Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

"The kids don't listen, so I have to repeat myself. I'm always repeating myself. You know, always saying the same thing more than once. I say it once, and then they make me say it again..."

Genetic Variation

PopGen ▷ Genetic Variation

Phenotypic Variation

Morphological
Color-Pattern
Behavioural

Genotypic Variation

Sequencing (e.g. Exon-Sequencing)
Genomic Markers (e.g. SNP)

Phenotypic plasticity
Epigenetics
Cultural inheritance

Phenotypic variation for individuals in a population is due to both **environmental** and **genetic factors**, as well as **interactions** and **correlations** between these factors (i.e. gene-environment and gene-gene interactions). The combined effect of all genetic variation, including possible interactions, can be referred to as the genotypic or genetic value or genetic architecture of a trait.

$$V_g = V_p + V_e + (V_p \cdot V_e)$$

$$V_g = V_p + V_e + (V_p \cdot V_e)$$

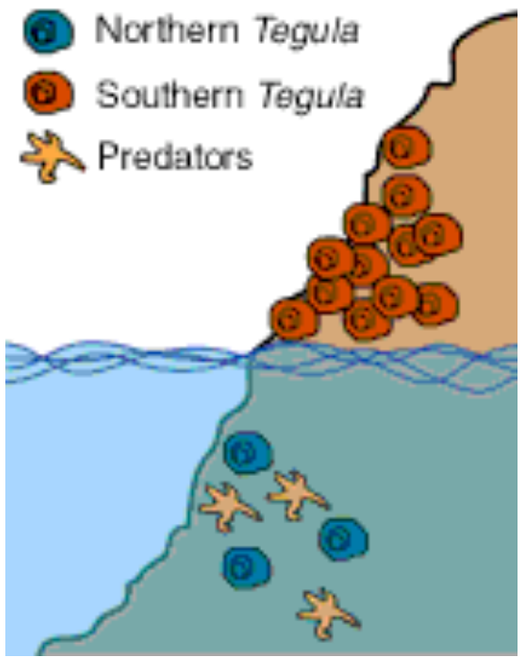
$$V_g \overset{V_e \approx 0}{=\!=} V_p$$

Common garden experiments are valuable for understanding how specific factors affect plant growth and development in a controlled environment. By minimising external variables, researchers can attribute observed changes to the manipulated factors with greater confidence.

$$V_e \approx 0$$

- Determine environment factors
- Understand (all) environment requirement
- Consider Interaction(s)

On rocky shores, animals have ranges that form clear spatial patterns. Some species live only in deep water, others only much higher up on the shore. One snail common to the California coast (*Tegula funebralis*) is found in both ranges. In southern California, Tegula live high up on the coast, while in northern California they live in deeper water.



PopGen ▷ Genetic Variation

Hypothesis:

Fawcett et al (1984) found that predators such as squid, starfish and crabs were more abundant in southern California than in northern California. Perhaps the intense predation in the south selected for snails living higher up the coast, out of reach of many predators. In the north, selection may not have been as strong, so snails were not selected to live high on the coast.

Experiment:

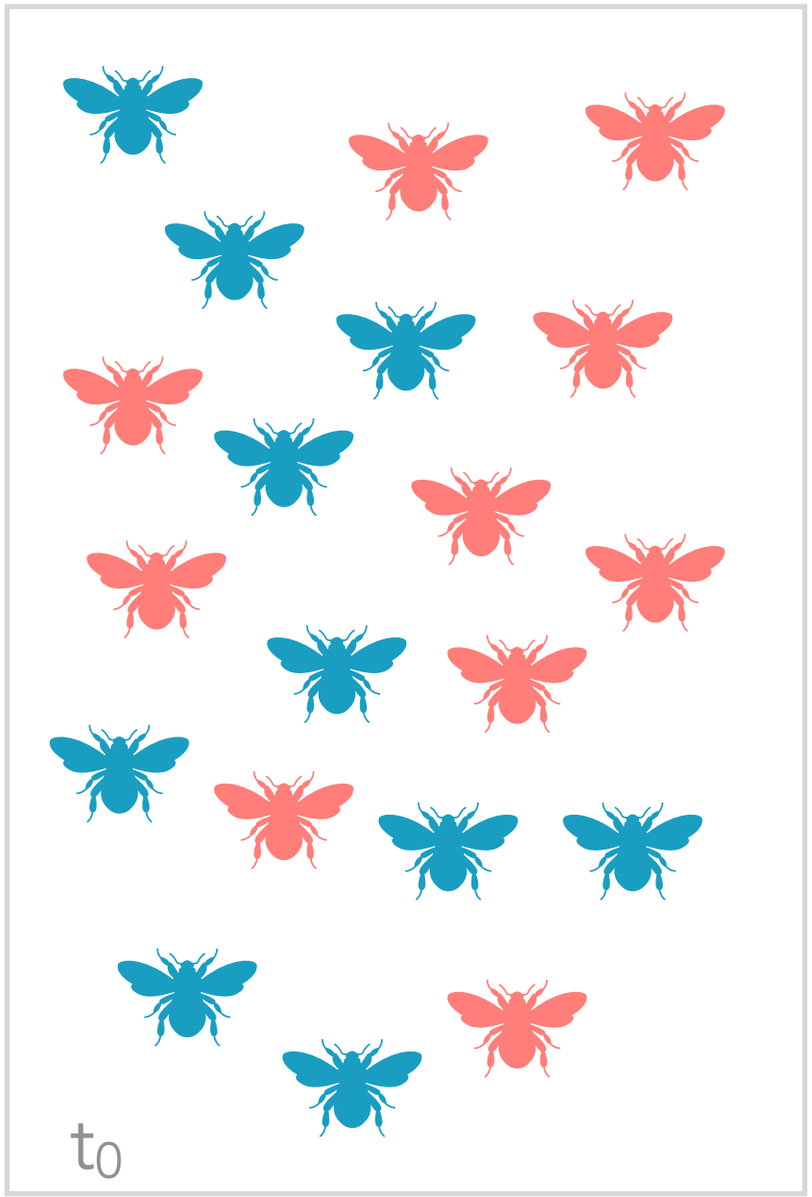
The hypothesis was tested by transplanting snails (**reciprocal transplantation or common garden experiments)**. Northern and southern snails were released into deep water and observed. When predators were present, all snails moved to higher ground (snails can probably sense the chemicals released by predators). But the southern snails moved further up the bank faster than the northern snails. Because the northern snails were slower and didn't get high enough, they were more likely to be eaten by predators.

What did this experiment show?

1. There is an innate difference between southern and northern snails (i.e. a difference that is not simply a function of being on a southern or northern coast). This difference is probably genetic (but we would need to do more experiments to be absolutely sure).
2. This difference may lead to a difference in survival. If predation is intense, snails that move higher and faster are more likely to survive.

