0.5 | 0 | | | 0.5 | 2 | 0.5 | -1 | -1 |
| | 10 | | 0 | 1 | 0 | 0.5 | 0.5 | | INV(A)*Lam | 0 | 0.5 | 1.5 | 0 | -1 |
| | 6 | | 0 | 0 | 1 | 0 | 0.5 | | | -1 | -1 | 0 | 2 | 0 |
| | 0 | | 0.5 | 0.5 | 0 | 1 | 0.25 | | | 0 | -1 | -1 | 0 | 2 |
| | | | 0 | 0.5 | 0.5 | 0.25 | 1 | | | | | | | |
| | | | | | | | | | | 2.5 | 0.5 | 0 | -1 | 0 |
| 3 | 1 | 0 | 1 | 1 | 0 | b | | 23 | | 0.5 | 2 | 0.5 | -1 | -1 |
| 1 | 2.5 | 0.5 | 0 | -1 | 0 | u1 | | 7 | | 0 | 0.5 | 2.5 | 0 | -1 |
| 0 | 0.5 | 2 | 0.5 | -1 | -1 | u2 | | 0 | | -1 | -1 | 0 | 3 | 0 |
| 1 | 0 | 0.5 | 2.5 | 0 | -1 | u3 | =' | 10 | | 0 | -1 | -1 | 0 | 2 |
| 1 | -1 | -1 | 0 | 3 | 0 | u4 | | 6 | | | | Z"Z+INV(A)*lam | | |
| 0 | 0 | -1 | -1 | 0 | 2 | u5 | | 0 | | | | | | |
| | | LHS | | | | | | RHS | | | | | | |
| | | | | | | | | | | | | U+BV | | |
| b | | 0.769231 | -0.46154 | -0.15385 | -0.38462 | -0.46154 | -0.26923 | 23 | | 7.846154 | | | | |
| u1 | | -0.46154 | 0.74359 | 0.025641 | 0.230769 | 0.410256 | 0.128205 | 7 | | -0.64103 | | 7.205128 | | |
| u2 | =' | -0.15385 | 0.025641 | 0.897436 | 0.076923 | 0.358974 | 0.487179 | 0 | | -0.4359 | | 7.410256 | | |
| u3 | | -0.38462 | 0.230769 | 0.076923 | 0.692308 | 0.230769 | 0.384615 | 10 | | 1.076923 | | 8.923077 | | |
| u4 | | -0.46154 | 0.410256 | 0.358974 | 0.230769 | 0.74359 | 0.294872 | 6 | =' | -0.97436 | | 6.871795 | | |
| u5 | | -0.26923 | 0.128205 | 0.487179 | 0.384615 | 0.294872 | 0.935897 | 0 | | 0.320513 | | 8.166667 | | |

# Extensions of Model

- Inclusion of Dominance and Epistasis
  - Dominance
    - Dominance effects are the result of interaction of alleles within a locus
    - Dominance relationship matrix needed
    - Reflects the probability that individuals have the same pair of alleles in common at a locus
  - Epistasis
    - Epistatic genetic effects are the result of interactions between alleles at different loci
    - Epistatic relationship matrix needed
    - Reflects the probability that individuals have the same pair of alleles in common at different loci (4 possible pairings of 2 alleles at 2 loci)
  - Useful in crossbreeding programs but generally not useful in pure breeding programs
    - An individual does not pass on dominance or epistatic effects (without inbreeding or cloning), which are a function of both parents
    - Exception is Additive x Additive epistasis which is a function of 2 alleles at different loci in the same gamete, but dissipates with recombination and/or segregation

## Estimation of Variances Using all Data in a Pedigree

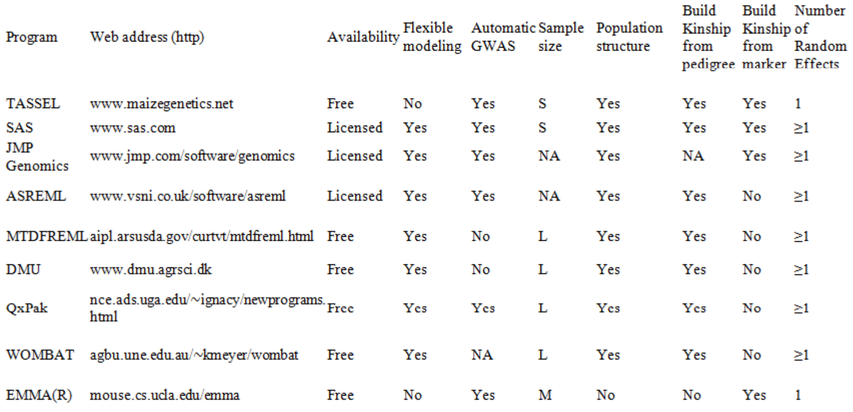
- REML
  - EM-REML iterative process whereby
    - A value is assumed for additive variance
    - Estimates of breeding values found
    - Additive variance V(A) is estimated as variance of breeding values V(A)=(u’A$^{-1}$u +stuff)/n
    - The new value of V(A) is substituted into the MME
    - Estimates of breeding values (u) are found
    - The process repeated until convergence
  - DF-REML work by trial and error finding a value of V(A) that maximize the likelihood

# Appendix 1

Software packages for estimating EBVs, Variance Components, GWAS and genomic selection

## Software engineering the mixed model for genome-wide association studies on large samples

http://bib.oxfordjournals.org/content/10/6/664/T1.expansion.html



| Program | Web address (http) | Availability | Flexible modeling | Automatic GWAS | Sample size | Population structure | Build Kinship from pedigree | Build Kinship from marker | Number of Random Effects |
|---|---|---|---|---|---|---|---|---|---|
| TASSEL | www.maizegenetics.net | Free | No | Yes | S | Yes | Yes | Yes | 1 |
| SAS | www.sas.com | Licensed | Yes | Yes | S | Yes | Yes | Yes | ≥1 |
| JMP Genomics | www.jmp.com/software/genomics | Licensed | Yes | Yes | NA | Yes | NA | Yes | ≥1 |
| ASREML | www.vsni.co.uk/software/asreml | Licensed | Yes | Yes | NA | Yes | Yes | No | ≥1 |
| MTDFREML | aipl.arsusda.gov/curtvt/mtdfreml.html | Free | Yes | No | L | Yes | Yes | No | ≥1 |
| DMU | www.dmu.agrsci.dk | Free | Yes | No | L | Yes | Yes | No | ≥1 |
| QxPak | nce.ads.uga.edu/~ignacy/newprograms.html | Free | Yes | Yes | L | Yes | Yes | No | ≥1 |
| WOMBAT | agbu.une.edu.au/~kmeyer/wombat | Free | Yes | NA | L | Yes | Yes | No | ≥1 |
| EMMA(R) | mouse.cs.ucla.edu/emma | Free | No | Yes | M | No | No | Yes | 1 |

# R packages

- QTL mapping
  - ***onemap*** – It is used to generate or rearrange genetic maps
  - ***rqtl*** – performs QTL mapping for bi-parental populations
  - **GAPIT** – most common package for Genome-Wide Association Mapping
- BLUP (Animal Model)
  - ***pedigree*** – Generates A matrix from sparse pedigree
  - ***MCMCglmm*** – Generalized Mixed Models incorporating pedigrees
  - ***pedigreemm*** - Fit mixed-effects models incorporating pedigrees
- Genomic Selection
  - ***rrBLUP*** – classic package to perform ridge regression BLUP and GBLUP
  - ***BGLR*** – whole genome regressions methods of genomic selection
  - ***randomForest*** – Random Forest Regression (non-parametric GS)
  - ***brnn*** – Bayesian Regularized Neural Network (non-parametric GS)
  - ***parallel*** – Allows the use of multiple cores for faster computation

Provided by Alencar Xavier (xaviera@purdue.edu)

## Appendix 2

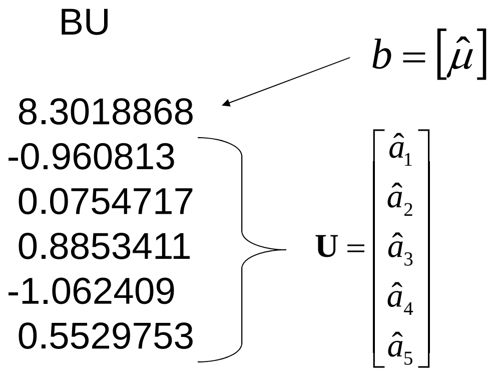
# SAS IML BLUP Programs

## For Examples 1 and 2

```
proc iml;
start main;

y={ 7,
    9,
   10,
    6,
    9};

  X={1,
     1,
     1,
     1,
     1};

A={1  0  0 .5   0,
     0  1  0 .5  .5,
     0  0  1  0  .5,
    .5 .5  0  1  .25,
     0 .5 .5 .25  1};
```

```
lam=1;

Z={1 0 0 0 0,
   0 1 0 0 0,
   0 0 1 0 0,
   0 0 0 1 0,
   0 0 0 0 1};

LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*Z+INV(A)#
   LAM));

RHS=(X`*Y)//(Z`*Y);
C=INV(HS);

BU=C*RHS;
print C BU;

finish main;
run;
quit;
```

# Estimates

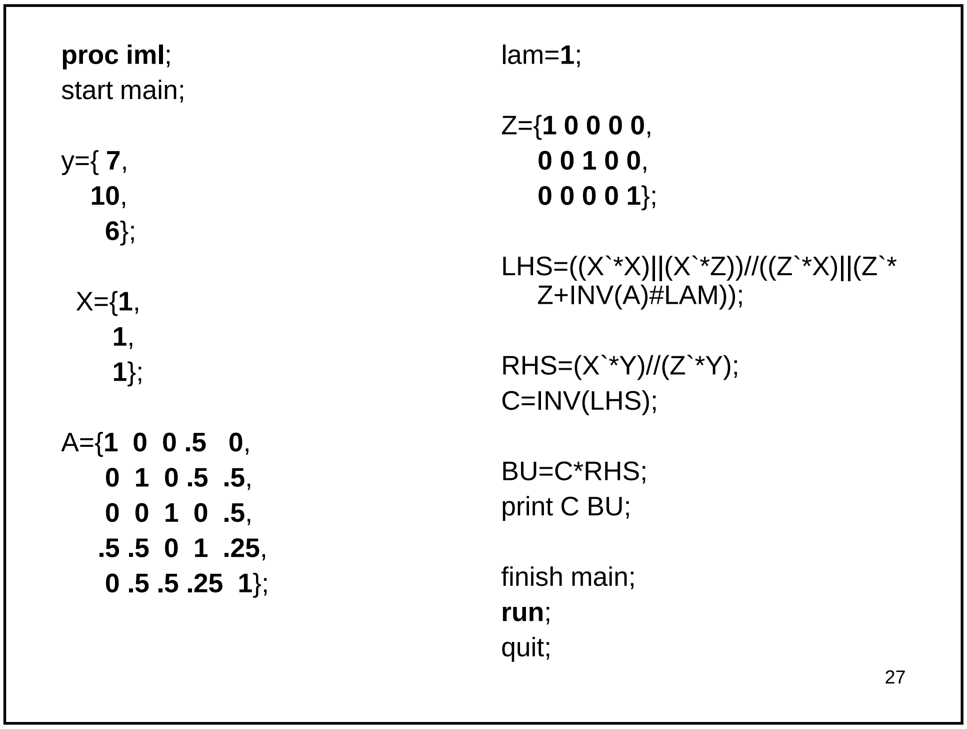


BU

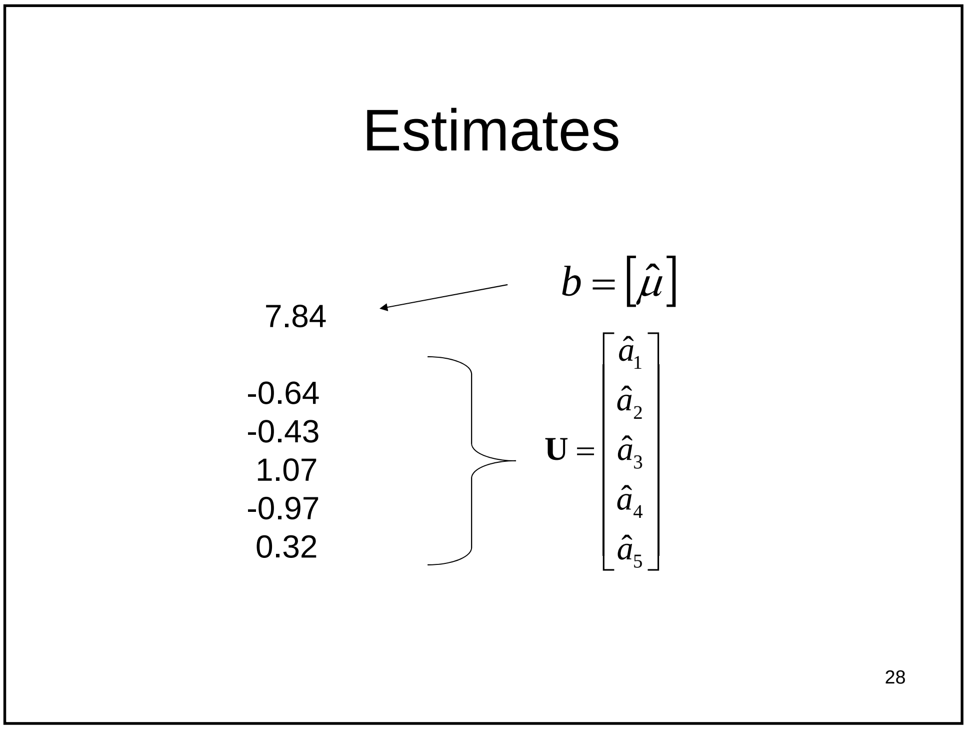
```
 8.3018868
-0.960813
 0.0754717
 0.8853411
-1.062409
 0.5529753
```

$$b = \left[\hat{\mu}\right]$$

$$\mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$



```
proc iml;
start main;

y={ 7,
   10,
    6};

  X={1,
     1,
     1};

A={1  0  0 .5   0,
   0  1  0 .5  .5,
   0  0  1  0  .5,
  .5 .5  0  1  .25,
   0 .5 .5 .25  1};

lam=1;

Z={1 0 0 0 0,
   0 0 1 0 0,
   0 0 0 0 1};

LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*
   Z+INV(A)#LAM));

RHS=(X`*Y)//(Z`*Y);
C=INV(LHS);

BU=C*RHS;
print C BU;

finish main;
run;
quit;
```



# Estimates

7.84

-0.64
-0.43
1.07
-0.97
0.32

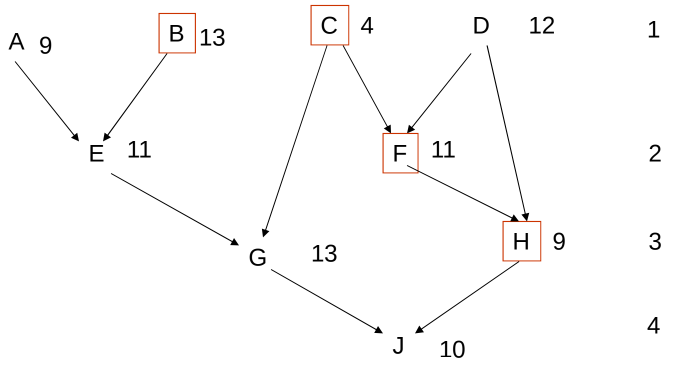
$$b = [\hat{\mu}]$$

$$\mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$

# Appendix 3

## Problems

# Problem 1



Find the best estimate of the genetic worth of each animal. Assume a heritability of .5. Work problem using either SAS IML or BLUPF90, also try REMLF90 to estimate variances

## Answer Problem 1

```
proc iml;
start main;

y={9,
   13,
   4,
   12,
   11,
   11,
   13,
    9,
   10};

X={1,
   1,
   1,
   1,
   1,
   1,
   1,
   1,
   1};
```

```
A={1     0      0      0      0.5  0     0.25   0      0.125,
   0     1      0      0      0.5  0     0.25   0      0.125,
   0     0      1      0      0    0.5  0.5    0.25   0.375,
   0     0      0      1      0    0.5  0      0.75   0.375,
   0.5   0.5    0      0      1    0    0.5    0      0.25,
   0     0      0.5    0.5    0    1    0.25   0.75   0.5,
   0.25  0.25   0.5    0      0.5  0.25 1      0.125  0.5625,
   0     0      0.25   0.75   0    0.75 0.125  1.25   0.6875,
   0.125 0.125  0.375  0.375  0.25 0.5  0.5625 0.6875 1.0625};

AINV=INV(A);
lam=1;

Z={1 0 0 0 0 0 0 0 0,
   0 1 0 0 0 0 0 0 0,
   0 0 1 0 0 0 0 0 0,
   0 0 0 1 0 0 0 0 0,
   0 0 0 0 1 0 0 0 0,
   0 0 0 0 0 1 0 0 0,
   0 0 0 0 0 0 1 0 0,
   0 0 0 0 0 0 0 1 0,
   0 0 0 0 0 0 0 0 1};

LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*Z+AINV#LAM));
RHS=(X`*Y)//(Z`*Y);
C=INV(LHS);
BU=C*RHS;
```

### Answer

```
       10.07
       -0.31
       1.689
       -2.28
       0.905
BU=    1.145
       -0.31
       0.564
       -0.19
       0.105
```

# Problem 1:

## BLUPF90 par.txt

```
# Example of single-trait animal model with one fixed effect
DATAFILE
phenotypes.txt
NUMBER_OF_TRAITS
1
NUMBER_OF_EFFECTS
2
OBSERVATION(S)
3
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBER_OF_LEVELS TYPE_OF_EFFECT [EFFECT NESTED]
1 9 cross
2 1 cross
RANDOM_RESIDUAL VALUES
1.0
RANDOM_GROUP
1
RANDOM_TYPE
add_animal
FILE
pedigree.txt
(CO)VARIANCES
1.0
```

