Lecture 5
Inbreeding and Crossbreeding

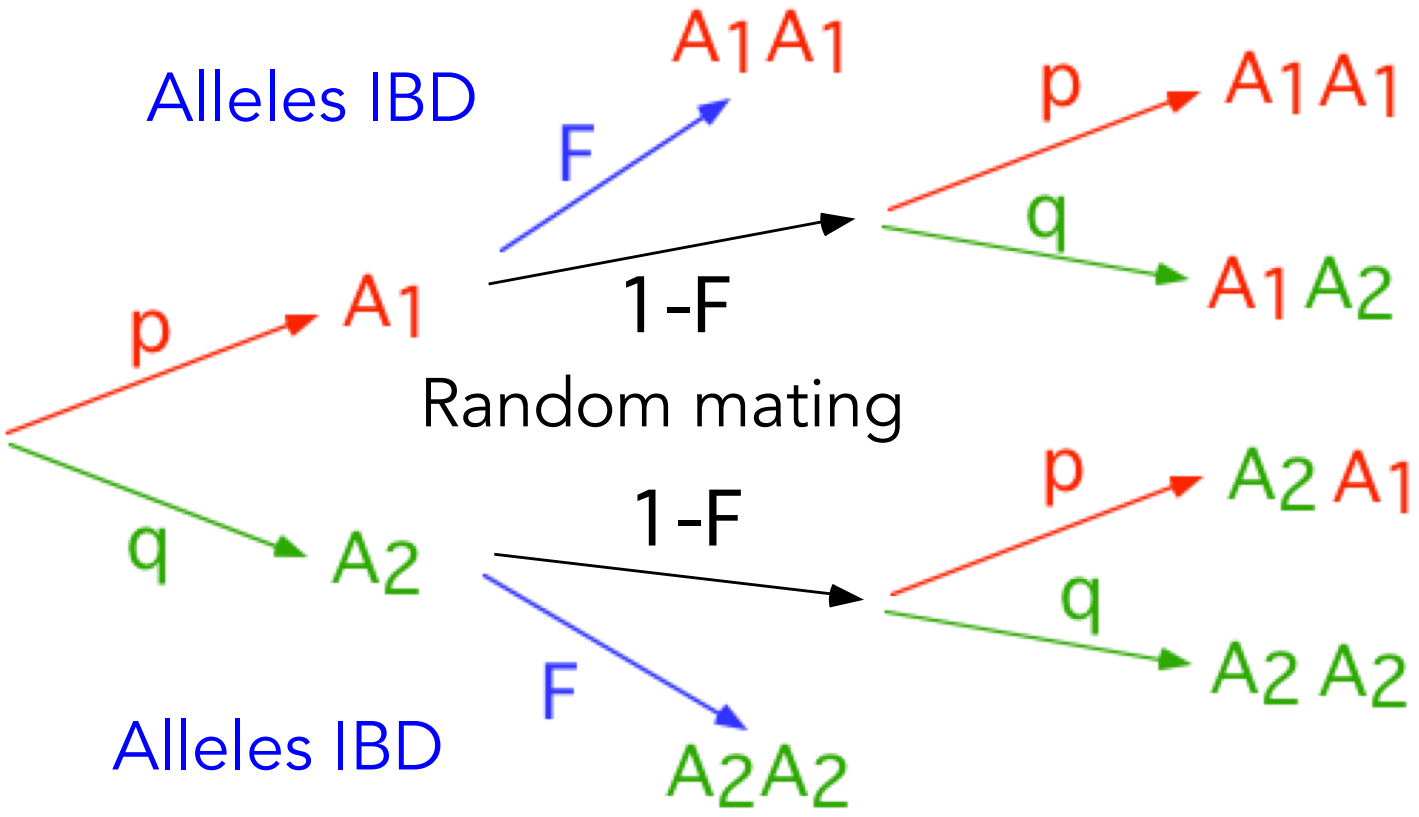
Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG (Module 9), Seattle
15 – 17 July 2019

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 μ_F under a level of inbreeding F is related to the mean μ_0 under random mating by

