BIOL 434/509: Population Genetics

Course web address:

http://www.zoology.ubc.ca/~whitlock/bio434/

Canvas: https://canvas.ubc.ca/courses/82642

Prof. Michael Whitlock
Department of Zoology
Office: 216 Biodiversity

whitlock@zoology.ubc.ca

Office hours: Mon. 2:30-3:30
and after class most days

on Zoom (see link on Canvas)

Textbook

Graham Coop

POPULATION AND QUANTITATIVE GENETICS

Available FREE at

[*https://github.com/cooplab/popgen-notes/blob/master/popgen_notes.pdf*](https://github.com/cooplab/popgen-notes/blob/master/popgen_notes.pdf)

Click the "Download" button there.

Evaluation

Final	50%
Mid-term	30%
Assignments	20%

Midterm

March 1 or 3, in class

Assignments

Available on Canvas

Due on Fridays at noon (or later -- see specific due time on Canvas)

Factors causing genotype frequency change

- Selection
- Mutation
- Random genetic drift
- Migration
- Recombination
- Non-random mating

Genetics terms

Gene

Multiple meanings:

- Any location in the genome
- Any functional element in the genome
- A protein coding region

Locus: A particular location in the genome

Genetics terms

Nucleotide/base pair

nucleotide is one of the basic building blocks of DNA (or RNA)

guanine, **adenine**, **cytosine** and **thymine**; abbreviated G, A, C, and T

base pair: pairs of GC or AT

Genetics terms

Allele

Two meanings:

- A gene sequence that has a set of particular properties (i.e., allele causing blue eyes or all sequences that contain a premature stop codon)

- The particular versions of DNA at a genetic locus in an individual. (“The individual inherited an allele from its mother and an allele from its father at this locus.”)

Genetic code

RNA codon table

1st position	2nd position				3rd position
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr stop stop	Cys Cys stop Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Amino Acids

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

synonymous mutation:

changes a codon to another codon without changing the amino acid

(e.g., GUU to GUC stays Valine)

nonsynonymous mutation:

changes a codon to another codon that *does* change the amino acid

(e.g., GUU to AUU causes a change from Valine to Isoleucine)

Neutral alleles:

Alleles are neutral if they have the same fitness as alternative alleles at the same locus.

Gamete:

A reproductive cell containing (usually) haploid copy of the genome.

Haploid / diploid

Haploid cells contain one copy of each chromosome in the genome. Diploid cells contain two copies of each chromosome.

Genetic descriptions of populations

Allele frequencies:

the relative frequency (proportion) of all alleles in that population which are that particular allele.

$$p,\ q,\ p_A,\ p_a$$

$$q = 1 - p$$

Allele frequency

Haploid example

24 individuals have a *C* at a particular locus;
12 have a *T* there.

$$p_C = \frac{24}{24 + 12} = 0.667$$

Allele frequency

Diploid example

Color variation caused by single nucleotide on the *mc1r* locus (melanocortin receptor 1).

AA	AG	GG
42	24	21

$$p_A = \frac{2(42) + 24}{2(42 + 24 + 21)} = 0.62$$

Genotype frequencies at one locus

Genotype frequency: frequency of some particular genotype

e.g., Kermode bears *mc1r*:

$$P_{AA} = \frac{42}{87} = 0.483$$

$$P_{AG} = \frac{24}{87} = 0.276 \qquad P_{GG} = \frac{21}{87} = 0.241$$

Calculating allele frequencies from genotype frequencies

$$p_A = \frac{2P_{AA} + P_{AG}}{2(P_{AA} + P_{AG} + P_{GG})} = P_{AA} + \frac{1}{2}P_{AG}$$

Always possible

Hardy-Weinberg

Predicts genotype frequencies from allele frequencies for a diploid autosomal locus

Assumptions

1. Mating occurs at random
2. Population size is very large
3. Migration is zero
4. Mutation is zero
5. Natural selection does not affect the gene in question

Why we care about Hardy-Weinberg

1. Helped prove dominance not main engine of evolution
2. Useful approximation
3. Helps estimate non-random mating (or other deviations from assumptions)

