Lecture 01: Introduction

UNE course:
The search for selection
3-- 7 Feb 2020

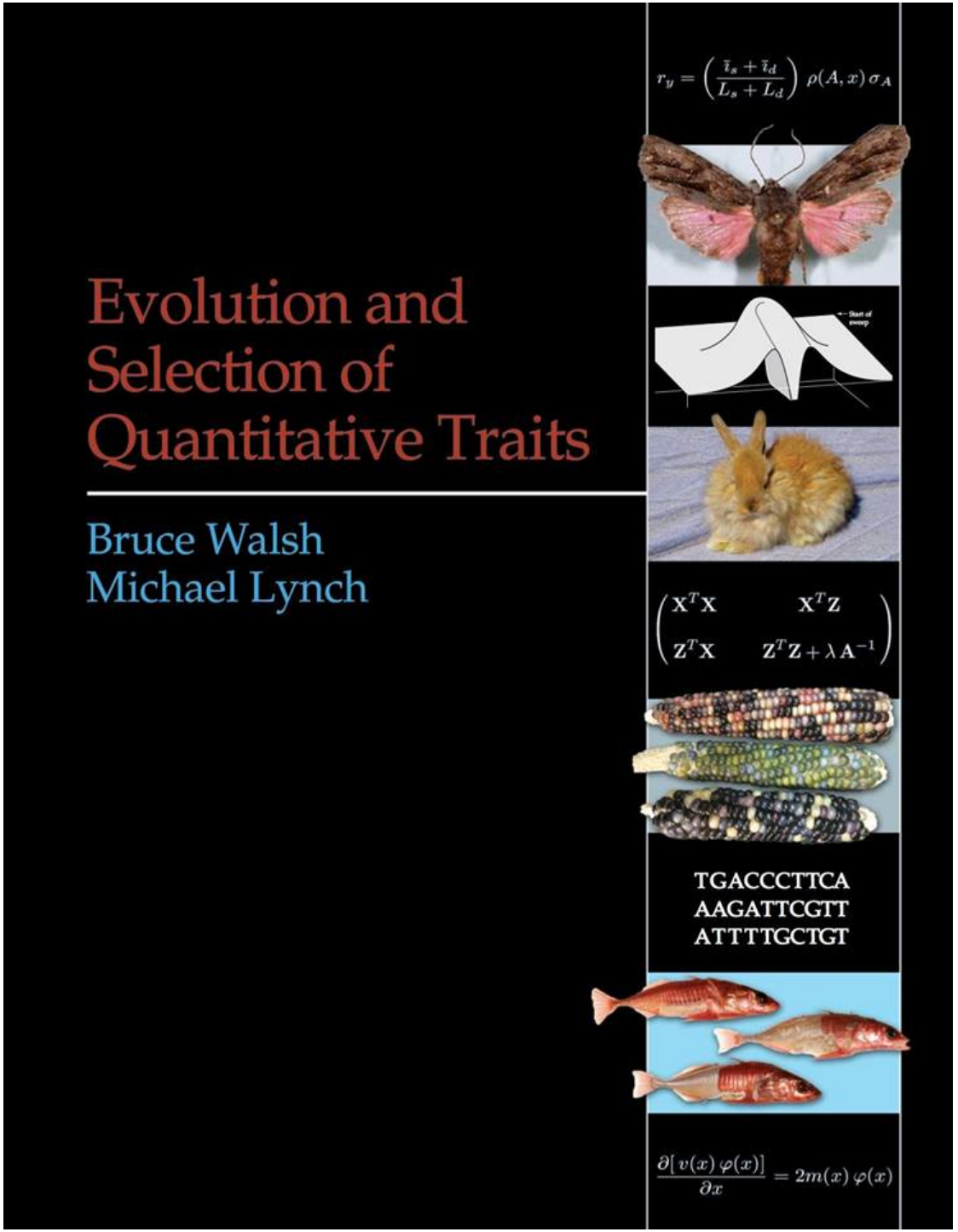
Bruce Walsh (University of Arizona)
jbwalsh@email.arizona.edu

Searching for signatures of selection

- Patterns of phenotypic divergence
  - Too little or too much divergence relative to drift
  - Examples: fossil data, temporal or spatial data on mean phenotypes, mRNA expression data
- Patterns in marker data associated with specific traits
  - A nonrandom pattern in QTL or GWAS effects
  - Example: Excess of “plus” QTL alleles for height
- Patterns in marker data (trait-independent)
  - Huge number of tests based on polymorphism data, divergence data, or both
  - Examples: Tajima’s D, SDS test, McDonald-Kreitman, HKA test
- Association between fitness data and trait phenotypes
  - Requires estimates of individual fitness (nontrivial)
  - Example: Lande-Arnold fitness regressions