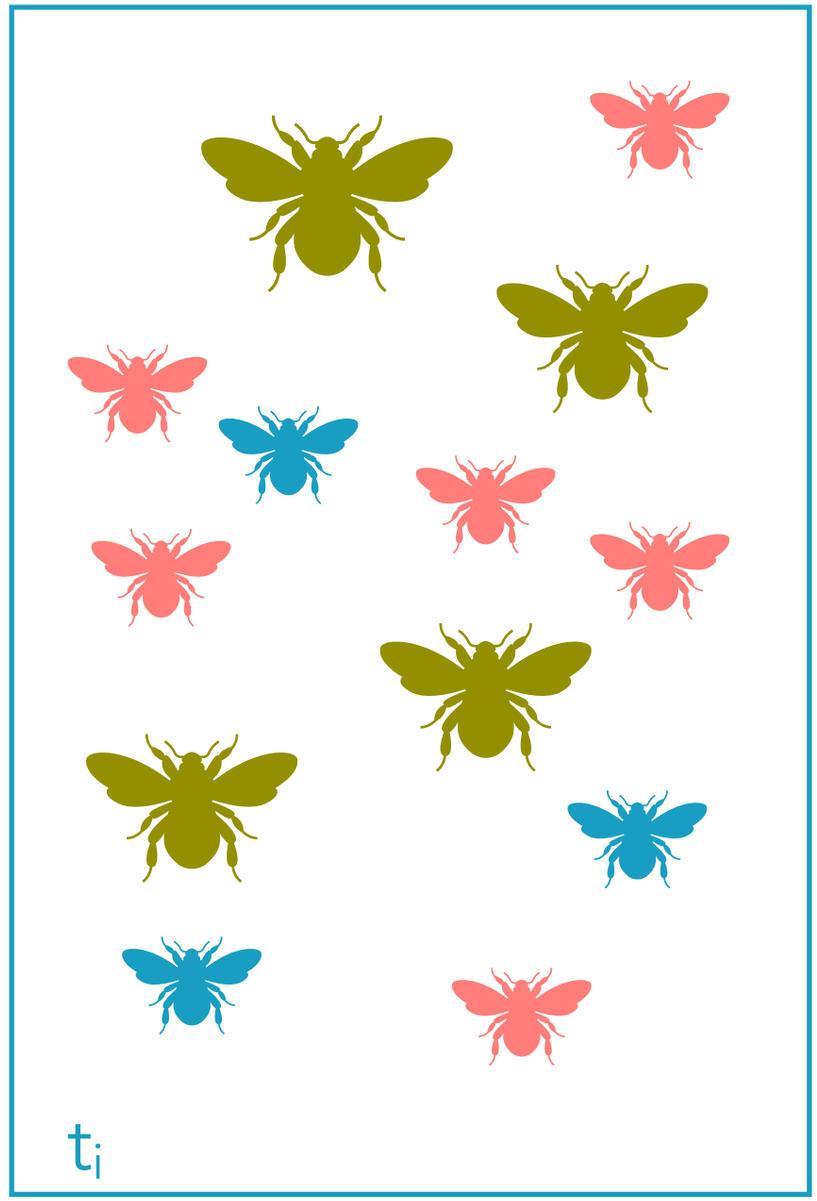
PopGen ▷ Genetic Variation

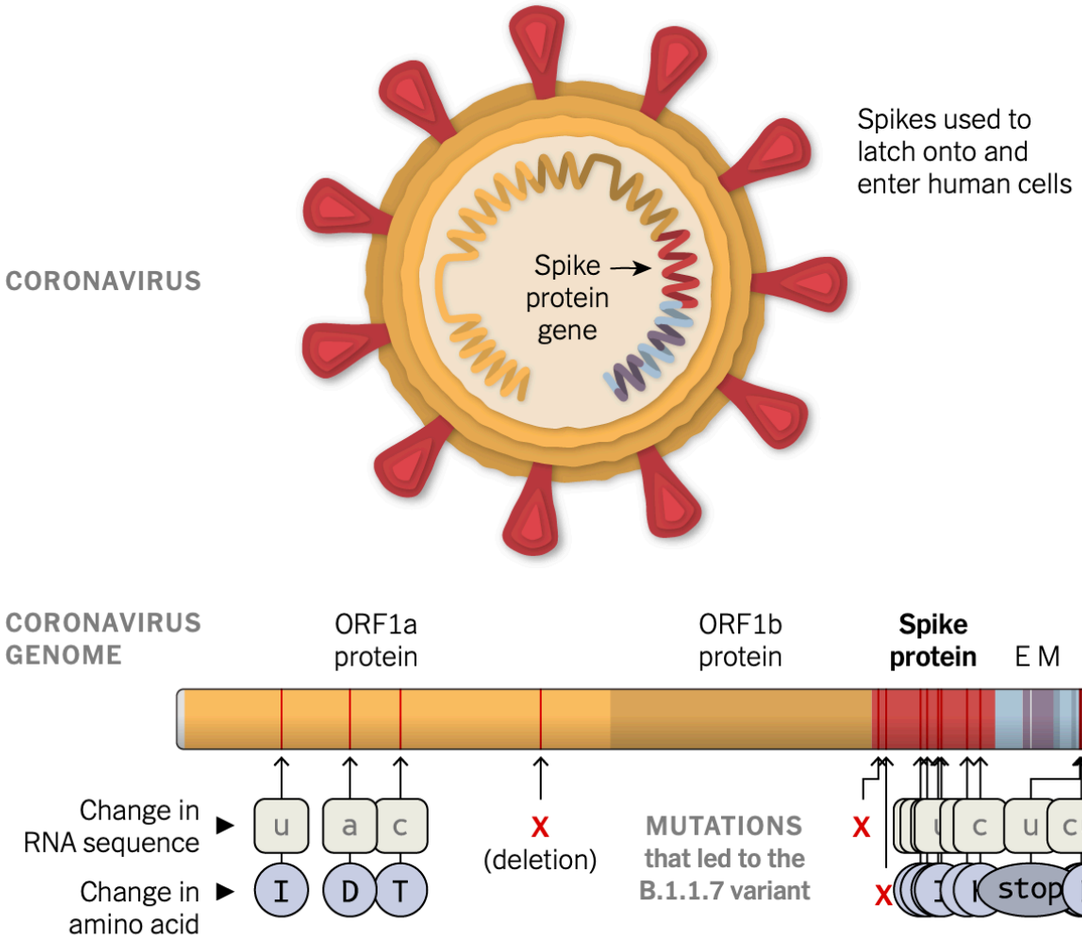
"Current" Phenotypic Variability



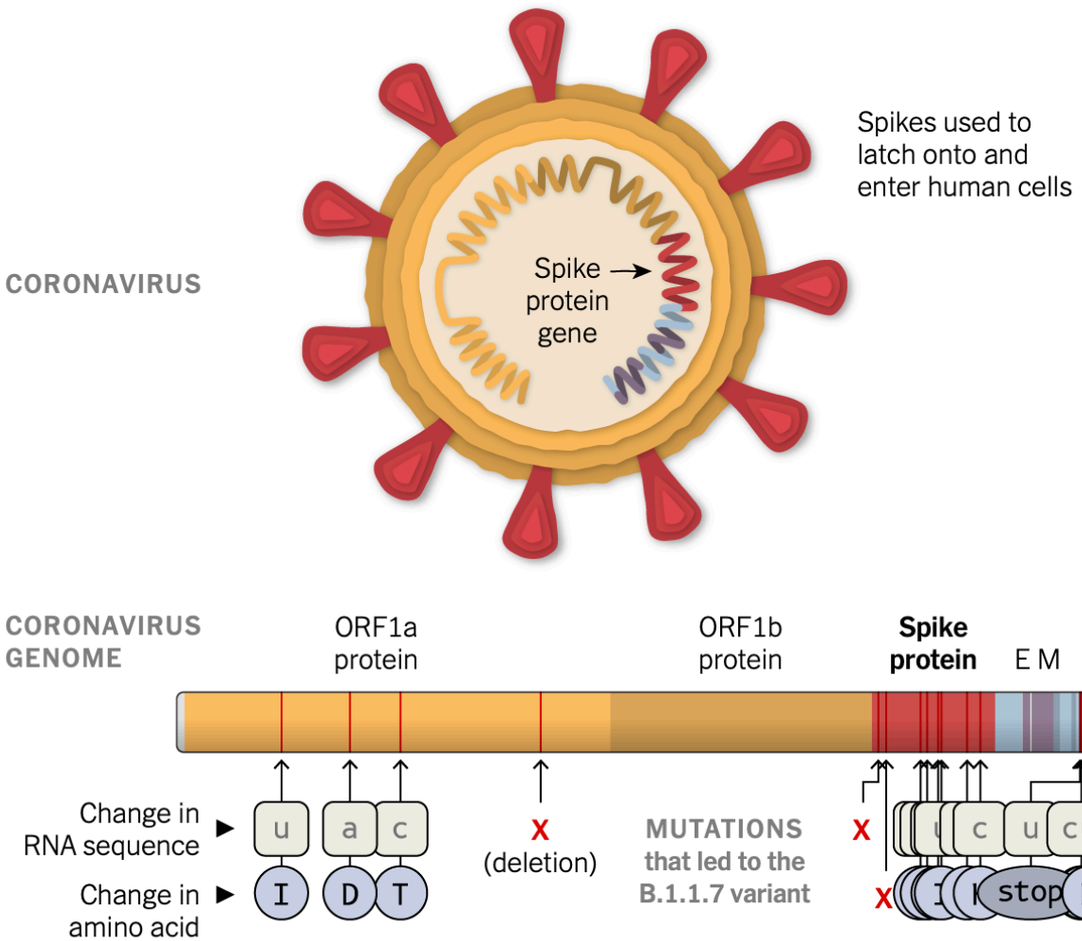
- Mutation
- Genetic Drift
- Selection
- Migration
- Non-Random Mating

"Future" Phenotypic Variability





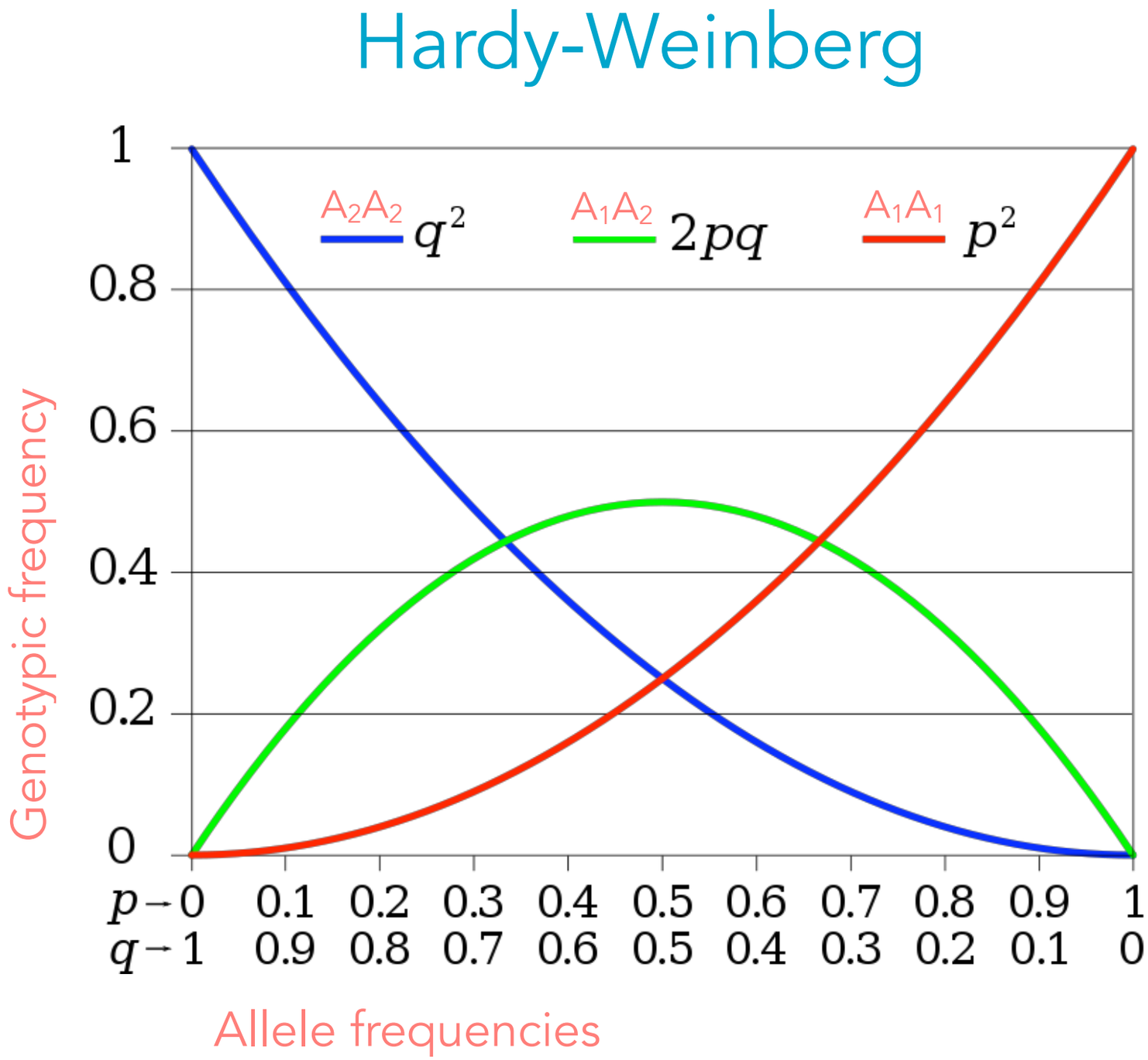
By Jonathan Corum | Source: Andrew Rambaut et al., Covid-19 Genomics Consortium U.K.



By Jonathan Corum | Source: Andrew Rambaut et al., Covid-19 Genomics Consortium U.K.

The focus is on relevant variants and therefore the figure is bias.

It is possible that mutations in other regions are selected against and therefore this conceived region have a low mutation rate.


