Lecture 5: Allelic Effects and Genetic Variances

Bruce Walsh lecture notes
Synbreed course
version 12 July 2013

Quantitative Genetics

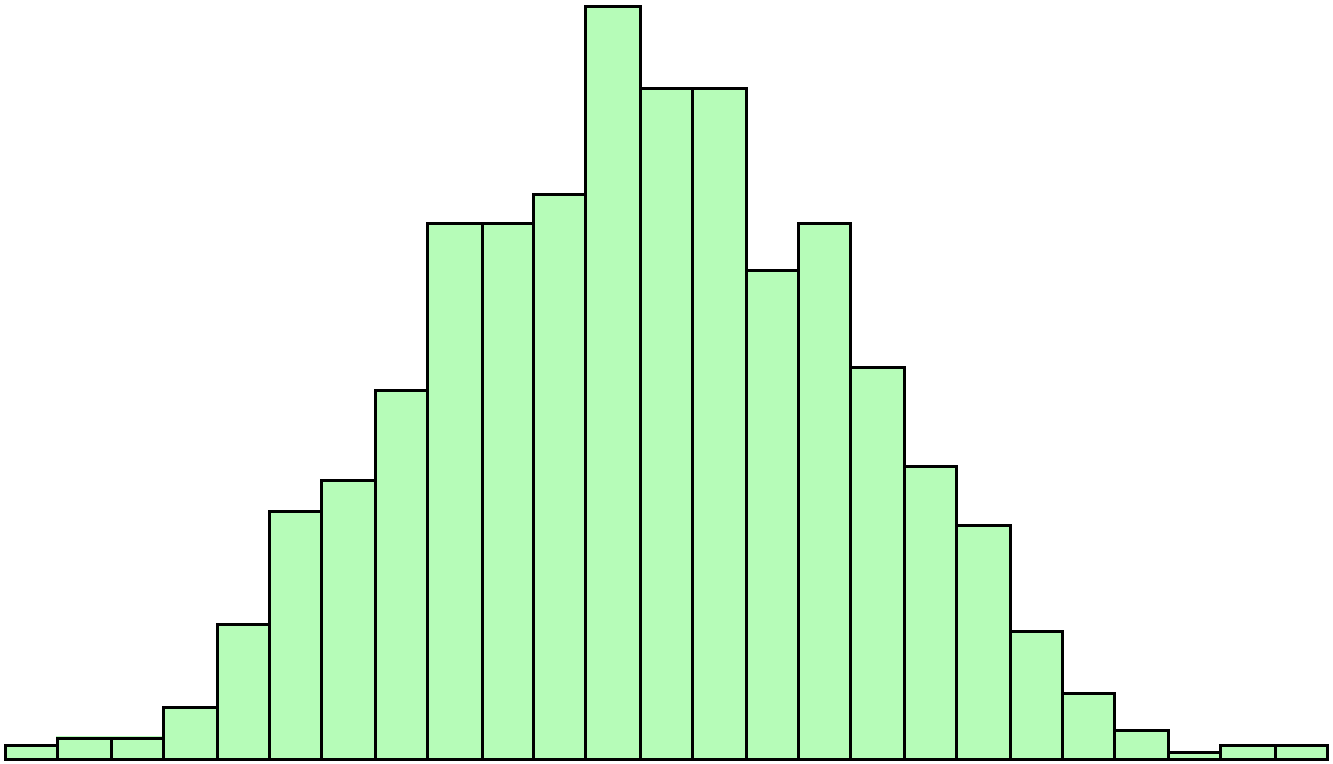
The analysis of traits whose variation is determined by both a number of genes and environmental factors

Phenotype is highly uninformative as to underlying genotype

Complex (or Quantitative) trait

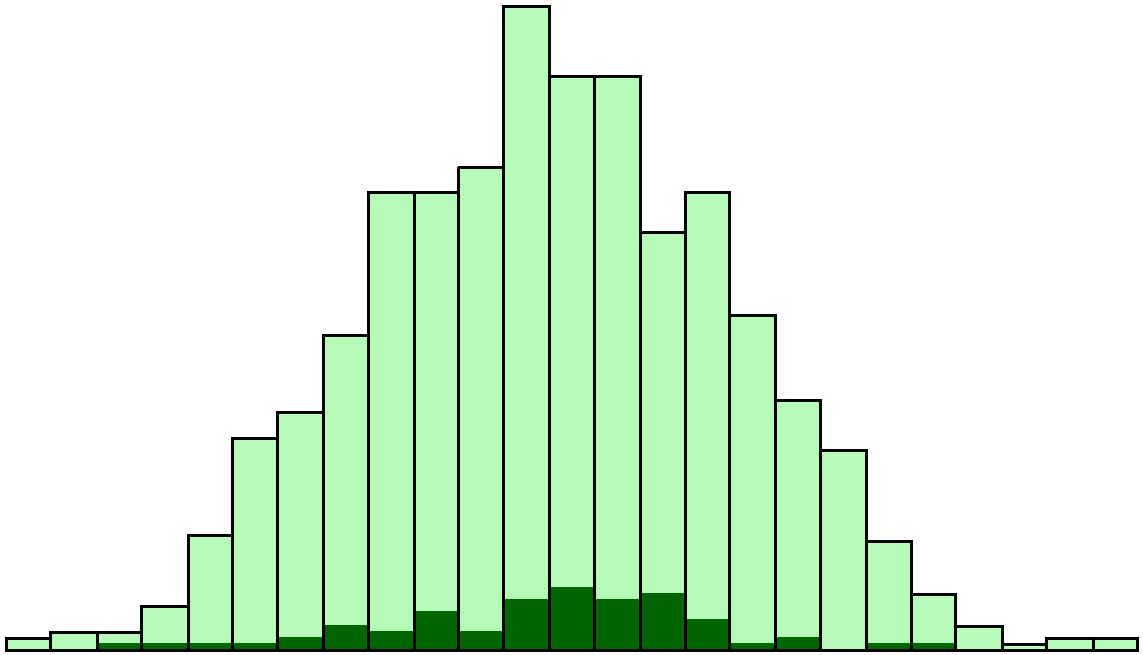
- No (apparent) simple Mendelian basis for variation in the trait
- May be a single gene strongly influenced by environmental factors
- May be the result of a number of genes of equal (or differing) effect
- Most likely, a combination of both multiple genes and environmental factors
- Example: Blood pressure, cholesterol levels
  - Known genetic and environmental risk factors
- Molecular traits can also be quantitative traits
  - mRNA level on a microarray analysis
  - Protein spot volume on a 2-D gel