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

For k loci, the change in mean is

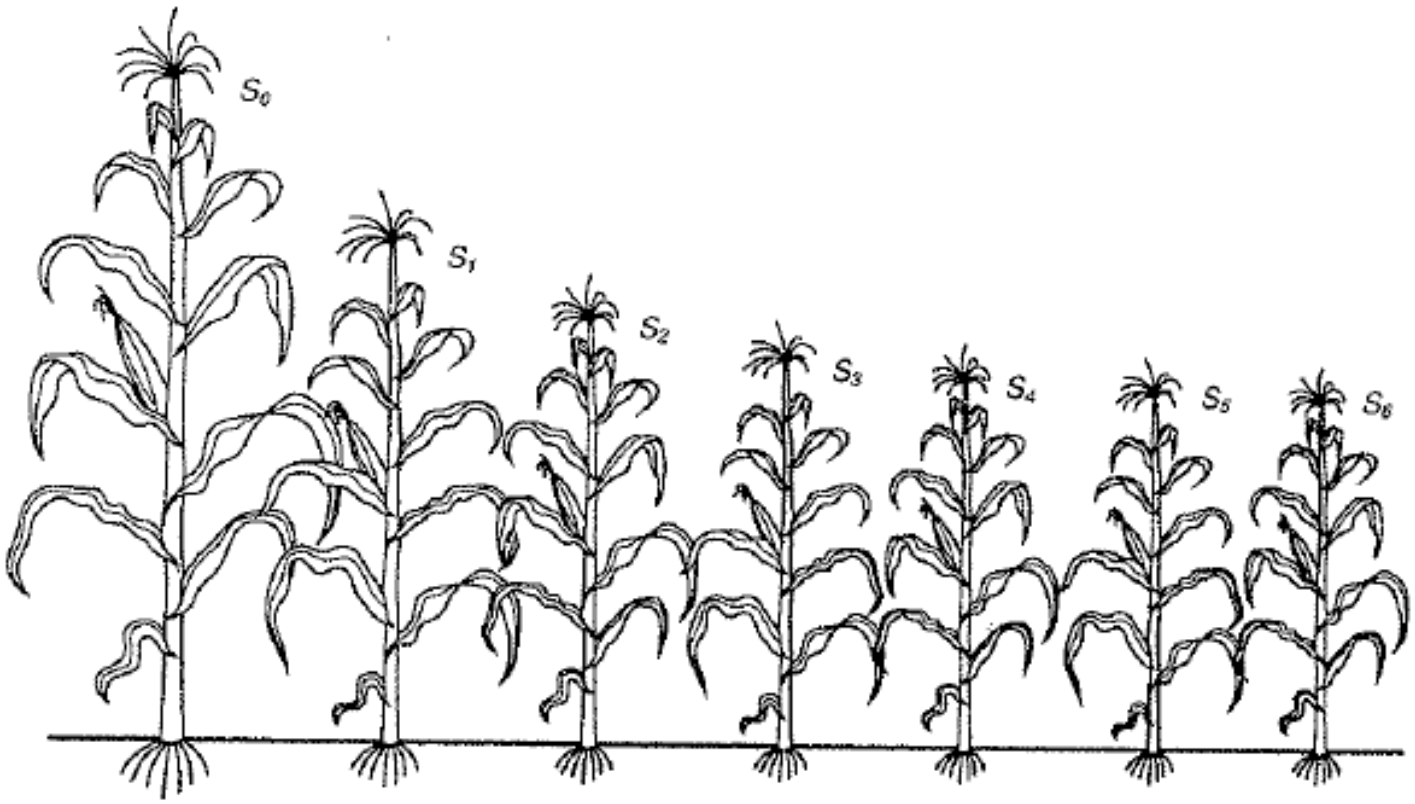
$$\mu_F = \mu_0 - 2F\sum_{i=1}^{k} p_i\, q_i\, d_i = \mu_0 - B\,F$$

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

- There will be a change of mean value if dominance is present (d not 0)
- For a single locus, if d > 0, inbreeding will decrease the mean value of the trait. If d < 0, inbreeding will increase the mean
- For multiple loci, a decrease (inbreeding depression) requires directional dominance --- dominance effects d_i tending to be positive.
- The magnitude of the change of mean on inbreeding depends on gene frequency, and is greatest when p = q = 0.5

Inbreeding Depression and Fitness traits

Inbreeding depression



Example for maize height

Fitness traits and inbreeding depression

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

Why do traits associated with fitness show inbreeding depression?