Hardy-Weinberg
A2A2 q^2
A1A2 $2pq$
A1A1 p^2
Genotypic frequency
1
0.8
0.6
0.4
0.2
0
p – 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
q – 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
Allele frequencies

The Hardy-Weinberg law for a locus with **2 alleles.**

$$(p+q)^2 = p^2 + 2pq + q^2 = 1$$

The **generalisation** of the Hardy-Weinberg law to **multiple alleles** requires no new ideas.

$$(p+q+r)^2 = p^2 + 2pq + q^2 + 2pr + 2qr + r^2$$

$p = frequency(A_1)$ $q = frequency(A_2)$ $r = frequency(A_3)$

The χ^2 test for 2-alleles and multiple alleles.

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs - \exp)^2}{\exp}$$

2 alleles

$$\chi^2 = \frac{\left(N_{11} - p^2 N\right)^2}{p^2 N} + \frac{\left(N_{12} - 2pqN\right)^2}{2pqN} + \frac{\left(N_{22} - q^2 N\right)^2}{q^2 N}$$

n alleles

$$\chi^2 = \sum_{i} \frac{\left(N_{ii} - {p_i}^2 N\right)^2}{{p_i}^2 N} + \sum_{i<j} \frac{\left(N_{ii} - 2 p_i p_j N\right)^2}{2 p_i p_j N}$$

Is the population at this locus at HW if $f(A_1) = p = 0.5$?

Is the population at this locus at HW if $f(A_1) = p = 0.5$?

We don’t know because we do not have enough information. We need information about the observed genotypes.

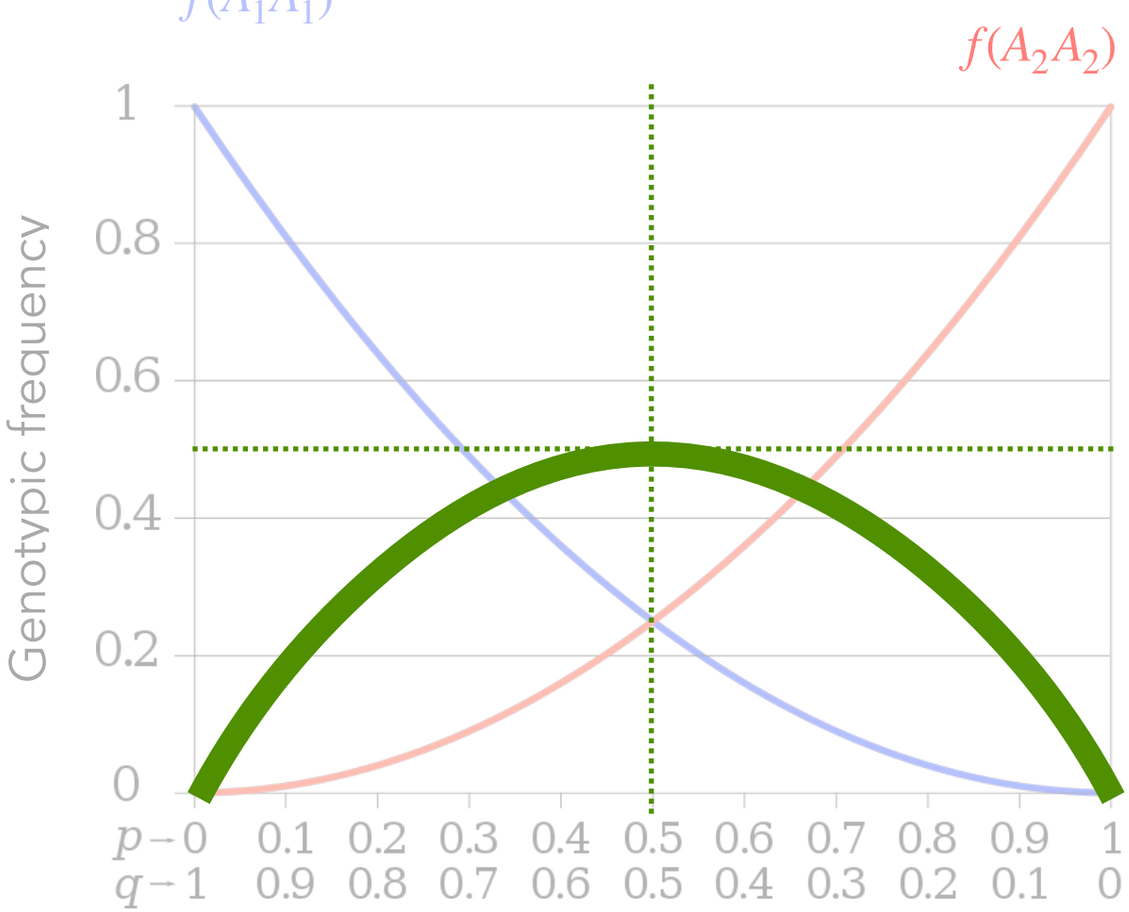
Is the population at this locus at HW if $f(A_1) = p = 0.5$?

All the genotypes in the population are heterozygote.

Is the population at this locus at HW if $f(A_1) = p = 0.5$?

All the genotypes in the population are heterozygote.