Mating table: Random union of gametes

Combination	Freq.	Offspring Frequencies		
		AA	AG	GG
A x A	p_A^2	1		
A x G	$p_A p_G$		1	
G x A	$p_G p_A$		1	
G x G	p_G^2			1
		p_A^2	$2\ p_A p_G$	p_G^2

Hardy-Weinberg frequencies: independent association of alleles

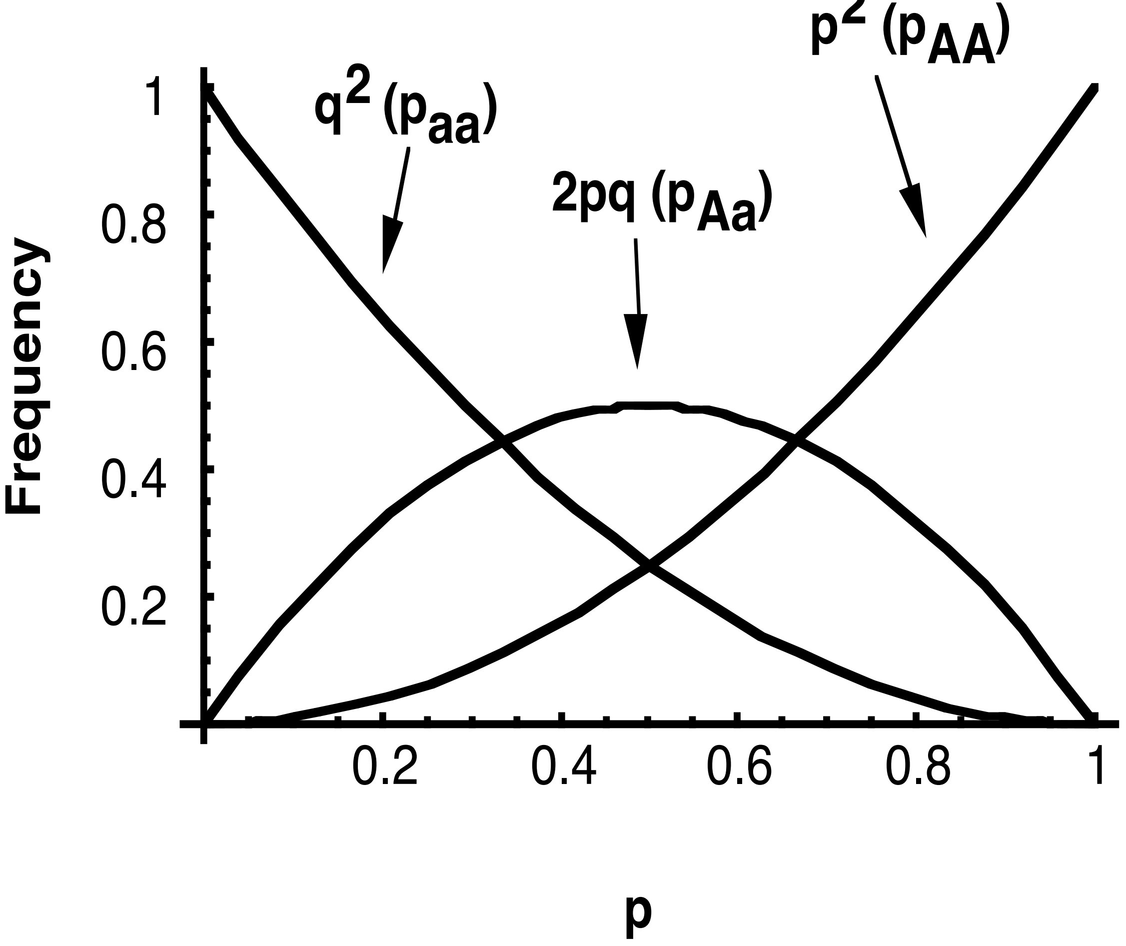
	A (frequency *p*)	*G* (q=1–p)
A	*AA* ($p_{AA}=p^2$)	*AG* (pq)
G	*AG* (pq)	*GG* (q^2)