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

Inbred depression in largely selfing lineages

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

Evolution of the Selfing Rate

- Automatic selection (the cost of outcrossing)
  - An allele that increases the selfing rate has a 50% advantage
  - Pollen discounting
- Selection for reproductive assurance
  - When population density is low, or pollinators rare, failure to outcross may occur
  - Baker's law: Colonizing species generally have the ability to self.

What stops all plants from being selfers?

Inbreeding depression. If fitness of selfed-produced offspring is less than 50% of that from outcrossed-produced

$$w(\eta) = \eta\,\overline{w}_1 + \frac{1}{2}(1-\eta)\,\overline{w}_0 + \frac{1}{2}(1-\overline{\eta})\,\overline{w}_0$$

$$\frac{\partial w(\eta)}{\partial \eta} = \overline{w}_1 - \frac{\overline{w}_0}{2}$$

Lande and Schemske (1985)

- As selfing rate increases, inbreeding load can decrease
  - If inbreeding largely due to recessive or partially recessive deleterious alleles, the mutation-selection equilibrium frequency decreases in selfers
  - As inbreeding load decreases, alleles that increase outcrossing rate are not favored
  - Hence, once largely selfing, very hard to revert

Marker-bases estimation of *f*

- Single-point estimators
  - Excess homozygosity, deficiency of heterozygotes
  - IBD vs correlation estimates
  - The idea of a reference population for allele frequencies
- Haplotype-based estimators
  - Runs of homozygosity (ROH)

Homozygosity estimators

Consider a focal individual, and let x_i denote the number of major alleles (B_i) at a biallelic locus i (alleles with frequency $p_i \geq 0.5$), where x_i equals 0, 1, or 2, for, respectively, the minor-allele homozygote ($b_i b_i$), the heterozygote ($B_i b_i$), or the major-allele homozygote ($B_i B_i$). Note that $x_i(2 - x_i)$ is only nonzero for a heterozygote. For a biallelic locus (such as almost all SNPs), the expected heterozygote frequency when an individual is inbred to level f becomes

$$\text{freq}(B_i b_i) = (1 - f)2p_i(1 - p_i) \tag{11.41a}$$

We can use this equation to obtain an estimate of f using multilocus data in several ways. First, one could simply average both sides over n markers in the focal individual, yielding the expectation

$$E\left[\frac{1}{n}\sum_{i=1}^{n} x_i(2 - x_i)\right] = (1 - f)\frac{1}{n}\sum_{i=1}^{n} 2p_i(1 - p_i)$$

which rearranges to give the (method-of-moments) estimator

$$\widehat{f}_{HOM,1} = 1 - \frac{\sum_{i=1}^{n} x_i(2 - x_i)}{\sum_{i=1}^{n} 2p_i(1 - p_i)} = 1 - \left(\frac{O[Het]}{E[Het]}\right) \tag{11.41b}$$

We can equivalently consider Equation 11.41b as an estimate based on the observed excess in homozygotes, as the reduction in heterozygotes results in an excess of homozygotes. Equation 11.41b can be rearranged (Purcell et al. 2007) to yield

$$\widehat{f}_{HOM,1} = \frac{O[Hom] - E[Hom]}{n - E[Hom]} \tag{11.41c}$$

as $O[Hom] + O[Het] = E[Hom] + E[Het] = n$.

Alternately, we can arrange Equation 11.41a as

$$f = 1 - \frac{\text{freq}(B_i b_i)}{2p_i(1-p_i)} \tag{11.42a}$$

which yields an alternative estimator,

$$\widehat{f}_{HOM,2} = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{x_i(2-x_i)}{2p_i(1-p_i)} \tag{11.42b}$$

Notice that his weights rarer heterozygotes more than the previous estimator, which weighted all equally.