Phenotypic distribution of a trait



Consider a specific locus influencing the trait

Values for QQ individuals shaded in dark green



For this locus, mean phenotype = 0.15, while overall mean phenotype = 0

Hence, it is very hard to distinguish the QQ individuals from all others simply from their phenotypic values

Goals of Quantitative Genetics

- Partition total trait variation into genetic (nature) vs. environmental (nurture) components
- Predict resemblance between relatives
  - If a sib has a disease/trait, what are your odds?
  - Selection response
  - Change in mean under inbreeding, outcrossing, assortative maing
- Find the underlying loci contributing to genetic variation
  - QTL -- quantitative trait loci
- Deduce molecular basis for genetic trait variation
- eQTLs -- expression QTLs, loci with a quantitative influence on gene expression
  - e.g., QTLs influencing mRNA abundance on a microarray

Dichotomous (binary) traits

Presence/absence traits (such as a disease) can (and usually do) have a complex genetic basis

Consider a disease susceptibility (DS) locus underlying a disease, with alleles D and d, where allele D significantly increases your disease risk

In particular, Pr(disease | DD) = 0.5, so that the penetrance of genotype DD is 50%

Suppose Pr(disease | Dd) = 0.2, Pr(disease | dd) = 0.05

dd individuals can rarely display the disease, largely because of exposure to adverse environmental conditions

dd individuals can give rise to phenocopies 5% of the time, showing the disease but not as a result of carrying the risk allele

If freq(d) = 0.9, what is Prob (DD | show disease) ?

$$\text{freq(disease)} = 0.1^2*0.5 + 2*0.1*0.9*0.2 + 0.9^2*0.05$$
$$= 0.0815 \quad \text{(Hardy-Weinberg assumption)}$$

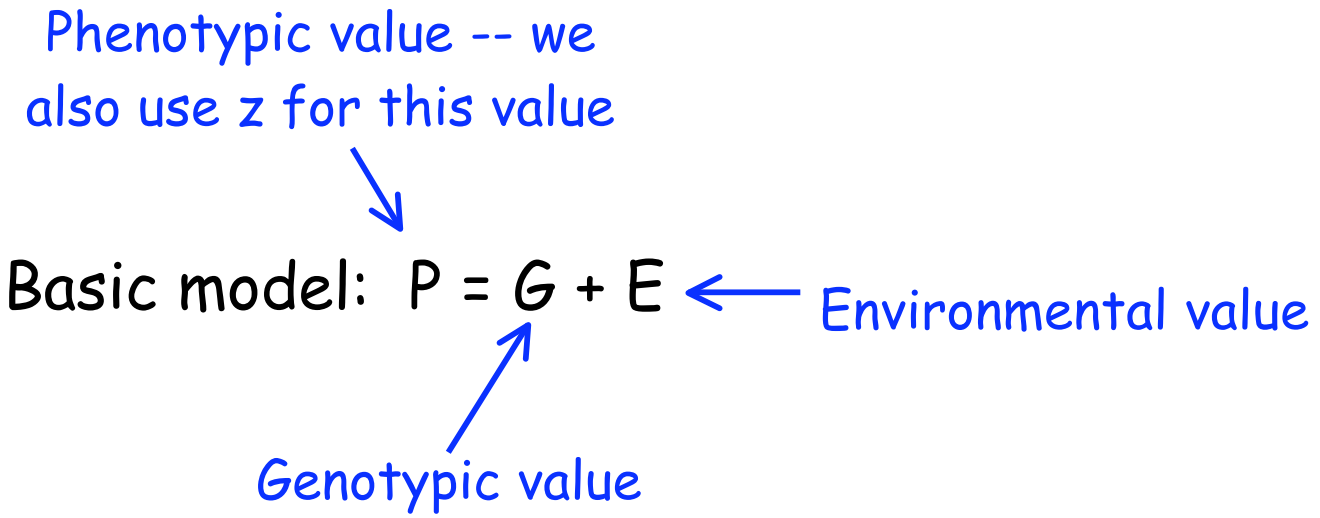
From Bayes' theorem,

$$\Pr(DD \mid \text{disease}) = \Pr(\text{disease} \mid DD)*\Pr(DD)/\text{Prob(disease)}$$
$$= 0.1^2*0.5 / 0.0815 = 0.06 \ (6\ \%)$$

Pr(Dd | disease) = 0.442, Pr(dd | disease) = 0.497

Thus about 50% of the diseased individuals are phenocopies

Basic model of Quantitative Genetics



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

Hence, genotypic values are functions of the environments experienced.