Observed genotypes: $f(A_1A_2) \rightarrow 100\,\%$

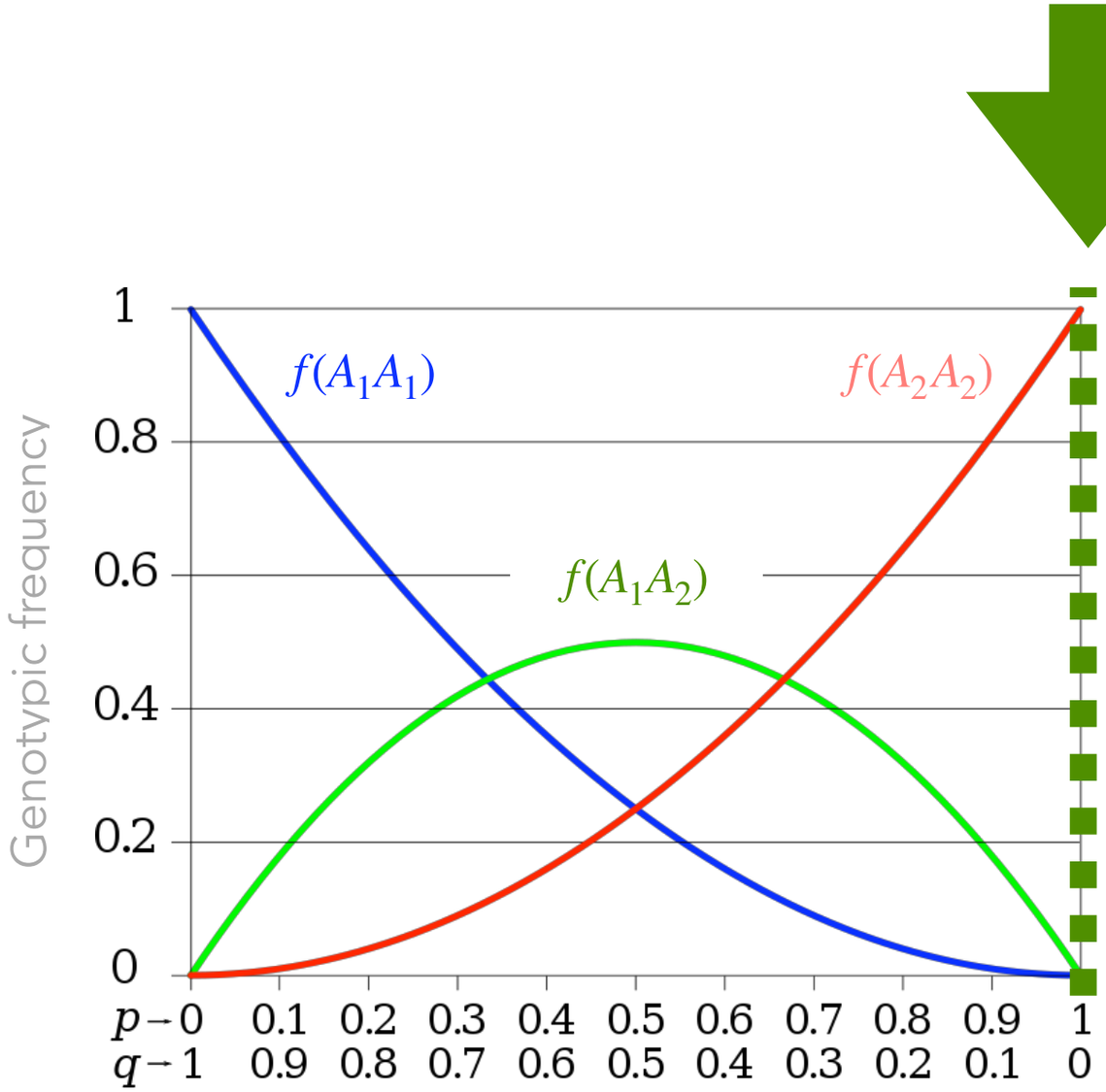


$f(A_1A_2)_{max} \Rightarrow 50\,\%$

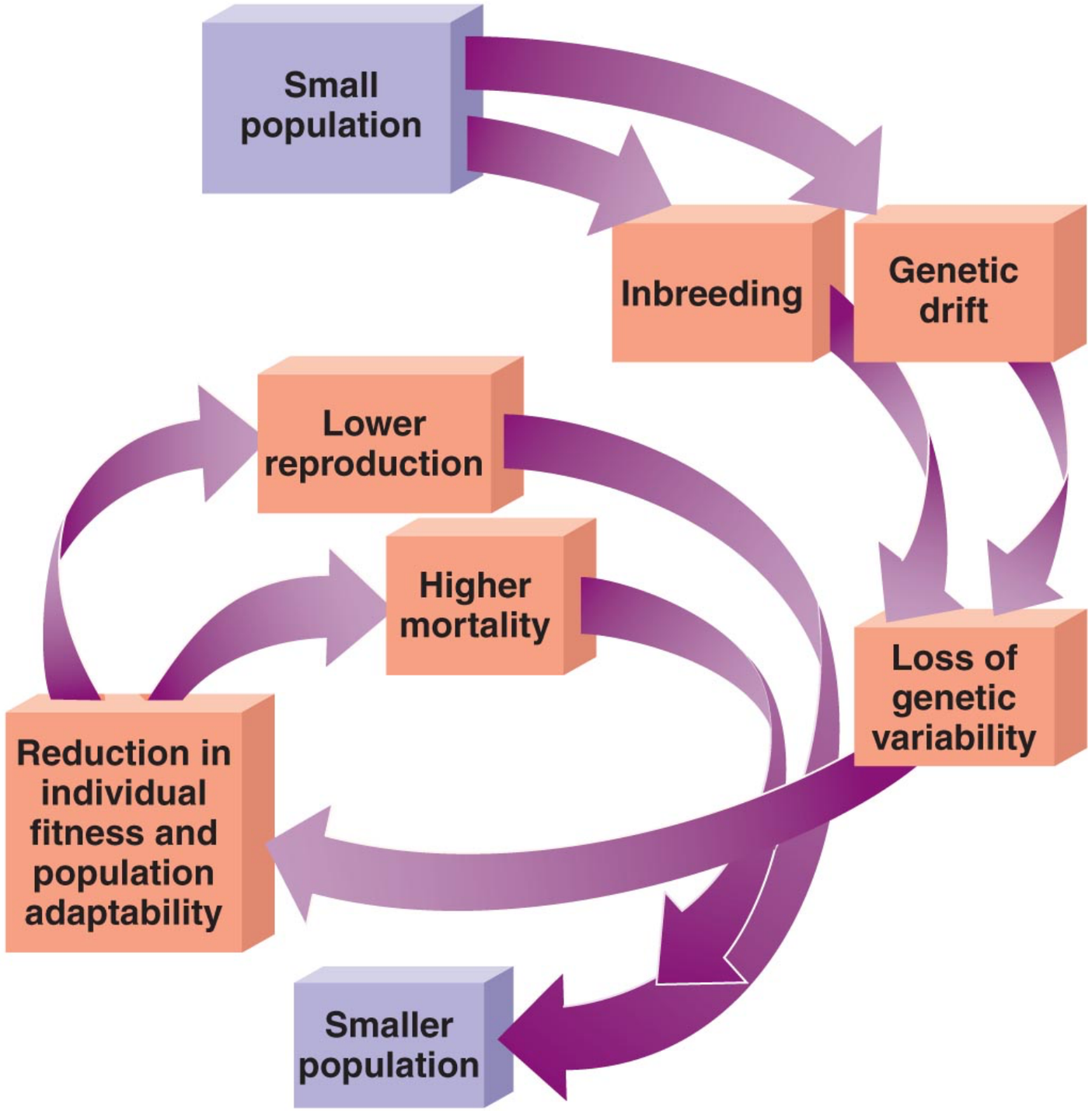
$f(A_1A_2)$

Is the population at this locus at HW if $f(A_1) = p = 1.0$?

Is the population at this locus at HW if $f(A_1) = p = 0$?



Genetic Drift

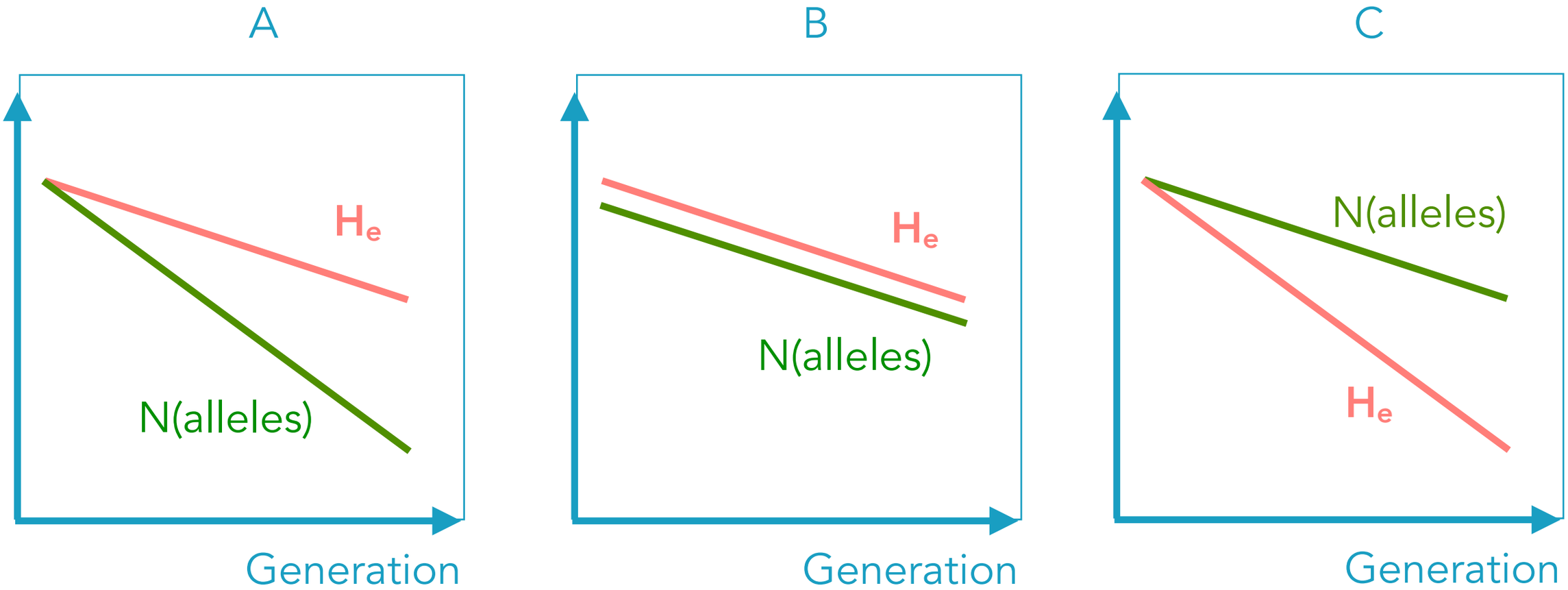


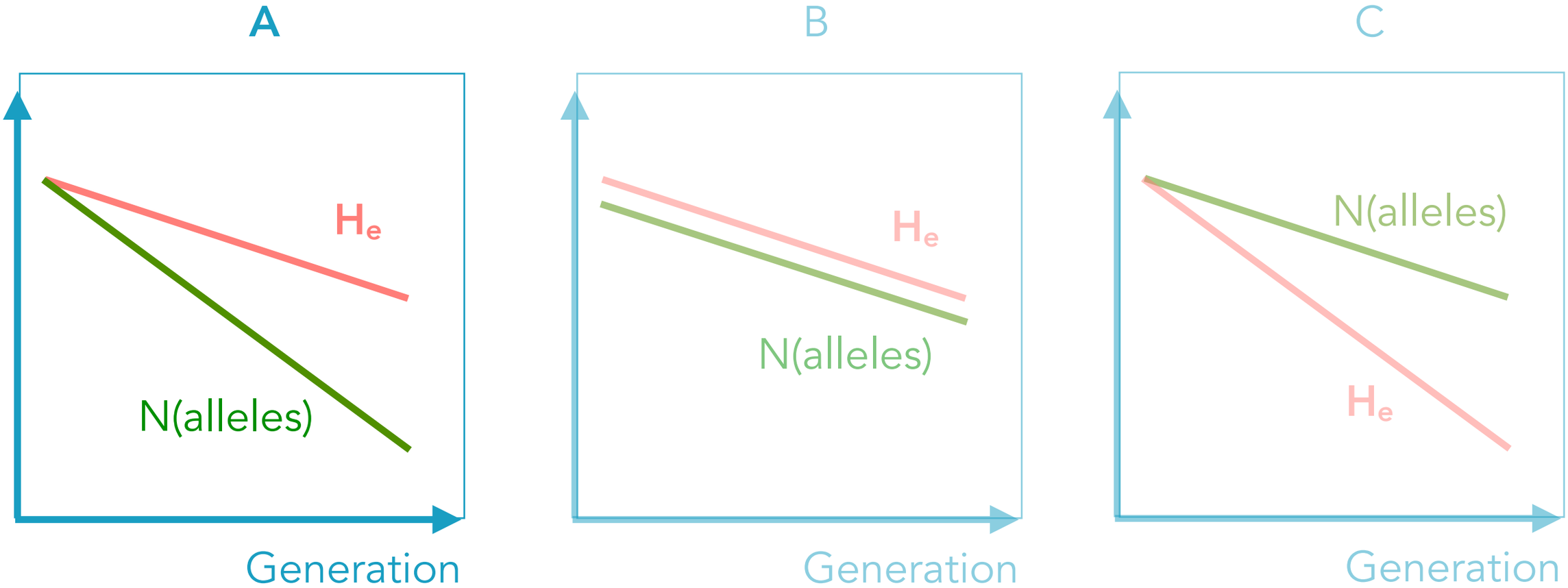
Extinction vortex is caused by a positive feedback loop (Gilpin and Soule, 1986).

Gilpin ME, Soulé ME (1986). "Minimum Viable Populations: Processes of Species Extinction". In M. E. Soulé. Conservation Biology: The Science of Scarcity and Diversity. Sinauer, Sunderland, Mass. pp. 19–34.

Genetic drift has several important effects on evolution:

1. **Drift reduces genetic variation in populations**, potentially reducing the ability of a population to evolve in response to new selective pressures.
2. Genetic drift acts faster and has **more drastic results in smaller populations**. This effect is particularly important in rare and endangered species.
3. Genetic drift tends to make different populations different from each other. It can contribute to speciation.


A
H_e
N(alleles)
Generation
B
H_e
N(alleles)
Generation
C
N(alleles)
H_e
Generation


A
H_e
N(alleles)
Generation
B
H_e
N(alleles)
Generation
C
N(alleles)
H_e
Generation

	Heterozygosity	Genetic Distance
Genetic Drift	?	?
Imbreeding	?	?
Mutation	?	?
Migration	?	?