Mating table:
Random mating of individuals

Pair	Freq.	Offspring Frequencies		
		AA	AG	GG
$AA \times AA$	p_{AA}^2	1		
$AA \times AG$	$2p_{AA}p_{AG}$	1/2	1/2	
$AA \times GG$	$2p_{AA}p_{GG}$		1	
$AG \times AG$	p_{AG}^2	1/4	1/2	1/4
$AG \times GG$	$2p_{AG}p_{GG}$		1/2	1/2
$GG \times GG$	p_{GG}^2			1
		p_A^2	$2p_Ap_G$	p_G^2

Note that HW does not require that the parents are already in HW frequency

Means that HW can be reached in a single generation


Frequency
1
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1
p
q^2 (p_{aa})
2pq (p_{Aa})
p^2 (p_{AA})

Hardy-Weinberg with more than two alleles

n alleles: $A_1, A_2, A_3, A_4, \ldots A_n$

frequencies: $p_1, p_2, p_3, p_4, \ldots p_n$

$$\sum_{i=1}^{n} p_i = 1$$

Frequency of $A_i A_i$ homozygote is p_i^2

Frequency of $A_i A_j$ heterozygote is $2p_ip_j$

Heterozygosity:

$$H = 1 - \sum p_i^2$$

Genotype frequencies at X-linked loci

For XY systems (like mammals and drosophila):

$$p = \frac{1}{3}p_{male} + \frac{2}{3}p_{female}$$

Change over generations:

$$p'_{male} = p_{female}$$

$$p'_{female} = \frac{1}{2}p_{male} + \frac{1}{2}p_{female}$$

$'$ notation indicates the next generation.

Genotype frequencies at X-linked loci

Genotype frequencies in females:

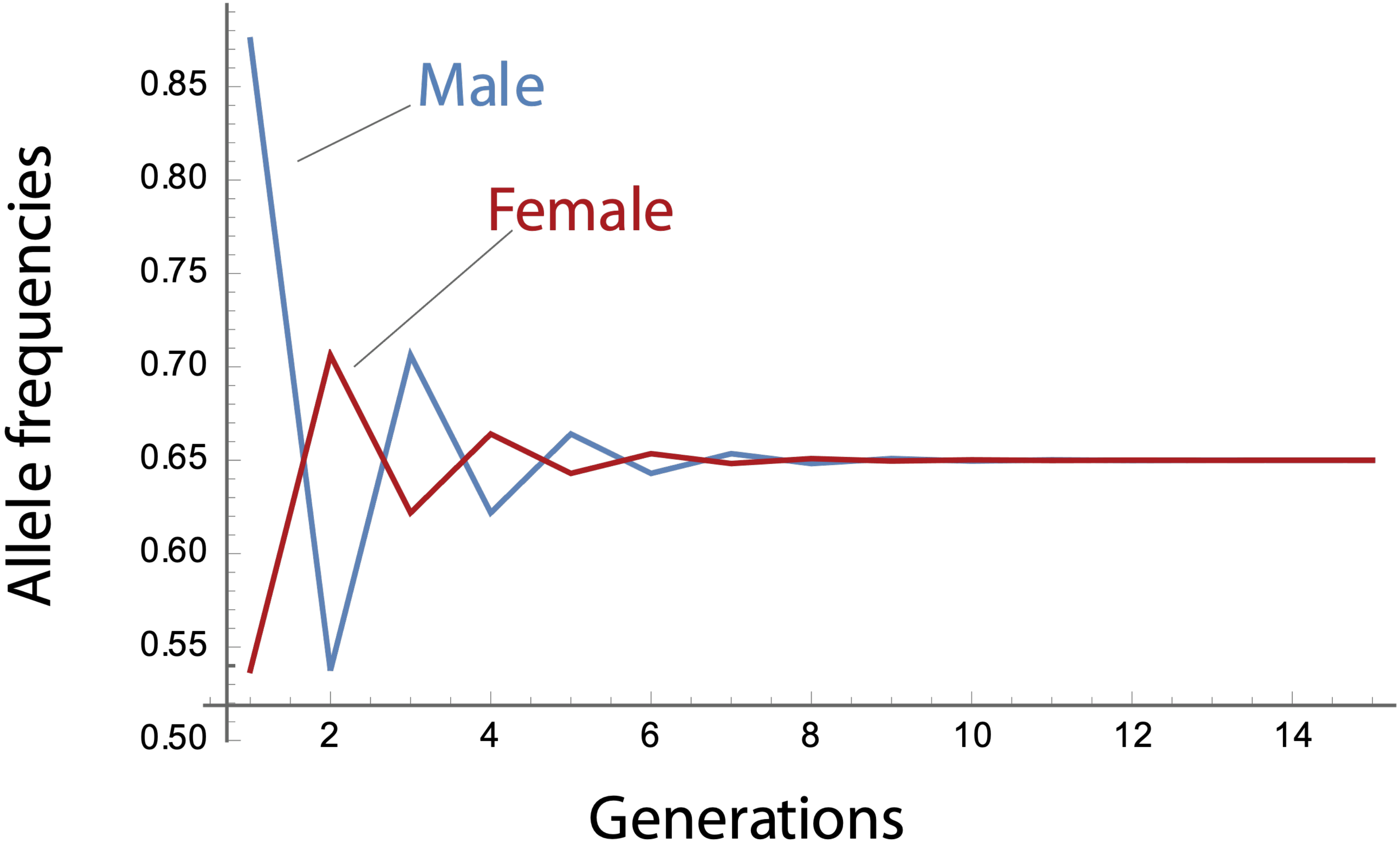
$$P'_{AA,female} = p_{A,male}\, p_{A,female}$$