# Problem 1:BLUPF90

pedigree.txt

```
1 0 0
2 0 0
3 0 0
4 0 0
5 2 1
6 3 4
7 3 5
8 6 4
9 8 7
```

phenotypes.txt

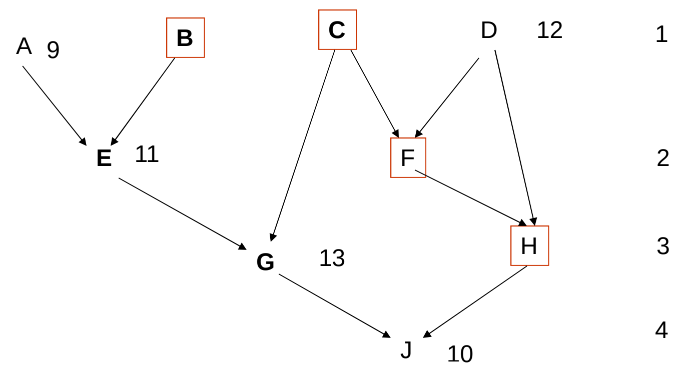
```
1 1  9
2 1 13
3 1  4
4 1 12
5 1 11
6 1 11
7 1 13
8 1  9
9 1 10
```

solutions

```
trait/effect level  solution
   1   1        1   -0.31070241
   1   1        2    1.68929759
   1   1        3   -2.28389748
   1   1        4    0.90530229
   1   1        5    1.14527490
   1   1        6   -0.31432205
   1   1        7    0.56530033
   1   1        8   -0.19149873
   1   1        9    0.09880650
   1   2        1   10.07738212
```

Genetic variance .00259
Residual 2.389

## Problem 2: Sex Limited Trait



Estimate breeding values for the males.
Assume a heritability of .5. Work problem using SAS and BLUPF90

## Answer Problem 2

```
proc iml;       A={1    0      0     0     0.5  0    0.25   0      0.125,
start main;       0     1      0     0     0.5  0    0.25   0      0.125,
                  0     0      1     0     0    0.5  0.5    0.25   0.375,
y={9,             0     0      0     1     0    0.5  0      0.75   0.375,
   12,            0.5   0.5    0     0     1    0    0.5    0      0.25,
   11,            0     0      0.5   0.5   0    1    0.25   0.75   0.5,
   13,            0.25  0.25   0.5   0     0.5  0.25 1      0.125  0.5625,
   10};           0     0      0.25  0.75  0    0.75 0.125  1.25   0.6875,
                  0.125 0.125  0.375 0.375 0.25 0.5  0.5625 0.6875 1.0625};

X={1,
   1,
   1,          AINV=INV(A);
   1,          lam=1;
   1};
               Z={1 0 0 0 0 0 0 0 0,
                  0 0 0 1 0 0 0 0 0,
                  0 0 0 0 1 0 0 0 0,
                  0 0 0 0 0 0 1 0 0,
                  0 0 0 0 0 0 0 0 1};

               LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*Z+AINV#LAM));
               RHS=(X`*Y)//(Z`*Y);
               C=INV(LHS);
               BU=C*RHS;
```

### Answer

```
      11.03
      -0.89
       0.247
       0.338
       0.307
BU=   -0.075
       0.206
       0.587
       0.023
      -0.102
```

# Problem 2:

## BLUPF90 par.txt

```
# Example of single-trait animal model with one fixed effect
DATAFILE
phenotypes.txt
NUMBER_OF_TRAITS
1
NUMBER_OF_EFFECTS
2
OBSERVATION(S)
3
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBER_OF_LEVELS TYPE_OF_EFFECT [EFFECT NESTED]
1 9 cross
2 1 cross
RANDOM_RESIDUAL VALUES
1.0
RANDOM_GROUP
1
RANDOM_TYPE
add_animal
FILE
pedigree.txt
(CO)VARIANCES
1.0
```

# Problem 2 BLUPF90

pedigree.txt

```
1 0 0
2 0 0
3 0 0
4 0 0
5 2 1
6 3 4
7 3 5
8 6 4
9 8 7
```

phenotypes.txt

```
1 1  9
4 1 12
5 1 11
7 1 13
9 1 10
```

solutions

```
trait/effect level  solution
  1    1        1    -0.89891190
  1    1        2     0.24640225
  1    1        3     0.34415584
  1    1        4     0.30835381
  1    1        5    -0.07985258
  1    1        6     0.21323271
  1    1        7     0.58932959
  1    1        8     0.03474903
  1    1        9    -0.14004914
  1    2        1    11.04422604
```

# Lecture 8
# QTL and Association mapping

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 13 – 15 July 2015

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

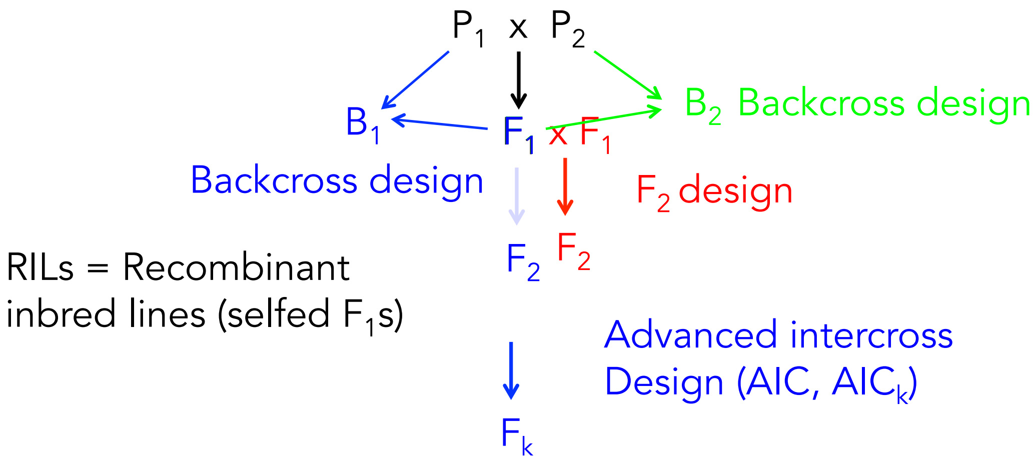
- QTL mapping tries to detect small (20-40 cM) chromosome segments influencing trait variation
  - Relatively crude level of resolution
- QTL mapping performed either using inbred line crosses or sets of known relatives
  - Uses the simple fact of an excess of parental gametes

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

$B_2$ Backcross design

$F_1$ x $F_1$

$F_2$ 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) = $(1-c)^2/4$

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) = $(1-c)^2$

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

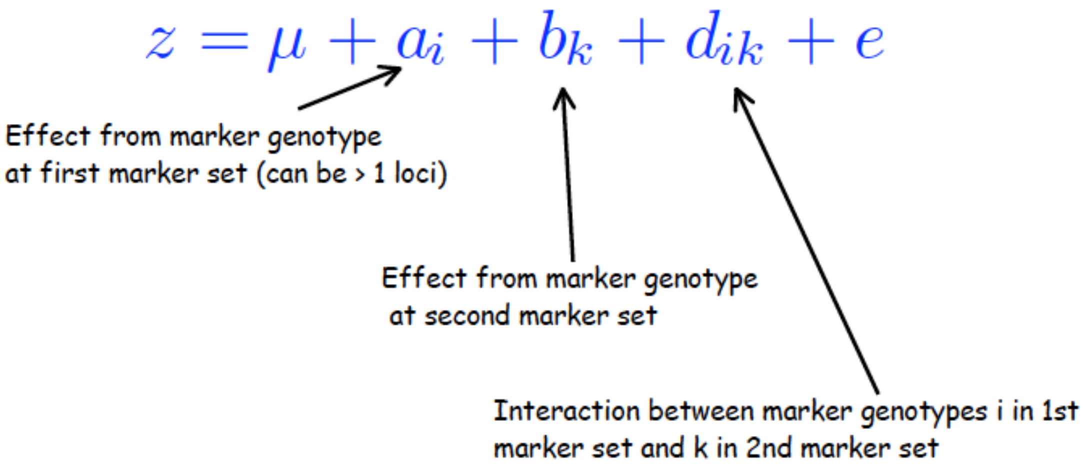
$$z_{ik} = \mu + b_i + e_{ik}$$

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)

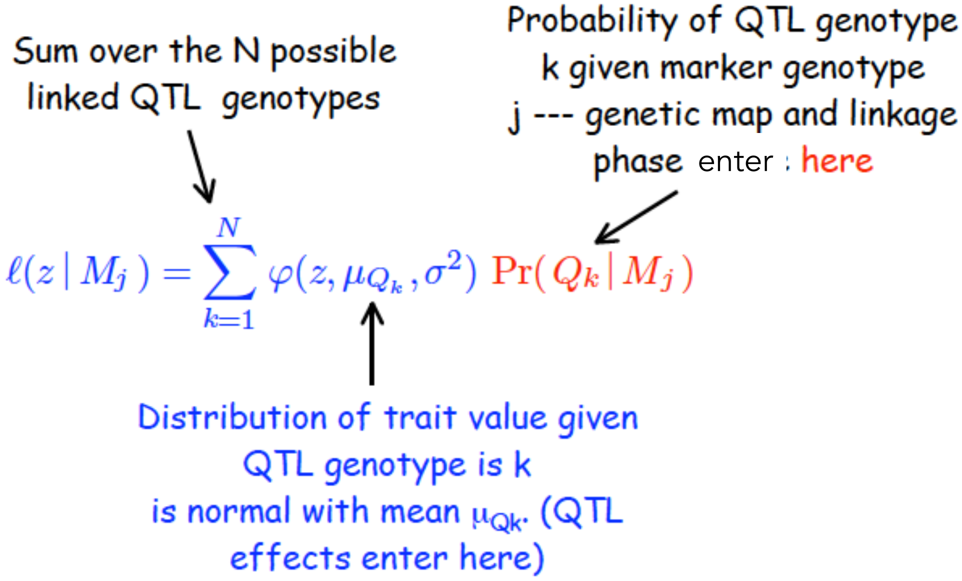
Basic likelihood function:

Trait value given marker genotype is type j

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

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

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

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

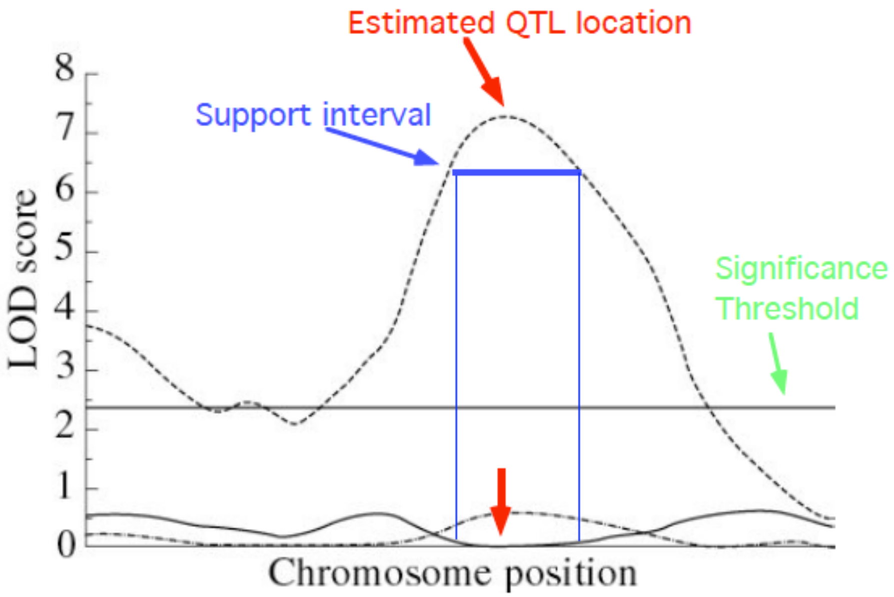
$$\mathrm{LR} = -2 \ln \frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z})}$$

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

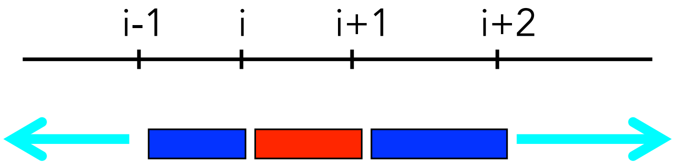
$$\mathrm{LOD}(c) = -\log_{10}\left[\frac{\max \ell_r(\mathbf{z})}{\max \ell(\mathbf{z}, c)}\right] = \frac{\mathrm{LR}(c)}{2 \ln 10} \simeq \frac{\mathrm{LR}(c)}{4.61}$$

A typical QTL map from a likelihood analysis



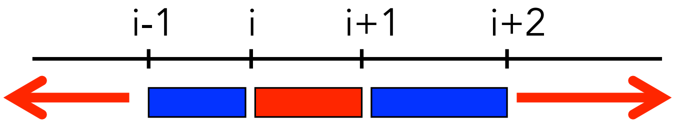
# Interval Mapping with Marker Cofactors

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

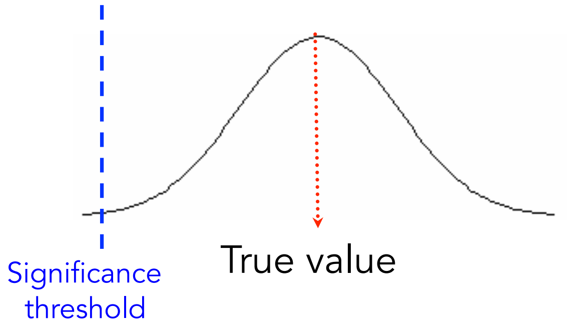
QTLs with low power of detection tend to have their effects *overestimated*, often very dramatically

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

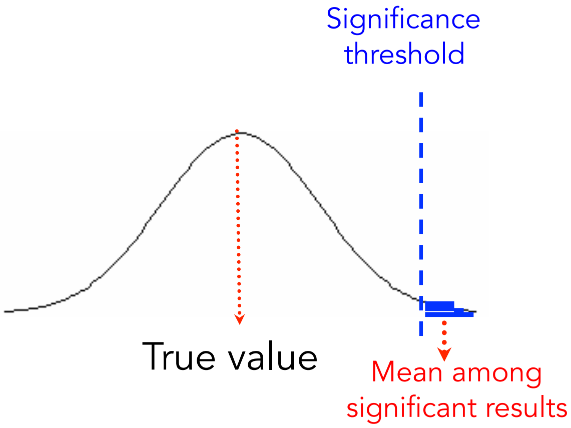
Also called the “winner’s curse” in the GWAS literature



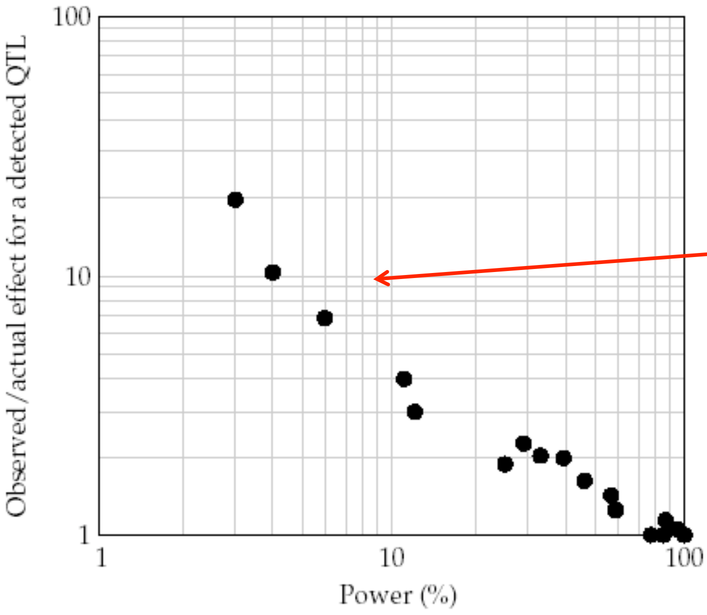
Distribution of
the realized value of an
effect in a sample

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

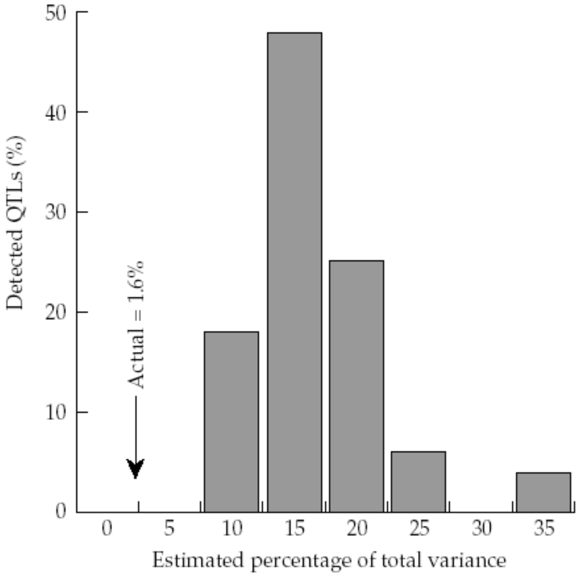


However, the mean of those declared significant is much larger than the true mean



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

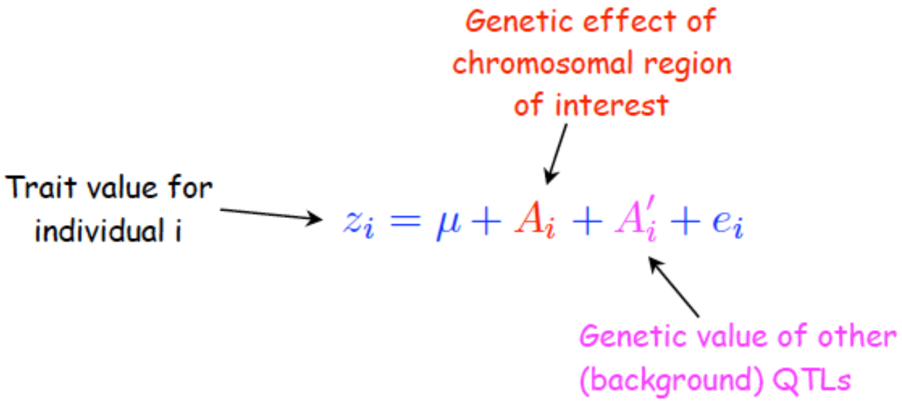
- Association mapping uses a set of very dense markers in a set of (largely) unrelated individuals
- Requires population level LD
- Allows for very fine mapping (1-20 kB)