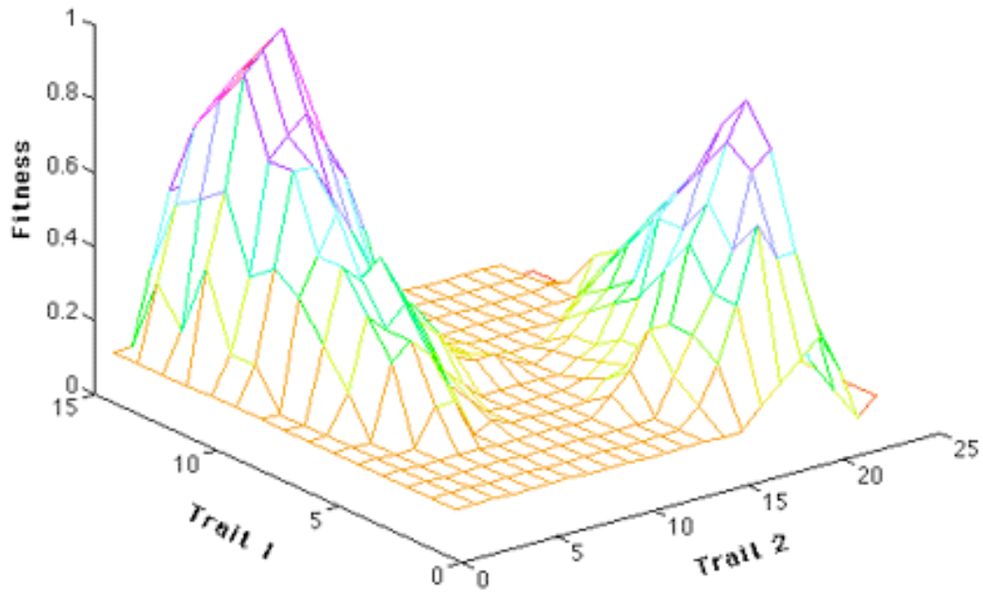
	Heterozygosity	Genetic Distance
Genetic Drift	↓	↑
Imbreeding	↓	↑
Mutation	↑	↑
Migration	↑ ↓	↑ ↓ no

"**The common conception of evolution focuses on change due to natural selection.** Natural selection is certainly an important mechanism of allele-frequency change, and it is the only mechanism that generates adaptation of organisms to their environments. **Other mechanisms, however, can also change allele frequencies, often in ways that oppose the influence of selection.** A nuanced understanding of evolution demands that we consider such mechanisms as genetic drift and gene flow, and that we recognize the error in assuming that selection will always drive populations toward the most well adapted state."

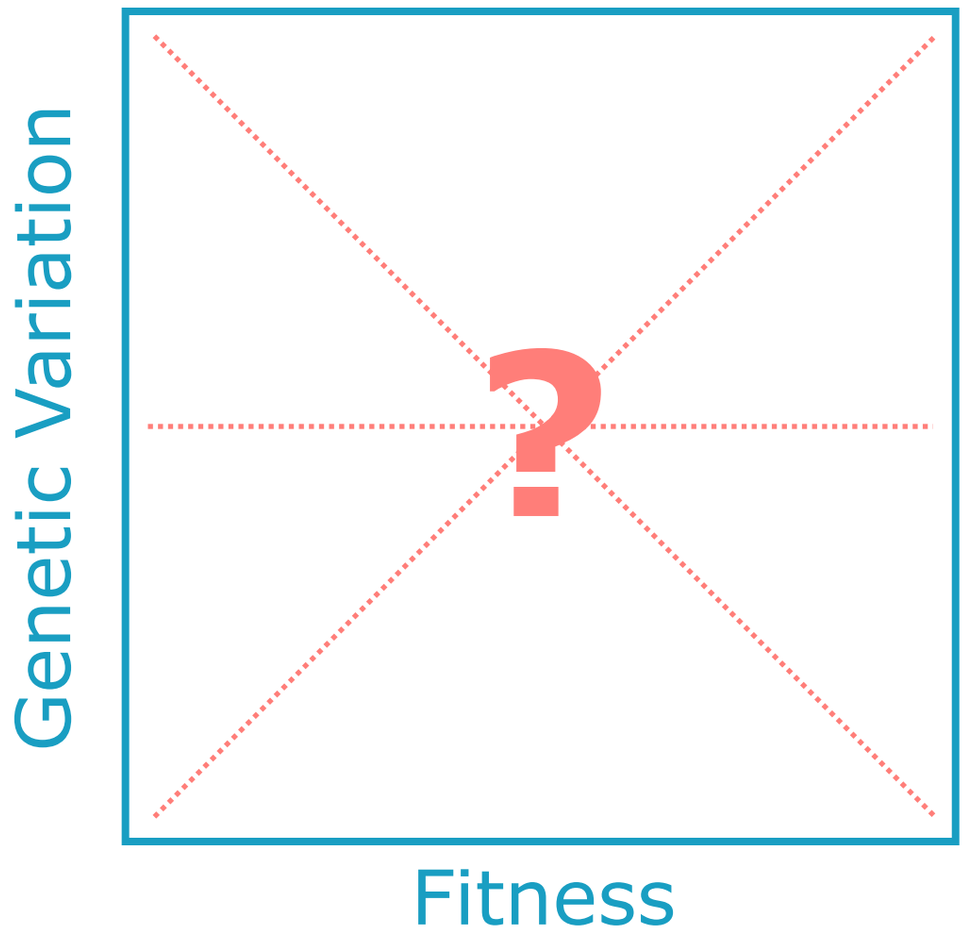
Source: Andrews, C. A. (2010) Natural Selection, Genetic Drift, and Gene Flow Do Not Act in Isolation in Natural Populations. Nature Education Knowledge 3(10):5

In a population, there may be just one coadapted gene complex, or there might be several different combinations of traits, each of which could have high fitness. This latter possibility gives rise to another concept: that of an **adaptive landscape**. An adaptive landscape is the description of the fitnesses of all possible combinations of different traits in a population. Adaptive landscapes are frequently represented graphically; fitness is plotted on a vertical axis and trait values for different genes are plotted on other axes. Combinations of traits that have high fitness thus appears as peaks, and combinations that have low fitness appear as valleys. Here is an example of an adaptive landscape:

This example shows two traits, and a situation in which there are two combinations of traits, the peaks in the graph, shown in purple, that have high fitness, while other combinations of the traits have low fitness. These peaks in the adaptive landscape can be called adaptive peaks; note that they are also combinations of different genetic traits that, together, have high fitness, so they are coadapted gene complexes. An adaptive peak and a co-adaptive gene complex are thus basically the same thing.



What is the connection between genetic variation and fitness?



N_e

"The number of breeding individuals in an idealised population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration."

Ronald Fisher and Sewall Wright

$$N_c \overset{?}{=} N_e$$

$N_c = N_e$

all individuals contribute to the allele pool

equal sex-ratios ($N_f = N_m$)

number of progeny (k=2, poisson distribution)

variance in offsprings (V_k=2, poisson distribution)

no fluctuation in population size

chromosome (organelle) linkage

The inbreeding coefficient (F) of an individual is the probability that an individual has two alleles at a locus that are identical by descent. It measures the amount of inbreeding by comparing the observed frequency of heterozygotes (H_o) in the population to the frequency expected under random mating - Hardy-Weinberg (H_e).

$$F = 1 - \frac{H_o}{H_e}$$

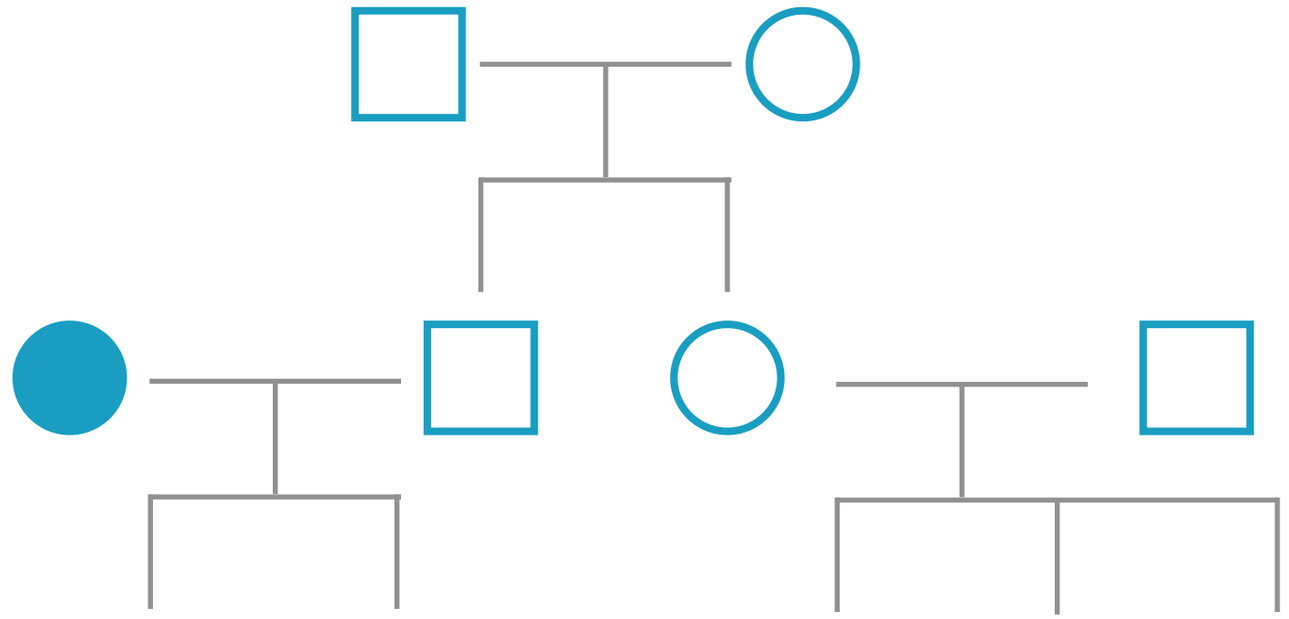
In a panmixic population the observed (H_o) and the expected frequency of heterozygosity is not significantly different.

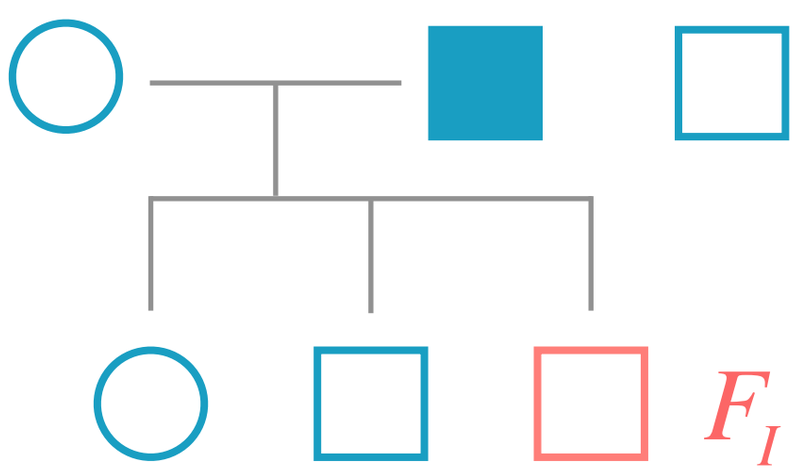
$$H_o \approx H_e \rightarrow F = 0$$

What factors could influence random-mating?