Walsh & Lynch
2018, aka Vol 2

Course covers
Chapters 8--10,
12, 29--30



A little about me

- BS in Mathematical Population Biology from UC Davis (Michael Turelli)
- PhD in Genetics from University of Washington (Joe Felsenstein)
- Post-doc at University of Chicago (Tom Nagylaki, Russ Lande)
- Been at University of Arizona for >30 years
  - Depts of Ecology & Evolutionary Biology, Public Health, Plant Science, Animal Science, Molecular & Cellular Biology

Research Interests

- Evolutionary biology
- Genetics of complex traits
- Statistical and mathematical modeling
- Animal and plant breeding (applied evolution)
- I've taught many different courses on quantitative genetics in 25 countries, covering
  - animal breeding
  - plant breeding
  - evolution
  - mathematical modeling
  - statistics
  - human genetics

Hobbies

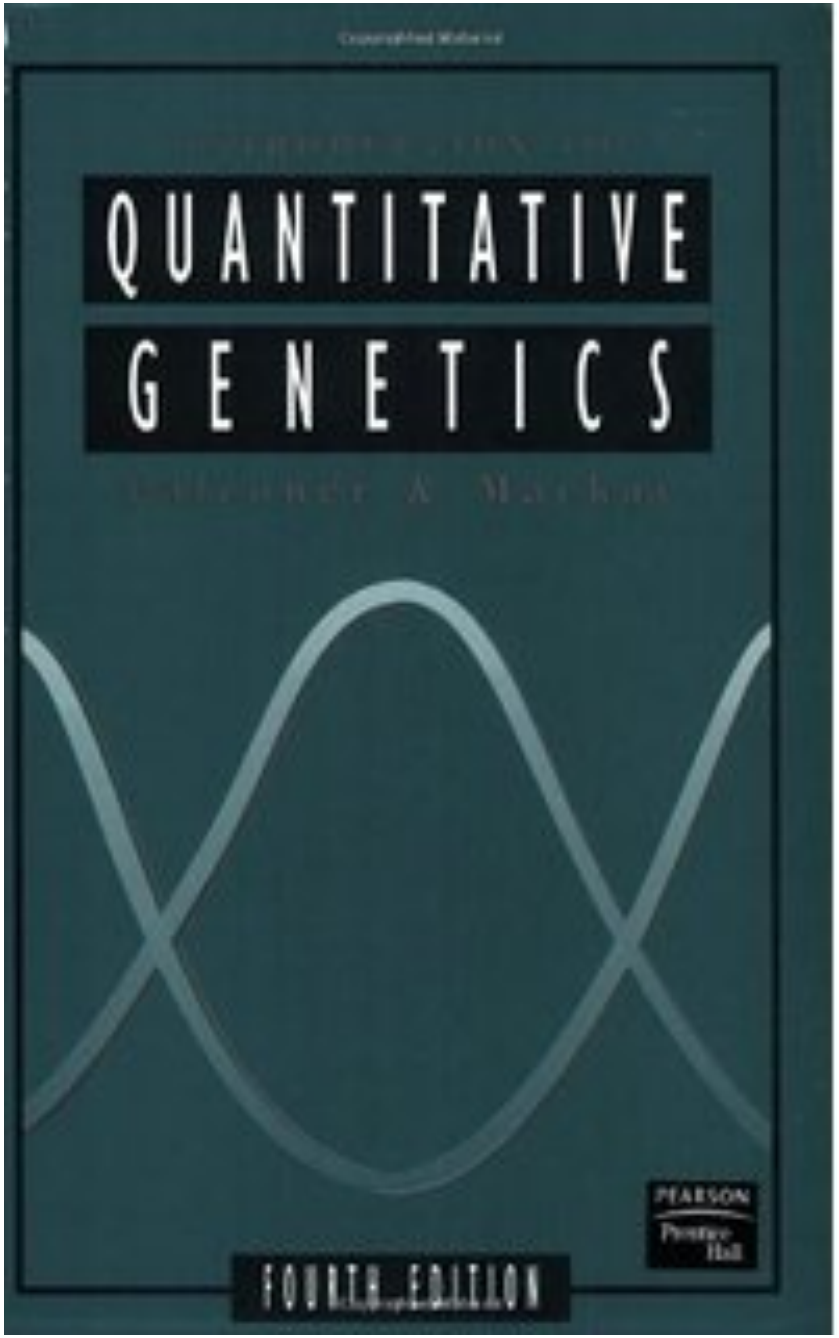
- I have an odd hobby, moth collecting

Time to tell the class about you!