Allelic-correlation estimators

IBD-based estimators effectively assume some ancestral population (typically unspecified) forms the reference (Wang 2014). An alternative was offered by Yang et al. (2011), who proposed an estimator based on the correlation among uniting gametes,

$$\widehat{f}_Y = \frac{1}{n}\sum_{i=1}^{n}\gamma_i, \quad \text{where} \quad \gamma_i = \frac{x_i^2 - (1+2p_i)x_i + 2p_i^2}{2p_i(1-p_i)} \tag{11.43a}$$

where the weights simplify to

$$\gamma_i = \begin{cases} (1-p_i)/p_i & \text{for } B_iB_i \;\; (x_i = 2) \\ -1 & \text{for } B_ib_i \;\; (x_i = 1) \\ p_i/(1-p_i) & \text{for } b_ib_i \;\; (x_i = 0) \end{cases} \tag{11.43b}$$

Runs of Homozygosity (ROH)

Single-point estimators of f use no positional information, averaging data from *individual markers*, rather than using *haplotypes*, and indeed often discarding some SNP data to avoid complications from LD. With the advent of dense SNP chips and whole-genome sequencing, the haplotype structure (i.e., LD) of SNPs can be used to obtain direct estimates of the fraction of the genome that is autozygous (Broman and Weber 1999; Chapman and Thompson 2003; McQuillan et al. 2008; Keller et al. 2011; Ceballos et al. 2018). The idea is to use **runs of homozygosity (ROHs)**, where a run is defined as a continuous DNA segment that is completely homozygous. This leads to the estimate of the fraction of the genome that is autozygous as

$$\widehat{f}_{ROH} = \frac{\text{total length of ROHs}}{\text{genome size}} \qquad (11.44)$$

Issues:
(i) Setting the threshold size
(ii) : Using physical vs genetic distances

Identity disequilibrium

- ID: a correlation among the identity states of loci
- If locus a is inbred, what does this tell us about locus b?
- Impact of linkage is somewhat minor. The expected ID between two fully linked loci is just twice the ID between unlinked loci under partial selfing
- g_2 metric and connection with variance in f

A more formal ID metric, g_2, was introduced by David et al. (2007). This parameter follows by considering the expectation that *pairs* of loci are jointly heterozygous within an individual, which can be written as

$$E[H_i H_j] = E[H_i]\, E[H_j]\, (1 + g_2) \tag{11.37a}$$

where H_i is a zero-one indicator variable of whether locus i is a heterozygote. If heterozygosity is uncorrelated over loci, then $E[H_i H_j] = E[H_i]\, E[H_j]$ and $g_2 = 0$. If the observation that one locus is heterozygous in an individual inflates the probability that other loci in that individual are also heterozygous, then $g_2 > 0$. David et al. proposed an approximate estimator of g_2, with an improved version offered by Stoffel et al. (2016). This metric is called g_2 because k-way disequilibrium metrics (g_k) were also proposed, with the expectation of joint heterozygosity at k loci within an individual being written as

$$E\left[\Pi_{i=1}^{k} H_i\right] = \left(\Pi_{i=1}^{k} E\left[H_i\right]\right) (1 + g_k) \tag{11.37b}$$

The connection between g_2 and the variance in inbreeding levels among individuals within in the population, $\sigma^2(f)$, was shown by David et al. (2007) to be

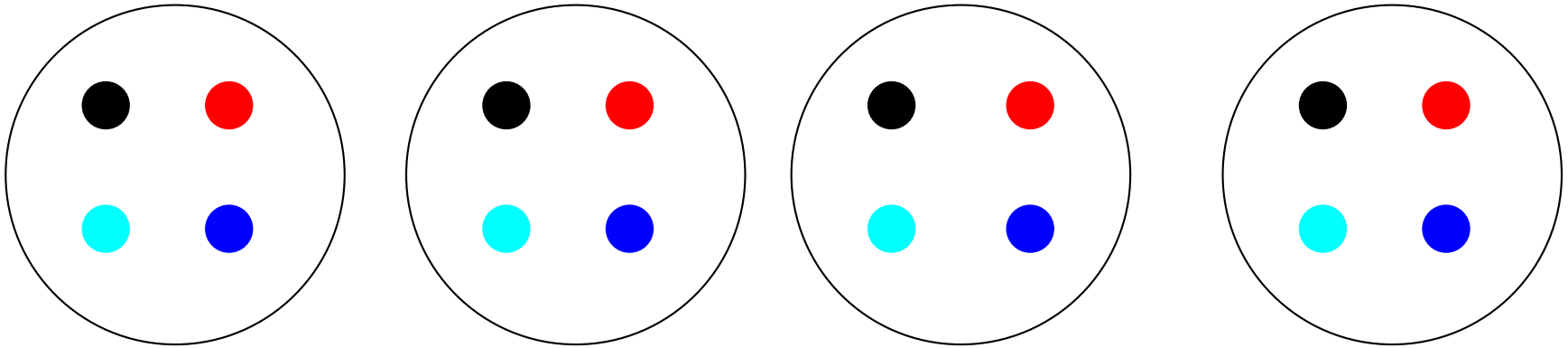
$$g_2 = \frac{\sigma^2(f)}{(1 - \overline{f})^2} \tag{11.37c}$$

where $\overline{f}$ is the average level of inbreeding in the population. In the absence of variance in the level of inbreeding, $g_2 = 0$ and there is no ID. Thus even a highly inbred population can *fail* to show ID if all individuals are *inbred to the same level*.