# 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

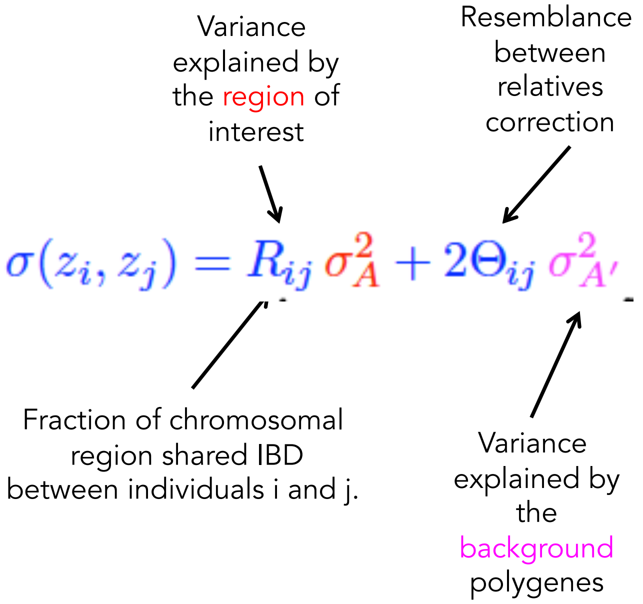
$$z_i = \mu + A_i + A_i' + e_i$$

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 $(1\text{-}0.0005)^t$ ~ 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%,

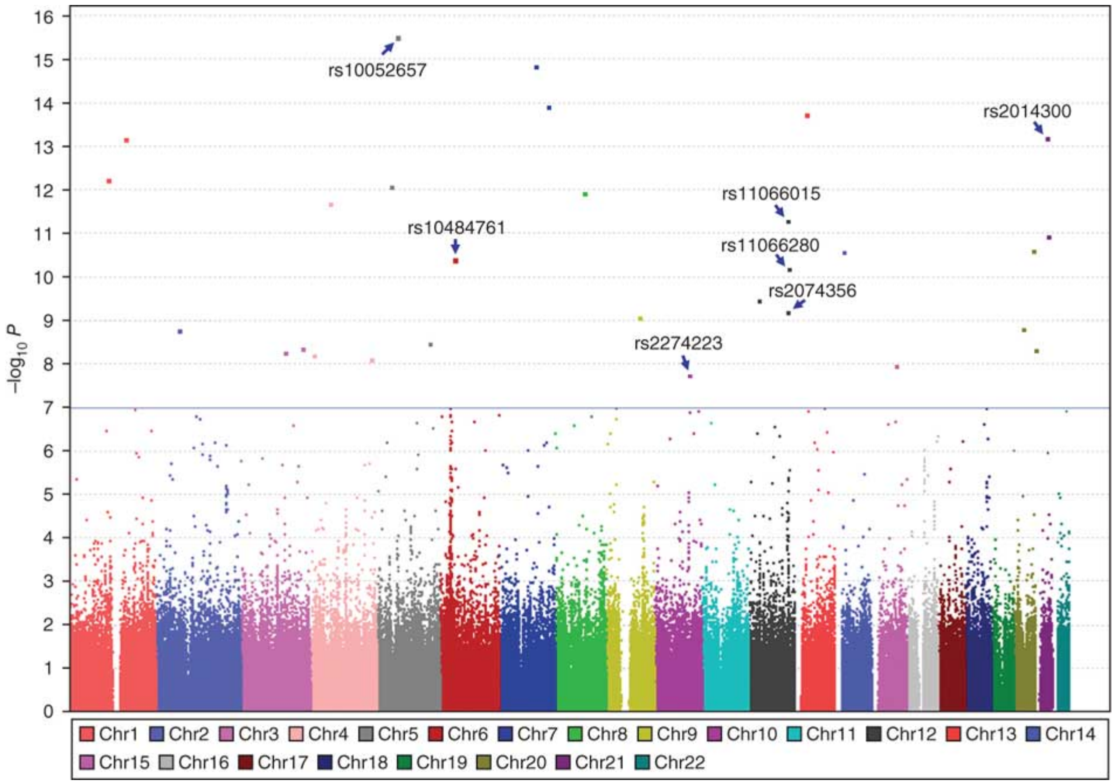
We thus have large samples and high disequilibrium, the recipe needed to detect linked QTLs of small effect

# Association mapping

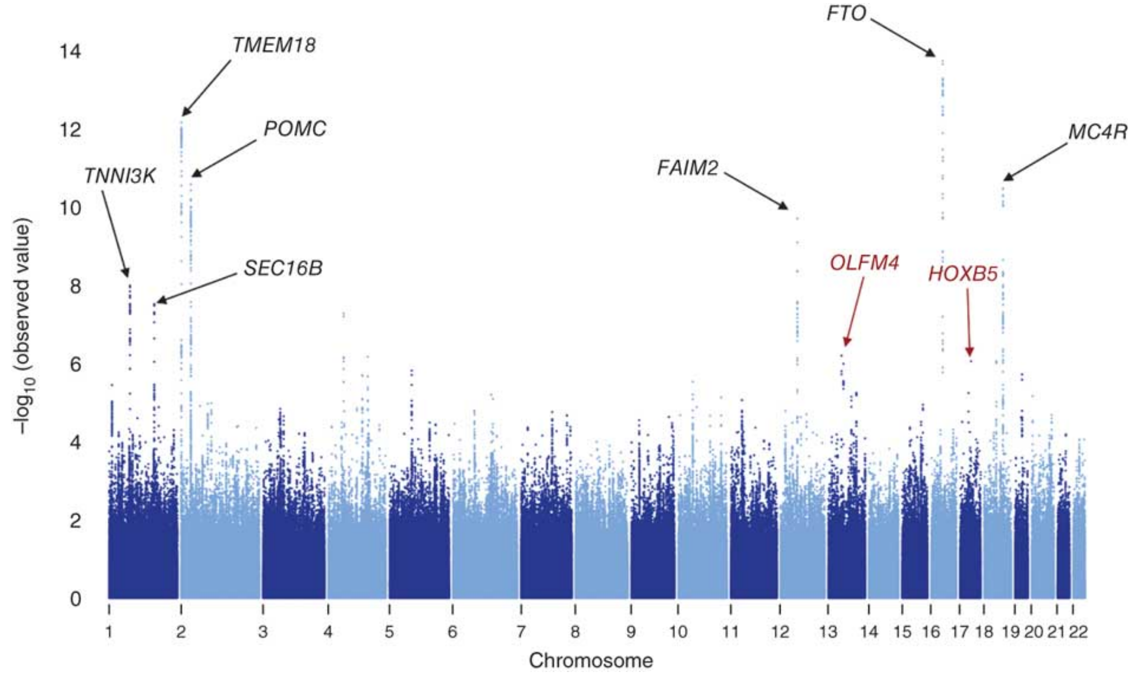
- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a 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
–log10 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



14
12
10
8
6
4
2
0
–log10 (observed value)
FTO
TMEM18
POMC
MC4R
TNNI3K
FAIM2
SEC16B
OLFM4
HOXB5
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.

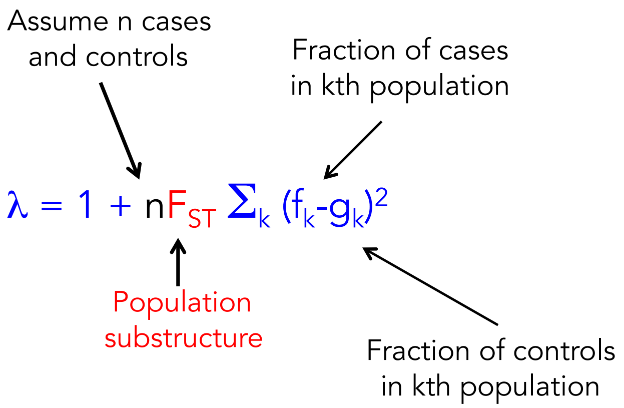
When population structure is present, the test statistic now follows a scaled $\chi^2$, so that if S is the test statistic, then $S/\lambda \sim \chi^2_1$ (so $S \sim \lambda\chi^2_1$)

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

$$\lambda = 1 + nF_{ST} \Sigma_k (f_k - g_k)^2$$

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).

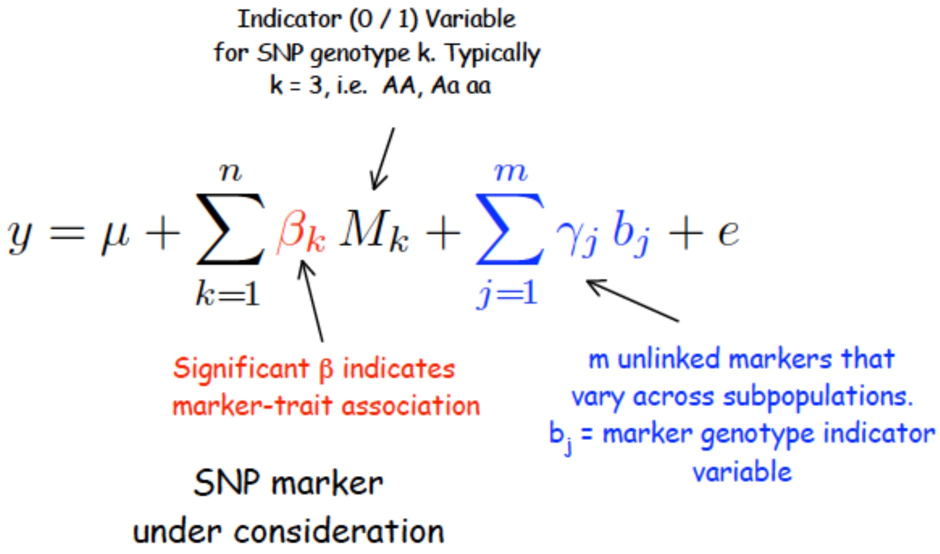
Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier

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.

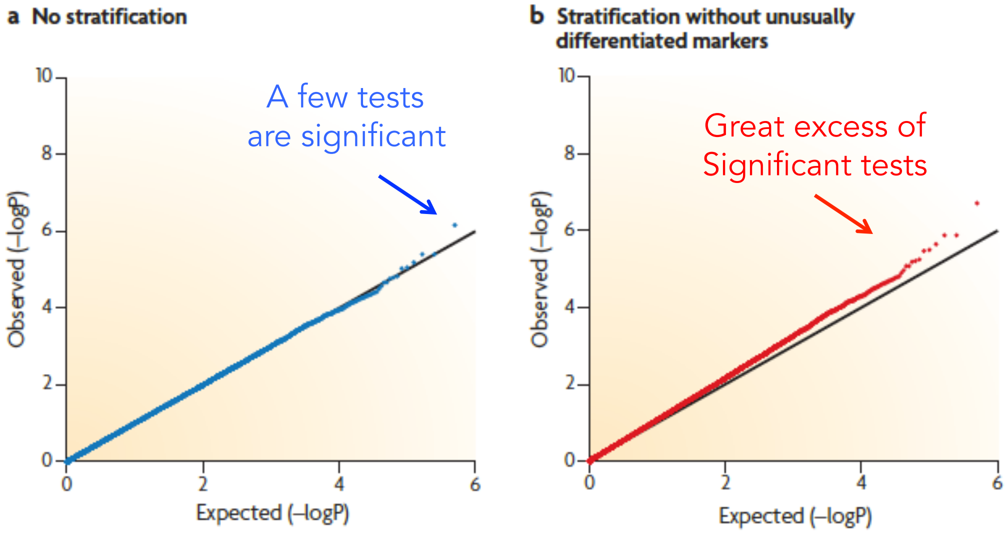
# GWAS Model diagnostics

## 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



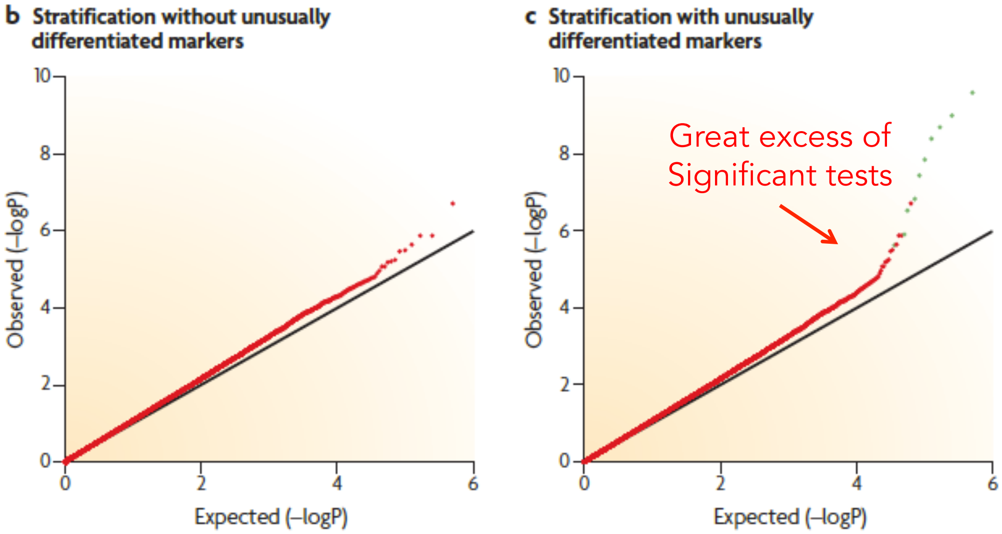
a No stratification

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

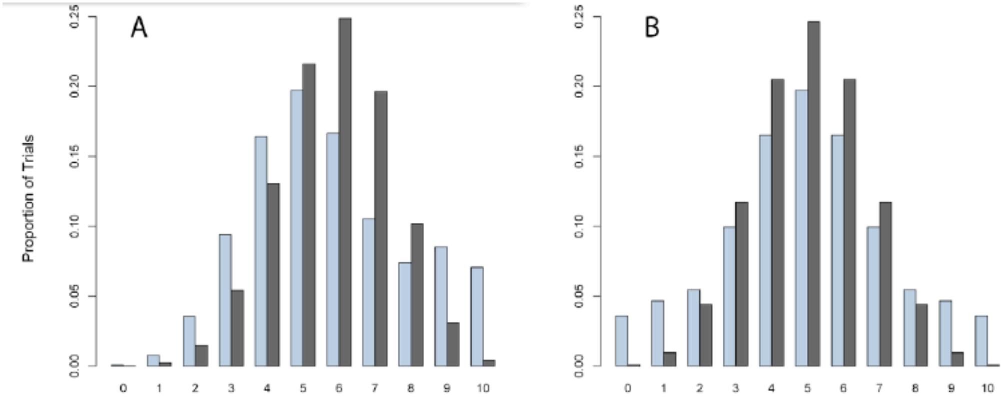
The case of the missing heritability

## 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, $p_1 = 1\text{-}p_0$ 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 ~ $\mathrm{Bin}(p_1,n_i)$
- 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 $n_i p_1(1\text{-}p_1)$

# C($\alpha$) test (cont).

- Suppose m variants in a region, test statistic is of the form
- $\Sigma_i (y_i - n_i p_1)^2 - n_i p_1 (1\text{-}p_1)$
- $y_i$ = 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

# Lecture 9: Using molecular markers to detect selection

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 13 – 15 July 2015

# Detecting selection

- Bottom line: looking for loci showing departures from the equilibrium neutral model
- What kinds of selection are of interest?
- Time scales and questions
- KEY POINTS
  - False positives very common
  - MOST selective events will not be detected
  - Those that are likely represent a rather biased sample

# Negative selection is common

- Negative (or purifying) selection is the removal of deleterious mutations by selection
- Leaves a strong signal throughout the genome
  - Faster substitution rates for silent vs. replacement codons
  - Comparative genomics equates strong sequence conservation (i.e., high negative selection) with strong functional constraints
  - The search for selection implies selection OTHER than negative

# Positive selection

- An allele increasing in frequency due to selection
  - Can either be a new mutation or a previously neutral/slightly deleterious allele whose fitness has changed due to a change in the environment.
  - Adaptation
- Balancing selection is when alternative alleles are favored by selection when rare
  - MHC, sickle-cell
- The "search for selection" is the search for signatures of positive, or balancing, selection

# Time scales of interest

- Ecological
  - An allele either currently undergoing selection or has VERY recently undergone selection
  - Detect using the nature of genetic variation within a population sample
  - Key: A SINGLE event can leave a signature
- Evolutionary
  - A gene or codon experiences REPEATED adaptive events over very long periods of time
  - Typically requires between-species divergence data
  - Key: Only informs us as to the long-term PATTERN of selection over a gene

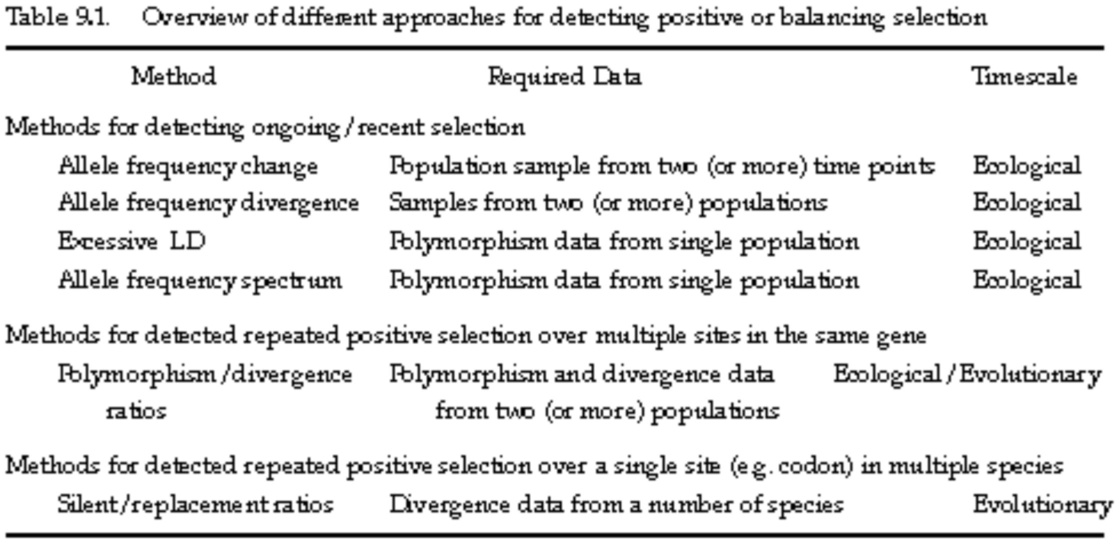


Table 9.1. Overview of different approaches for detecting positive or balancing selection

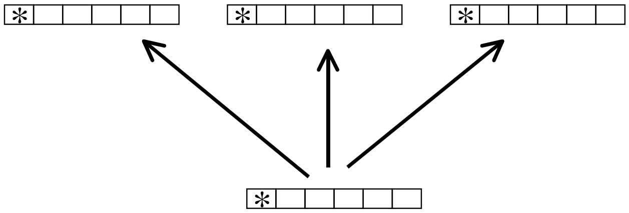
| Method | Required Data | Timescale |
|---|---|---|
| Methods for detecting ongoing / recent selection | | |
| Allele frequency change | Population sample from two (or more) time points | Ecological |
| Allele frequency divergence | Samples from two (or more) populations | Ecological |
| Excessive LD | Polymorphism data from single population | Ecological |
| Allele frequency spectrum | Polymorphism data from single population | Ecological |
| Methods for detected repeated positive selection over multiple sites in the same gene | | |
| Polymorphism / divergence ratios | Polymorphism and divergence data from two (or more) populations | Ecological / Evolutionary |
| Methods for detected repeated positive selection over a single site (e.g. codon) in multiple species | | |
| Silent/replacement ratios | Divergence data from a number of species | Evolutionary |

# Biased scan for selection

- Current/very recent selection at a single site requires rather strong selection to leave a signature.
  - Small shifts in allele frequencies at multiple sites unlikely to leave signatures
  - Very small time window (~0.1 Ne generations) to detect such an event once it has occurred.
- Recurrent selection
  - Phylogenic comparisons:  Multiple substitution events at the same CODON required for a signal
  - OK for "arms-race" genes, likely not typical

- Recurrent selection at sites OVER a gene
  - Comparing fixed differences between two species with the observed levels of polymorphism
  - Requires multiple substitutions at different codons (i.e., throughout the gene) for any signal
  - Hence, a few CRITICAL adaptive substitutions can occur in a gene and not leave a strong enough signal to detect
  - Power depends on the number of adaptive substitutions over the background level of neutral substitutions

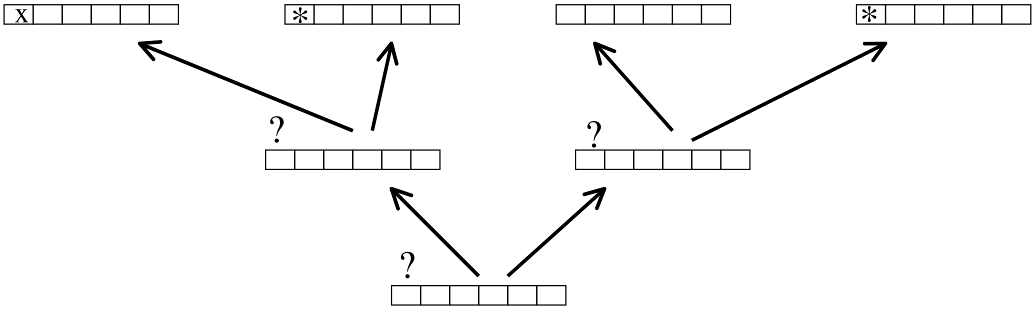
Sample of a gene from several individuals in the same population



Ongoing, or recent, selection

Detecting ongoing selection within a population. Requires a population sample, in which we look for inconsistencies of the pattern of variation from the equilibrium neutral model. Can detect on-going selection in a single region, influencing the pattern of variation at linked neutral loci.