What factors could influence random-mating?

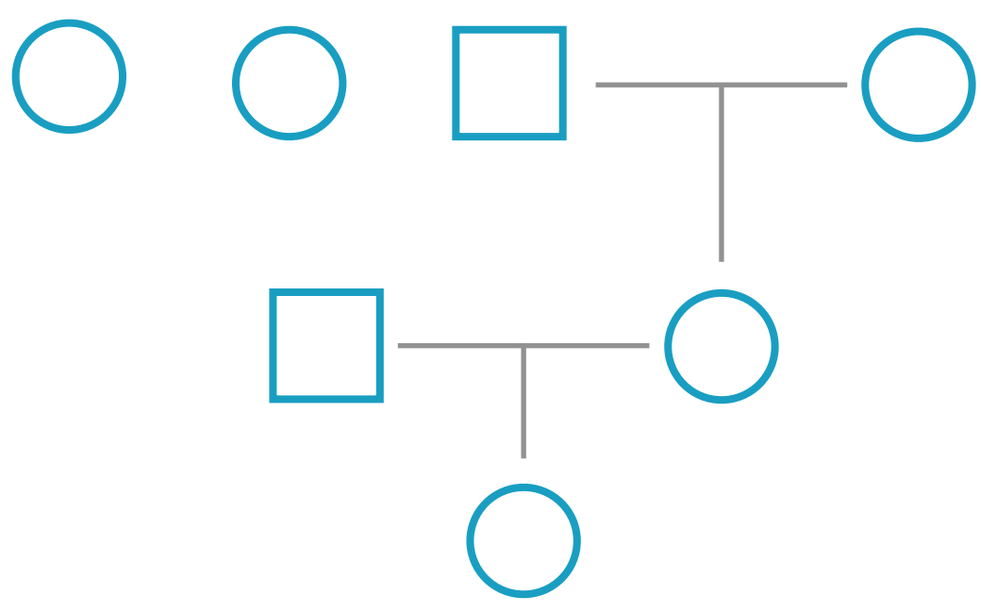
- Distance
- Availability
- Genetic composition
- Genetic compatibility
- Behavioural differences / incompatibility





$F_I = ?$

inbreeding coefficient



In the irreversible mutation model the allele frequency decreases over time depending on the mutation rate.

$$p_t = p_0(1-\mu)^t$$

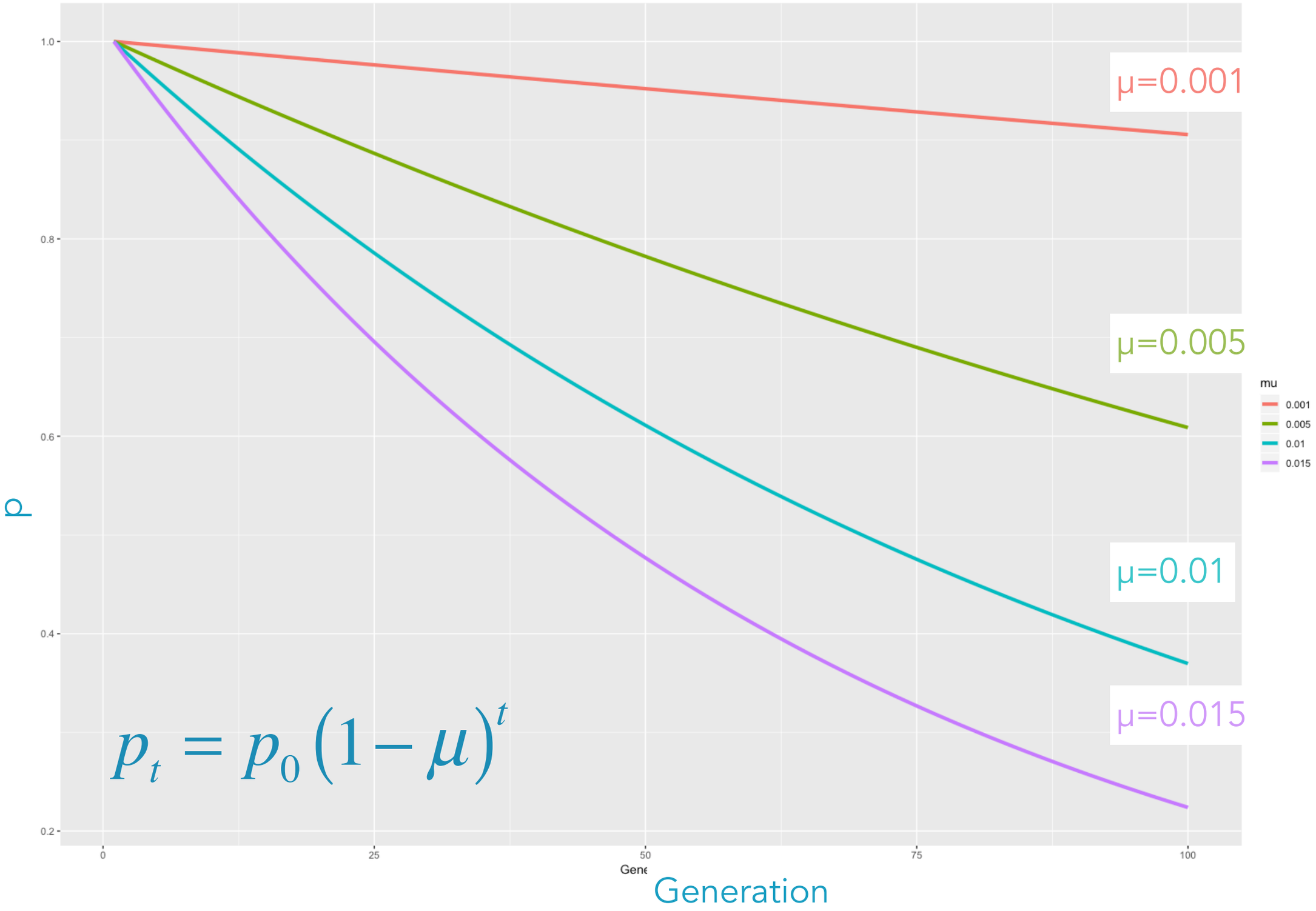
p_t : frequency of allele A after t generations
p_0 : starting frequency of allele A
μ : muation rate

The mutation rate in an "ideal" population equals zero and the allele frequency does not change from one generation to the next.

$$\mu = 0 \rightarrow p_t = p_0$$

```r
mutation_rates <- c(0.001, 0.005, 0.01, 0.015)

res.mutation <- data.frame(
  mu = rep(mutation_rates, each = 100),
  Generation = rep(1:100, times = 4),
  p = NA
)

for (mu in mutation_rates) {
  py <- 1
  res.mutation$p[res.mutation$mu == mu] <- py
  for (t in 2:100) {
    p.0 <-
      res.mutation$p[res.mutation$mu == mu &
                       res.mutation$Generation == (t - 1)]
    p.1 <- p.0 * (1 - mu)
    res.mutation$p[res.mutation$mu == mu &
                     res.mutation$Generation == t] <- p.1
  }
}
res.mutation$mu <- factor(res.mutation$mu)
ggplot(data = res.mutation, aes(x = Generation, y = p)) +
       geom_line(aes(colour = mu), size = 1.5)
```

PopGen ▷ Mutation



$$p_t = p_0(1-\mu)^t$$

In the irreversible mutation model the allele frequency decreases over time depending on the mutation rate.

$$p_t = p_0 e^{-\mu t}$$

p_t : frequency of allele A after t generations
p_0 : starting frequency of allele A
μ : muation rate
e : Euler's number

The number **e** is an important numbers in mathematics.

$$e = 2.718281828459045235 3...$$

There are different ways to calculate the value of **e**, but none of them ever give an exact answer, because **e** is irrational and its digits go on forever without repeating.

$$e = (1 + \frac{1}{n})^n$$

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + ...$$

Euler number in R:

```
exp(1)
```

exponential function

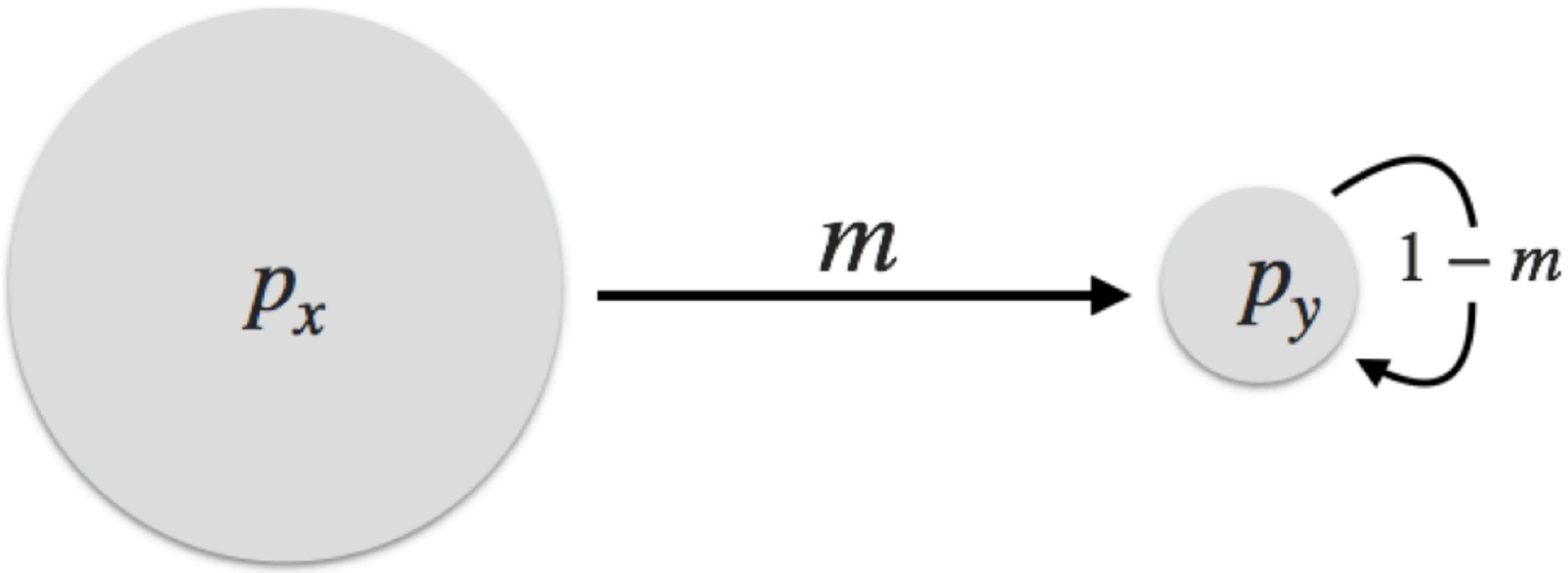
```
sprintf("%.10f", exp(1))
```

Do it yourself:

```
e <- sum(1/factorial(0:100))
```

Island Mainland Model

An island-mainland model is a simple formulation of a population model. There is a large mainland with allele frequency (p_x) and a small island population with allele frequency p_y. For each generation, a fraction m of the alleles from the mainland arrive on the island. The island population is composed of the fraction of alleles that stay on the island (1-m) and the ones that newly arrive from the mainland (m).



