Lecture 1
Hardy-Weinberg equilibrium and key forces affecting gene frequency

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

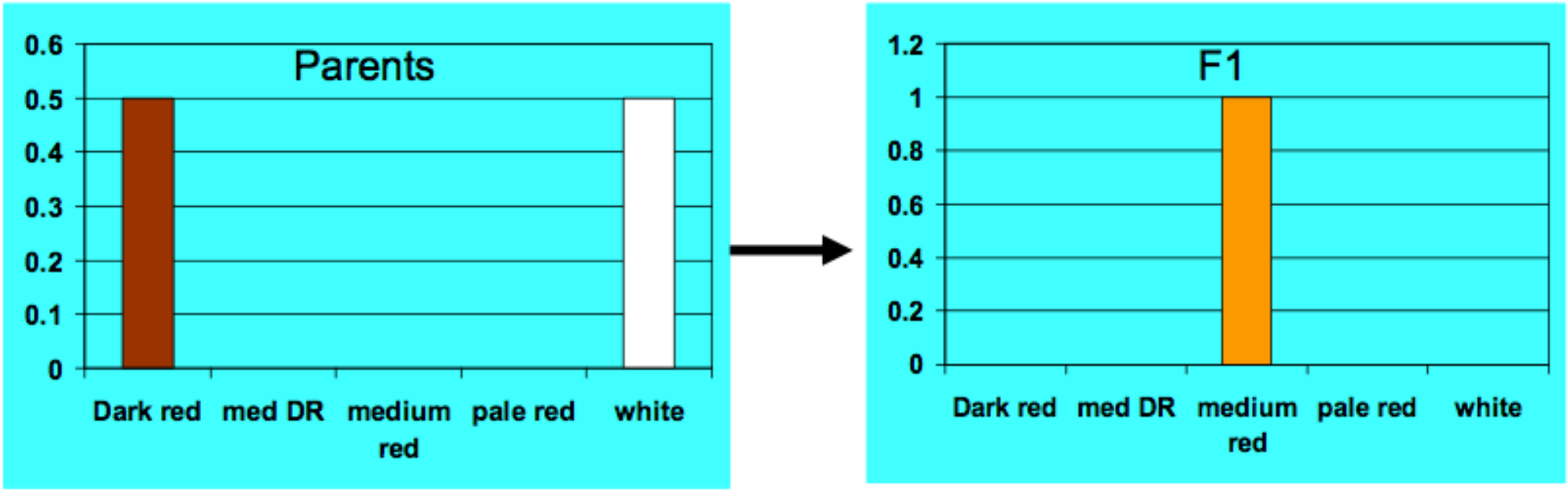
Outline

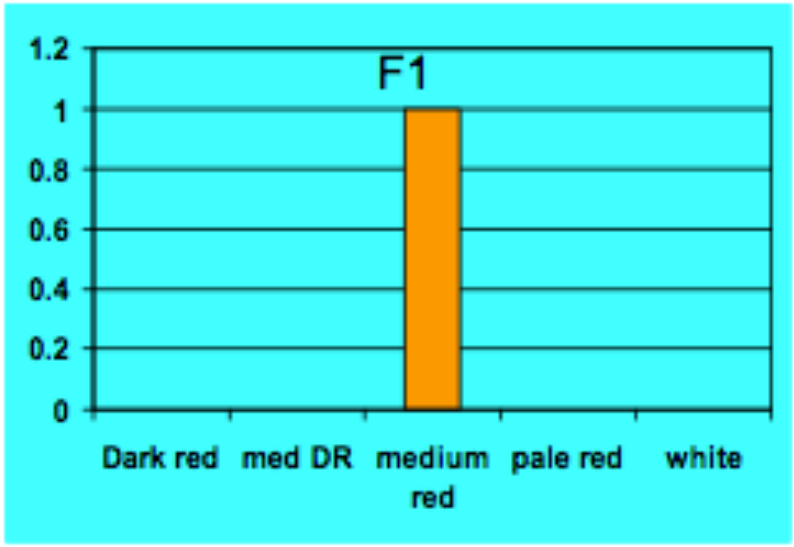
- Genetics of complex traits
- Stability of distributions over time
- Hardy-Weinberg
- Multilocus Hardy-Weinberg
- Population Structure
- Selection

Mendelian basis of complex traits

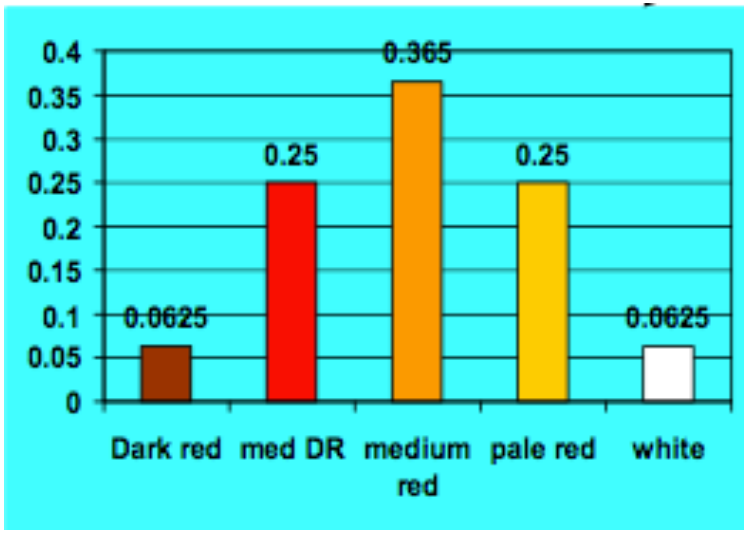
- Classic experiment of Nilsson-Ehle (1908) on wheat color
- "Simple" traits (green vs. yellow peas, etc.) had a single-gene basis
- Do complex traits have a different genetic basis?
  - Notion of blending inheritance (offspring = blended average of parents)