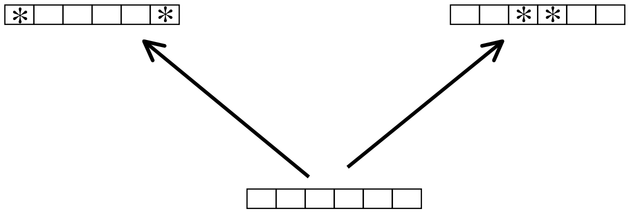
Sample of a gene over several species



Divergence data on a phylogeny.
Repeated positive selection at the same site

A phylogenic comparison of a sequence over a group of species is done on a codon-by-codon basis, looking for those with a higher replacement than site rate. Requires MULTIPLE substitutions at the same codon over the tree

Fixed differences between two species



Positive selection occurring over
multiple sites within the gene

Comparison of divergence data for a pair of species. Requires a background estimate of the expected divergence from fixation of neutral sites, which is provided from the polymorphism data (I'll cover this shortly).

# Key points

- Methods for detecting selection
  - Are prone to false-positives
    - The rejection of the null (equilibrium neutral model) can occur for reasons other the positive/balancing selection, such as changes in the population size
  - Are under-powered
    - Most selection events likely missed
  - Detect only specific types of selection events
    - Ongoing moderate to strong events
    - Repeated adaptive substitutions in a few codons over a phylogeny
    - Repeated adaptive substitutions over all sites in a gene

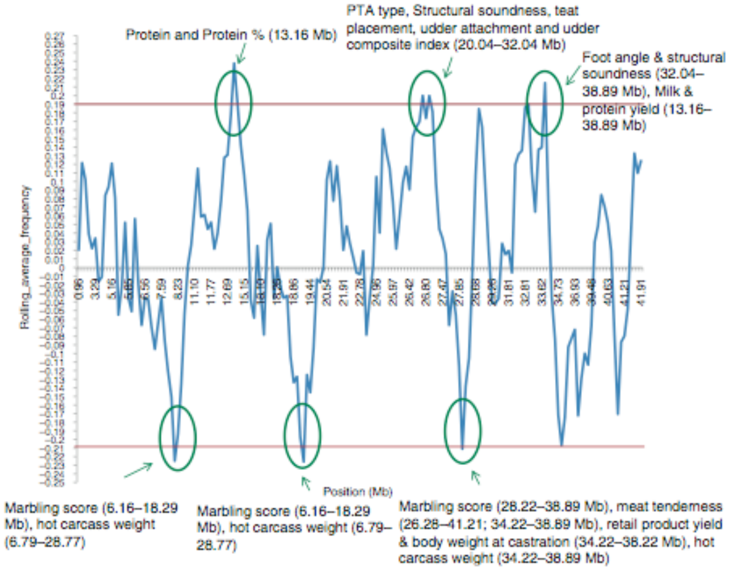
# Detecting on-going selection

- Excessive allele frequency change/divergence
- Selective Sweeps
  - Reduction in polymorphism around a selected site
- Shifts in the allele frequency spectrum
  - i.e., too many rare alleles
- Allelic age inconsistencies
  - Allele too common relative to its age
  - Excessive LD in a common allele

# Excess allele frequency change

- Logically, most straightforward
- Need estimates of $N_e$, time
- Need two (or more) time points
- Generally weak power unless selection strong or time between sampling long
- Example: Divergence between breeds selected for different goals

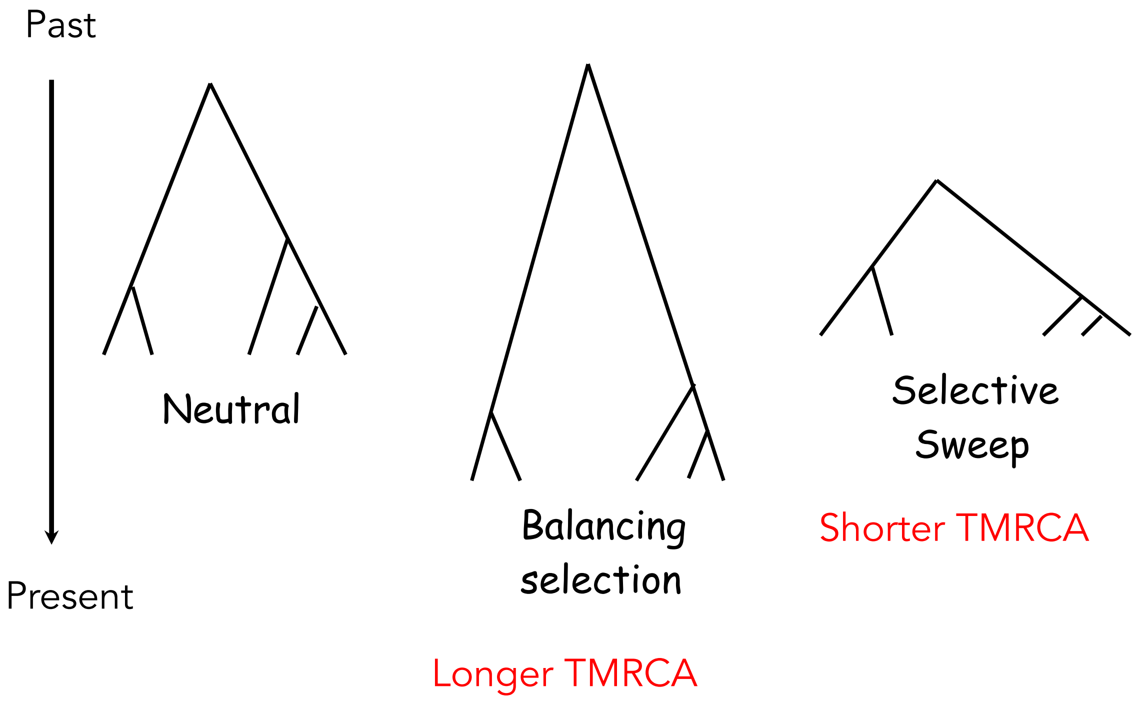
**Example 9.1.** Angus and Holstein represent breeds of *Bos taurus* that have been selected, respectively, for beef and milk production. As such, might would expect allele frequency differences between the breeds, some of which represent differential selection on milk and beef traits. Prasad et al. (2008) uses 355 SNP markers on chromosome 19 (BT19) and another 175 SNPs on chromosome 29 (BT29) to search for significant allele frequency differences between these breeds. They used a five marker sliding window, computing the difference between the mean allele frequency in Holsteins and the mean frequency in Angus. Significantly positive values indicate potential alleles selected for milk production, while significant negatives values suggests alleles potentially selected for beef production. Figure 9.1 shows the result for chromosome 19. The authors used a permutation test to access the significance, with the species label for any given marker randomly assigned, and the difference for each five-marker window scored, generating an empirical distribution under the null hypothesis of breed-effects. Deviations above the upper significance line show alleles at a significantly higher frequency in Holsteins and deviations below the lower significance line indicates alleles that are significantly more frequency in Angus. The authors were able to relate these locations to locations of QTLs for various milk and beef production traits. Example 9.8 discusses Hayes et al. (2008), who also examine allele frequency differences between these two breeds.


Protein and Protein % (13.16 Mb)
PTA type, Structural soundness, teat placement, udder attachment and udder composite index (20.04–32.04 Mb)
Foot angle & structural soundness (32.04–38.89 Mb), Milk & protein yield (13.16–38.89 Mb)
Rolling_average_frequency
0.27
0.19
0.1
0
-0.1
-0.2
-0.25
0.96
3.29
5.16
6.56
7.59
8.23
11.10
11.77
12.69
15.15
18.86
19.44
20.54
21.91
22.78
24.96
25.97
26.42
26.80
27.47
27.85
28.68
31.81
32.81
33.62
34.73
36.93
40.63
41.21
41.91
Position (Mb)
Marbling score (6.16–18.29 Mb), hot carcass weight (6.79–28.77)
Marbling score (6.16–18.29 Mb), hot carcass weight (6.79–28.77)
Marbling score (28.22–38.89 Mb), meat tenderness (26.28–41.21; 34.22–38.89 Mb), retail product yield & body weight at castration (34.22–38.22 Mb), hot carcass weight (34.22–38.89 Mb)

Five-marker window scans of difference between Holstein & Angus breeds (dairy vs. beef selection)

# Selective sweeps

- Classic visual tool to look for potential sites under selection
  - Common approach in the search for domestication genes
- Positive selection reduces Ne for linked sites
  - Reduces TMRCA and hence variation
- Balancing selection increases Ne for linked sites
  - Increases TMRCA and hence increase variation



Past

Present

Neutral

Balancing selection

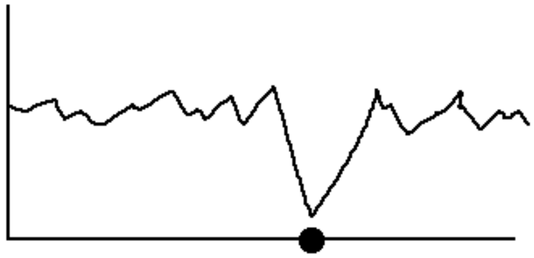
Longer TMRCA

Selective Sweep

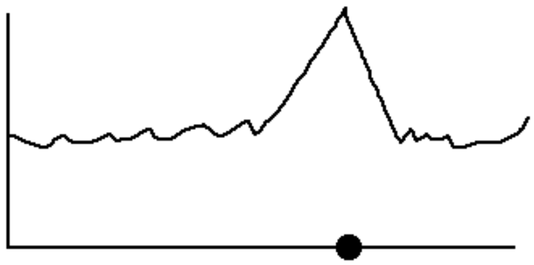
Shorter TMRCA

# Scanning for Sweeps

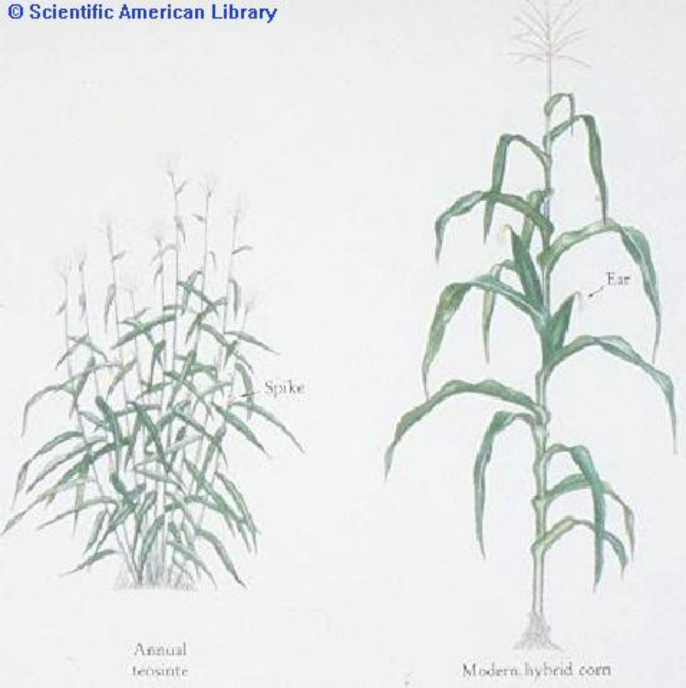
- Use a sliding window to look at variation along a chromosome (or around a candidate gene)
- Decrease (with respect to some standard) consistent with linked site under recent/ ongoing positive selection
- Increase consistent with balancing selection



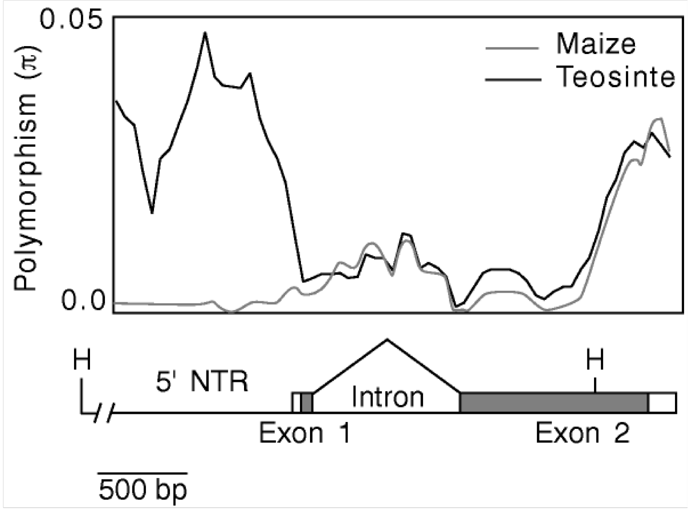
Signal of positive selection, OR reduction in mutation rate



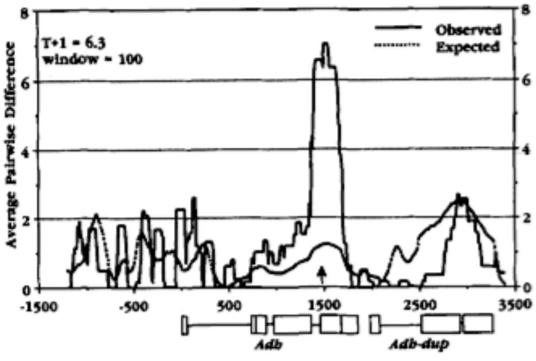
Signal of balancing selection, OR increase in mutation rate



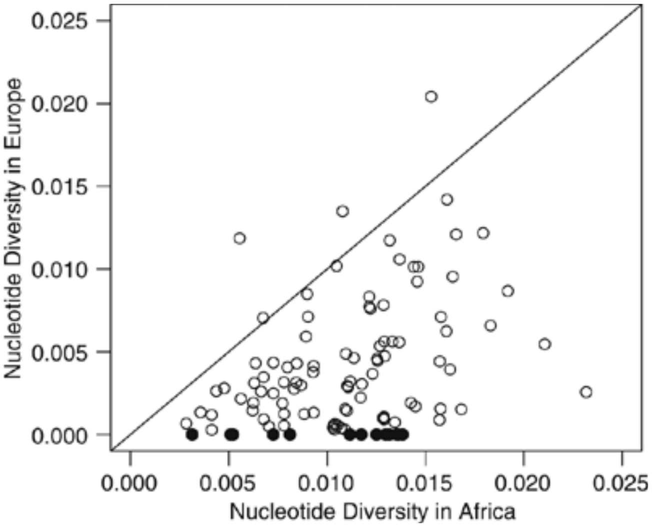
Domestication: Maize vs. teosinte



*tb1* in maize. Used teosinte as a control for expected background levels of variation



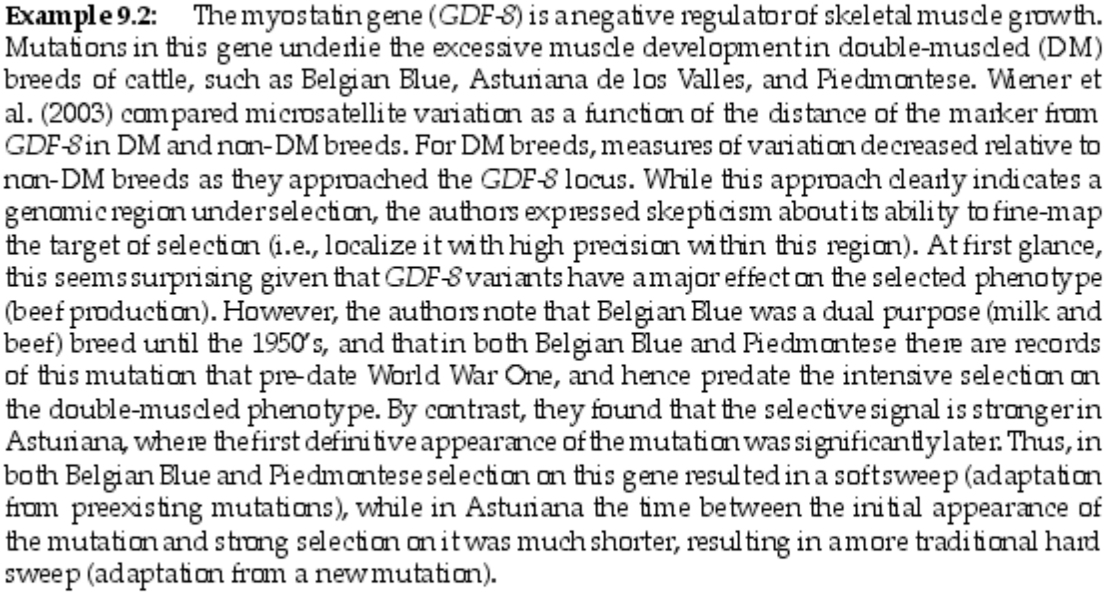
ADH in Drosophila.  Strong candidate for balancing selection of the Fast and Slow alleles, due to a single aa replacement at the location marked by the arrow