$$P'_{Aa,female} = p_{A,male}\, p_{a,female} + p_{a,male}\, p_{A,female}$$

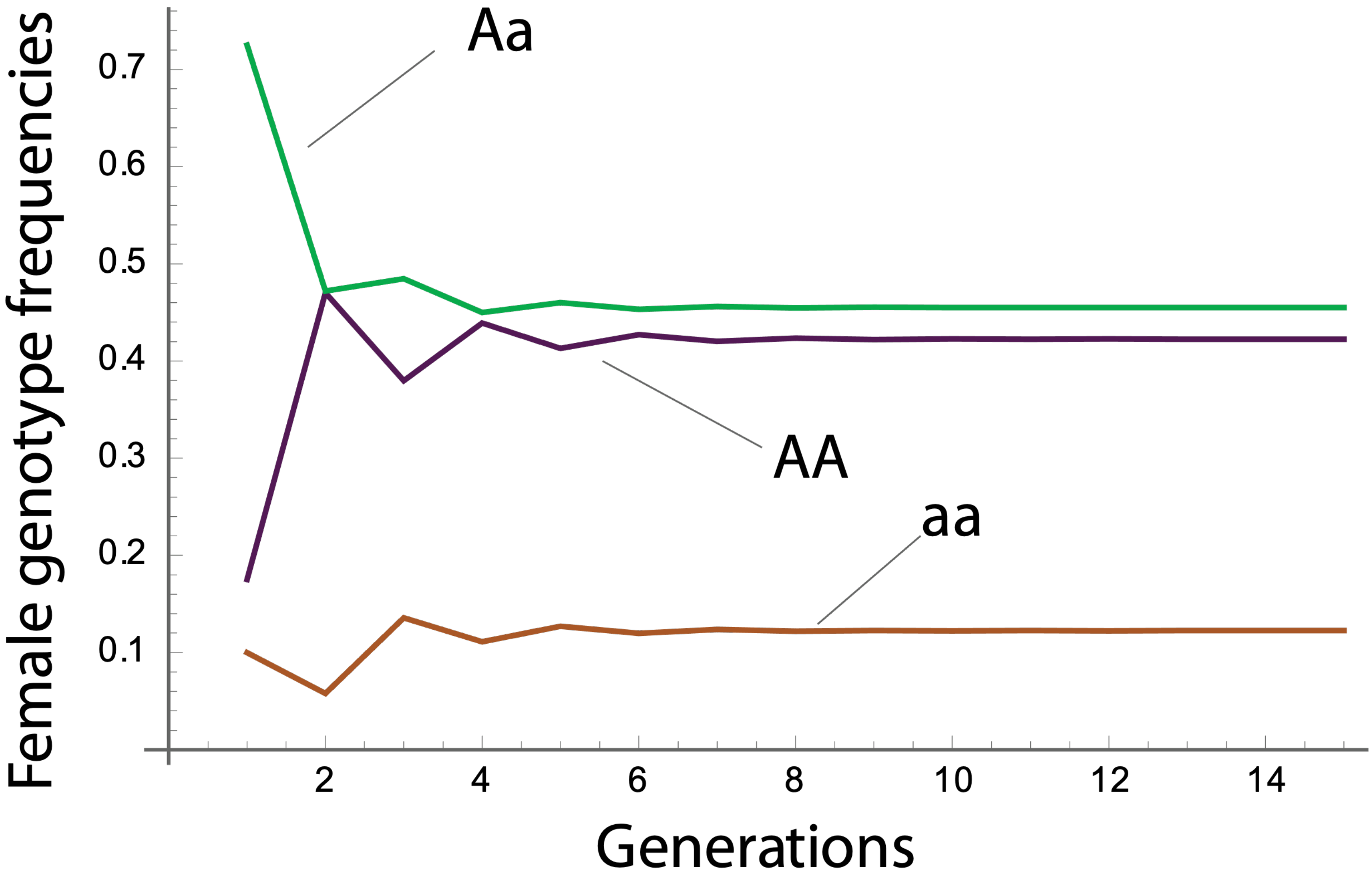
$$P'_{aa,female} = p_{a,male}\, p_{a,female}$$