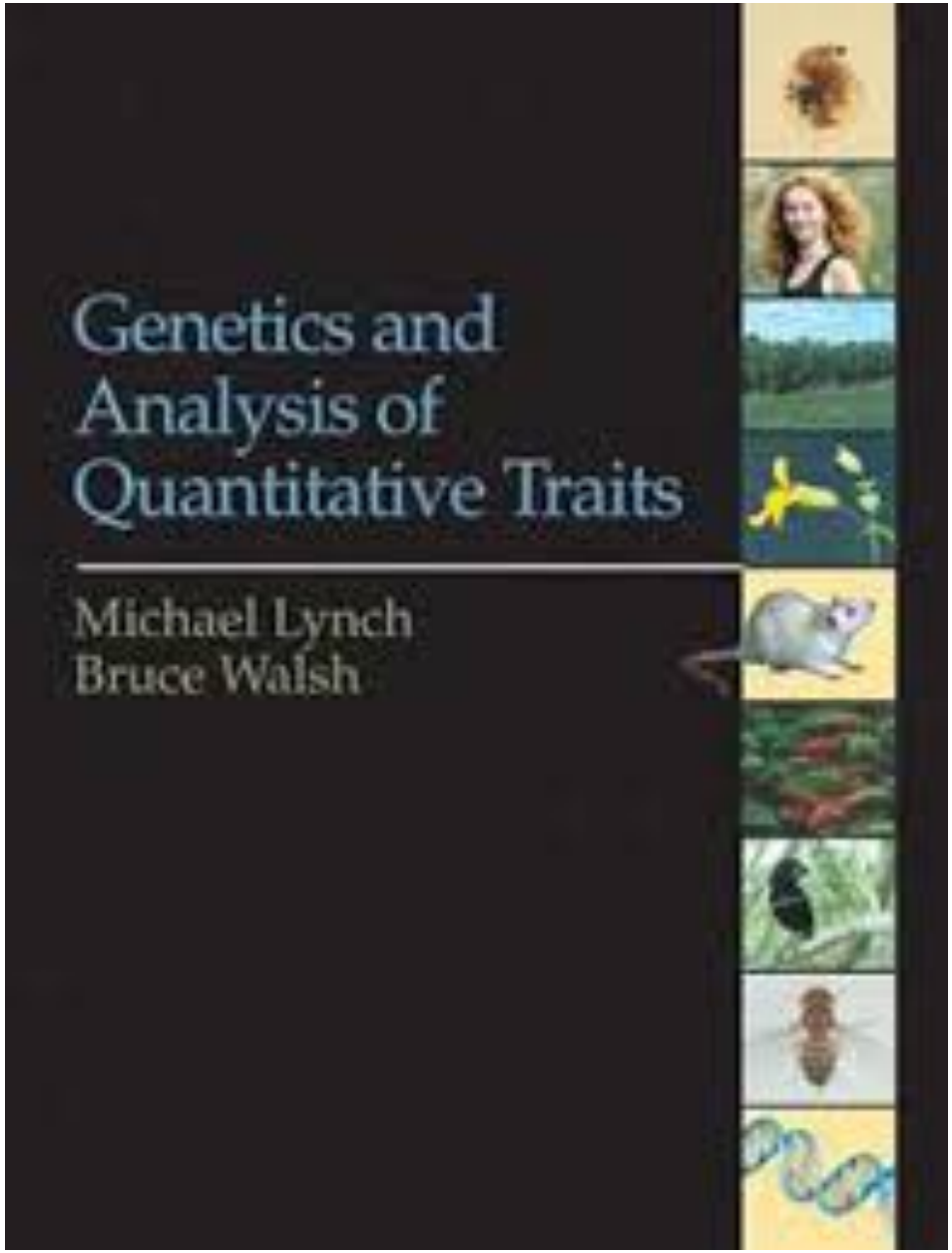
- Name
- Where you are located and your current status (faculty, student, post-Doc, etc.)
- Current research interest
- Why you are taking this class
- One topic you would like addressed in the class
- A fun fact about

Background: Quantitative genetics

- Fisher's variance decomposition
- Additive genetic variance and heritability
- Inbreeding and effective population size
- Mutational variance
- Additive variance under drift and mutation
- The breeder's equation and selection response



Falconer



Lynch & Walsh (1998)
(aka Vol 1)

Basic model of Quantitative Genetics

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

Basic model: P = G + E ← Environmental value

Genotypic value

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

Hence, genotypic values are functions of the environments experienced.