Scan of *Drosophila* genes in Africa (source population) and Europe (recently founded population).  Less diversity in Europe, but some loci (filled circles) strong candidates for a sweep

Double-muscle cattle:
Belgian blue

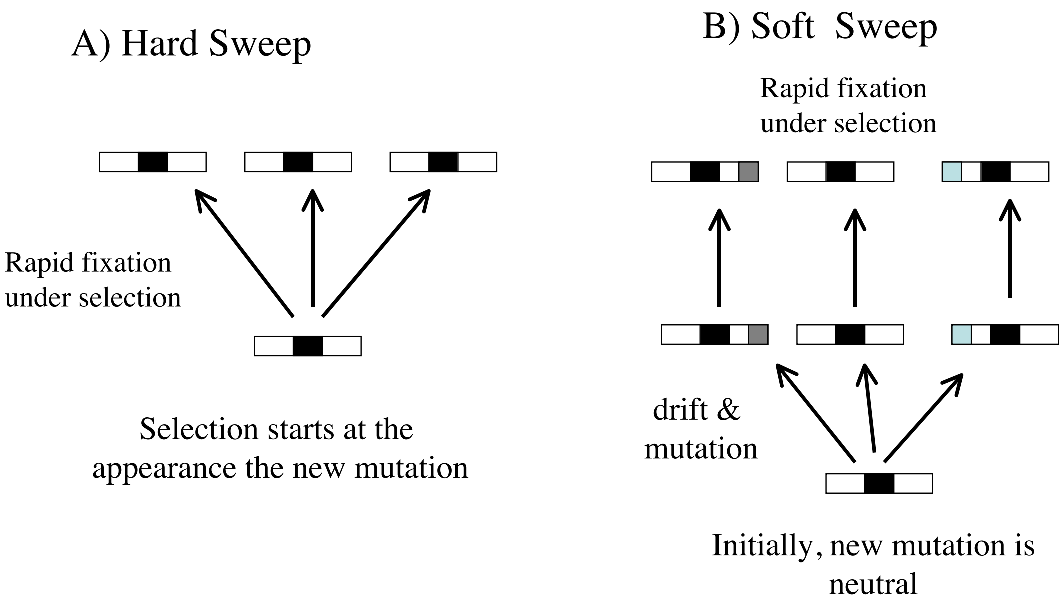
Reduction in microsatellite copy number variance often used



**Example 9.2:** The myostatin gene (*GDF-8*) is a negative regulator of skeletal muscle growth. Mutations in this gene underlie the excessive muscle development in double-muscled (DM) breeds of cattle, such as Belgian Blue, Asturiana de los Valles, and Piedmontese. Wiener et al. (2003) compared microsatellite variation as a function of the distance of the marker from *GDF-8* in DM and non-DM breeds. For DM breeds, measures of variation decreased relative to non-DM breeds as they approached the *GDF-8* locus. While this approach clearly indicates a genomic region under selection, the authors expressed skepticism about its ability to fine-map the target of selection (i.e., localize it with high precision within this region). At first glance, this seems surprising given that *GDF-8* variants have a major effect on the selected phenotype (beef production). However, the authors note that Belgian Blue was a dual purpose (milk and beef) breed until the 1950's, and that in both Belgian Blue and Piedmontese there are records of this mutation that pre-date World War One, and hence predate the intensive selection on the double-muscled phenotype. By contrast, they found that the selective signal is stronger in Asturiana, where the first definitive appearance of the mutation was significantly later. Thus, in both Belgian Blue and Piedmontese selection on this gene resulted in a soft sweep (adaptation from preexisting mutations), while in Asturiana the time between the initial appearance of the mutation and strong selection on it was much shorter, resulting in a more traditional hard sweep (adaptation from a new mutation).

# Issues with sweeps

- Need sufficient background variation before selection for a strong signal
  - Strong domestication event (e.g. sorghum) can remove most variation over entire genome
  - Inbreeding greatly reduces variation
- The signal persists for only a short time
  - ~ 0.1 Ne generations
  - Distance for effects roughly 0.01 s/c
- Sweep region often asymmetric around target site
- Hard sweeps can be detected, soft sweeps leave (at best) a weak signal



A) Hard Sweep

Rapid fixation under selection

Selection starts at the appearance the new mutation

B) Soft Sweep

Rapid fixation under selection

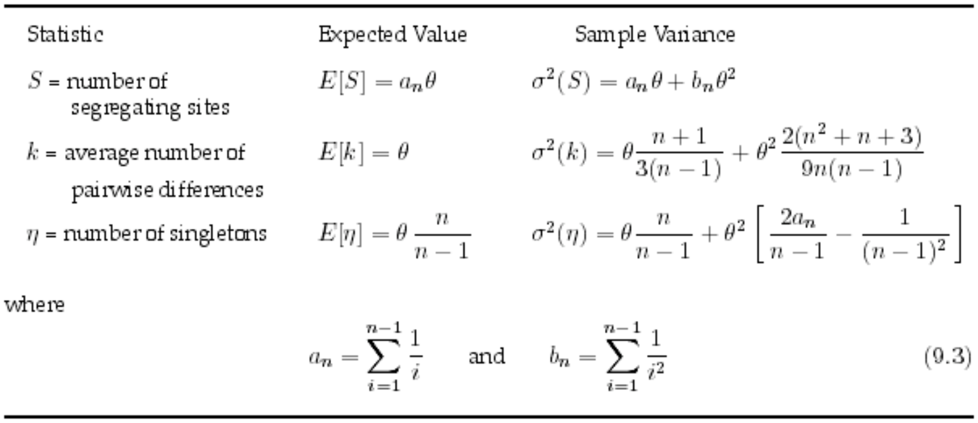
drift & mutation

Initially, new mutation is neutral

# Site frequency spectrum tests

- A large collection of tests based on comparing different measures of variation at a target site within a population sample
- Tajima's D is the classic
- Problem: significant result from either selection OR changes in population size/ structure (drift, mutation NOT at equilibrium)

Under the equilibrium neutral model, multiple ways to estimate $\theta = 4N_eu$ using different metrics of variation



| Statistic | Expected Value | Sample Variance |
|---|---|---|
| $S$ = number of segregating sites | $E[S] = a_n\theta$ | $\sigma^2(S) = a_n\theta + b_n\theta^2$ |
| $k$ = average number of pairwise differences | $E[k] = \theta$ | $\sigma^2(k) = \theta\dfrac{n+1}{3(n-1)} + \theta^2\dfrac{2(n^2+n+3)}{9n(n-1)}$ |
| $\eta$ = number of singletons | $E[\eta] = \theta\,\dfrac{n}{n-1}$ | $\sigma^2(\eta) = \theta\dfrac{n}{n-1} + \theta^2\left[\dfrac{2a_n}{n-1} - \dfrac{1}{(n-1)^2}\right]$ |

where

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i} \quad \text{and} \quad b_n = \sum_{i=1}^{n-1}\frac{1}{i^2} \tag{9.3}$$

$$\widehat{\theta}_S = \frac{S}{a_n}, \qquad \widehat{\theta}_k = k, \qquad \widehat{\theta}_\eta = \frac{n-1}{n}\,\eta$$

All should be consistent if model holds.

# Tajima’s D

$$D = \frac{\widehat{\theta}_k - \widehat{\theta}_S}{\sqrt{\alpha_D S + \beta_D S^2}}$$

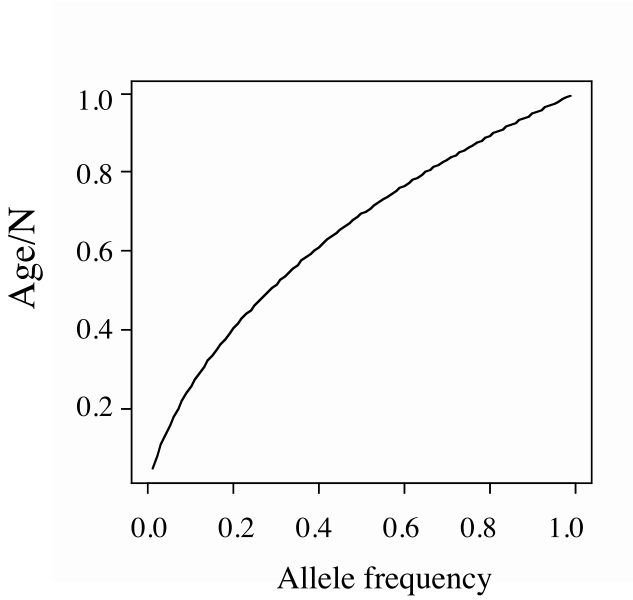
$$\alpha_D = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) - \beta_D$$

$$\beta_D = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right)$$

Negative value:  excess number of rare alleles consistent with either positive selection OR expanding population size

Positive value:  excess number of common alleles consistent with either balancing selection OR Population subdivision

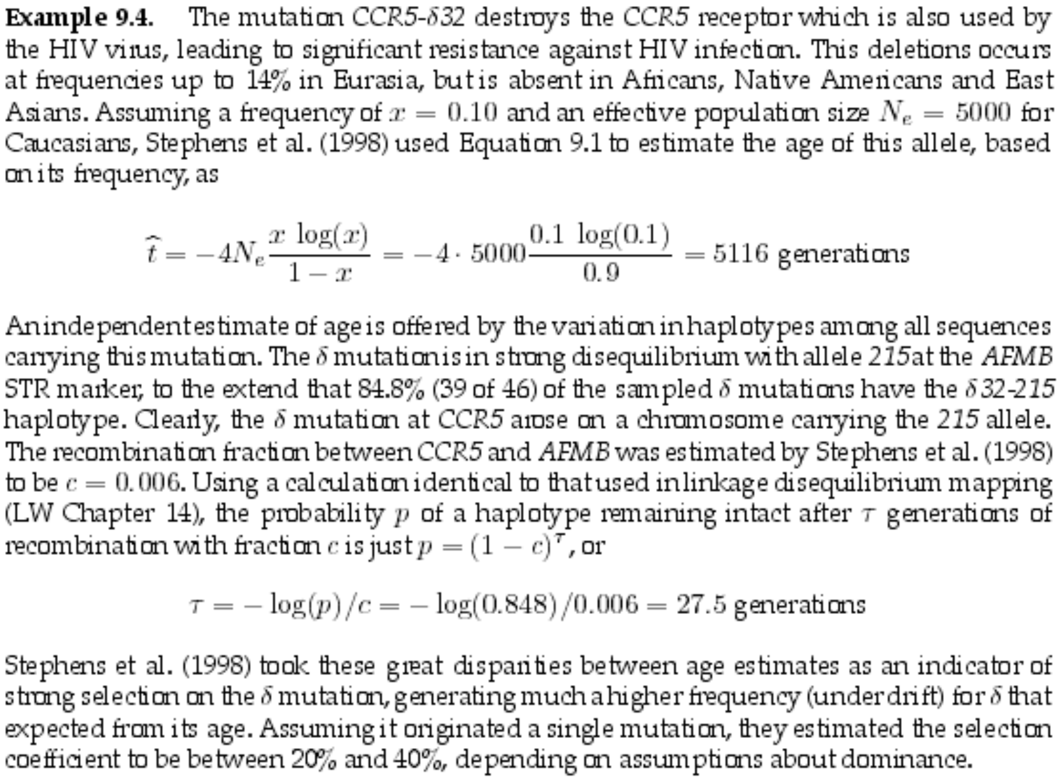
# Consistency of allelic age



$$E(t) = -4N\frac{x}{1-x}\ln(x)$$

Under drift, a common allele is an old allele

Common alleles should not be young



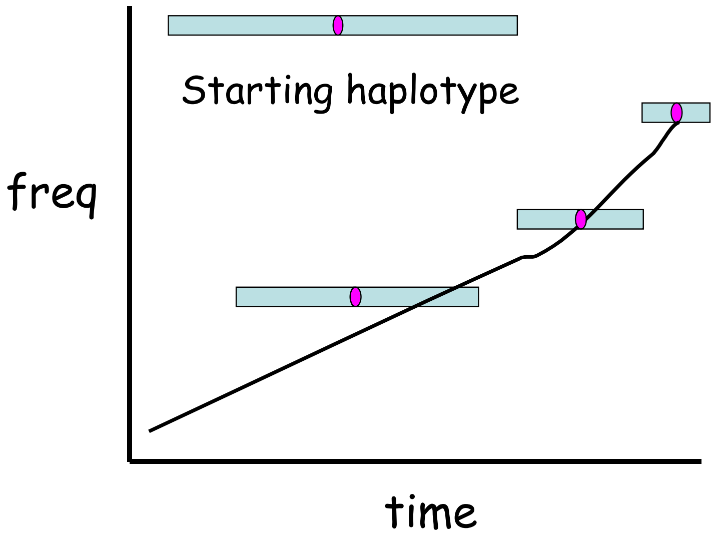
**Example 9.4.** The mutation *CCR5-δ32* destroys the *CCR5* receptor which is also used by the HIV virus, leading to significant resistance against HIV infection. This deletions occurs at frequencies up to 14% in Eurasia, but is absent in Africans, Native Americans and East Asians. Assuming a frequency of $x = 0.10$ and an effective population size $N_e = 5000$ for Caucasians, Stephens et al. (1998) used Equation 9.1 to estimate the age of this allele, based on its frequency, as

$$\widehat{t} = -4N_e \frac{x\ \log(x)}{1-x} = -4 \cdot 5000 \frac{0.1\ \log(0.1)}{0.9} = 5116 \text{ generations}$$

An independent estimate of age is offered by the variation in haplotypes among all sequences carrying this mutation. The $\delta$ mutation is in strong disequilibrium with allele *215* at the *AFMB* STR marker, to the extend that 84.8% (39 of 46) of the sampled $\delta$ mutations have the *δ32-215* haplotype. Clearly, the $\delta$ mutation at *CCR5* arose on a chromosome carrying the *215* allele. The recombination fraction between *CCR5* and *AFMB* was estimated by Stephens et al. (1998) to be $c = 0.006$. Using a calculation identical to that used in linkage disequilibrium mapping (LW Chapter 14), the probability $p$ of a haplotype remaining intact after $\tau$ generations of recombination with fraction $c$ is just $p = (1-c)^\tau$, or

$$\tau = -\log(p)/c = -\log(0.848)/0.006 = 27.5 \text{ generations}$$

Stephens et al. (1998) took these great disparities between age estimates as an indicator of strong selection on the $\delta$ mutation, generating much a higher frequency (under drift) for $\delta$ that expected from its age. Assuming it originated a single mutation, they estimated the selection coefficient to be between 20% and 40%, depending on assumptions about dominance.



Common alleles should have short haplotypes under drift -- longer time for recombination to act

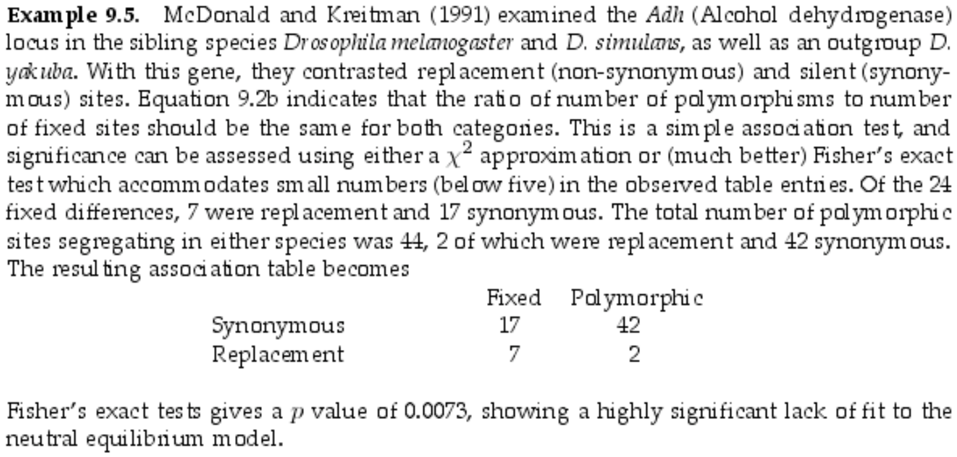
Common alleles with long haplotypes --- good signal for selection, rather robust to demography

# Joint polymorphism-divergence tests

- HKA, McDonald-Kreitman (MK) tests
  - MK test is rather robust to demographic issues
- Require polymorphism data from one (or more) species, divergence data btw species
- Look at ratio of divergence to polymorphism

$$H_i = 4N_e\mu_i, \qquad d_i = 2t\mu_i$$

$$\frac{H_i}{d_i} = \frac{4N_e\mu_i}{2t\mu_i} = \frac{2N_e}{t}$$



**Example 9.5.** McDonald and Kreitman (1991) examined the *Adh* (Alcohol dehydrogenase) locus in the sibling species *Drosophila melanogaster* and *D. simulans*, as well as an outgroup *D. yakuba*. With this gene, they contrasted replacement (non-synonymous) and silent (synonymous) sites. Equation 9.2b indicates that the ratio of number of polymorphisms to number of fixed sites should be the same for both categories. This is a simple association test, and significance can be assessed using either a $\chi^2$ approximation or (much better) Fisher's exact test which accommodates small numbers (below five) in the observed table entries. Of the 24 fixed differences, 7 were replacement and 17 synonymous. The total number of polymorphic sites segregating in either species was 44, 2 of which were replacement and 42 synonymous. The resulting association table becomes

| | Fixed | Polymorphic |
|---|---|---|
| Synonymous | 17 | 42 |
| Replacement | 7 | 2 |

Fisher's exact tests gives a $p$ value of 0.0073, showing a highly significant lack of fit to the neutral equilibrium model.