Basic model of Quantitative Genetics

Basic model: $P = G + E$

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

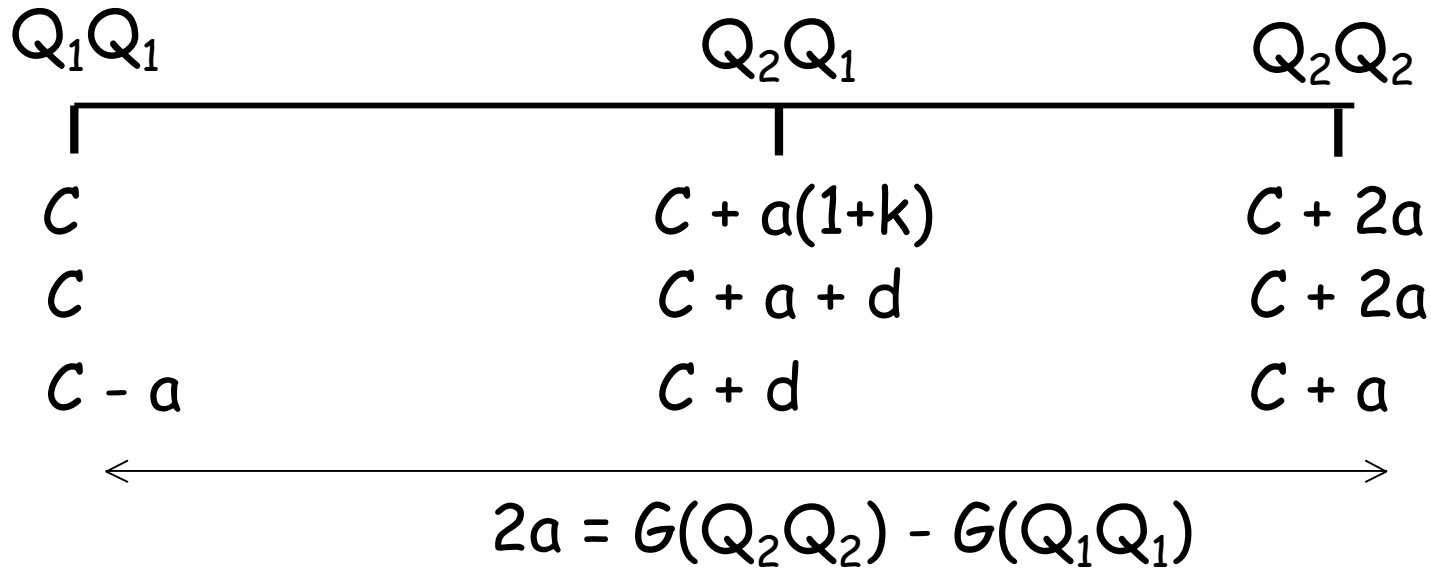
G = average value of an inbred line over a series of environments

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

The transmission of genotypes versus alleles

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

Contribution of a locus to a trait



d measures dominance, with d = 0 if the heterozygote is exactly intermediate to the two homozygotes