Variance Changes Under Inbreeding

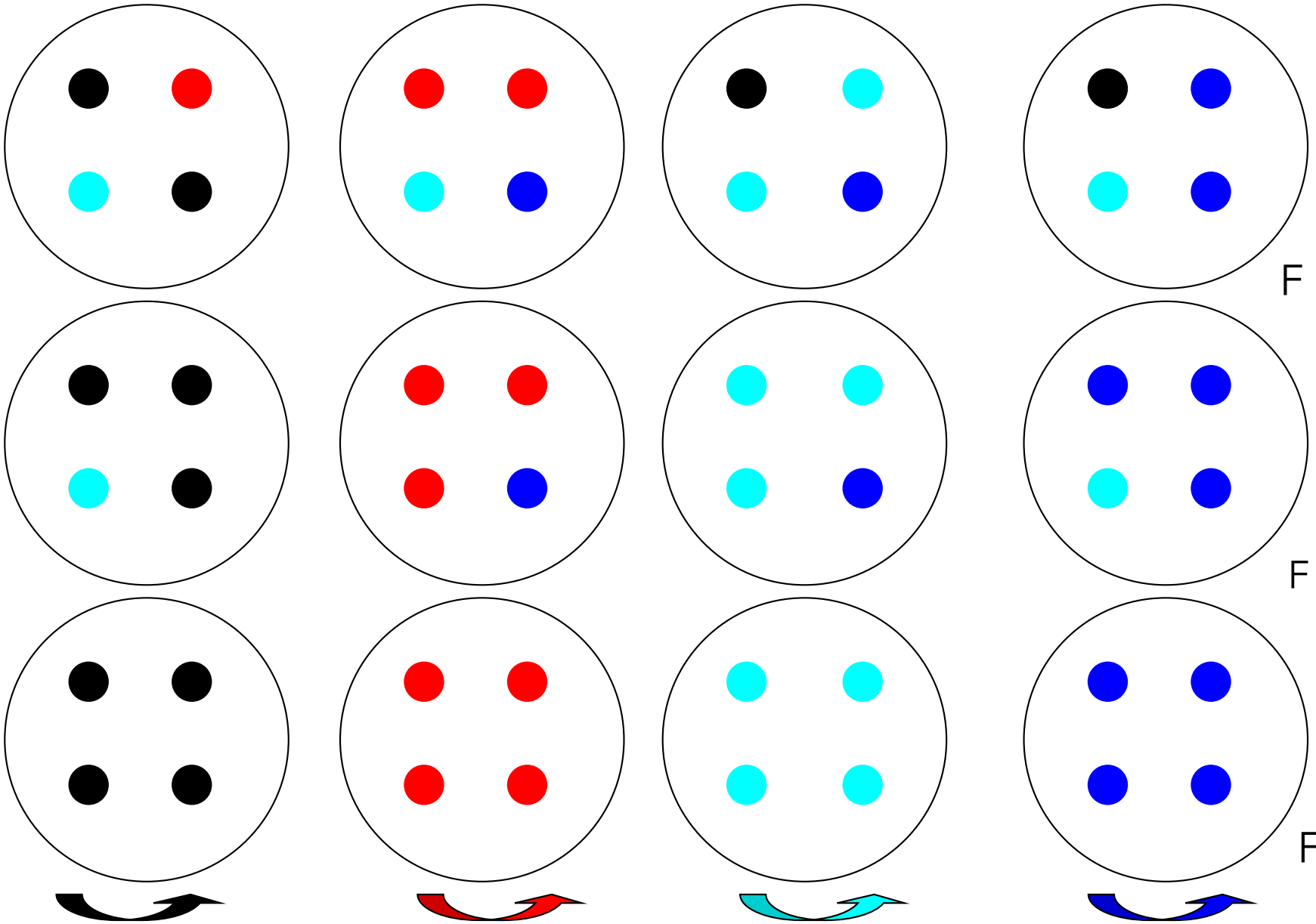
Inbreeding reduces variation within each population