F_1 in a cross of dark red pure line x white pure line seems to support blending

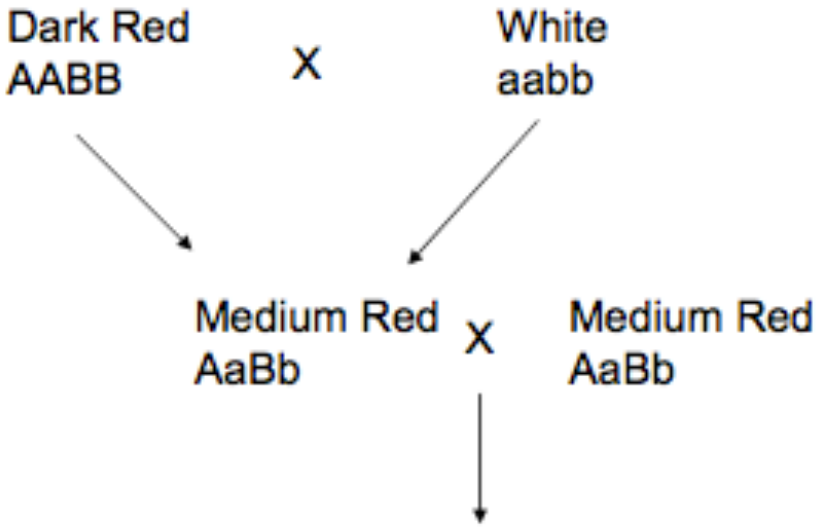




However, "outbreak of variation" in the F_2 rules out blending



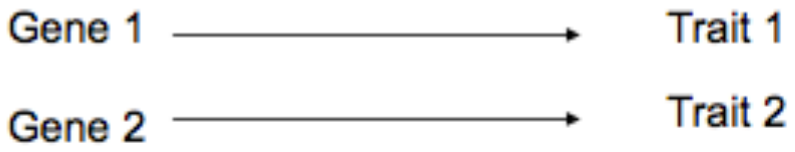
Hypothesis: 2 loci acting independently and cumulatively on one trait?



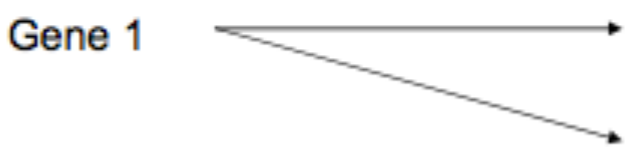
Dark Red	Medium Dark Red	Medium Red	Pale Red	White
AABB	AABb AaBB	AaBb AAbb aaBB	Aabb aaBb	aabb
1/16	4/16	6/16	4/16	1/16

Gene Effects

Usual Mendelian Concept



Simple Traits



Pleiotropy

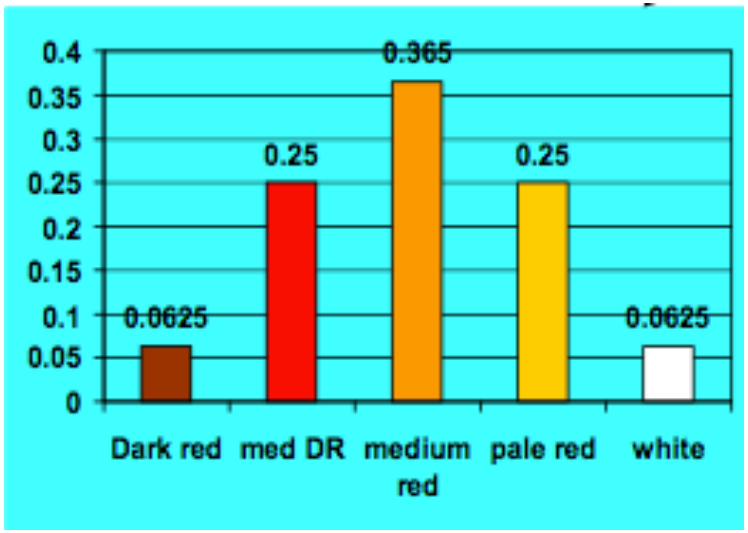
Genetic Correlation Between Traits

Polygenic Trait

Stability of the phenotypic distribution over time

Stability of the phenotype distribution

The parental lines, F1, and F2 all differ from each other. What happens to the distribution of F2 trait values in the F3, F4, Fx?



Case 1: random mating

- Suppose the F2 are randomly mated. What are the genotype frequencies in the following generation?
- These are given by the Hardy-Weinberg theorem.
- If p = freq(A) and q = freq(a), then
  - freq(AA) = p^2
  - freq(Aa) = 2pq
  - freq(aa) = q^2

- Here freq(A) = freq(a) = ½, and freq(B) = freq(b) = ½. Assuming the A and B loci are unlinked, then independent assortment gives
  - Freq(dark red) = Freq(AABB) = freq(AA)*freq(BB) = (1/4) (1/4) = 0.0625
  - Freq(white) = freq(aabb) = freq(aa)*freq(bb) = 0.0625
  - Freq(med red) = freq(AAbb or AaBb or aaBB)
    - = (1/4)*(1/4) + (1/2)*(1/2) + (1/4)*(1/4) = 0.375
- Hence, the distribution of phenotypes in the F3 is the same as the F2. What about in the F4? F5?

Case 2: Inbred lines

- Suppose instead that each F2 is used to form an inbred line, and continually selfed over many generations. What happens to the distribution after complete selfing?
- Now each locus is a homozygote, with Freq(AA) = freq(aa) = freq(BB) = freq(bb) = ½
  - AABB = dark red (25%)
  - AAbb, aaBB = medium red (50%)
  - aabb = white (25%)

During selfing

- During selfing, an AA or aa line only produces AA /aa. However, an Aa line has probablity ¼: ½ : ¼ of producing AA : Aa : aa
- Hence, after one generation of selfing
  - Freq(AA) = Freq(AA | parent AA) + Freq(AA | parent Aa) = 1*(1/4) + (1/4)*(1/2) = 3/8
  - Freq(aa) = 3/8, freq(Aa) = 1/4
  - Same for the B locus
- Resulting phenotypic (seed color) frequencies are
  - Freq(dark red) = Freq(AABB) = freq(AA)*freq(BB) = (3/8) (3/8) = 0.1406
  - Freq(white) = freq(aabb) = freq(aa)*freq(bb) = 0.1406
  - Freq(med red) = freq(AAbb or AaBb or aaBB)
    - = (3/8)*(3/8) + (2/8)*(2/8) + (3/8)*(3/8) = 0.344

Hardy-Weinberg

Importance of HW

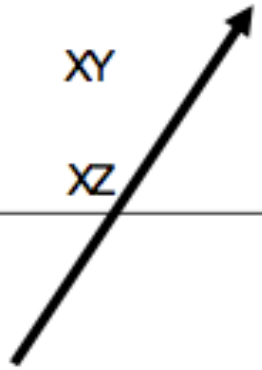
- HW states that the distribution of genotypes in a population are stable under random mating, provided no
  - Drift (i.e., pop size is large)
  - Migration (i.e., no input of individuals from other populations/breeding programs)
  - Selection (no forces to systemically change allele frequencies)

Derivation of the Hardy-Weinberg result

- Consider any population, where
  - Freq(AA) = X
  - Freq(Aa) = Y
  - Freq(aa) = Z
  - freq(A) = p = freq(AA) + (1/2) freq(Aa) = X + ½ Y
- What happens in the next generation from random mating?