Basic model of Quantitative Genetics

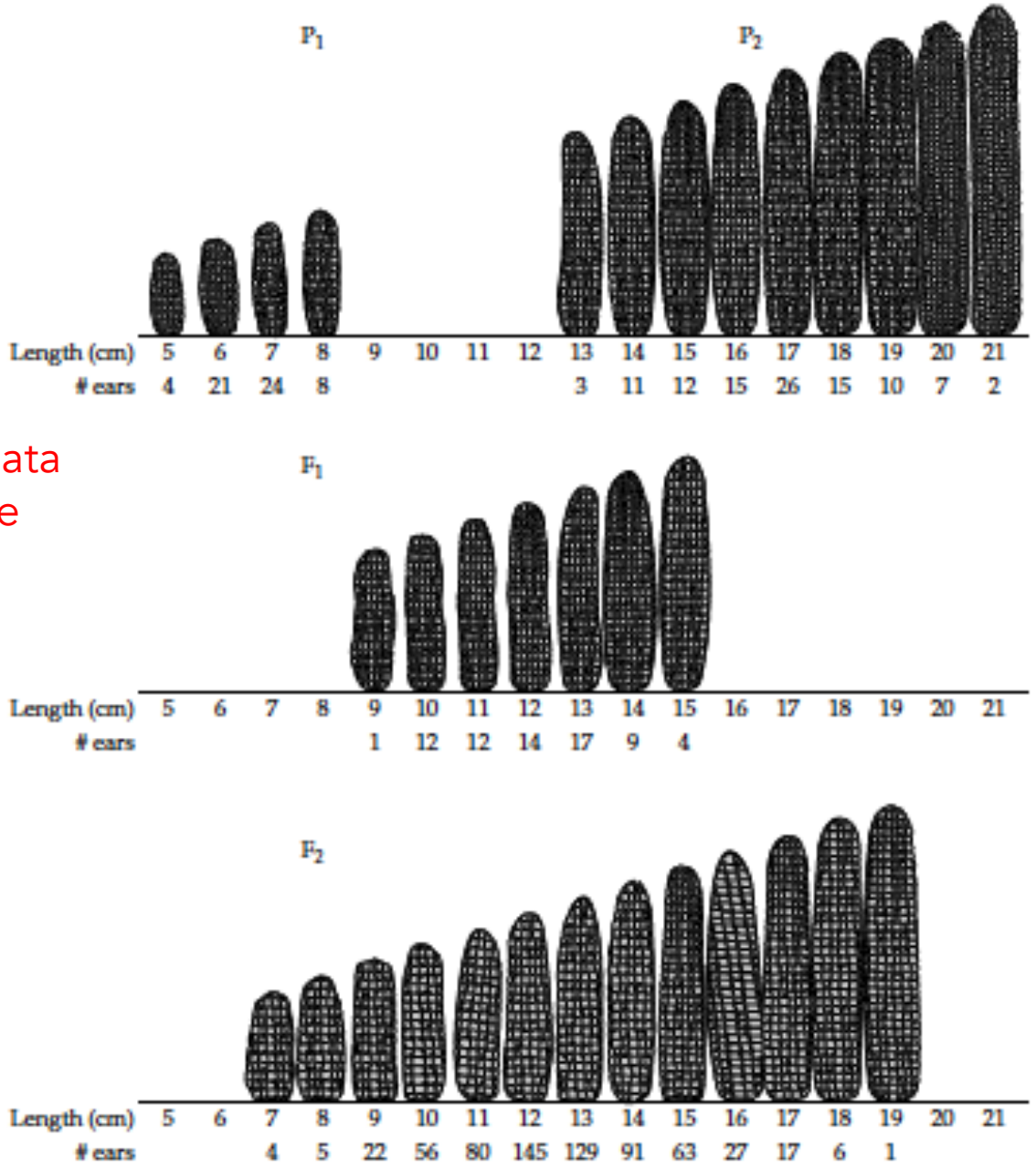
Basic model: P = G + E

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

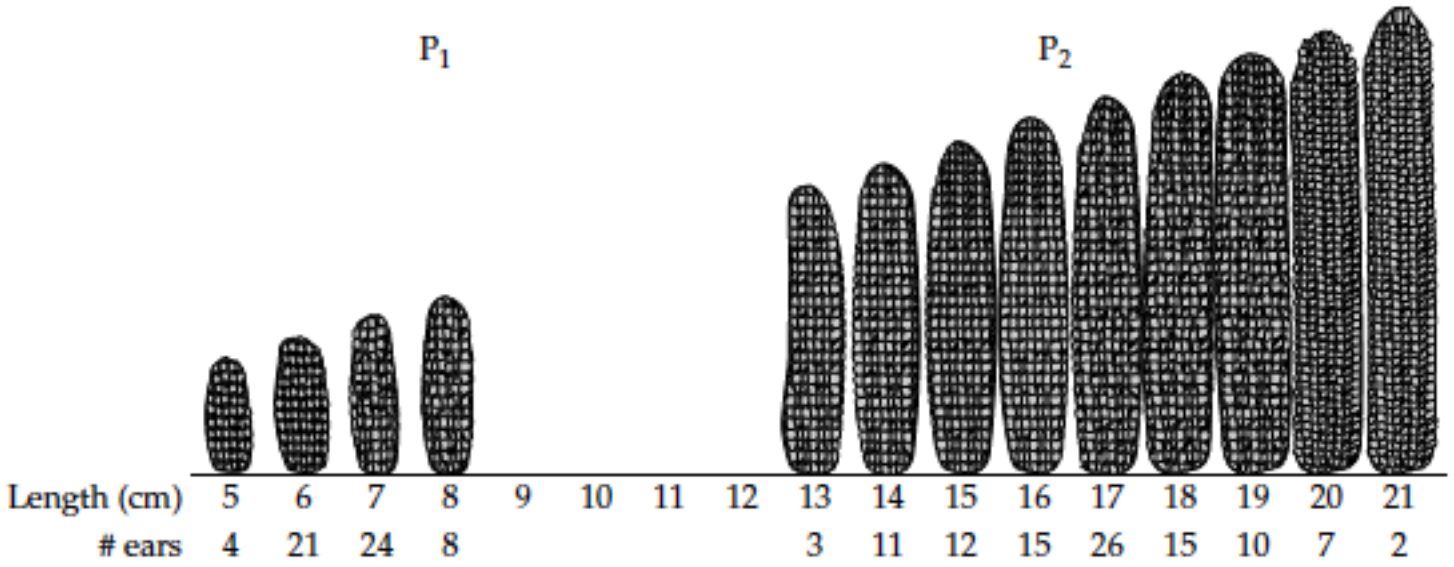
G = average value of an inbred line over a series of environments