$$d = ak = G(Q_1Q_2) - [G(Q_2Q_2) + G(Q_1Q_1)]/2$$

k = d/a is a scaled measure of the dominance

Example: Apolipoprotein E & Alzheimer's

Genotype	ee	Ee	EE
Average age of onset	68.4	75.5	84.3

$2a = G(EE) - G(ee) = 84.3 - 68.4 \rightarrow a = 7.95$

$ak = d = G(Ee) - [\, G(EE)+G(ee)]/2 = -0.85$

$k = d/a = -0.10$ Only small amount of dominance

Example: Booroola (B) gene

Genotype	bb	Bb	BB
Average Litter size	1.48	2.17	2.66

$2a = G(BB) - G(bb) = 2.66 - 1.46 \rightarrow a = 0.59$

$ak = d = G(Bb) - [G(BB) + G(bb)]/2 = 0.10$

$k = d/a = 0.17$

Population means: Random mating

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

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

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

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

Contribution from homozygotes

Contribution from heterozygotes

Population means: Inbred cross F_2

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

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

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

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

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

Population means: RILs from an F_2

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

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

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

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

The average effect of an allele

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

Random mating

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

Suppose parent contributes A

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

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

α_A = Mean(A transmitted) - μ = **q[a + d(q-p)]**

Random mating

Now suppose parent contributes a