Frequency of matings

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



Random Mating=independence

Genotype frequencies in next generation

		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

Freq(AA) = $1 * X^2 + ½ * 2XY + (1/4)\, Y^2 = (X + ½\, Y)^2 = p^2$.

Freq(aa) = $1 * Z^2 + ½ * 2YZ + (1/4)\, Y^2 = (Z + ½\, Y)^2 = q^2$.

What about the next generation?

Possible Matings	Frequency of Mating	Expected Frequency of Offspring 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

Freq(AA) = $1 * p^4 + \frac{1}{2} * 4p^3q + (1/4)\ 4p^2q^2 = p^2\,(p+q)^2 = p^2$.

Genotype frequencies unchanged

Hardy-Weinberg

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

After one generation of random mating, genotype frequencies remain unchanged and are given by HW proportions

Assuming random mating, no migration, drift, or selection, then allele frequencies remain unchanged

More generally, for any number of alleles, freq(A_iA_i) = p_i^2, freq(A_iA_j) = $2p_ip_j$.

Hybridization

- Hardy-Weinberg assumes allele frequencies are the same in both sexes. If not, then after one generation of random mating, the frequencies of autosomal alleles is the same in both sexes, and HW is obtained on the second generation
- Suppose Freq(A in males) = p_m, Freq(A in females) = p_f. Average allele frequency $p = (p_m + p_f)/2$.
- In generation one,
  - Freq(AA) = $p_m * p_f$ which is different from p^2 if p_m & p_f differ
  - Freq(Aa) = $p_m (1-p_f) + (1-p_m) p_f$

Example

- Cross females from a pop where $p_f = 0.4$ with males from a pop where $p_m = 0.6$. Average frequency = 0.5.
  - Under random-mating, freq(Aa) = 0.5
  - Here, Freq(Aa) = $p_m (1-p_f) + (1-p_m) p_f$ = 0.4*0.4 + 0.6*0.6 = 0.52
  - Hence, with crosses between populations where allele frequencies differ, we see an excess of heterozygotes.
  - Excess in F_1, Hardy-Weinberg values in F_2.
  - Implications for persistence of heterosis.

Crosses vs. synthetics

- In a cross, males and females are always from different populations. Example of nonrandom mating!
- In a synthetic, all individuals are randomly-mated, therefore F_2 is in HW
- Example: equal mix of P_1 X P_2
  - In a synthetic, 25% of crosses are P_1 X P_1, 50% P_1 x P_2, 25% P_2 x P_2.

Multi-locus Hardy-Weinberg

Multi-locus HW

- When following multiple loci, we need to considers gametes, rather than alleles
  - For example, an AaBb parent gives four distinct gametes AB, Ab, aB, ab
  - While allele frequencies do not change under random mating, gamete frequencies can.
  - Concept of linkage disequilibrium

Genotypic frequencies under HW

- Under multi-locus HW,
  - Freq(AABB) = Freq(AA)*Freq(BB)
  - i.e., can use single-locus HW on each locus, and then multiply the results
- When D is non-zero (LD is present), cannot use this approach
  - Rather, must follow gametes

Linkage Disequilibrium

- Under linkage equilibrium, the frequency of gametes is the product of allele frequencies,
  - e.g. Freq(AB) = Freq(A)*Freq(B)
  - A and B are independent of each other
- If the linkage phase of parents in some set or population departs from random (alleles not independent) , linkage disequilibrium (LD) is said to occur
- The amount D_{AB} of disequilibrium for the AB gamete is given by
  - D_{AB} = Freq(AB) gamete - Freq(A)*Freq(B)
  - D > 0 implies AB gamete more frequent than expected
  - D < 0 implies AB less frequent than expected

The Decay of Linkage Disequilibrium

The frequency of the AB gamete is given by

$$\text{freq}(AB) = \text{freq}(A)\,\text{freq}(B) + D_{AB}$$

LE value

Departure from LE

If recombination frequency between the A and B loci is c, the disequilibrium in generation t is