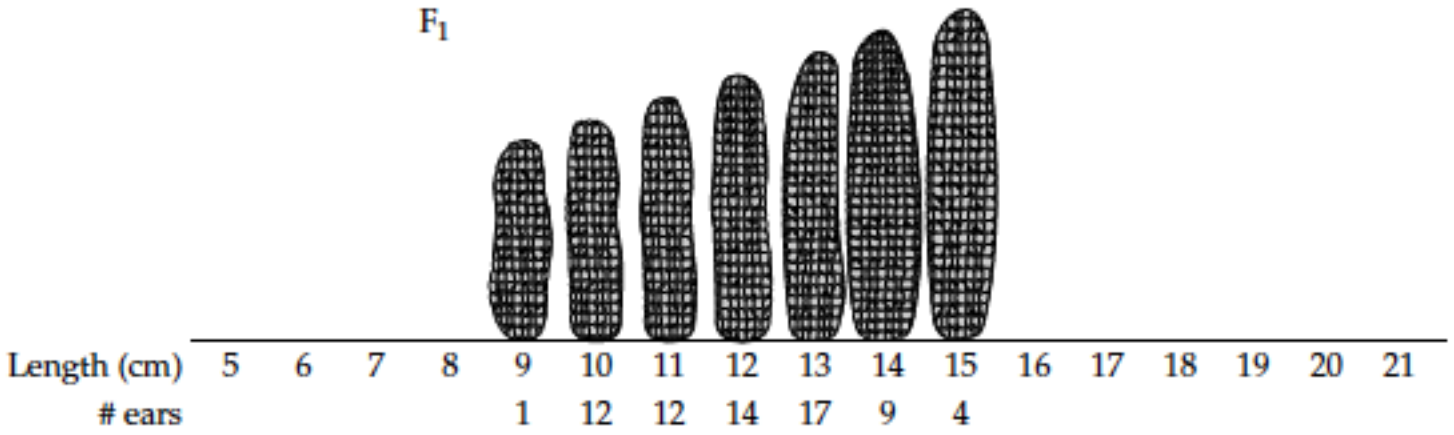
G x E interaction --- The performance of a particular genotype in a particular environment differs from the sum of the average performance of that genotype over all environments and the average performance of that environment over all genotypes. Basic model now becomes P = G + E + GE

East (1911) data on US maize crosses

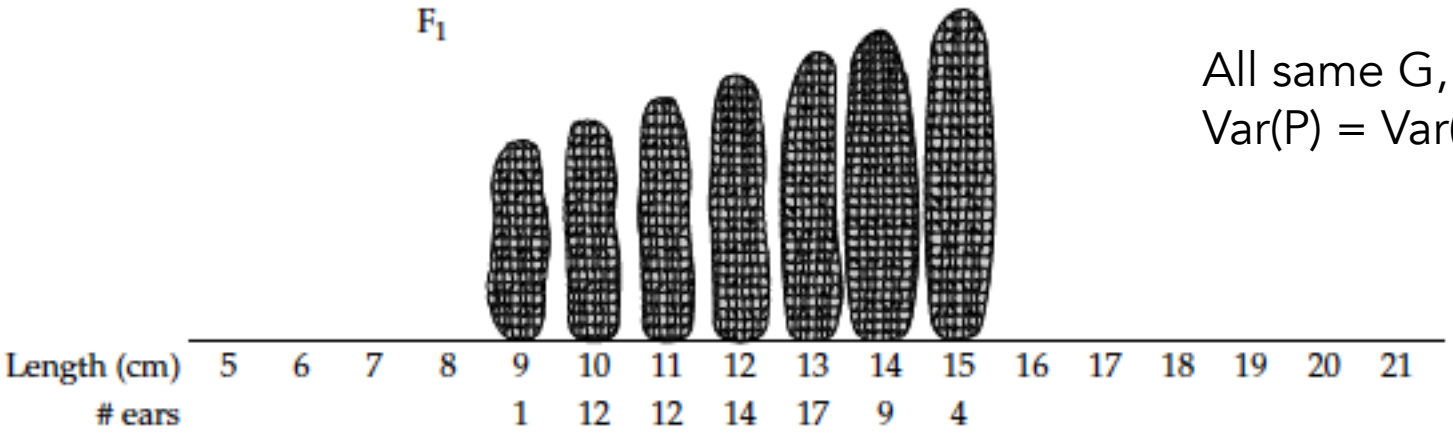


Same G, Var(P) = Var(E)