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

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

α_a = Mean(a transmitted) - μ = **-p[a + d(q-p)]**

α, the average effect of an allelic substitution

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

Dominance deviations

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

Fisher's (1918) Decomposition of *G*

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

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

Consider the genotypic value G_{ij} resulting from an A_iA_j individual

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

α_i = average contribution to genotypic value for allele i

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

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

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

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

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

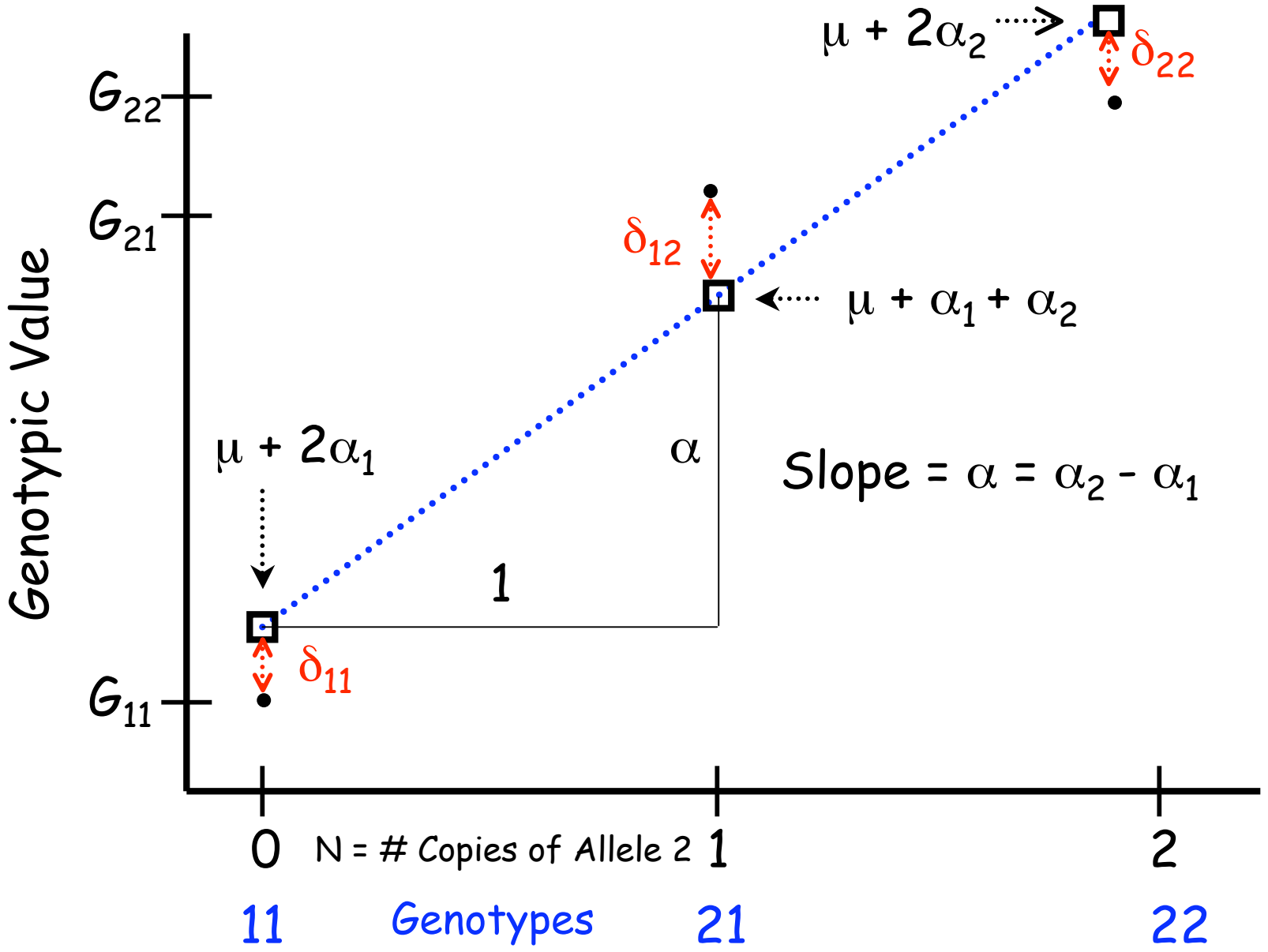
$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$$

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

Dominance deviations --- the difference (for genotype A_iA_j) between the genotypic value predicted from the two single alleles and the actual genotypic value, namely any interactions (dominance) between the two alleles

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

This decomposition is a regression of G



Fisher's decomposition is a Regression

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

Predicted value

Residual error

A notational change clearly shows this is a regression,

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

Independent (predictor) variable N = # of A_2 alleles

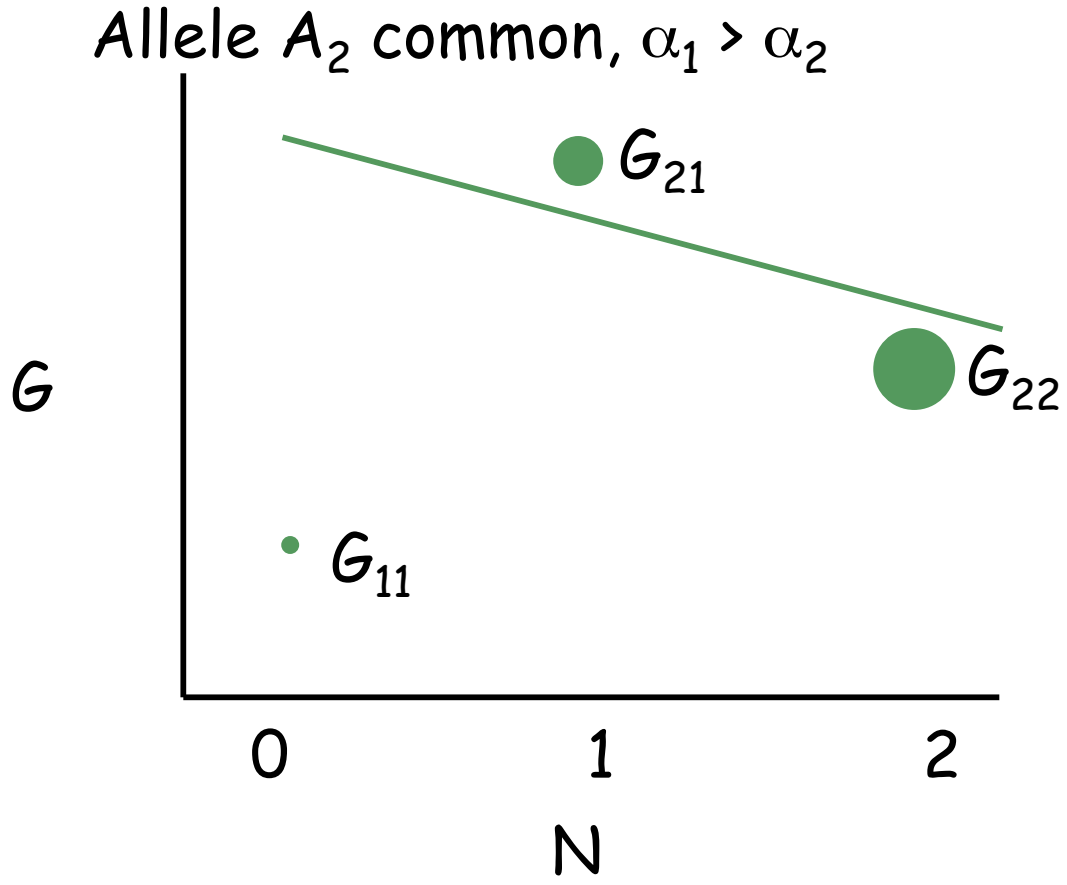
Note that the slope α_2 - α_1 = α, the average effect of an allelic substitution