$$D(t) = D(0)(1-c)^t$$

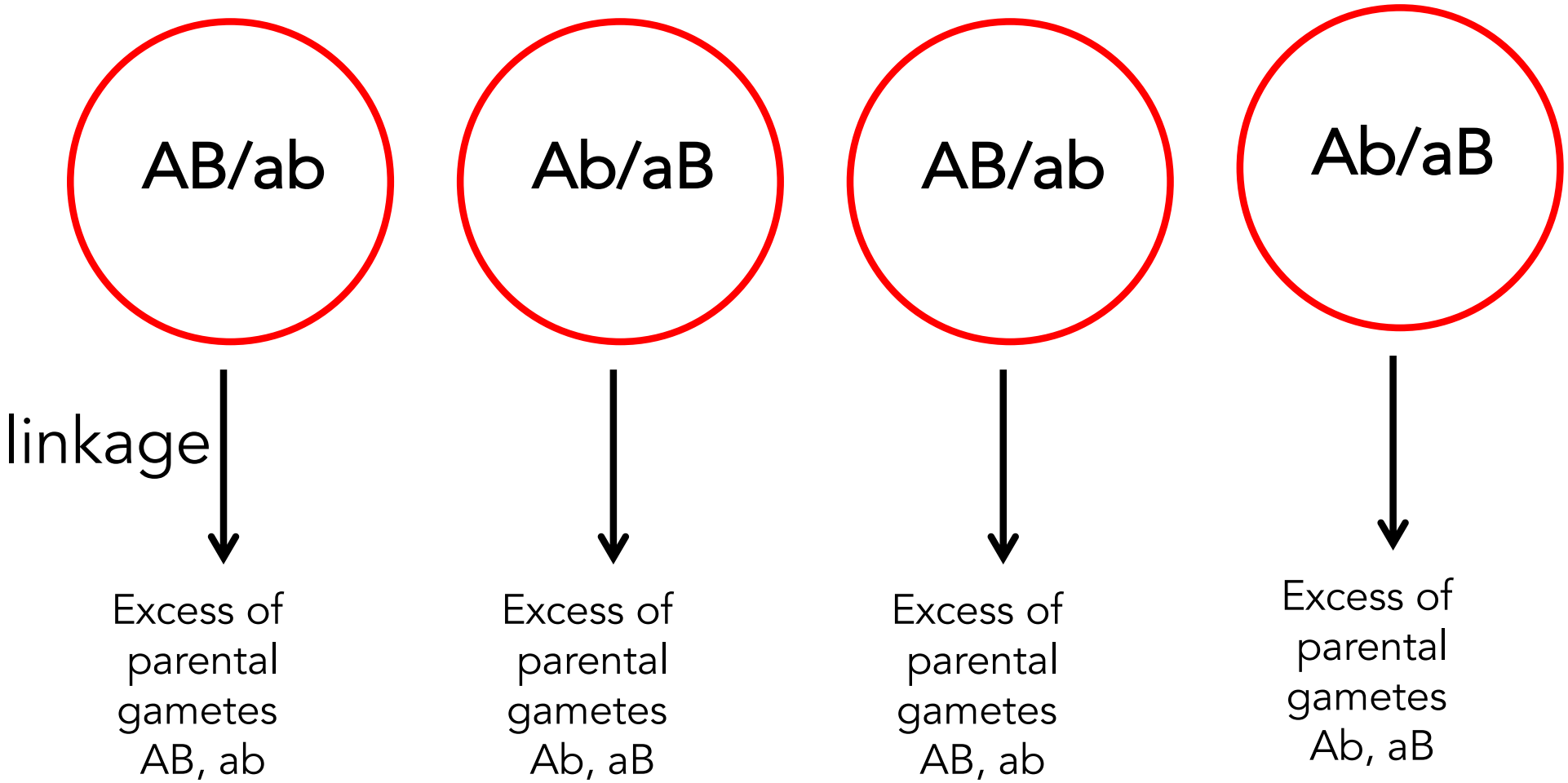
Initial LD value

Note that D(t) -> zero, although the approach can be slow when c is very small

Dynamics of D

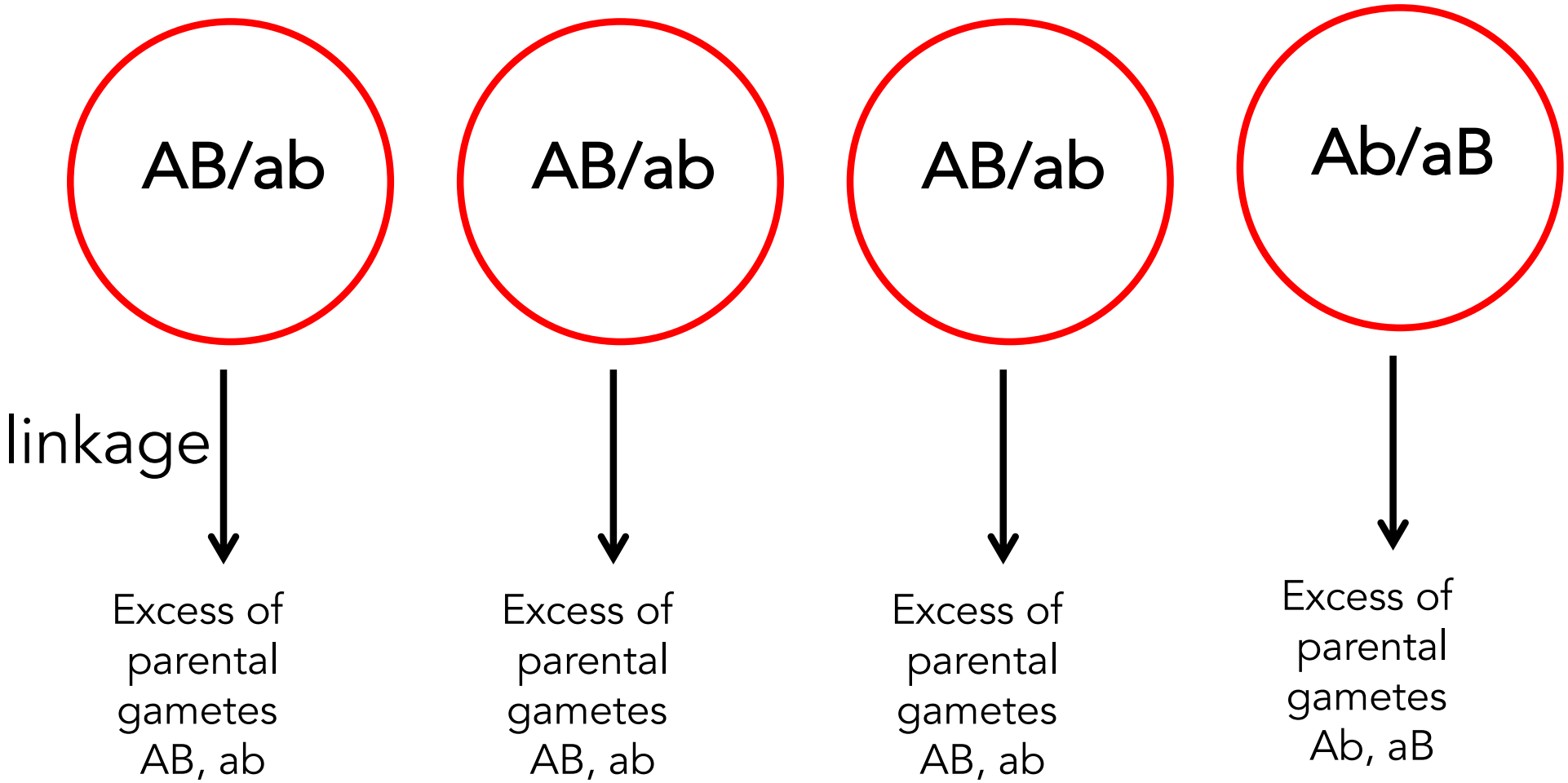
- Under random mating in a large population, allele frequencies do not change. However, gamete frequencies do if there is any LD
- The amount of LD decays by (1-c) each generation
  - $D(t) = (1-c)^t D(0)$
- The expected frequency of a gamete (say AB) is
  - Freq(AB) = Freq(A)*Freq(B) + D
  - Freq(AB in gen t) = Freq(A)*Freq(B) + $(1-c)^t D(0)$

No LD: random distribution of linkage phases



Pool all gametes: AB, ab, Ab, aB equally frequent

With LD, nonrandom distribution of linkage phase



Pool all gametes: Excess of AB, ab due to an excess of AB/ab parents

Example

- Suppose Freq(A) = 0.4, freq(B) = 0.3, D = 0.1
- Freq(AB) gamete is freq(A)*freq(B) + D
  - Freq(AB) = 0.4*0.3 + 0.1 = 0.22
- Freq(AABB) = Freq(AB)*Freq(AB) = 0.22^2 = 0.0484
- At multilocus HW,
  - Freq(AABB) = Freq(AA)*freq(BB) = $0.4^2*0.3^2$= 0.0192
- Suppose c = 0.2. In next generation,
  - D(1) = (1-0.2)*D(0) = 0.8*0.1 = 0.08,
  - Freq(AB) – 0.20; freq(AABB) = 0.04

Population structure

Population Structure

Populations often show structure, with an apparently single random-mating population instead consisting of a collection of several random-mating subpopulations

Suppose there are n subpopulations, and let w_k be the probability that an random individual is from population k

Let p_{ik} denote the frequency of allele A_i in subpopulation k.