Island Mainland Model

$$p_1 = \underbrace{p_0(1-m)}_{\text{alleles that did not migrate}} + \underbrace{mp_m}_{\text{new alleles}}$$

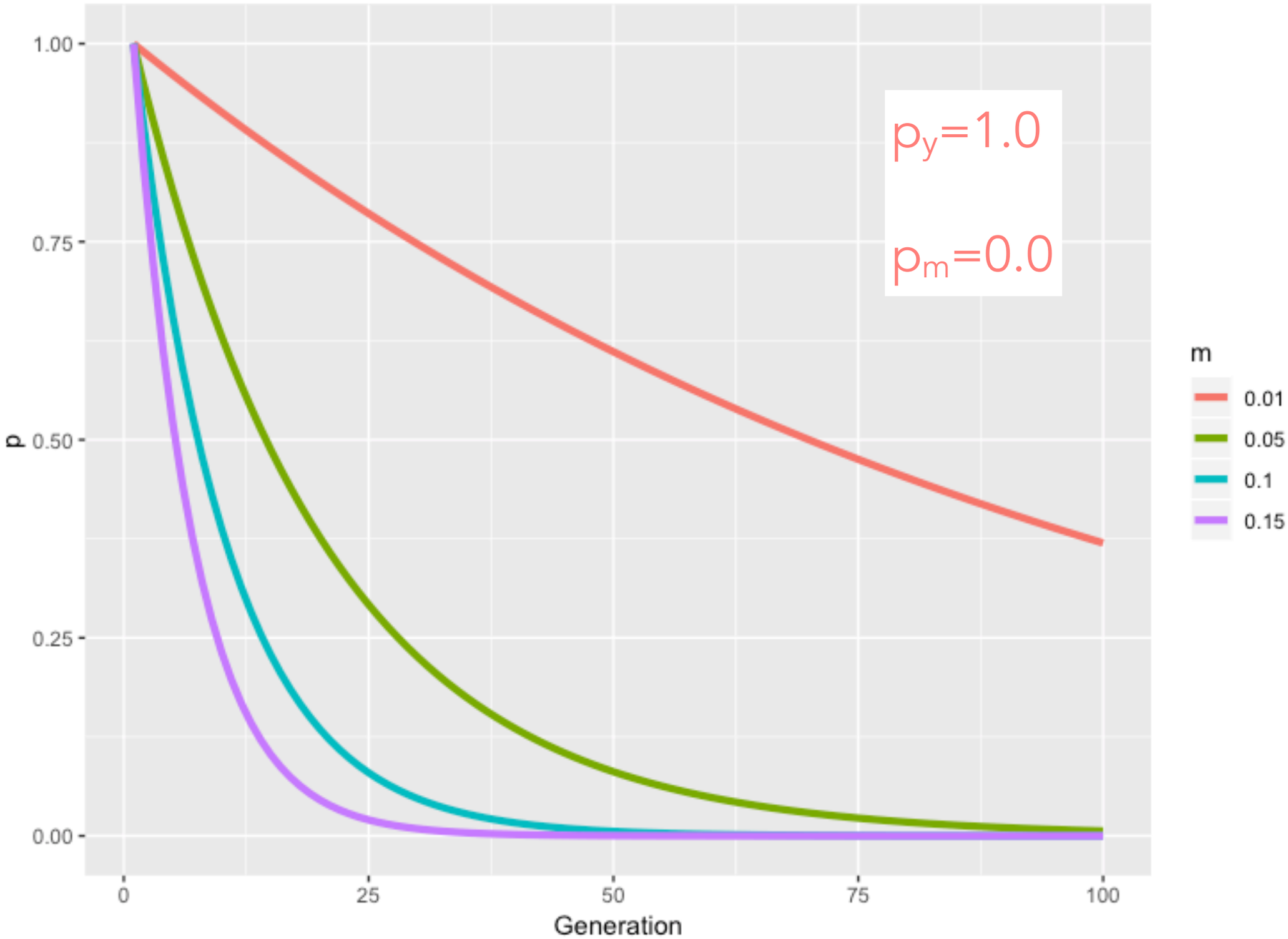
m: proportion of migrants (migration rate)

$\Rightarrow$ (1-m): proportion of non-migrants

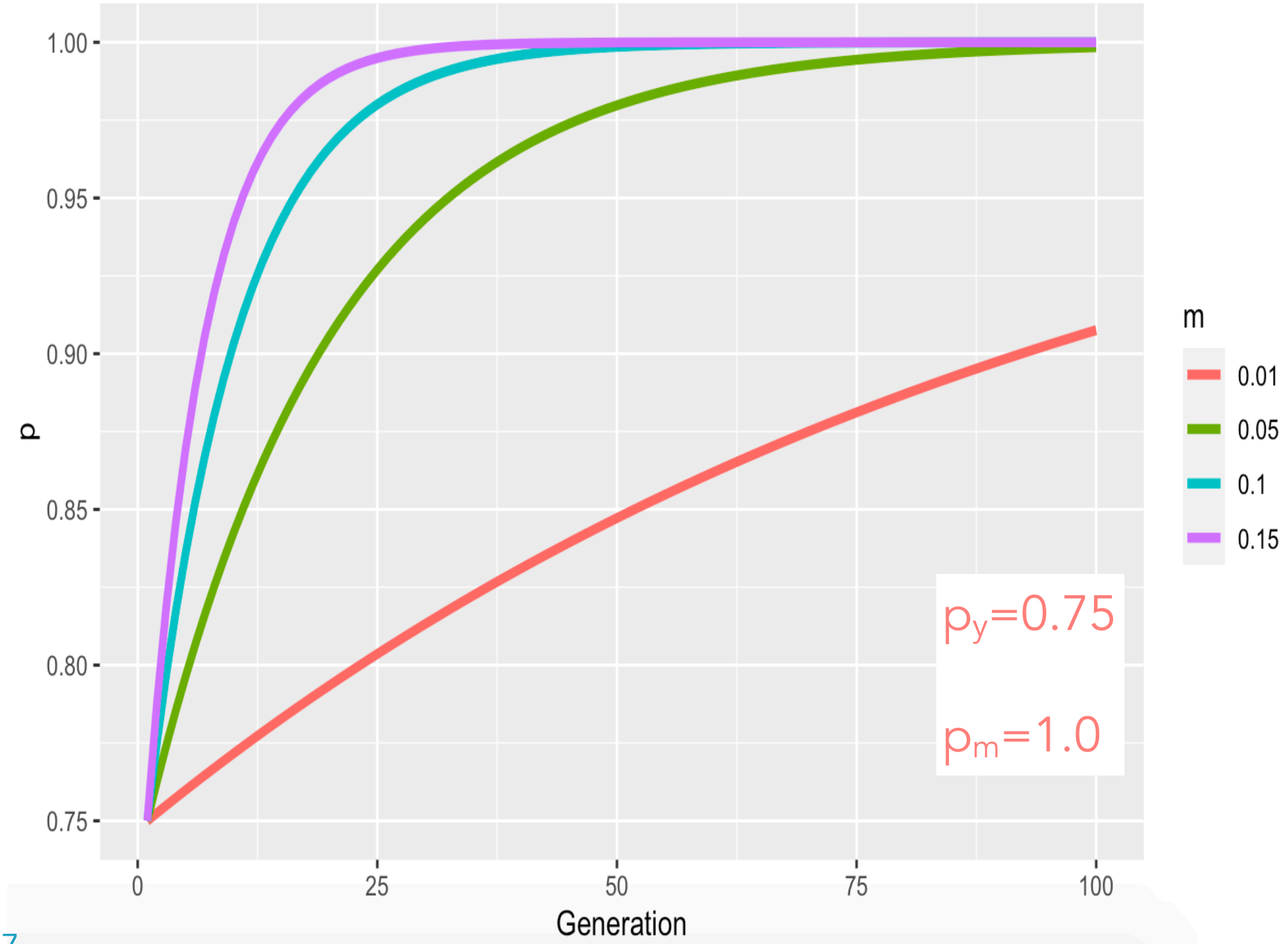
$$m = 0 \rightarrow p_1 = p_0$$

```
migration_rates <- c(.01, .05, .10, .15)

res.migration <- data.frame(
  m = rep(migration_rates, each = 100),
  Generation = rep(1:100, times = 4),
  p = NA
)

for (m in migration_rates) {
  py <- 1
  pm <- 0 # frequency mainland
  res.migration$p[res.migration$m == m] <- py
  for (t in 2:100) {
    p.0 <-
      res.migration$p[res.migration$m == m &
                        res.migration$Generation == (t - 1)]
    p.1 <- (1 - m) * p.0 + pm * m
    res.migration$p[res.migration$m == m &
                      res.migration$Generation == t] <- p.1
  }
}

res.migration$m <- factor(res.migration$m)
ggplot(data = res.migration, aes(x = Generation, y = p)) +
  geom_line(aes(color = m), size = 1.5)
```

PopGen ▷ Migration (Gene flow)