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

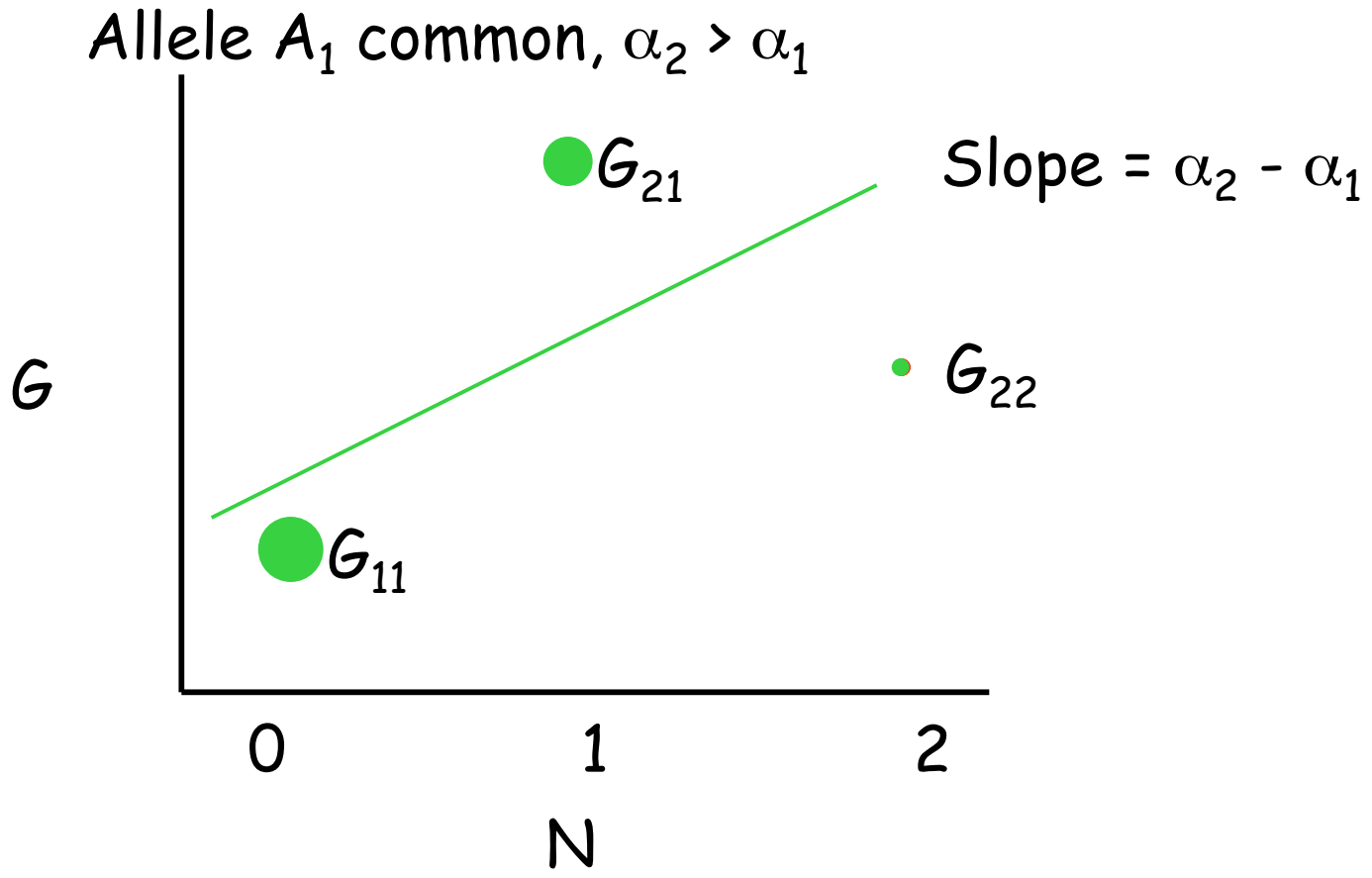
Intercept Regression slope

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

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

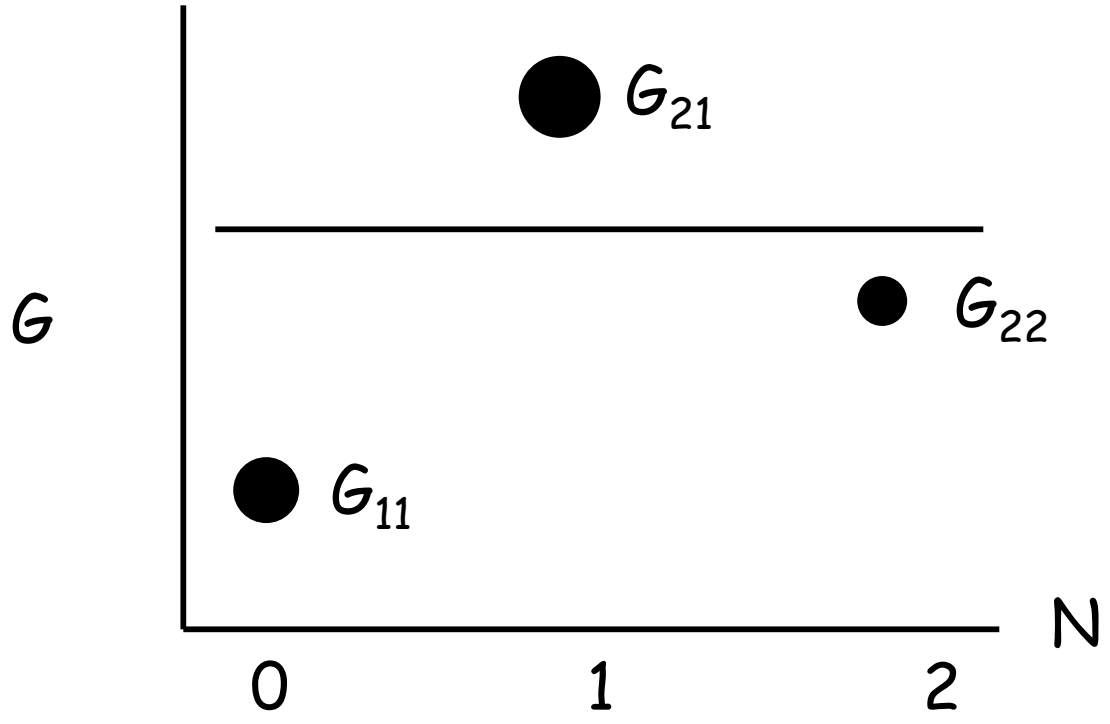


The size of the circle denotes the weight associated with that genotype. While the genotypic values do not change, their frequencies (and hence weights) do.



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

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



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

Consider a diallelic locus, where p_1 = freq(Q_1)

Genotype	Q_1Q_1	Q_2Q_1	Q_2Q_2
Genotypic value	0	a(1+k)	2a

Mean $\mu_G = 2p_2\, a(1 + p_1 k)$

Allelic effects

$$\alpha_2 = p_1 a\,[\,1 + k\,(p_1 - p_2)\,]$$

$$\alpha_1 = -\,p_2 a\,[\,1 + k\,(p_1 - p_2)\,]$$

Dominance deviations $\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j$

Average Effects and Additive Genetic Values

The α values are the average effects of an allele

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

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

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

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

A is called the Breeding value or Additive genetic value

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

Why all the fuss over A?

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

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

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

Genetic Variances

Writing the genotypic value as

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

The genetic variance can be written as

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

This follows since

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

As Cov(α,δ) = 0 (under random mating)

Genetic Variances

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

Additive Genetic Variance
(or simply Additive Variance)

Dominance Genetic Variance
(or simply Dominance Variance)

Hence, total genetic variance = additive + dominance variances,

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

Key concepts (so far)