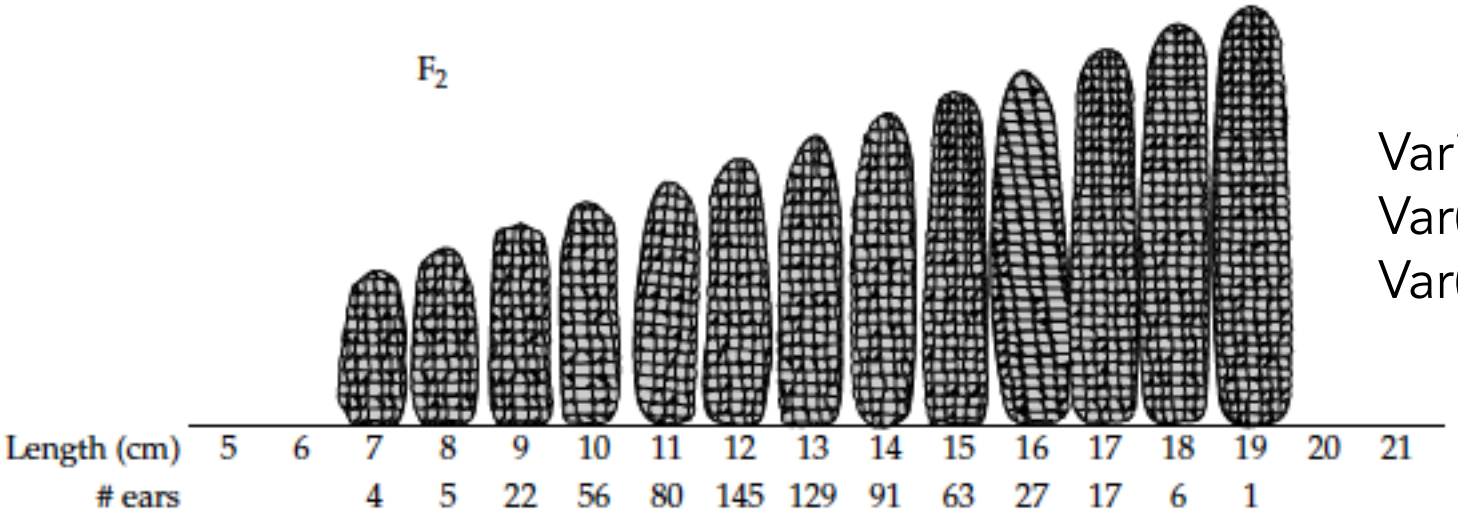




Each sample (P_1, P_2, F_1) has same G, all variation in P is due to variation in E



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



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

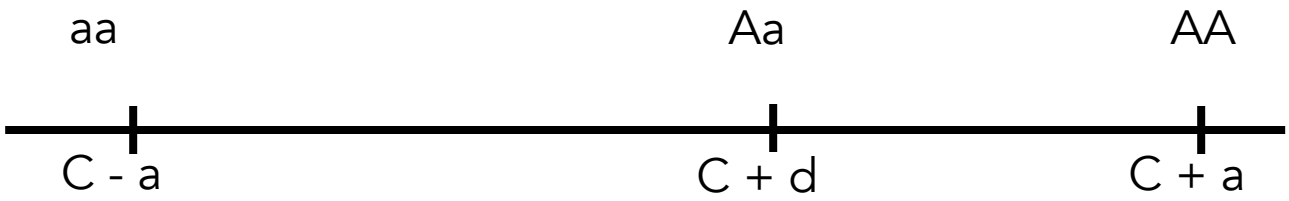
$Var(F_2) > Var(F_1)$ due to Variation in G

The transmission of genotypes versus alleles

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

Genotypic values

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



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

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

The average effect of an allele

- The average effect α_A of an allele **A** is defined by the difference between offspring that gets allele **A** 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)]

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

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

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

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

A is called the Breeding value or the Additive genetic value

Key features of breeding values

- The expected mean deviation of an offspring is simply the average of the breeding values of its parents
- Because breeding values are deviations from the population mean, the expected breeding value of a random individual is zero
- If a (say) male is crossed to a number of random, unrelated females, the expected mean deviation in his offspring is simply BV/2. Hence, can easily estimate a BV via crosses (BLUP is a generalization of this idea, see WL Chapters 19, 20)