Hardy-Weinberg equilibrium is not reached in a single generation with sex chromosomes, if $p_{male} \neq p_{female}$.

Allele frequency change on X chromosome



Genotype frequency change on X chromosome -Females



Genotype frequencies and linkage disequilibrium

The **gamete frequency** is the relative frequency of a particular haplotype.

e.g., one locus has alleles **A** and **a**

second locus has alleles **B** and **b**:

four possible gametes: **AB**, **Ab**, **aB**, and **ab**.

P_{Ab} is the frequency of gametes that carry **A** at the first locus and **b** at the second locus.

2-locus genotype frequencies

If the alleles at the two loci are *independent*, then the gamete frequency should be the product of its allele frequencies.

$$P_{Ab} = p_A p_b$$

If alleles are independent between loci, two-locus genotypes can be found by multiplying genotype frequencies at the two locus.

E.g.,

$$P_{Aabb} = 2p_A p_a \times p_b^2$$

Associations between loci:

Linkage disequilibrium—the statistical association of alleles at different loci

Badly named:

- Doesn’t require physical linkage
- Isn’t necessarily out of equilibrium

Associations between loci: Linkage disequilibrium

Measured with *D*:

$$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$$

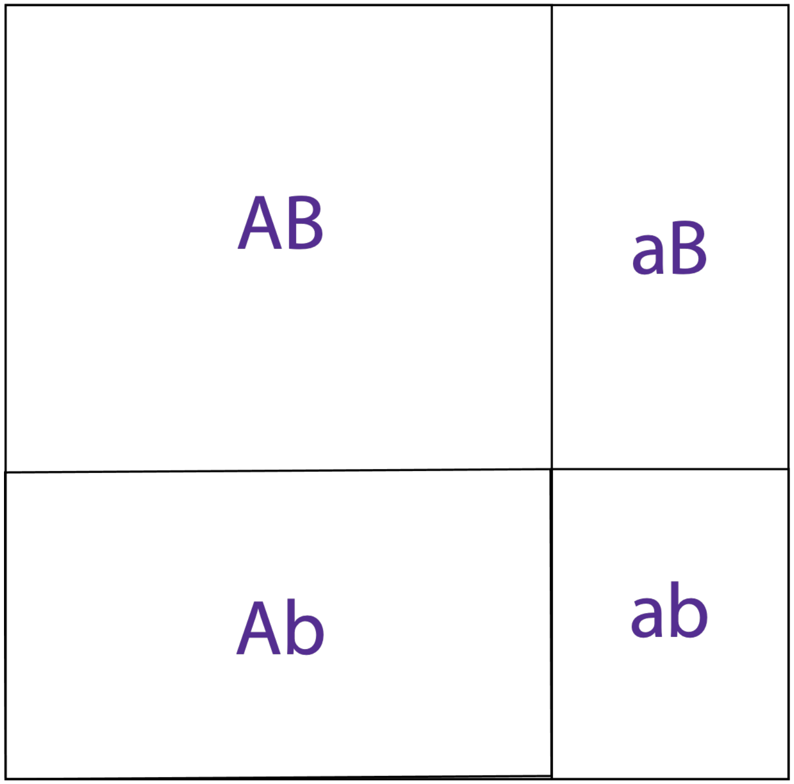
Equivalent to:

$$D = P_{AB} - p_A p_B$$

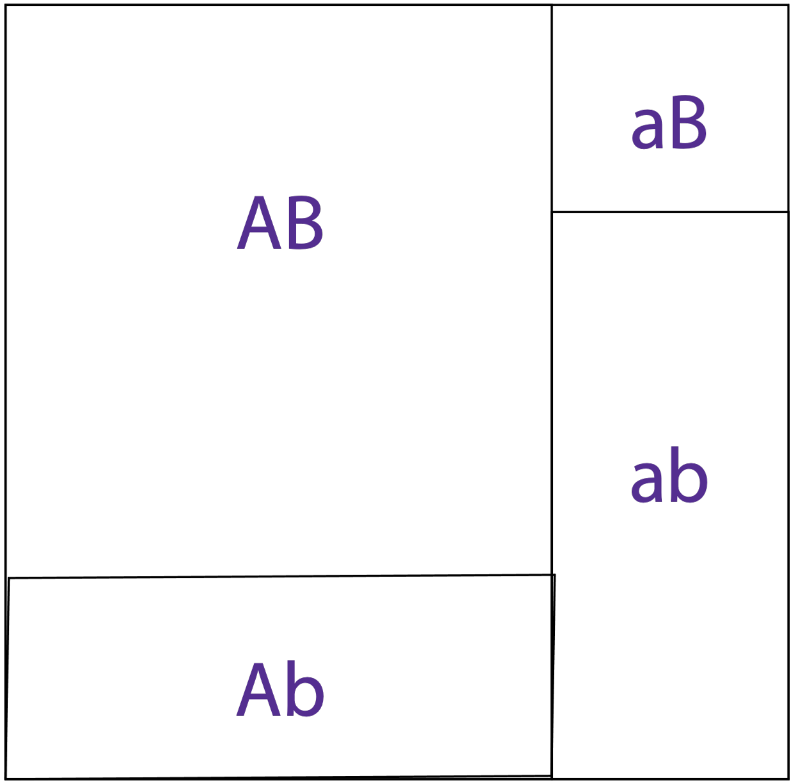
Gamete frequencies with a given *D*:

$$\begin{aligned} P_{AB} &= p_A p_B + D \\ P_{Ab} &= p_A p_b - D \\ P_{aB} &= p_a p_B - D \\ P_{ab} &= p_a p_b + D \end{aligned}$$

$D = 0$



$D = 0.1$



$$p_A = 0.7; p_B = 0.6$$

$$P_{AB} = 0.42$$

$$P_{AB} = 0.52$$

Associations between loci:
Linkage disequilibrium

D is the covariance of the alleles among individuals.

We sometimes measure LD as the correlation between alleles:

$$Correlation = \frac{D}{\sqrt{p_A p_a p_B p_b}}$$

Note: Sometimes this correlation is expressed with an r, but r is also used for the recombination rate between the loci.

Measures of variability: *H*

H is the heterozygosity – the fraction of individuals which are heterozygotes.

It can refer to a locus, a gene region, or all measured loci in a population.

H

H can either refer to the actual observed proportion of heterozygotes, or (often) the expected heterozygosity given Hardy-Weinberg expectations.

Expected heterozygosity at a locus is $H = 2pq$.

(more generally, $H = 1 - \sum p_i^2$)

Expected heterozygosity over L loci is simply the average H for each locus:

$$H = \frac{\sum H_i}{L}$$

H is often calculated using only polymorphic loci.