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

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

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

Since $E[\alpha] = 0$,

$Var(\alpha) = E[(\alpha - \mu_a)^2] = E[\alpha^2]$

One locus, 2 alleles: $\sigma_A^2 = 2p_1 p_2 a^2[1 + k(p_1 - p_2)]^2$

Dominance alters additive variance

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

Dominance variance

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

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

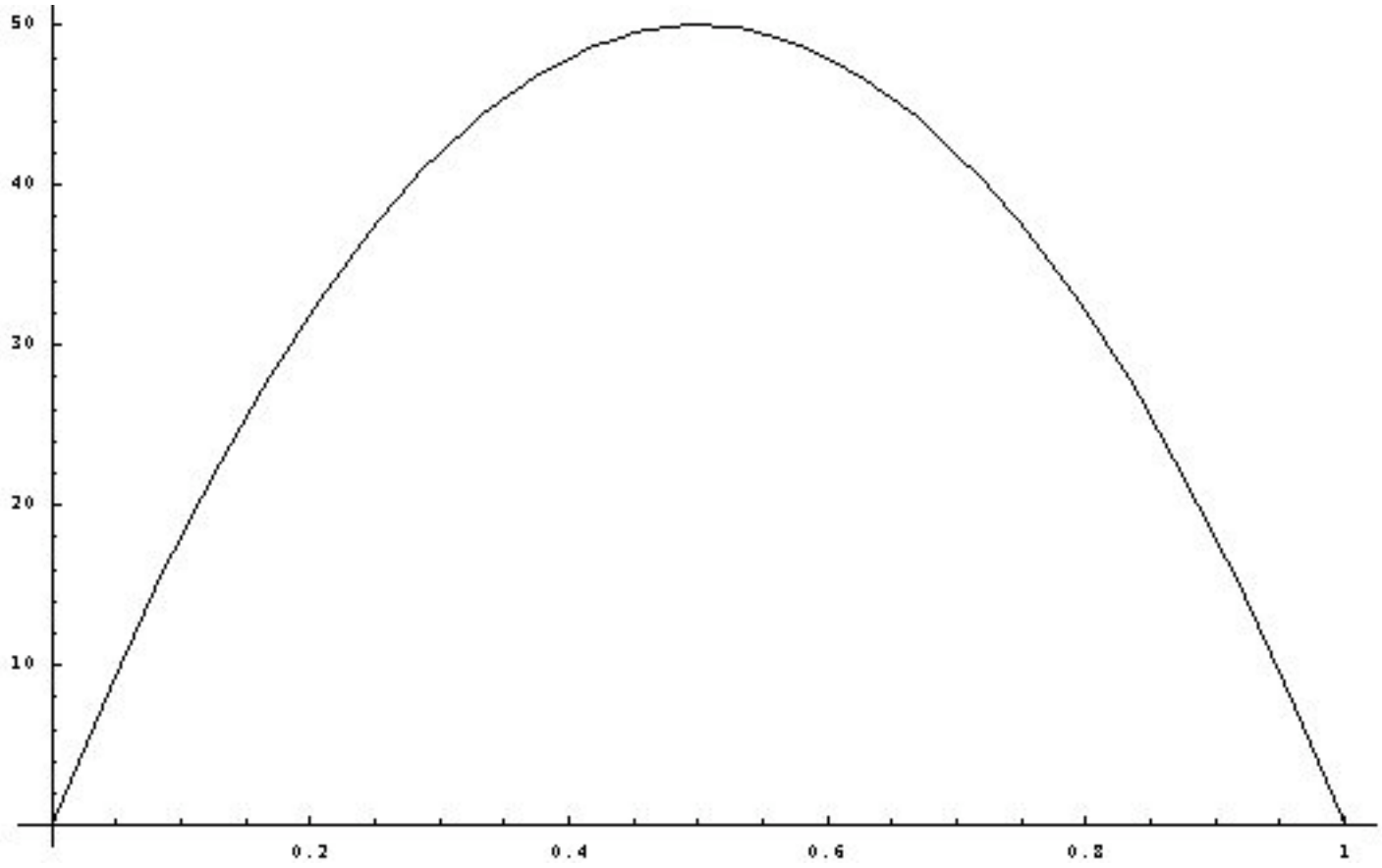
Equals zero if k = 0

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

This is a symmetric function of allele frequencies

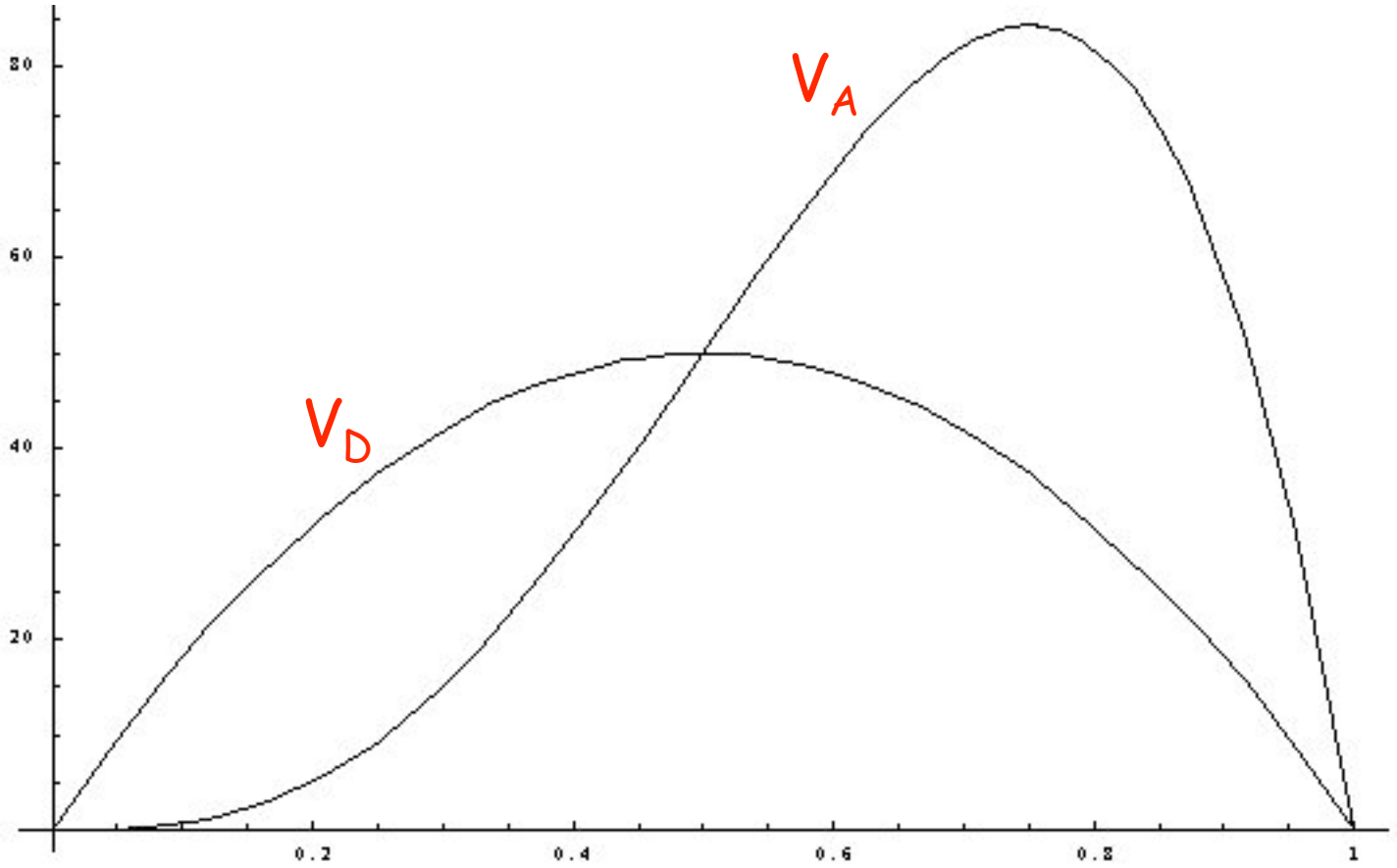
Can also be expressed in terms of d = ak

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



Allele frequency, p

Complete dominance (k = 1)