The overall frequency of allele A_i is

$$p_i = \sum_{k=1}^{n} w_k * p_{ik}$$

The frequency of A_iA_i in the population is just

$$\text{freq}(A_iA_i) = \sum_{k=1}^{n} w_k p_{ik}^2$$

Expressed in terms of the population frequency of A_i,

$$\begin{aligned}\text{freq}(A_iA_i) &= p_i^2 - \left(p_i^2 - \sum_{k=1}^{n} w_k * p_{ik}^2 \right) \\ &= p_i^2 + \text{Var}(p_i)\end{aligned}$$

Thus, unless the allele has the same frequency in each population ($\text{Var}(p_i) = 0$), the frequency of homozygotes exceeds that predicted from HW

Similar logic gives the frequency of heterozygotes as

$$\text{freq}(A_iA_j) = 2p_ip_j + \text{Cov}(p_i, p_j)$$

Hence, when the population shows structure,
homozygotes are more common
than predicted from HW, while heterozygotes can be more (or less) common than expected under HW, as the covariance could be zero, positive, or negative

Population structure also generates disequilibrium

Again suppose there are k subpopulations, each in linkage equilibrium

The population frequency of A_iB_j gametes is

$$\mathrm{Freq}(A_iB_j) = \sum_{k=1}^{n} w_k * pA_{ik} * pB_{jk}$$

The population-wide disequilibrium becomes

$$D_{ij} = \mathrm{Freq}(A_iB_j) - \mathrm{Freq}(A_i) * \mathrm{Freq}(B_j)$$
$$= \sum_{k]=1}^{n} w_k * p_{A_{ik}} * p_{B_{jk}} - \left(\sum_{k=1}^{n} w_k * p_{Aik}\right)\left(\sum_{k=1}^{n} w_k * p_{Bik}\right)$$

Consider the simplest case of k = 2 populations

Let p_i be the frequency of A_i in population 1,
$p_i + \delta_i$ in population 2.

Likewise, let q_j be the frequency of B_j in population 1,
$q_j + \delta_j$ in population 2.

The expected disequilibrium becomes

$$D_{ij} = \delta_i * \delta_j * [\, w_1 (1 - w_1) \,]$$

Here, w_1 is the frequency of population 1

F_{ST}, a measure of population structure

- One measure of population structure is given by Wright's F_{ST} statistic (also called the fixation index)
- Essentially, this is the fraction of genetic variation due to between-population differences in allele frequencies
- Changes in allele frequencies can be caused by evolutionary forces such as genetic drift, selection, and local adaptation
- Consider a biallelic locus (A, a). If p denotes overall population frequency of allele A,
  - then the overall population variance is p(1-p)
  - $Var(p_i)$ = variance in p over subpopulations
  - $F_{ST} = Var(p_i)/[p(1-p)]$

Example of F_{ST} estimation

Population	Freq(A)
1	0.1
2	0.6
3	0.2
4	0.7

Assume all subpopulations contribute equally to the overall metapopulation

Overall freq(A) = p = (0.1 + 0.6 + 0.2 + 0.7)/4 = 0.4

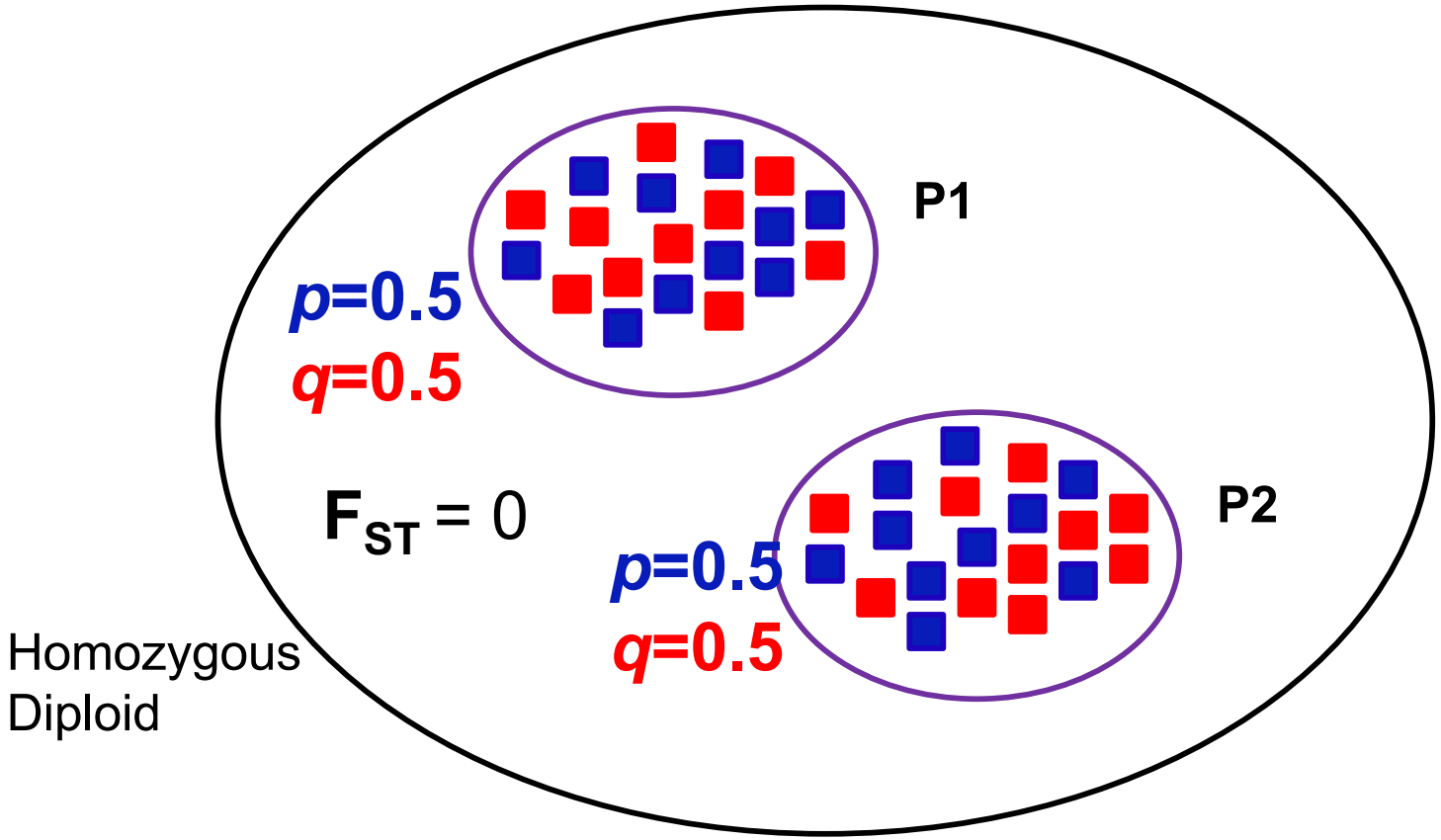
$Var(p_i) = E(p_i^2) - [E(p_i)]^2 = E(p_i^2) - p^2$