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



F = 0

Between-group variance increases with F



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	$(1-F)\ V_A$	0	V_A
Total	$(1+F)\ V_A$	$2V_A$	V_A

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

Line Crosses: Heterosis

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

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

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

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

Expected levels of heterosis

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

The expected amount of heterosis becomes

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

- Heterosis depends on dominance: d = 0 = no inbreeding depression and no Heterosis. As with inbreeding depression, directional dominance is required for heterosis A X A epistasis can also generate heterosis, while epistasis must include D for inbreeding depression (e.g., A X D, D X D)

- 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	55×10^6	13×10^6 ha
Sorghum	48	40	19	13×10^6	9×10^6 ha
Sunflower	60	50	30	7×10^6	6×10^6 ha
Rice	12	30	4	15×10^6	6×10^6 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

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

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

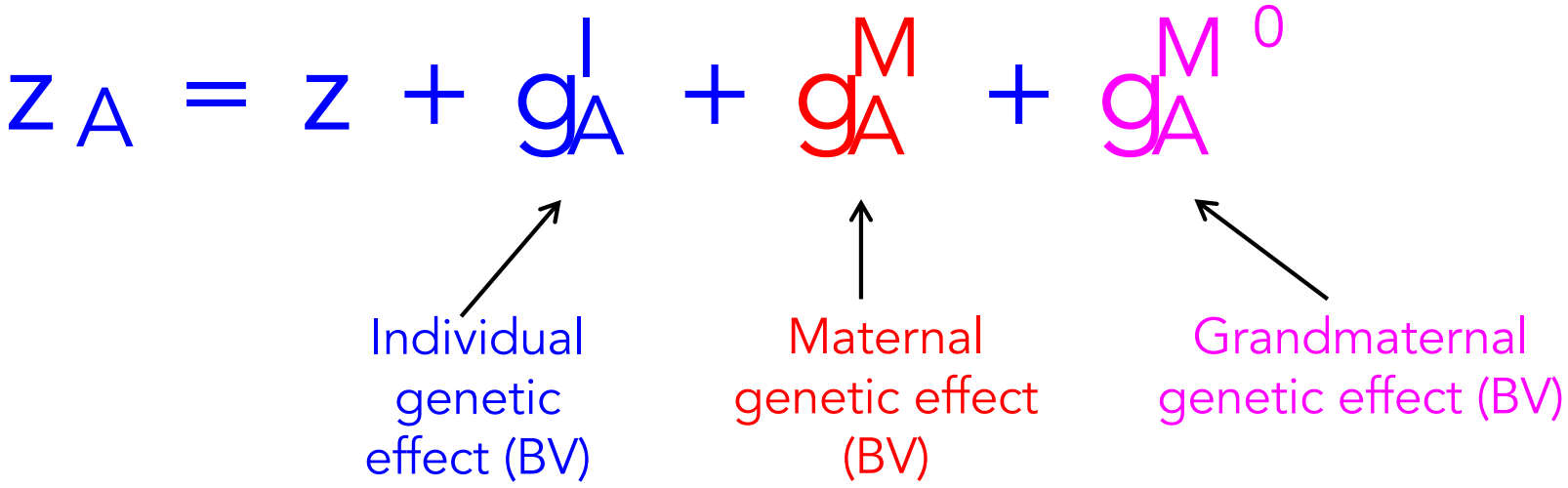
- Individual heterosis
  - enhanced performance in a hybrid individual
- Maternal heterosis
  - enhanced maternal performance (such as increased litter size and higher survival rates of offspring)
  - Use of crossbred dams
  - Maternal heterosis is often comparable, and can be greater than, individual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