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 $Cov(\alpha,\delta) = 0$

Genetic Variances

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

Additive Genetic Variance
(or simply Additive Variance)

Dominance Genetic Variance
(or simply dominance variance)

Hence, total genetic variance = additive + dominance variances,

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

Key concepts (so far)

- 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 details

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

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

$= 2a^2p(1-p)$ when k = 0

Dominance variance

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

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

Equals zero if k = 0

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

This is a symmetric function of allele frequencies

Can also be expressed in terms of d = ak

Can estimate genetic variances from sets of relatives

- The slope of a parent-offspring regression is simply
  - [Var(A)/2]/ Var(z) = $h^2/2$
  - Where h^2 is the (narrow-sense) heritability
- More generally, the phenotypic covariance between known sets of relatives can be used to estimate genetic variances
  - See Falconer (basics) or Lynch and Walsh (all the gory stuff) for details

Drift

Sewall Wright

R. A. Fisher

Genetic Drift

Random sampling of 2N gametes to form the N individuals making up the next generation results in changes in allele frequencies.

This process, originally explored by Wright and Fisher, is called Genetic Drift.

Suppose there are currently i copies of allele A, so that freq(A) = p = i/(2N)

What is the probability that, following a generation of random sampling, the freq of A is j/(2N)?

Wright-Fisher model (or process) of drift

This probability follows binominal sampling,

$$\Pr(i \text{ copies} \rightarrow j \text{ copies}) = \frac{N!}{(N-j)!j!}\left(\frac{i}{N}\right)^{j}\left(\frac{N-i}{N}\right)^{N-j}$$

p = i/N 1- p

Hence, if the current allele frequency is p, the expected allele frequency in the next generation is also p, but with sampling variance p(1-p)/(2N)

Thus, when N is large, the changes in allele frequency over any generation are expected to be rather small

However, the cumulative effects of generations of such sampling are very considerable.

Eventually, any random allele will either be lost from the population or fixed (frequency one).

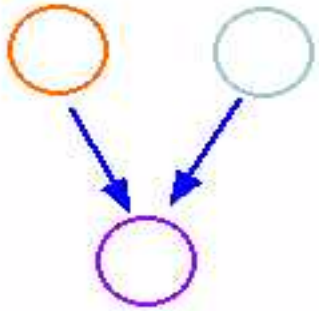
If the allele has initial frequency p, then

$$\Pr(\text{Fixation}) = p$$

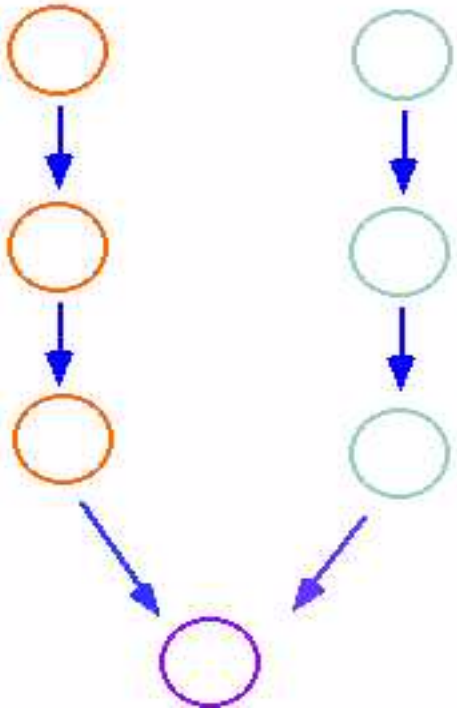
$$\Pr(\text{loss}) = 1 - p$$

The expected time to fixation is on order of 4N generations.

pick 2 random alleles (sequences)



Probability they share common ancestor in last generation = 1/2N



Expected number of generations to reach the common ancestor (sequence) is 2N generations

Impact of drift

- Random change in allele frequencies results in loss of genetic variation over time as alleles become fixed
- "power" (i.e., time scale) of drift is
  - ~ 1/(2N) per generation
  - 2N is the expected time back to the common ancestor for two random sequences (the coalescence time)
  - 4N generations is the expected time back to the common ancestor for the entire population

Loss of heterozygosity under drift

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t$$

$$H_t \simeq H_0 e^{-t/(2N)}$$

$$t = -2N \ln(H_t/H_0)$$

Loss of additive variation under drift

- More complicated with non-additive gene effects, see WL Chapters 11 and 23
- With only additive effects (d = 0, no epistasis), then $Var(A) = a^2 * H$, where H is the heterozygosity.
- Hence, over time, drift removes any additive variance
  - $Var(A,t) = Var(A,0)[\ 1 - 1/(2N)\ ]^t$

Actual vs. effective population size

- The actual size of a population N is far less important than the effective number of breeding individuals, or the effective population size, N_e, which sets the strength of drift.
- N_e replaces N in expressions for the power of drift

Effective Population size, N_e

When the population is not ideal (changes over time, unequal sex ratio, uneven contribution from individuals), we can still compute an effective population size N_e which gives the size of an ideal population that behaves the same as our population

We will consider N_e under

- population bottlenecks
- unequal sex ratio
- unequal contribution for all individuals

All the details in WL Chapter 3

N_e under varying population size

If the actual population size varies over time, the effective population size is highly skewed towards the smallest value

If the populations sizes have been N(1), N(2), …, N(k), the effective population size is given by the harmonic mean

$$N_e = \frac{k}{\sum_{i=1}^{k} \frac{1}{N(i)}}$$

Suppose the population sizes are 10000, 10000, 10000, 100.