$Var(p_i) = [(0.1^2 + 0.6^2 + 0.2^2 + 0.7^2)/4] - 0.4^2 = 0.065$

Total population variance = $p(1-p) = 0.4(1-0.4) = 0.24$

Hence, $F_{ST} = Var(p_i) /[p(1-p)] = 0.065/0.24 = 0.27$

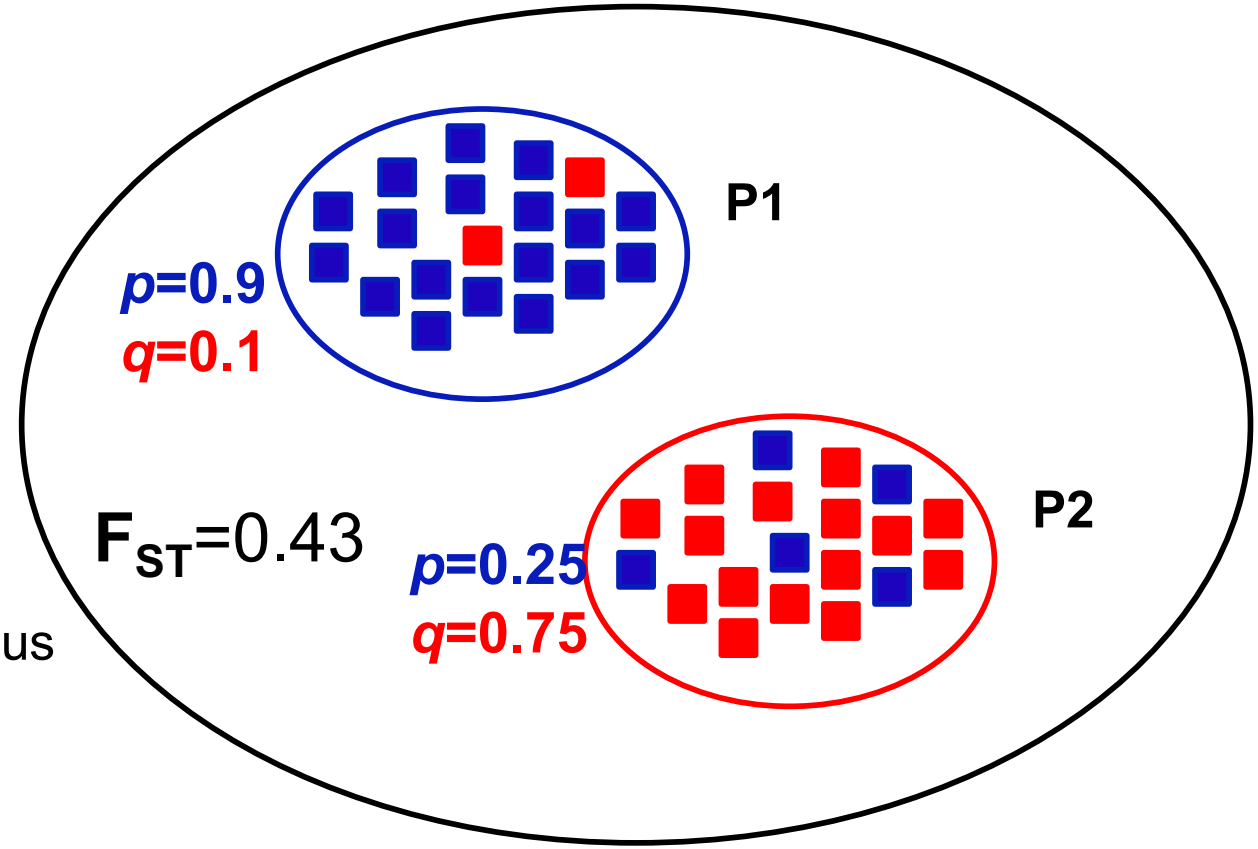
Graphical example of F_{ST}



No population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

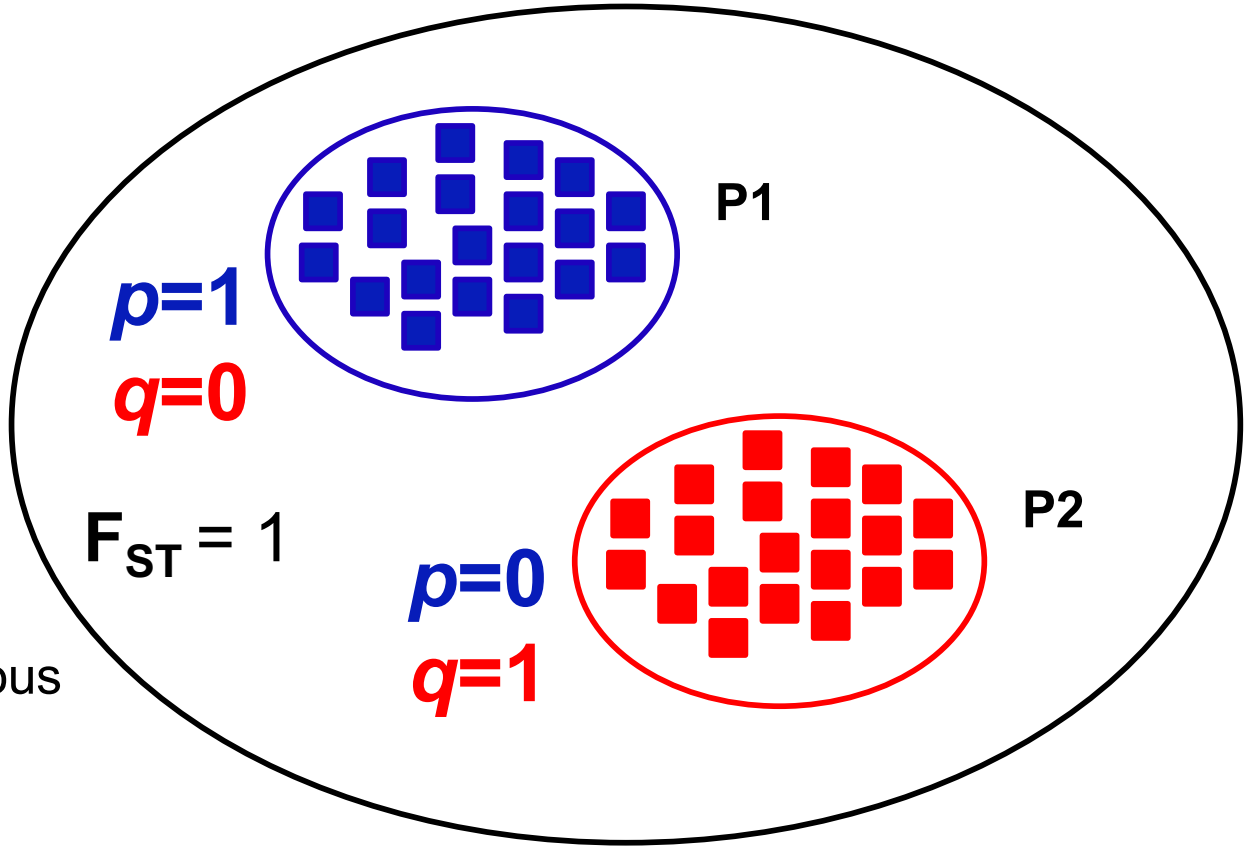
Graphical example of F_{ST}



Strong population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

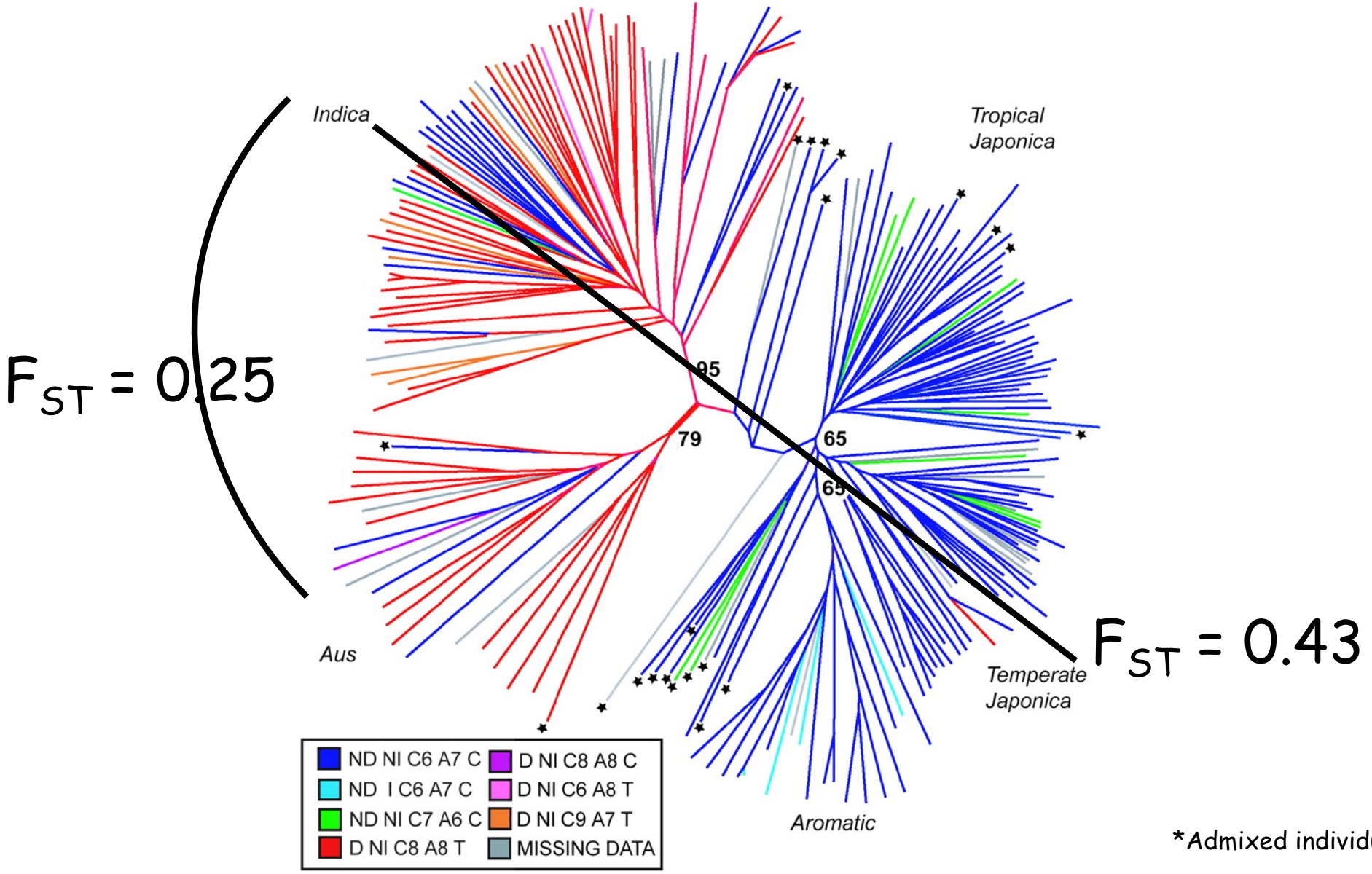
Graphical example of F_{ST}



Complete population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

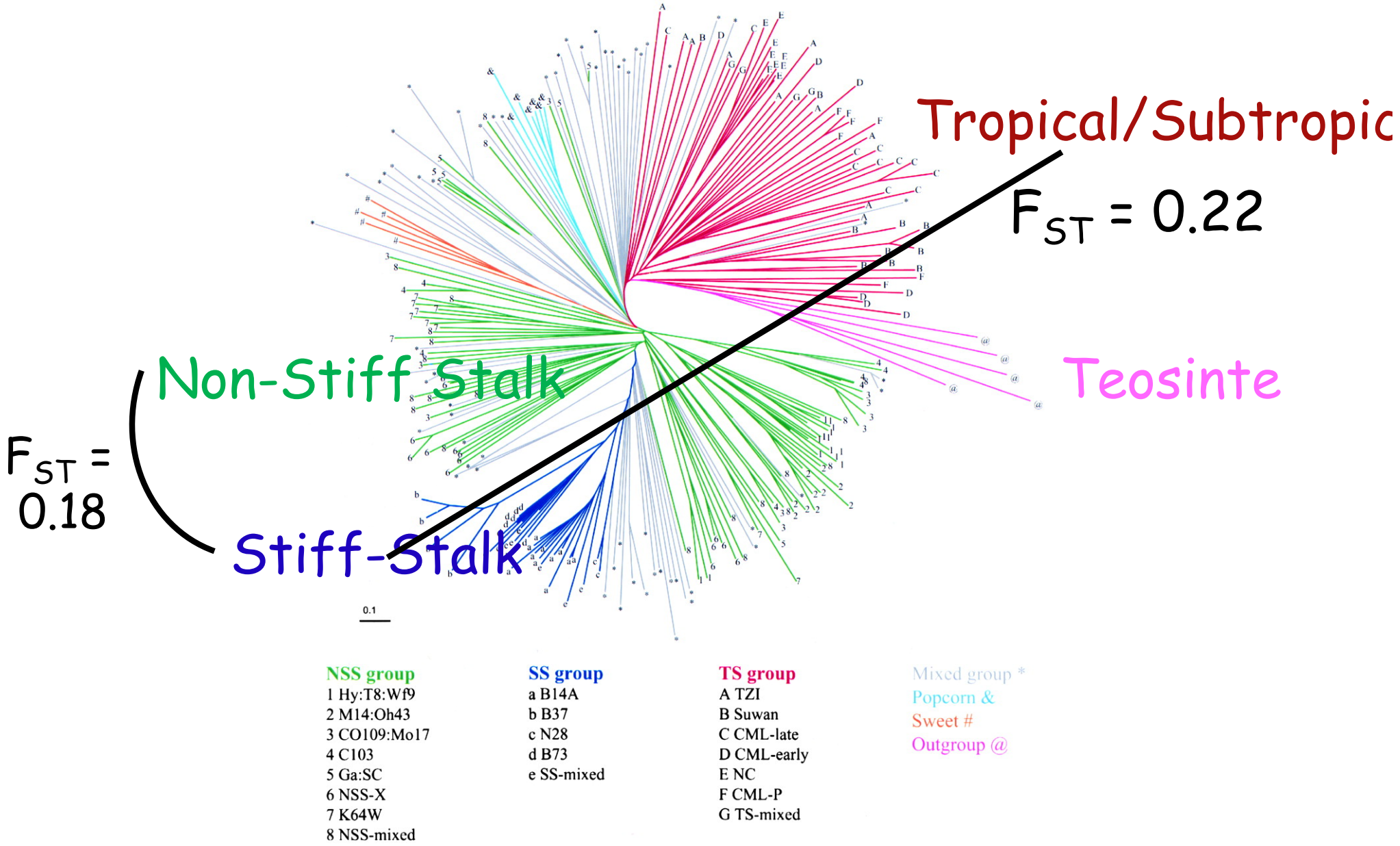
Rice population structure



Unrooted neighbor-joining tree based on C.S. Chord (Cavalli-Sforza and Edwards 1967) based on 169 nuclear SSRs. The key relates the color of the line to the chloroplast haplotype based on ORF100 and PS-ID sequences.

Garris et al. 2005. Genetics 169:1631-1638