Cool feature: can estimate # of adaptive substitutions

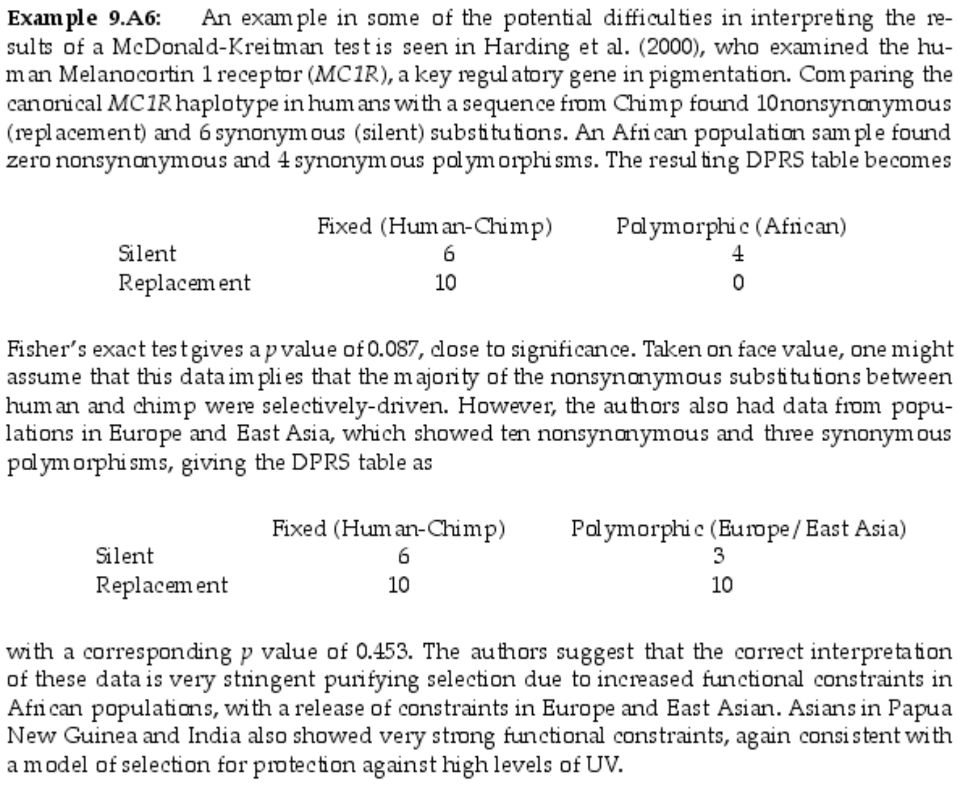
= 7 - 17(2/42) = 6

Robust to most demographic issues

However, replacement polymorphic sites can overestimate neutral rate due to deleterious alleles segregating

# Strengths and weaknesses

- Only detects a pattern of adaptive substitutions at a gene.
  - Require multiple events to have any power
  - Can’t tell which replacements were selectively-driven
- MK test robust to many demographic issues, but NOT fool-proof
  - Any change in the constraints between processes generating polymorphisms and processes generating divergence can be regarded as evidence for selection



**Example 9.A6:** An example in some of the potential difficulties in interpreting the results of a McDonald-Kreitman test is seen in Harding et al. (2000), who examined the human Melanocortin 1 receptor (*MC1R*), a key regulatory gene in pigmentation. Comparing the canonical *MC1R* haplotype in humans with a sequence from Chimp found 10nonsynonymous (replacement) and 6 synonymous (silent) substitutions. An African population sample found zero nonsynonymous and 4 synonymous polymorphisms. The resulting DPRS table becomes

| | Fixed (Human-Chimp) | Polymorphic (African) |
|---|---|---|
| Silent | 6 | 4 |
| Replacement | 10 | 0 |

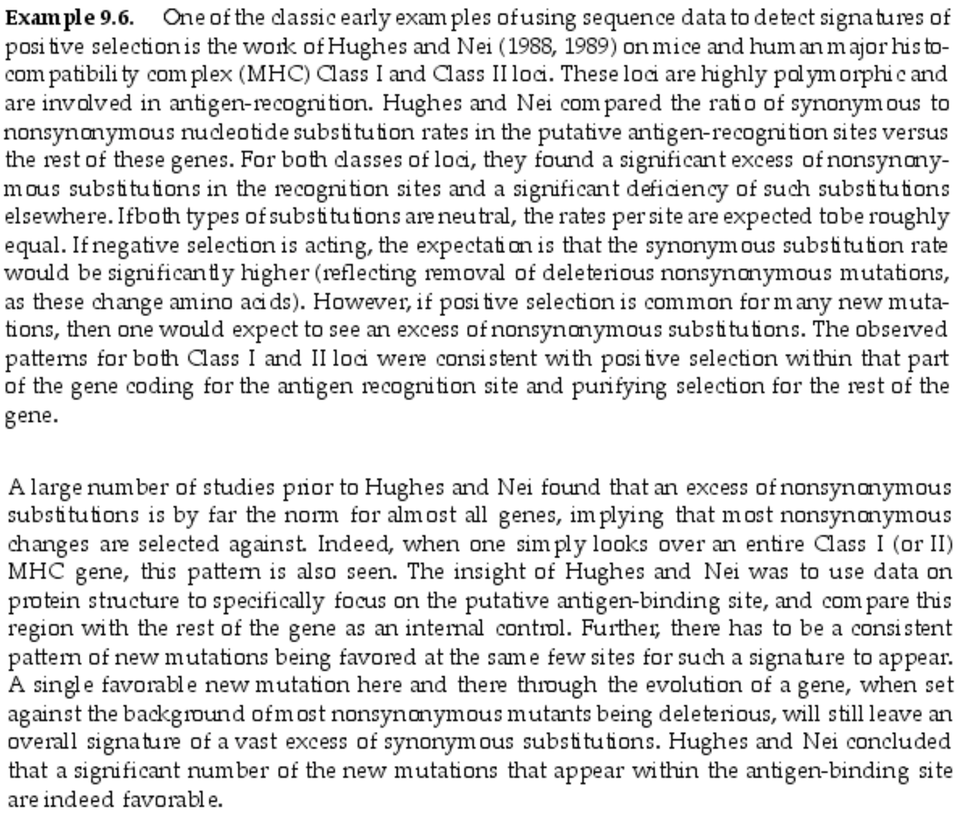
Fisher’s exact test gives a $p$ value of 0.087, close to significance. Taken on face value, one might assume that this data implies that the majority of the nonsynonymous substitutions between human and chimp were selectively-driven. However, the authors also had data from populations in Europe and East Asia, which showed ten nonsynonymous and three synonymous polymorphisms, giving the DPRS table as

| | Fixed (Human-Chimp) | Polymorphic (Europe/East Asia) |
|---|---|---|
| Silent | 6 | 3 |
| Replacement | 10 | 10 |

with a corresponding $p$ value of 0.453. The authors suggest that the correct interpretation of these data is very stringent purifying selection due to increased functional constraints in African populations, with a release of constraints in Europe and East Asian. Asians in Papua New Guinea and India also showed very strong functional constraints, again consistent with a model of selection for protection against high levels of UV.

# $K_A/K_s$ tests

- THE classic test for selection, requiring gene sequences over a known phylogeny
  - – $K_A$ = replacement substitution rate
  - – $K_s$ = silent substitution rate
    - Neutral proxy
  - – $\omega = K_A/K_s$
- $\omega > 1$: positive selection.
  - – Problem: most codons have $K_s > K_A$, so that even with repeated adaptive substitutions throughout a gene, signal still swamped.



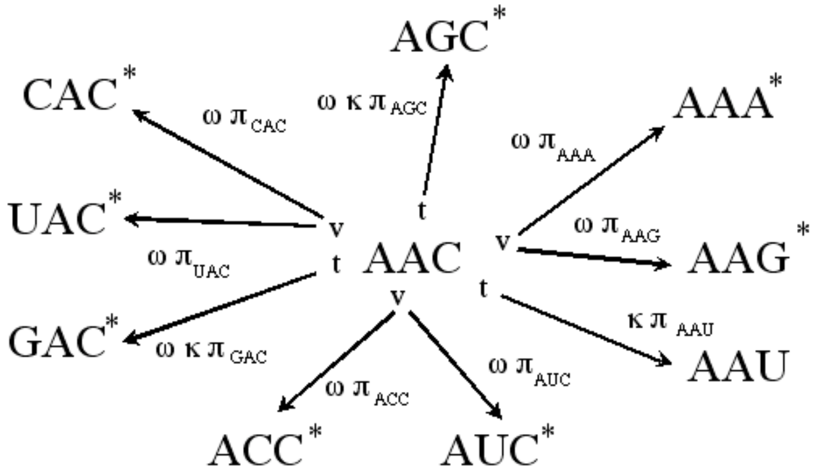
**Example 9.6.** One of the classic early examples of using sequence data to detect signatures of positive selection is the work of Hughes and Nei (1988, 1989) on mice and human major histocompatibility complex (MHC) Class I and Class II loci. These loci are highly polymorphic and are involved in antigen-recognition. Hughes and Nei compared the ratio of synonymous to nonsynonymous nucleotide substitution rates in the putative antigen-recognition sites versus the rest of these genes. For both classes of loci, they found a significant excess of nonsynonymous substitutions in the recognition sites and a significant deficiency of such substitutions elsewhere. If both types of substitutions are neutral, the rates per site are expected to be roughly equal. If negative selection is acting, the expectation is that the synonymous substitution rate would be significantly higher (reflecting removal of deleterious nonsynonymous mutations, as these change amino acids). However, if positive selection is common for many new mutations, then one would expect to see an excess of nonsynonymous substitutions. The observed patterns for both Class I and II loci were consistent with positive selection within that part of the gene coding for the antigen recognition site and purifying selection for the rest of the gene.

A large number of studies prior to Hughes and Nei found that an excess of nonsynonymous substitutions is by far the norm for almost all genes, implying that most nonsynonymous changes are selected against. Indeed, when one simply looks over an entire Class I (or II) MHC gene, this pattern is also seen. The insight of Hughes and Nei was to use data on protein structure to specifically focus on the putative antigen-binding site, and compare this region with the rest of the gene as an internal control. Further, there has to be a consistent pattern of new mutations being favored at the same few sites for such a signature to appear. A single favorable new mutation here and there through the evolution of a gene, when set against the background of most nonsynonymous mutants being deleterious, will still leave an overall signature of a vast excess of synonymous substitutions. Hughes and Nei concluded that a significant number of the new mutations that appear within the antigen-binding site are indeed favorable.

# Codon-based models

- The way around this problem is to analyze a gene on a codon-by-codon basis
  - Such codon-based models assign all (nonstop) codons a value from 1 to 61
  - A model of transition probabilities between all one-nucleotide transitions is constructed
  - Maximum likelihood used to estimate parameters
  - Model with $\omega = 1$ over all codons contrasted with a model where $\omega > 1$ at some (unspecified) set of codons.

$$q_{ij} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega\pi_j & \text{for a nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases} \quad \text{for } 1 \le i, j \le 61$$



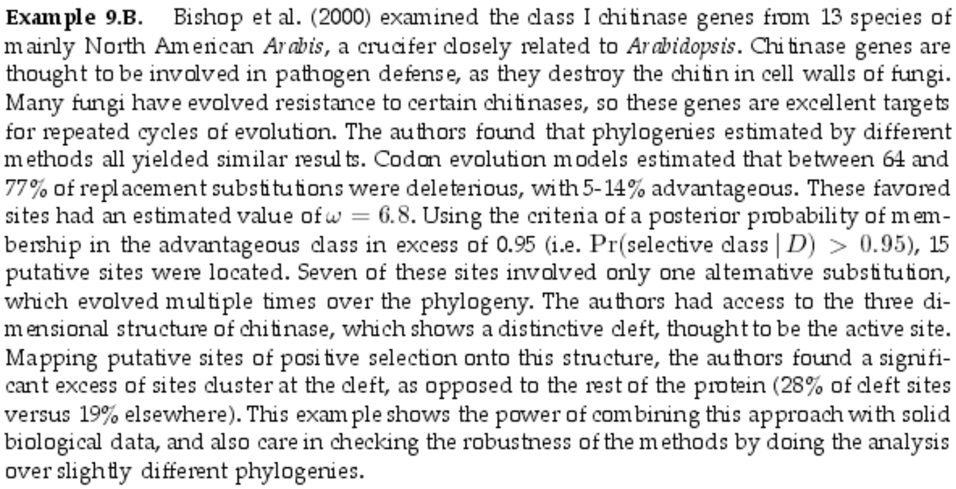
Model easily expanded to allow for several classes of codons

$$q_{ij}^{(k)} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega^{(k)}\pi_j & \text{for a nonsynonymous transversion} \\ \omega^{(k)}\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases}$$

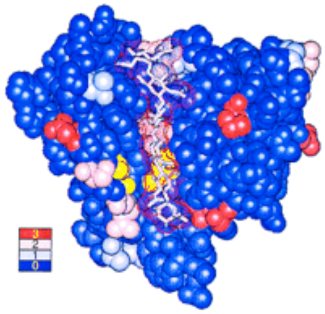
$$\omega^{(k)} = \begin{cases} 0 & \text{deleterious class} \\ 1 & \text{neutral class} \\ \omega > 1 & \text{positively-selected class} \end{cases}$$

Can use Baye's theorem to assign posterior probabilities that a given codon is in a given class (i.e., localize sites of repeated positive selection

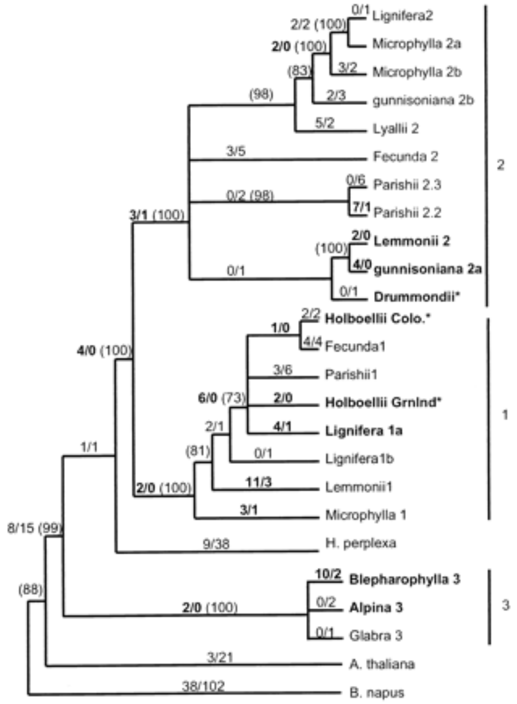
$$\Pr(\text{class } i \mid D) = \frac{\Pr(D \mid \omega_i)\Pr(\text{class } i)}{\Pr(D)} = \frac{\Pr(D \mid \omega_i)\Pr(\text{class } i)}{\sum_{i=1}^{k} \Pr(D \mid \omega_i)\Pr(\text{class } i)}$$



**Example 9.B.** Bishop et al. (2000) examined the class I chitinase genes from 13 species of mainly North American *Arabis*, a crucifer closely related to *Arabidopsis*. Chitinase genes are thought to be involved in pathogen defense, as they destroy the chitin in cell walls of fungi. Many fungi have evolved resistance to certain chitinases, so these genes are excellent targets for repeated cycles of evolution. The authors found that phylogenies estimated by different methods all yielded similar results. Codon evolution models estimated that between 64 and 77% of replacement substitutions were deleterious, with 5-14% advantageous. These favored sites had an estimated value of $\omega = 6.8$. Using the criteria of a posterior probability of membership in the advantageous class in excess of 0.95 (i.e. $\Pr(\text{selective class} \mid D) > 0.95$), 15 putative sites were located. Seven of these sites involved only one alternative substitution, which evolved multiple times over the phylogeny. The authors had access to the three dimensional structure of chitinase, which shows a distinctive cleft, thought to be the active site. Mapping putative sites of positive selection onto this structure, the authors found a significant excess of sites cluster at the cleft, as opposed to the rest of the protein (28% of cleft sites versus 19% elsewhere). This example shows the power of combining this approach with solid biological data, and also care in checking the robustness of the methods by doing the analysis over slightly different phylogenies.