Individual vs. Maternal Heterosis in Sheep traits

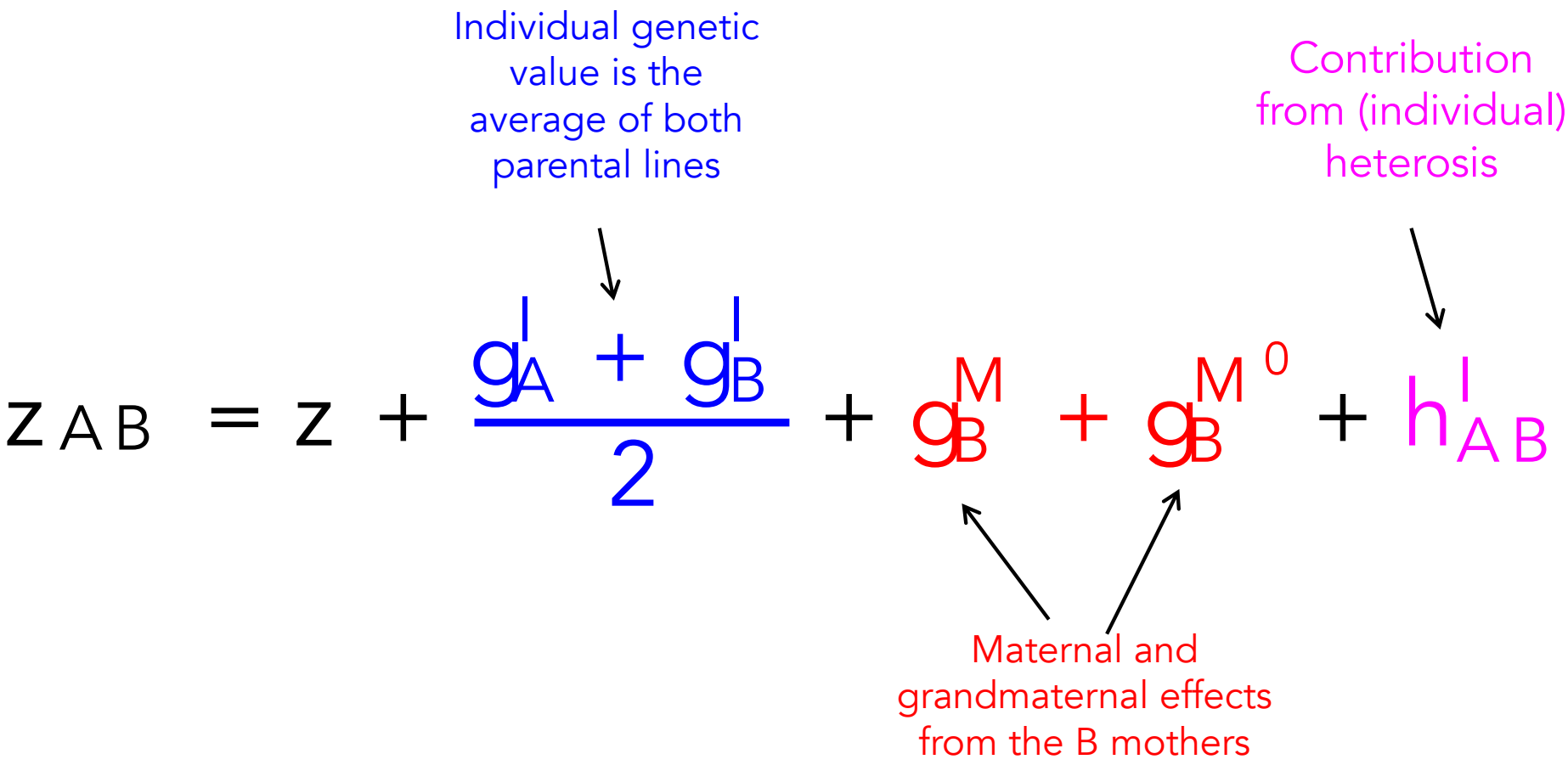
Trait	Individual H	Maternal H	total
Birth weight	3.2%	5.1%	8.3%
Weaning weight	5.0%	6.3%	11.3%
Birth-weaning survival	9.8%	2.7%	12.5%
Lambs reared per ewe	15.2%	14.7%	29.9%
Total weight lambs/ewe	17.8%	18.0%	35.8%
Prolificacy	2.5%	3.2%	5.7%

Estimating the Amount of Heterosis in Maternal Effects

Contributions to mean value of line A



Consider the offspring of an A sire and a B dam



$$z_{AB} = z + \frac{g^I_A + g^I_B}{2} + g^M_B + g^{M^0}_B + h^I_{AB}$$

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

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