N_e becomes 399

If 10000 is replaced by 10^9, N_e becomes 400

N_e under unequal sex ratios

When there are different number of males (N_m) and females (N_f), the effective population size is skewed towards the rarer sex

$$N_e = \frac{4N_m * N_f}{N_m + N_f}$$

Suppose we used 2 pollen plants to fertilize a 1000 seed plants. What is N_e in this case?

N_e = (4*2*1000)/(2 + 1000) = 8

N_e under unequal individual contributions

Not all individuals contribute equally to the next generation. What effect does this have on N_e?

Let σ^2_0 be the variance in offspring number for individuals in the population, then

$$N_e \simeq \frac{2N}{\sigma_O^2/2 + 1}$$

If contributions follow a Poisson distribution with a mean of 2 offspring per parent (male + female replace each other), then $\sigma^2_0 = 2$, and $N_e = N$

N_e under unequal individual contributions (cont)

$$N_e \simeq \frac{2N}{\sigma_O^2/2 + 1}$$

If all individuals contribute EXACTLY the same number of offspring, $\sigma^2_0 = 0$, and $N_e = 2N$, so that the effective pop size is twice the actual size

Mutational variance

- What keeps additive variance from being driven to zero by drift? New mutation!
- Quantitative-trait variation is a function of (1) the mutation rate and (2) the mutational effects (details WL Chapter 28)
- Var(m) = $2Nu\ Var(a_m)$
- Var(m)/Var(E) is often called the mutational heritability, h^2_m
- This is typically around 10^{-3}

Table 12.1 Estimates of the mutational heritability for a variety of organisms and characters.

Species	Character	h_m^2	Reference
Drosophila melanogaster	Abdominal bristle number	0.0035	See text
	Sternopleural bristle number	0.0043	See text
	Enzyme activities	0.0022	Clark et al. 1995b Harada 1995
	Ethanol resistance	0.0009	Weber and Diggins 1990
	Body weight	0.0047	Clark et al. 1995b
	Wing dimensions	0.0020	Santiago et al. 1992
	Viability	0.0003	Mukai 1964 Mukai et al. 1972 Cardellino and Mukai 1975 Ohnishi 1977
Tribolium castaneum	Pupal weight	0.0091	Goodwill and Enfield 1971
Daphnia pulex	Life-history traits	0.0017	Lynch 1985
Mouse	Lengths of limb bones	0.0234	Bailey 1959
	Mandible measures	0.0231	Festing 1973
	Skull measures	0.0052	Carpenter et al. 1957 Deol et al. 1957 Hoi-Sen 1972
	6-week weight	0.0034	Caballero et al. 1995
Arabidopsis thaliana	Life-history traits	0.0039	Schultz et al. (in prep.)
Maize	Plant size	0.0112	Russell et al. 1963
	Reproductive traits	0.0073	Russell et al. 1963
Rice	Plant size	0.0030	Oka et al. 1958
	Reproductive traits	0.0028	Sakai and Suzuki 1964
Barley	Life-history traits	0.0002	Cox et al. 1987

Table from LW

Mutation-drift equilibrium

- Drift removes variation, mutation introduces it
- Equilibrium additive variance (under neutrality) is $2N_eVar(m)$
- A delicate issue is the assumed mutational effects model
  - $a' = a + \alpha$, where α is the mutation effect.
  - This is called the incremental or Brownian-motion model (Hill and Lynch)
  - Depends on history: previous effect value (a)

Other models

- House-of-cards, HOC (Kingman)
  - $a' = \alpha$
  - Independent of history (current value)
- Regression model (Zeng and Cockerham)
  - $a' = \tau\, a + \alpha$
  - τ = impact of history. $\tau = 0$ recovers HOC, $\tau = 1$ recovers incremental
- All the gory details in WL Chapter 28

The breeder's equation

- The response to selection, R, (the across -generation change in the mean) is related to the within-generation change in the mean (S) by the breeder's equation
  - $R = h^2S$
  - Chapters 6, 13-17, and 20 in WL cover this in detail
  - The selection intensity $i = S/\sigma_z$, i.e., the within -generation change in standard deviations
  - The selection gradient $\beta = S/\sigma_z^2$,
  - The breeder's equation can be written as
    - $R = \beta \sigma_A^2$, which is called the Lande Equation