Maize population structure



Phylogenetic tree for 260 inbred lines using the log-transformed proportion of shared alleles distance

Liu et al. 2003. Genetics 165:2117-2128

Flint-Garcia et al. 2005. Plant J. 144:1054-1064

Selection

One locus with two alleles

Genotype	AA	Aa	aa
Frequency (before selection)	p^2	$2p(1-p)$	$(1-p)^2$
Fitness	W_{AA}	W_{Aa}	W_{aa}
Frequency (after selection)	$\frac{p^2 W_{AA}}{\overline{W}}$	$\frac{2p(1-p) W_{Aa}}{\overline{W}}$	$\frac{(1-p)^2 W_{aa}}{\overline{W}}$

Where $\overline{W} = p^2 W_{AA} + 2p(1-p) W_{Aa} + (1-p)^2 W_{aa}$

is the mean population fitness, the fitness of an random individual, e.g. $\overline{W} = E[W]$

The new frequency p′ of A is just
freq(AA after selection) + (1/2) freq(Aa after selection)

$$p' = \frac{p^2 W_{AA} + p(1-p) W_{Aa}}{\overline{W}} = p \frac{p W_{AA} + (1-p) W_{Aa}}{\overline{W}}$$

The fitness rankings determine the ultimate fate of an allele

If $W_{AA} \geq W_{Aa} > W_{aa}$, allele A is fixed, a lost