Class I Chitinase *(Arabis)*



# Strengths and weaknesses

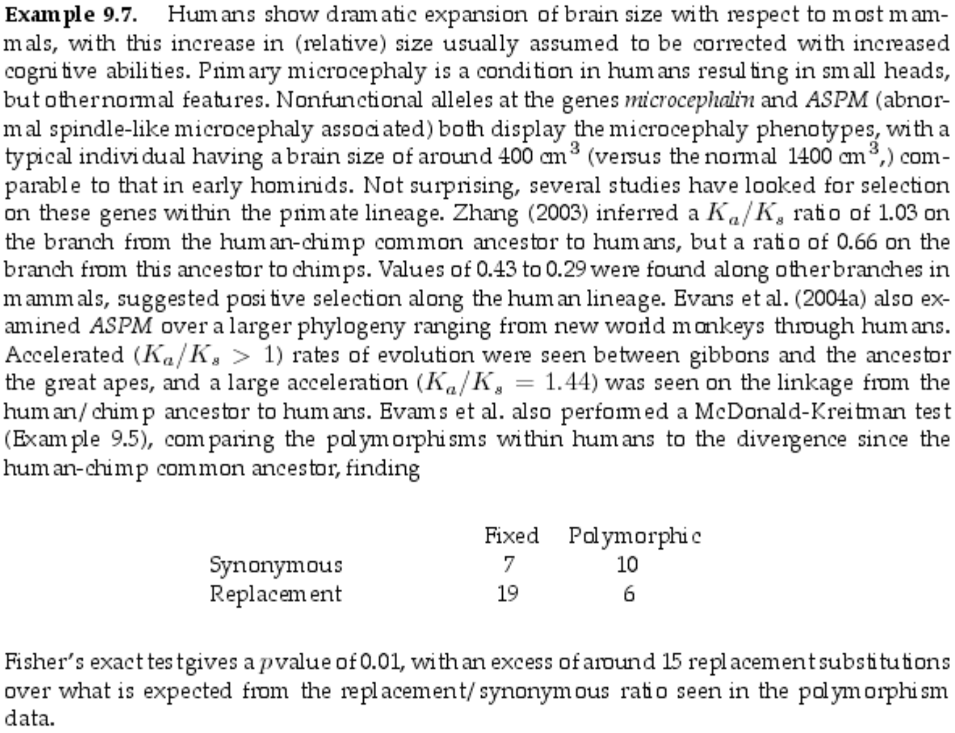
- Strengths
  - Can assign repeated selection to SPECIFIC codons
  - Requires only single sequences for each species
- Weaknesses:
  - Models can be rather delicate
  - Can only detect repeated selection at particular codons, NOT throughout a gene

The spandrels of San Marco (Gould and Lewontin 1979)

Very elaborate structure DOES not imply function nor adaptation

# Structure vs. function

- Molecular biologists are largely conditioned to look for function through structure
- Problem:  elaborate structures can serve little function
- Cannot simply assume an adaptive explanation because the structure is complex

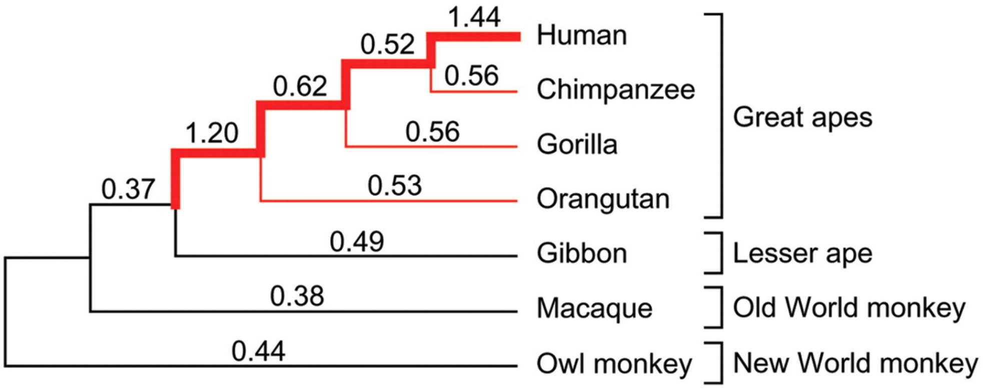


**Example 9.7.** Humans show dramatic expansion of brain size with respect to most mammals, with this increase in (relative) size usually assumed to be corrected with increased cognitive abilities. Primary microcephaly is a condition in humans resulting in small heads, but other normal features. Nonfunctional alleles at the genes *microcephalin* and *ASPM* (abnormal spindle-like microcephaly associated) both display the microcephaly phenotypes, with a typical individual having a brain size of around 400 cm$^3$ (versus the normal 1400 cm$^3$,) comparable to that in early hominids. Not surprising, several studies have looked for selection on these genes within the primate lineage. Zhang (2003) inferred a $K_a/K_s$ ratio of 1.03 on the branch from the human-chimp common ancestor to humans, but a ratio of 0.66 on the branch from this ancestor to chimps. Values of 0.43 to 0.29 were found along other branches in mammals, suggested positive selection along the human lineage. Evans et al. (2004a) also examined *ASPM* over a larger phylogeny ranging from new world monkeys through humans. Accelerated ($K_a/K_s > 1$) rates of evolution were seen between gibbons and the ancestor the great apes, and a large acceleration ($K_a/K_s = 1.44$) was seen on the linkage from the human/chimp ancestor to humans. Evams et al. also performed a McDonald-Kreitman test (Example 9.5), comparing the polymorphisms within humans to the divergence since the human-chimp common ancestor, finding

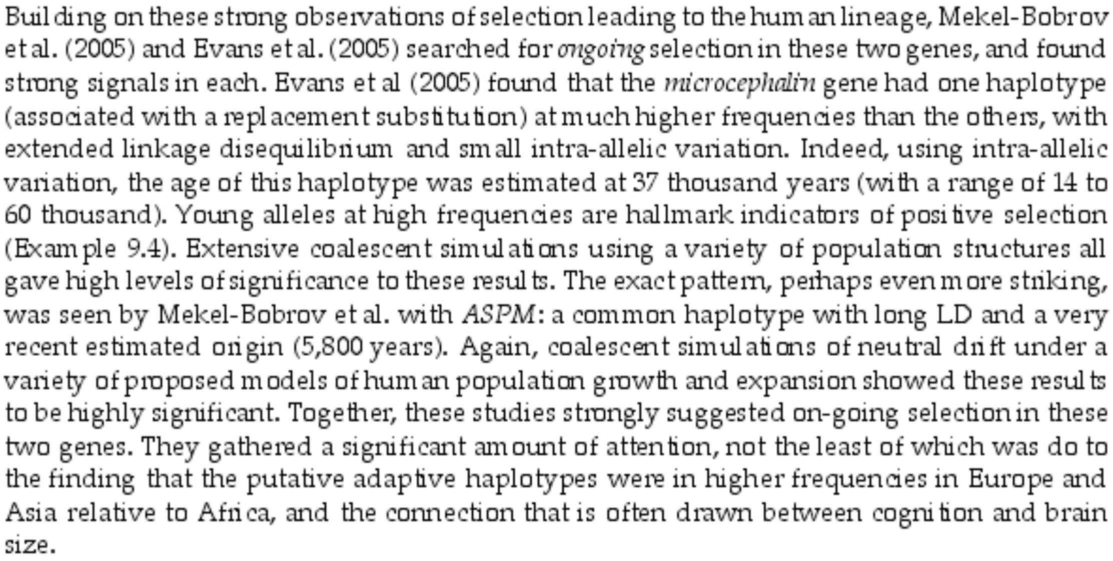
| | Fixed | Polymorphic |
|---|---|---|
| Synonymous | 7 | 10 |
| Replacement | 19 | 6 |

Fisher's exact test gives a $p$ value of 0.01, with an excess of around 15 replacement substitutions over what is expected from the replacement/synonymous ratio seen in the polymorphism data.

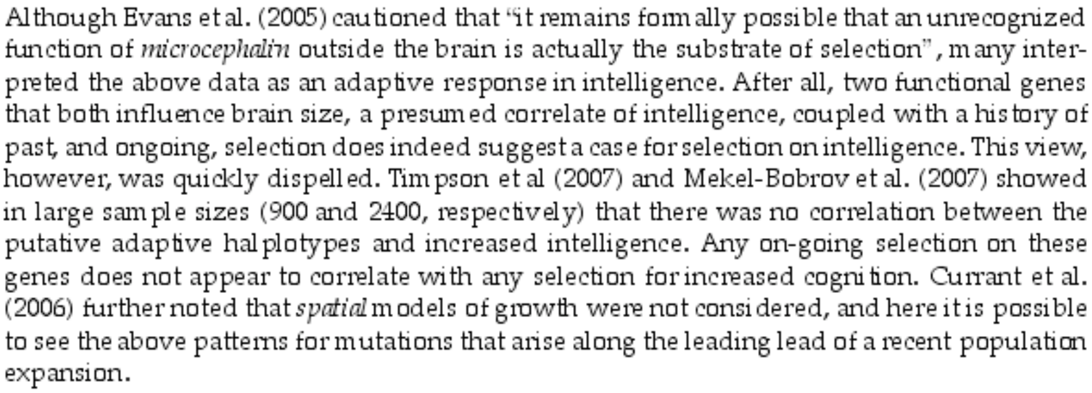
ω values shown on braches



ASPM



Building on these strong observations of selection leading to the human lineage, Mekel-Bobrov et al. (2005) and Evans et al. (2005) searched for *ongoing* selection in these two genes, and found strong signals in each. Evans et al (2005) found that the *microcephalin* gene had one haplotype (associated with a replacement substitution) at much higher frequencies than the others, with extended linkage disequilibrium and small intra-allelic variation. Indeed, using intra-allelic variation, the age of this haplotype was estimated at 37 thousand years (with a range of 14 to 60 thousand). Young alleles at high frequencies are hallmark indicators of positive selection (Example 9.4). Extensive coalescent simulations using a variety of population structures all gave high levels of significance to these results. The exact pattern, perhaps even more striking, was seen by Mekel-Bobrov et al. with *ASPM*: a common haplotype with long LD and a very recent estimated origin (5,800 years). Again, coalescent simulations of neutral drift under a variety of proposed models of human population growth and expansion showed these results to be highly significant. Together, these studies strongly suggested on-going selection in these two genes. They gathered a significant amount of attention, not the least of which was do to the finding that the putative adaptive haplotypes were in higher frequencies in Europe and Asia relative to Africa, and the connection that is often drawn between cognition and brain size.



Although Evans et al. (2005) cautioned that "it remains formally possible that an unrecognized function of *microcephalin* outside the brain is actually the substrate of selection", many interpreted the above data as an adaptive response in intelligence. After all, two functional genes that both influence brain size, a presumed correlate of intelligence, coupled with a history of past, and ongoing, selection does indeed suggest a case for selection on intelligence. This view, however, was quickly dispelled. Timpson et al (2007) and Mekel-Bobrov et al. (2007) showed in large sample sizes (900 and 2400, respectively) that there was no correlation between the putative adaptive halplotypes and increased intelligence. Any on-going selection on these genes does not appear to correlate with any selection for increased cognition. Currant et al. (2006) further noted that *spatial* models of growth were not considered, and here it is possible to see the above patterns for mutations that arise along the leading lead of a recent population expansion.

# Lecture 10 Genomic Selection

## References

Goddard (2008) Genomic selection: prediction of accuracy and maximization of long term response Genetica DOI 10.1007/s10709-008-9308-0

NejatiJavaremi, A., C. Smith, and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science **75**: 1738-1745.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science **91**: 4414-4423

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819-1829

# Genomic Relationship Matrix

- Assumes
  - Dense markers evenly spaced across the genome
  - Assumes markers are in LD with QTL affecting trait(s) of interest
  - Alike in State (AIS) alleles were at one time a result of a single mutation, thus IBD when traced back in evolutionary time
  - Each marker accounts for an equal proportion of genetic variance (infinitesimal model)
  - Genetic Effects are Normally Distributed

# Compute (AIS) relationship matrix (G)

$$TA_k = 2\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{4}$$

$TA_k$=total allelic relationship at k$^{th}$ locus
$TA_k$=2x coefficient of relationship
(Malecot. 1948)

$$G^*_{xy} = \frac{\sum_{k=1}^{L} TA_k}{L} = \frac{2\sum_{k=1}^{L}\left(\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{4}\right)}{L}$$

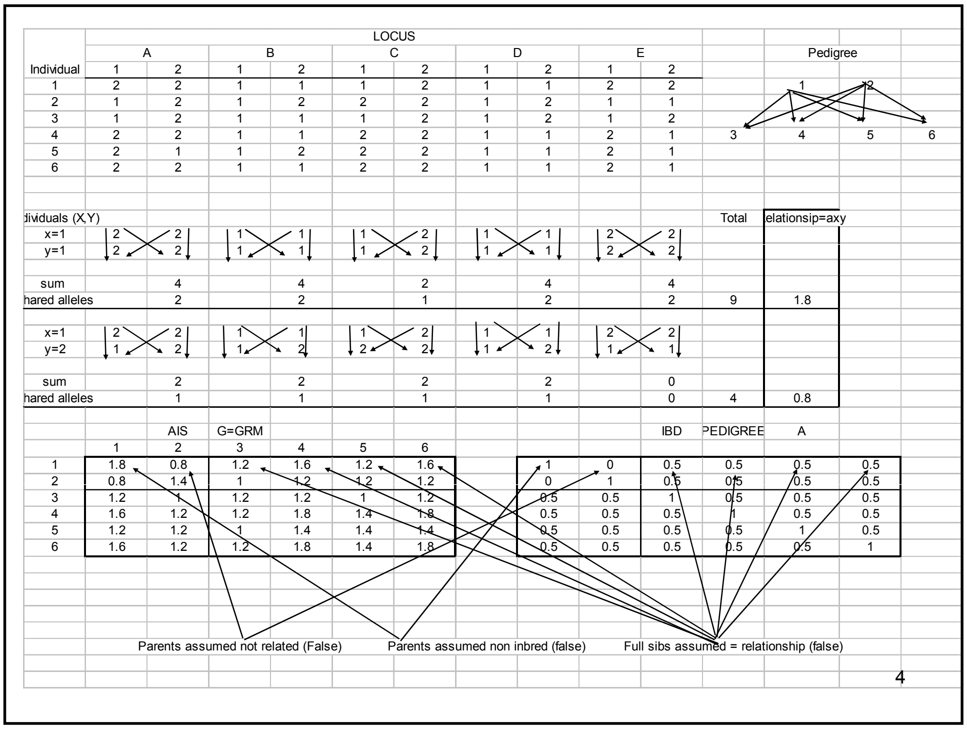
$$\mathbf{G} = \sigma^2_{A*}\mathbf{G}^*$$

$$\sigma^2_{A*}$$

Is the additive genetic variance associated with the markers for the trait

$$\sigma^2_{A*} < \sigma^2_A$$

Note: with low marker density the markers may not capture any genetic variance


Pedigree
1
2
3
4
5
6

|  | LOCUS |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | A |  | B |  | C |  | D |  | E |  |
| Individual | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 |
| 3 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 |
| 4 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 |
| 5 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 |
| 6 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 |

| dividuals (X,Y) |  |  |  |  |  |  |  |  |  |  | Total | elationsip=axy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x=1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |  |  |
| y=1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |  |  |
| sum |  | 4 |  | 4 |  | 2 |  | 4 |  | 4 |  |  |
| hared alleles |  | 2 |  | 2 |  | 1 |  | 2 |  | 2 | 9 | 1.8 |
| x=1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |  |  |
| y=2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 |  |  |
| sum |  | 2 |  | 2 |  | 2 |  | 2 |  | 0 |  |  |
| hared alleles |  | 1 |  | 1 |  | 1 |  | 1 |  | 0 | 4 | 0.8 |

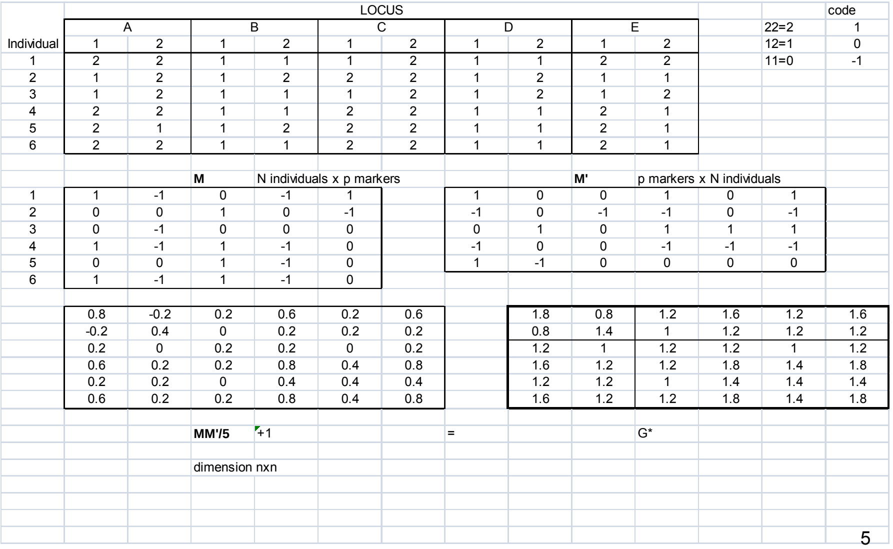
|  |  | AIS | G=GRM |  |  |  |  |  |  | IBD | PEDIGREE | A |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 |  |  |  |  |  |  |  |
| 1 | 1.8 | 0.8 | 1.2 | 1.6 | 1.2 | 1.6 |  | 1 | 0 | 0.5 | 0.5 | 0.5 | 0.5 |
| 2 | 0.8 | 1.4 | 1 | 1.2 | 1.2 | 1.2 |  | 0 | 1 | 0.5 | 0.5 | 0.5 | 0.5 |
| 3 | 1.2 | 1 | 1.2 | 1.2 | 1 | 1.2 |  | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0.5 |
| 4 | 1.6 | 1.2 | 1.2 | 1.8 | 1.4 | 1.8 |  | 0.5 | 0.5 | 0.5 | 1 | 0.5 | 0.5 |
| 5 | 1.2 | 1.2 | 1 | 1.4 | 1.4 | 1.4 |  | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 0.5 |
| 6 | 1.6 | 1.2 | 1.2 | 1.8 | 1.4 | 1.8 |  | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1 |

Parents assumed not related (False)

Parents assumed non inbred (false)

Full sibs assumed = relationship (false)

# G* Computed Directly from M



| | | | | | LOCUS | | | | | | | | code |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | B | | C | | D | | E | | | 22=2 | 1 |
| Individual | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | | 12=1 | 0 |
| 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | | 11=0 | -1 |
| 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | | | |
| 3 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | | | |
| 4 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | | | |
| 5 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | | | |
| 6 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | | | |
| | | | | | | | | | | | | | |
| | | | **M** | N individuals x p markers | | | | | **M'** | p markers x N individuals | | | |
| 1 | 1 | -1 | 0 | -1 | 1 | | 1 | 0 | 0 | 1 | 0 | 1 | |
| 2 | 0 | 0 | 1 | 0 | -1 | | -1 | 0 | -1 | -1 | 0 | -1 | |
| 3 | 0 | -1 | 0 | 0 | 0 | | 0 | 1 | 0 | 1 | 1 | 1 | |
| 4 | 1 | -1 | 1 | -1 | 0 | | -1 | 0 | 0 | -1 | -1 | -1 | |
| 5 | 0 | 0 | 1 | -1 | 0 | | 1 | -1 | 0 | 0 | 0 | 0 | |
| 6 | 1 | -1 | 1 | -1 | 0 | | | | | | | | |
| | | | | | | | | | | | | | |
| | 0.8 | -0.2 | 0.2 | 0.6 | 0.2 | 0.6 | | 1.8 | 0.8 | 1.2 | 1.6 | 1.2 | 1.6 |
| | -0.2 | 0.4 | 0 | 0.2 | 0.2 | 0.2 | | 0.8 | 1.4 | 1 | 1.2 | 1.2 | 1.2 |
| | 0.2 | 0 | 0.2 | 0.2 | 0 | 0.2 | | 1.2 | 1 | 1.2 | 1.2 | 1 | 1.2 |
| | 0.6 | 0.2 | 0.2 | 0.8 | 0.4 | 0.8 | | 1.6 | 1.2 | 1.2 | 1.8 | 1.4 | 1.8 |
| | 0.2 | 0.2 | 0 | 0.4 | 0.4 | 0.4 | | 1.2 | 1.2 | 1 | 1.4 | 1.4 | 1.4 |
| | 0.6 | 0.2 | 0.2 | 0.8 | 0.4 | 0.8 | | 1.6 | 1.2 | 1.2 | 1.8 | 1.4 | 1.8 |
| | | | | | | | | | | | | | |
| | | | **MM'/5** | +1 | | | = | | G* | | | | |
| | | | | | | | | | | | | | |
| | | | dimension nxn | | | | | | | | | | |

# Coding
# Genomic Relationship Matrix

| Genotype | Frequency | 012 | 101 | centered |
|---|---|---|---|---|
| AA | $(p_i)^2$ | 2 | 1 | $2\text{-}2p_i$ |
| Aa | $2p_i(1\text{-}p_i)$ | 1 | 0 | $1\text{-}2p_i$ |
| aa | $(1\text{-}p_i)^2$ | 0 | -1 | $2p_i$ |
| Mean | 1 | $2p_i$ | $(1\text{-}2p_i)$ | 0 |

Does it make a difference?