Allele frequency, p

Epistasis

The two-locus decomposition allowing for all possible interactions is given by

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

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

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

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

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

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

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

Effects and Variance when using a testor

- A common design in plant breeding is to cross members from a population to a testor to generate a testcross.
  - Testor can be either an inbred line or an outcrossing population
  - Often from a different heteroic group from the population being tested
  - Often testor is an elite genotype
- The average effect of an allele in a testcross, its variance, and its additive (General combining ability, GCA) and interaction (Specific combining ability, SCA) effects all follow in analogous fashion to previous results for crosses within a population

The average effect of an allele in a testcross

- The concept of the average effect of an allele when crossed within its population is easily extended to the average effect of an allele when crossed to a testor.
  - Called the testcross average effect.
- The average effect of allele X in this testcross, α_x^T, is defined as difference between the mean value of offspring getting this allele from the population versus the mean value of a random offspring from this cross
  - Will turn out to be a function of the frequencies of alleles in both the tested and the testor population.

Mean value for a testcross

Suppose the frequency of *A* is p in the population and p^T in the testor (with q and q^T similarly defined for a).

testor

Parental line

	A (p^T)	a (q^T)
A (p)	pp^T *C* + a	pq^T *C* + d
a (q)	qp^T *C* + d	qq^T *C* - a

Mean of cross = $C + a(pp^T - qq^T) + d(pq^T + qp^T)$

Average testcross mean in a series of RILs

- Slide 17 gave an expression for the expected average performance from a series of RILs formed by crossing two populations.
- A similar expression exists for the average testcross performance for a series of RILs from a cross of A x B
  - Mean = (1/2) μ_A^T + (1/2) μ_B^T, namely the average of the testcross means for A and B
  - More generally (since lines can, by chance, give an unequal contribution of alleles),
    - Mean = $\pi_A \mu_A^T + \pi_B \mu_B^T$, where $\pi_A = (1- \pi_B)$ is the fraction of alleles from A in the sample if RILS
    - Can use molecular markers to estimate the π_x directly. Here π_x is the fraction of SNP alelles from line x.

α_A^T, testcross effect of allele A

Suppose parent contributes A

Allele from testor parent	Probability	Genotype	Value
A	p^T	AA	C + a
a	q^T	Aa	C + d

Mean(A transmitted) = $p^T(C + a) + q^T(C + d) = C + p^Ta + q^Td$

α_A^T = Mean(A transmitted) - μ = $q[a + d(q^T-p^T)]$

Likewise,

α_a^T = Mean(a transmitted) - μ = $-p[a + d(q^T-p^T)]$

α^T, the average testcross effect of an allelic substitution

- $\alpha^T = \alpha_A{}^T - \alpha_a{}^T$ is the average testcross effect of an allelic substitution, the change in mean trait value when an *a* allele in a random testcrossed individual is replaced by an *A* allele
  - $\alpha^T = a + d(q^T - p^T)$. Note that this is independent of the allele frequencies in the parental population, and depends ONLY on the testor allele frequencies (p^T, q^T).
    - $\alpha_A{}^T = q\alpha^T$, $\alpha_a{}^T = -p\alpha^T$, and $E(\alpha_x{}^T) = 0$

Testcross variance

- Just as the additive genetic variance was the population variance in the sum of the average effects of an allele, the testcross variance is variance in the average testcross effects of a random allele
  - $Var(A^T) = Var(\alpha_x^T) = Var(\alpha_x^T)$
  - $Var(\alpha_x^T) = p\,(\alpha_A^T)^2 + q\,(\alpha_a^T)^2 =$
  - $p(q[a + d(q^T-p^T)])^2 + q(-p[a + d(q^T-p^T)])^2$
- $= pq[a + d(q^T-p^T)]^2$
  - Hence, $Var(\alpha_x^T) = pq[a + d(q^T-p^T)]^2$

GCS and SCA

- Consider a cross between individuals from population 1 and population 2
- Let $\mu_{1 \times 2}$ denote the average value for all of these crosses, and let G_{ij} be the average genotypic value of an individual from a cross from individual (or line) i in population one and individual (or line) j from population two.
- Analogous to Fisher's decomposition, we can write this in terms of two additive effects and one interaction effect.

$$G_{ij} = \mu_{1\times2} + \alpha_i^2 + \alpha_j^1 + \delta_{ij}^{12}$$

α_i^2 is the testcross average effect for allele i (more generally an allele from individual i) when tested using population 2 as a testor, with α_j^1 similarly defined for allele j (from pop 2) using one as the testor

δ_{ij}^{12} is the interaction between allele i from and allele j in the testcross of 1 and 2

The sum over all loci of the α_i^2 values is the general combining ability (*GCA*) of line i when crossed to line 2 (note these are cross-specific)!

The sum of the δ is the specific combining ability (*SCA*)

$$G_{ij} = \mu + GCA_i^2 + GCA_j^1 + SCA_{ij}^{12}$$

The superscripts denoting the population in which the allele is being tested is often suppressed

The *GCA* is akin to the breeding value from one parent, but now it is the testcross value of that parent

The predicted mean of a particular cross is the sum of the two *GCA*s for those individuals/lines

As with average effects and dominance deviations, these are only defined with respect to a particular reference set of crosses (i.e., lines from Pop 1 X lines from pop 2)

Within-population crosses vs. testors

	Within-pop	testor
Allelic effects	α	α^T
Additive transmitting factor	Breeding value A	GCA
Predicting offspring mean	$A_1/2 + A_2/2$	$GCA_1 + GCA_2$
Nonadditive component	Dominance value	SCA
Genetic Variances	Var(A), Var(D)	Var(GCA), Var(SCA)