If $W_{Aa} > W_{AA}, W_{aa}$, selection maintains both A & a

Overdominant selection

General expression for selection with n alleles

Let p_i = freq(A_i), W_{ij} = fitness A_iA_j

$$p_i' = p_i \frac{W_i}{\overline{W}}, \qquad W_i = \sum_{j=1}^{n} p_j W_{ij}, \qquad \overline{W} = \sum_{i=1}^{n} p_i W_i$$

W_i = marginal fitness of allele A_i

$\overline{W}$ = mean population fitness = $E[W_i] = E[W_{ij}]$

If $W_i > \overline{W}$, allele A_i increases in frequency

If a selective equilibrium exists, then $W_i = \overline{W}$ for all segregating alleles.

- Suppose fitnesses are 1: 1.2:1.4 for the genotypes qq: Qq:QQ and p =freq(Q)=0.2

	qq	qQ	QQ
Freq	$0.8^2 = 0.64$	$2*0.8*0.2 = 0.32$	$0.2^2 = 0.04$
Fitness	1	1.2	1.4
Freq*fit	0.64	0.384	0.056

Mean fitness = 0.64 + 0.384 + 0.056 = 1.08

Freq(Qq after selection) = 0.384/1.08 = 0.356

Freq(QQ after selection) = 0.04/1.08 = 0.037

New freq (Q) = (1/2)* 0.356 + 0.037 = 0.215