GBLUP, NO
ssGBLUP, Yes G matrix needs to scaled the same as A matrix

Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. Journal Of Dairy Science **92**: 4656-4663.
Chen, C., I. Misztal, I. Aguilar, A. Legarra, and W. Muir, 2011 Effect of different genomic relationship matrices on accuracy and scale. Journal Of Animal Science **89**: 2673-2679.

# Mixed Model Equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

Simplifications If $\mathbf{R} = \mathbf{I}\sigma_e^2$

$\mathbf{G} = \sigma_{A*}^2 \mathbf{MM'}/L$

**M** (n individuals x p markers)
**M**(n,p)M'(p,n)
**MM'**(n,n)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\dfrac{\sigma_e^2}{\sigma_{A*}^2}\left(\mathbf{MM}'/L\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

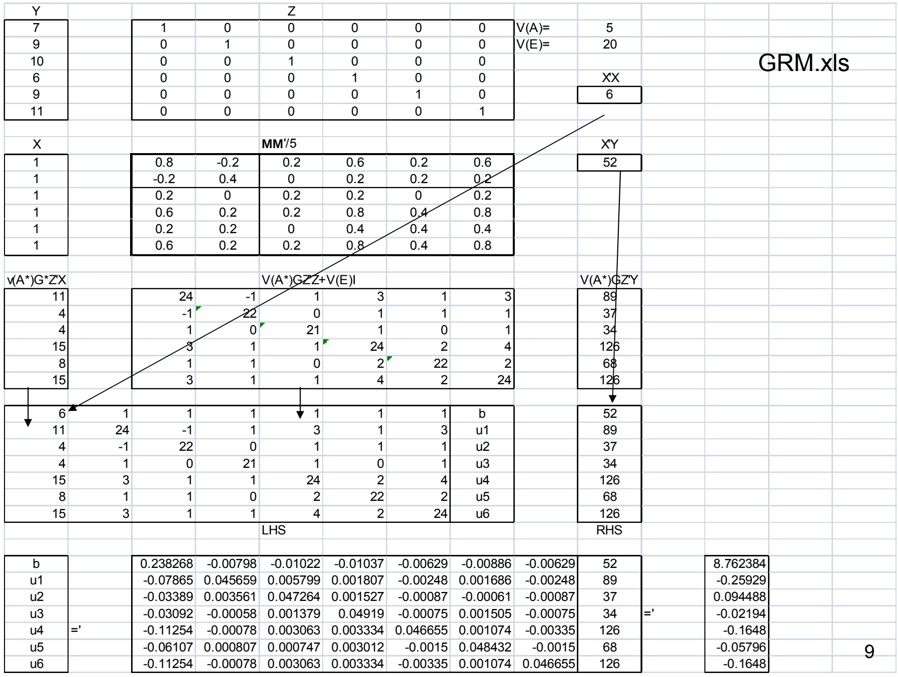
# Example

$$\mathbf{Y} = \begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \\ 11 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad b = [\mu_0] \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & -1 & 1 & -1 & 0 \end{bmatrix} \quad \mathbf{MM'}/L = \begin{bmatrix} .8 & -.2 & .2 & .6 & .2 & .6 \\ -.2 & .4 & 0 & .2 & .2 & .2 \\ .2 & 0 & .2 & .2 & 0 & .2 \\ .6 & .2 & .2 & .8 & .4 & .8 \\ .2 & .2 & 0 & .4 & .4 & .4 \\ .6 & .2 & .2 & .8 & .4 & .8 \end{bmatrix}$$

$\sigma_A^2 = 10$
$\sigma_{A*}^2 = 5$
$\sigma_\varepsilon^2 = 20$

Note, only ½ the additive genetic variance was captured by the markers



GRM.xls

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | | | | Z | | | | | | | |
| 7 | | 1 | 0 | 0 | 0 | 0 | 0 | V(A)= | 5 | | |
| 9 | | 0 | 1 | 0 | 0 | 0 | 0 | V(E)= | 20 | | |
| 10 | | 0 | 0 | 1 | 0 | 0 | 0 | | | | |
| 6 | | 0 | 0 | 0 | 1 | 0 | 0 | | X'X | | |
| 9 | | 0 | 0 | 0 | 0 | 1 | 0 | | 6 | | |
| 11 | | 0 | 0 | 0 | 0 | 0 | 1 | | | | |
| | | | | | | | | | | | |
| X | | | | MM'/5 | | | | | X'Y | | |
| 1 | | 0.8 | -0.2 | 0.2 | 0.6 | 0.2 | 0.6 | | 52 | | |
| 1 | | -0.2 | 0.4 | 0 | 0.2 | 0.2 | 0.2 | | | | |
| 1 | | 0.2 | 0 | 0.2 | 0.2 | 0 | 0.2 | | | | |
| 1 | | 0.6 | 0.2 | 0.2 | 0.8 | 0.4 | 0.8 | | | | |
| 1 | | 0.2 | 0.2 | 0 | 0.4 | 0.4 | 0.4 | | | | |
| 1 | | 0.6 | 0.2 | 0.2 | 0.8 | 0.4 | 0.8 | | | | |
| | | | | | | | | | | | |
| v(A*)G*Z'X | | | | V(A*)GZ'Z+V(E)I | | | | | V(A*)GZ'Y | | |
| 11 | | 24 | -1 | 1 | 3 | 1 | 3 | | 89 | | |
| 4 | | -1 | 22 | 0 | 1 | 1 | 1 | | 37 | | |
| 4 | | 1 | 0 | 21 | 1 | 0 | 1 | | 34 | | |
| 15 | | 3 | 1 | 1 | 24 | 2 | 4 | | 126 | | |
| 8 | | 1 | 1 | 0 | 2 | 22 | 2 | | 68 | | |
| 15 | | 3 | 1 | 1 | 4 | 2 | 24 | | 126 | | |
| | | | | | | | | | | | |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | b | | 52 | | |
| 11 | 24 | -1 | 1 | 3 | 1 | 3 | u1 | | 89 | | |
| 4 | -1 | 22 | 0 | 1 | 1 | 1 | u2 | | 37 | | |
| 4 | 1 | 0 | 21 | 1 | 0 | 1 | u3 | | 34 | | |
| 15 | 3 | 1 | 1 | 24 | 2 | 4 | u4 | | 126 | | |
| 8 | 1 | 1 | 0 | 2 | 22 | 2 | u5 | | 68 | | |
| 15 | 3 | 1 | 1 | 4 | 2 | 24 | u6 | | 126 | | |
| | | | | LHS | | | | | RHS | | |
| | | | | | | | | | | | |
| b | | 0.238268 | -0.00798 | -0.01022 | -0.01037 | -0.00629 | -0.00886 | -0.00629 | 52 | | 8.762384 |
| u1 | | -0.07865 | 0.045659 | 0.005799 | 0.001807 | -0.00248 | 0.001686 | -0.00248 | 89 | | -0.25929 |
| u2 | | -0.03389 | 0.003561 | 0.047264 | 0.001527 | -0.00087 | -0.00061 | -0.00087 | 37 | | 0.094488 |
| u3 | | -0.03092 | -0.00058 | 0.001379 | 0.04919 | -0.00075 | 0.001505 | -0.00075 | 34 | =' | -0.02194 |
| u4 | =' | -0.11254 | -0.00078 | 0.003063 | 0.003334 | 0.046655 | 0.001074 | -0.00335 | 126 | | -0.1648 |
| u5 | | -0.06107 | 0.000807 | 0.000747 | 0.003012 | -0.0015 | 0.048432 | -0.0015 | 68 | | -0.05796 |
| u6 | | -0.11254 | -0.00078 | 0.003063 | 0.003334 | -0.00335 | 0.001074 | 0.046655 | 126 | | -0.1648 |

# Equivalent Model
# Estimation of Marker effects

$$\begin{bmatrix} \mathbf{X'}_{1,N}\ \mathbf{X}_{N,1} & \mathbf{X}_{1,N}{}'\mathbf{M}_{N,p} \\ \mathbf{M'}_{p,N}\ \mathbf{X'}_{N,1} & \mathbf{M}'_{p,N}\mathbf{M}_{N,p} + \frac{\sigma_e^2}{\sigma_g^2}\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{1,1} \\ \mathbf{g}_{p,1} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}_{1,N}\ \mathbf{Y}_{N,1} \\ \mathbf{M'}_{p,N}\ \mathbf{Y}_{N,1} \end{bmatrix}$$

Each Marker effects is solved for

Assumption depends on method

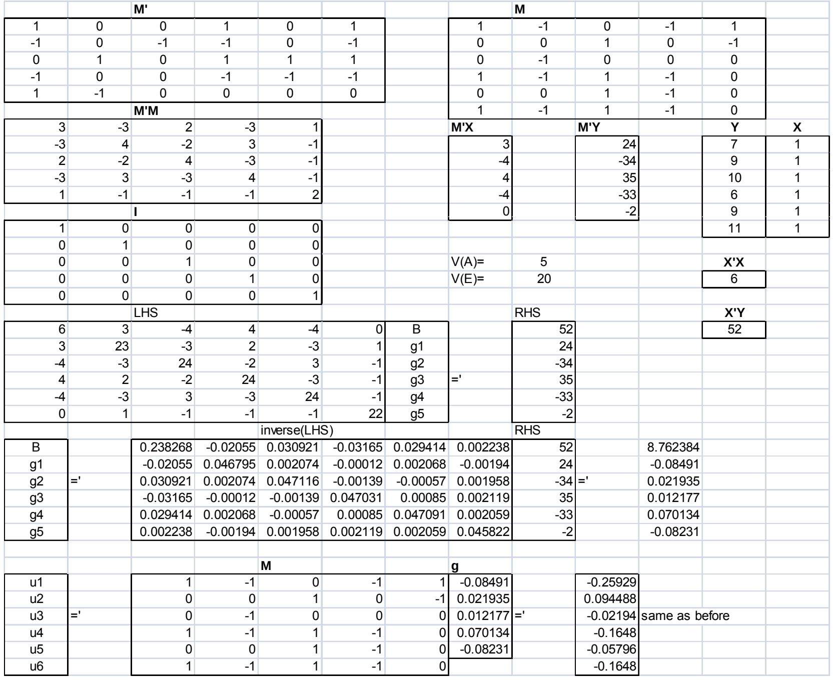
1) (GBLUP, ssGBLUP) Genetic variance associated with each marker is equal $\sigma_g^2 = \left(\frac{\sigma_{A*}^2}{L}\right)$

2) (Bayes A) sampled from a t distribution

3) (Bayes B and Bayes C π) from a mixture of distributions (null and t)

$$\hat{u}_i = GEBV_i = \mathbf{Mg} = \sum_j M_{ij}\hat{g}_j$$

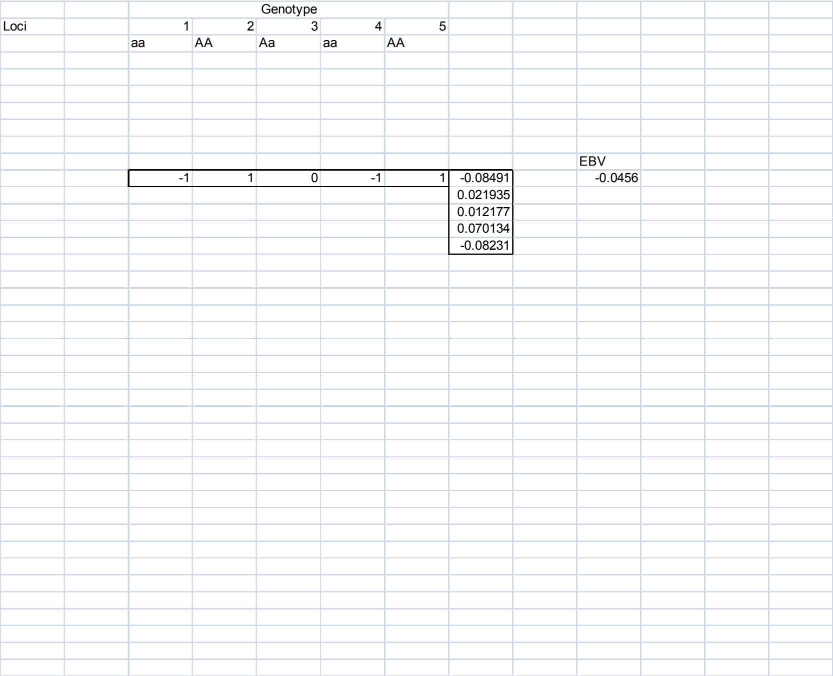
Marker effects BLUP.xls



| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **M'** | | | | | | **M** | | | | |
| 1 | 0 | 0 | 1 | 0 | 1 | | 1 | -1 | 0 | -1 | 1 | |
| -1 | 0 | -1 | -1 | 0 | -1 | | 0 | 0 | 1 | 0 | -1 | |
| 0 | 1 | 0 | 1 | 1 | 1 | | 0 | -1 | 0 | 0 | 0 | |
| -1 | 0 | 0 | -1 | -1 | -1 | | 1 | -1 | 1 | -1 | 0 | |
| 1 | -1 | 0 | 0 | 0 | 0 | | 0 | 0 | 1 | -1 | 0 | |
| | | **M'M** | | | | | 1 | -1 | 1 | -1 | 0 | |
| 3 | -3 | 2 | -3 | 1 | | | **M'X** | | **M'Y** | | **Y** | **X** |
| -3 | 4 | -2 | 3 | -1 | | | 3 | | 24 | | 7 | 1 |
| 2 | -2 | 4 | -3 | -1 | | | -4 | | -34 | | 9 | 1 |
| -3 | 3 | -3 | 4 | -1 | | | 4 | | 35 | | 10 | 1 |
| 1 | -1 | -1 | -1 | 2 | | | -4 | | -33 | | 6 | 1 |
| | | **I** | | | | | 0 | | -2 | | 9 | 1 |
| 1 | 0 | 0 | 0 | 0 | | | | | | | 11 | 1 |
| 0 | 1 | 0 | 0 | 0 | | | | | | | | |
| 0 | 0 | 1 | 0 | 0 | | | V(A)= | 5 | | | **X'X** | |
| 0 | 0 | 0 | 1 | 0 | | | V(E)= | 20 | | | 6 | |
| 0 | 0 | 0 | 0 | 1 | | | | | | | | |
| | | LHS | | | | | | RHS | | | **X'Y** | |
| 6 | 3 | -4 | 4 | -4 | 0 | B | | 52 | | | 52 | |
| 3 | 23 | -3 | 2 | -3 | 1 | g1 | | 24 | | | | |
| -4 | -3 | 24 | -2 | 3 | -1 | g2 | | -34 | | | | |
| 4 | 2 | -2 | 24 | -3 | -1 | g3 | =' | 35 | | | | |
| -4 | -3 | 3 | -3 | 24 | -1 | g4 | | -33 | | | | |
| 0 | 1 | -1 | -1 | -1 | 22 | g5 | | -2 | | | | |
| | | | | inverse(LHS) | | | | RHS | | | | |
| B | | 0.238268 | -0.02055 | 0.030921 | -0.03165 | 0.029414 | 0.002238 | 52 | | 8.762384 | | |
| g1 | | -0.02055 | 0.046795 | 0.002074 | -0.00012 | 0.002068 | -0.00194 | 24 | | -0.08491 | | |
| g2 | =' | 0.030921 | 0.002074 | 0.047116 | -0.00139 | -0.00057 | 0.001958 | -34 | =' | 0.021935 | | |
| g3 | | -0.03165 | -0.00012 | -0.00139 | 0.047031 | 0.00085 | 0.002119 | 35 | | 0.012177 | | |
| g4 | | 0.029414 | 0.002068 | -0.00057 | 0.00085 | 0.047091 | 0.002059 | -33 | | 0.070134 | | |
| g5 | | 0.002238 | -0.00194 | 0.001958 | 0.002119 | 0.002059 | 0.045822 | -2 | | -0.08231 | | |
| | | | | | | | | | | | | |
| | | | | **M** | | | **g** | | | | | |
| u1 | | 1 | -1 | 0 | -1 | 1 | -0.08491 | | -0.25929 | | | |
| u2 | | 0 | 0 | 1 | 0 | -1 | 0.021935 | | 0.094488 | | | |
| u3 | =' | 0 | -1 | 0 | 0 | 0 | 0.012177 | =' | -0.02194 | same as before | | |
| u4 | | 1 | -1 | 1 | -1 | 0 | 0.070134 | | -0.1648 | | | |
| u5 | | 0 | 0 | 1 | -1 | 0 | -0.08231 | | -0.05796 | | | |
| u6 | | 1 | -1 | 1 | -1 | 0 | | | -0.1648 | | | |

# Example

missing phenotypes but know genotypes and marker effects following training =pure genomic selection



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Genotype | | | | | | | |
| Loci | | 1 | 2 | 3 | 4 | 5 | | | |
| | | aa | AA | Aa | aa | AA | | | |
| | | | | | | | | | EBV |
| | | -1 | 1 | 0 | -1 | 1 | -0.08491 | | -0.0456 |
| | | | | | | | 0.021935 | | |
| | | | | | | | 0.012177 | | |
| | | | | | | | 0.070134 | | |
| | | | | | | | -0.08231 | | |

# Problems and relation to Association Analysis

- Admixture
  - Major problem
  - False Positives
  - Spurious Correlations
  - Correlation does not mean Causation
- Partial Solution
  - Use Igenstrat to correct for structure
  - Use Structure to correct for structure
  - Does not correct for phase
- Application : economics
  - 2 stage
  - Use Dense SNP genotyping (60k) on all selected male **parents**
  - Use low density genotyping (512) on all selection **candidates**
  - Impute genotypes of female breeders
- Use in Humans to determine disease risk
  - Use dense SNP chip for predicton of “risk” or “merit”
  - Don’t worry about which markers are most predictive, Use them all
  - Solves “missing heritability” issues