Measures of variability: π

π : proportion of sites that differ between two randomly chosen homologous DNA sequences.

DNA sequences

CGATT
AAATG
AAATG

Three gene copies

DNA sequences

CGATT

AAATG

AAATG

⬆⬆⬆⬆⬆

Five sites

DNA sequences

CGATT

AAATG

AAATG

Three SNPs

SNP = Single Nucleotide Polymorphism

Molecular diversity, π

(a.k.a. nucleotide diversity, virtual heterozygosity)

`CGATT`
`AAATG`
`AAATG`

π = The average fraction of sites that differ between two randomly sampled sequences

Molecular diversity, π

(A.k.a. nucleotide diversity, virtual heterozygosity)

CGATT

AAATG

AAATG

On average, the number of differences between the sequences is:

$$(3 + 3 + 0) / 3 = 2$$

There are 5 sites. So the average fraction of sites that differ between a pair of sequences is:

$$\pi = 2 / 5 = 0.4$$

Measures of variability: π

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Individual 6
G	G	G	G	G	G
T	T	T	T	T	T
A	A	A	A	A	A
T	T	T	T	T	T
C	C	T	T	T	C
C	C	C	C	C	C
C	C	C	C	C	C
C	C	C	C	C	C
G	A	G	G	G	G
T	G	T	T	T	T
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
A	A	A	A	A	A
G	G	G	G	G	G
T	T	T	T	T	T
C	C	C	C	C	C
T	T	T	T	T	T

21 nucleotides
(out of 397 measured)
Alcohol dehydrogenase (*ADH*)
Drosophila simulans

McDonald, J. H. and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652.

Measures of variability: π

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Individual 6
G	G	G	G	G	G
T	T	T	T	T	T
A	A	A	A	A	A
T	T	T	T	T	T
C	C	T	T	T	C
C	C	C	C	C	C
C	C	C	C	C	C
C	C	C	C	C	C
G	A	G	G	G	G
T	G	T	T	T	T
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
A	A	A	A	A	A
G	G	G	G	G	G
T	T	T	T	T	T
C	C	C	C	C	C
T	T	T	T	T	T

21 nucleotides
(out of 397 measured)
Alcohol dehydrogenase (*ADH*)
Drosophila simulans

Individual 1 differs from individual 2 at 2 sites

Measures of variability: π

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Individual 6
G	G	G	G	G	G
T	T	T	T	T	T
A	A	A	A	A	A
T	T	T	T	T	T
C	C	T	T	T	C
C	C	C	C	C	C
C	C	C	C	C	C
C	C	C	C	C	C
G	A	G	G	G	G
T	G	T	T	T	T
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
A	A	A	A	A	A
G	G	G	G	G	G
T	T	T	T	T	T
C	C	C	C	C	C
T	T	T	T	T	T

21 nucleotides
(out of 397 measured)
Alcohol dehydrogenase (*ADH*)
Drosophila simulans

Individual 1 differs from individual 3 at 1 site

Measures of variability: π

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Individual 6
G	G	G	G	G	G
T	T	T	T	T	T
A	A	A	A	A	A
T	T	T	T	T	T
C	C	T	T	T	C
C	C	C	C	C	C
C	C	C	C	C	C
C	C	C	C	C	C
G	A	G	G	G	G
T	G	T	T	T	T
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
T	T	T	T	T	T
C	C	C	C	C	C
C	C	C	C	C	C
A	A	A	A	A	A
G	G	G	G	G	G
T	T	T	T	T	T
C	C	C	C	C	C
T	T	T	T	T	T

21 nucleotides
(out of 397 measured)
Alcohol dehydrogenase (*ADH*)
Drosophila simulans

The 15 possible pairs differ on average by 1.267 sites

$$\pi = \frac{1.267}{397} = 0.032$$

Genetic variance

Important measure of evolutionary potential

Genetic variance in terms of the number of a specific allele carried by individuals is (assuming Hardy-Weinberg equilibrium):

$$V_G = 2pq$$

If one allele has an effect on a phenotype (relative to the other allele) of α, then the genetic variance contributed to that phenotype caused by that locus is

$$V_G = 2pq\alpha^2$$