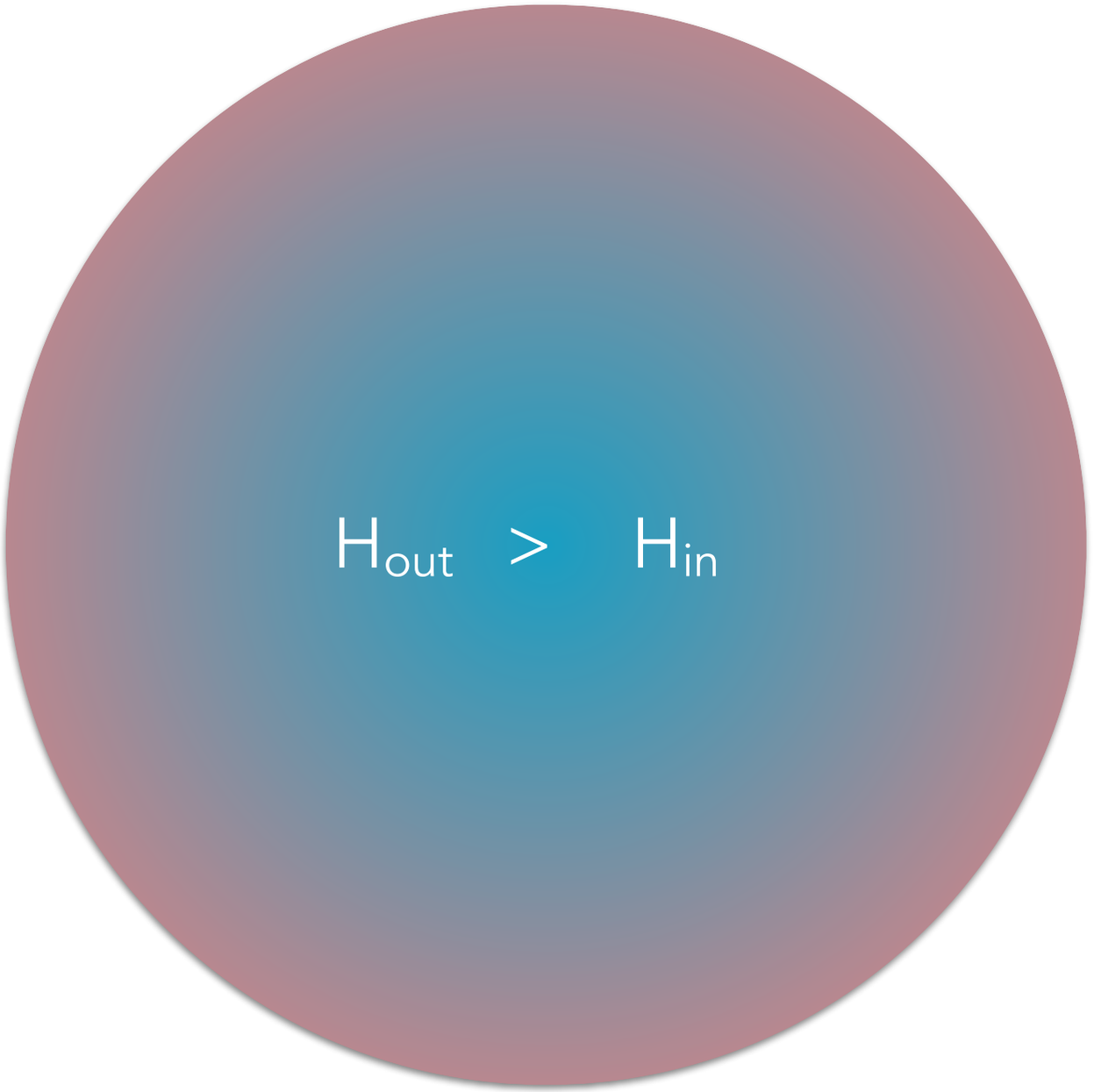


PopGen ▷ Migration (Gene flow)



"Populations inhabiting novel environments, such as at **range limits**, should be considered collectively for their potential to produce novel, adaptive genetic combinations. In the race to facilitate species tracking of rapidly changing climates, prescriptive gene flow among range limit populations may represent a form of genetic rescue that increases the evolutionary capacity of range limit populations to respond to rapidly changing selective regimes."

Sexton JP, Strauss SY, & Rice KJ (2011). Gene flow increases fitness at the warm edge of a species' range.
Proceedings of the National Academy of Sciences of the United States of America, 108
(28), 11704-9 PMID: 21709253


H_{out} > H_{in}

Can mutations be use to test for gene flow?

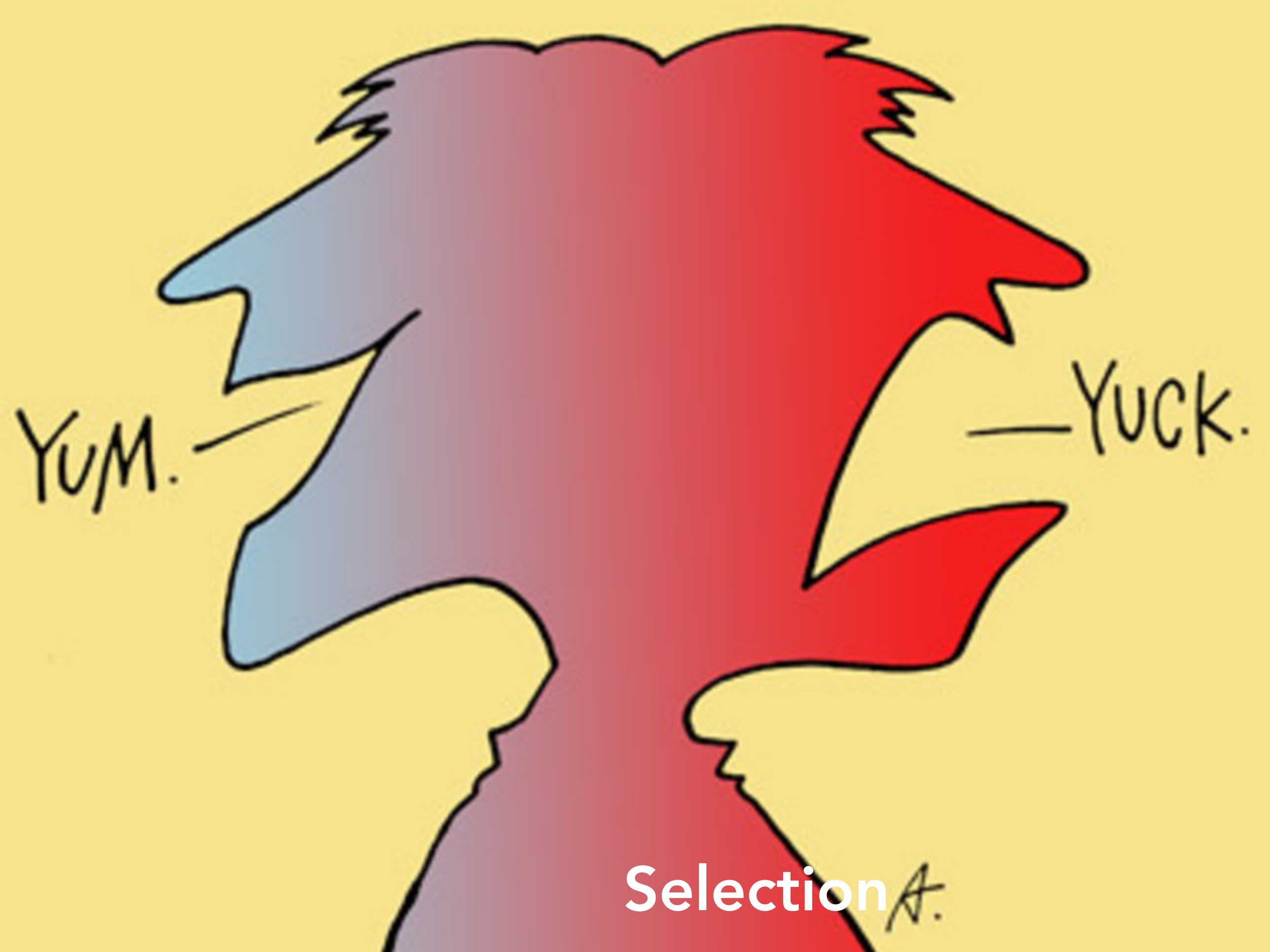
Evolution, 39(1), 1985, pp. 53–65

RARE ALLELES AS INDICATORS OF GENE FLOW

MONTGOMERY SLATKIN
Department of Zoology, NJ-15, University of Washington, Seattle, WA 98195

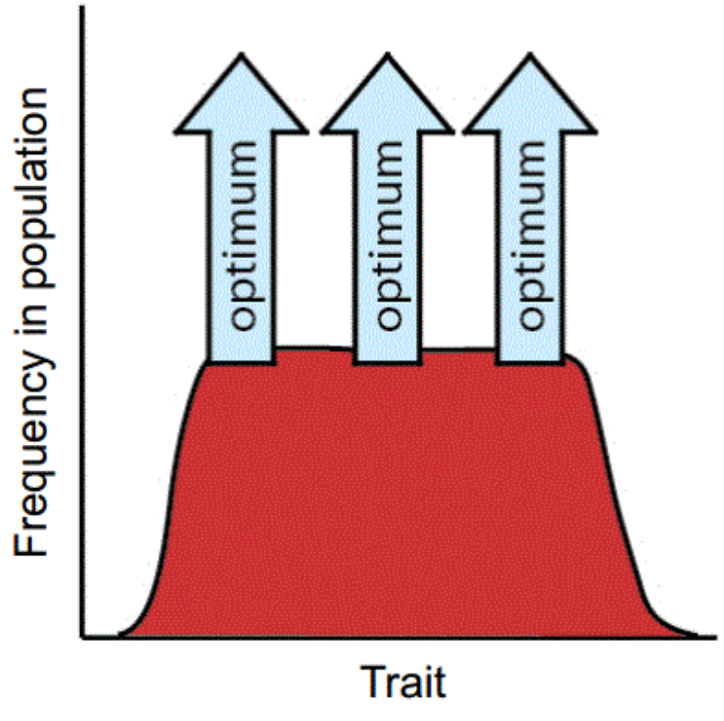
Summary.—A method for estimating the average level of gene flow in a subdivided population, as measured by the average number of migrants exchanged between local populations, Nm, is presented. The results from a computer simulation model show that the logarithm of Nm is approximately linearly related to the logarithm of the average frequency of private alleles, $\bar{p}(1)$, in a sample of alleles from the population. It is shown that this result is relatively insensitive to changes in parameters of the model other than Nm and the number of individuals sampled per population. The dependence of the value of $\bar{p}(1)$ on the numbers of individuals sampled provides a rough way to correct for differences in sample size.

This method was applied to data from 16 species, showing that estimated values of Nm range from much greater than 1 to less than 0.1. These results confirm the qualitative analysis of Slatkin (1981). This method was also applied to subsamples from a population to show how to measure the extent of isolation of local populations.


YUM.
—YUCK.
Selection A.

Sometimes **variability increases** when a trait lost its importance for survival. That has happened to some extent with our receptors for smell. We have about 1,000 genes that encode our smelling ability but only about half of them work - a different half in each person.

Darwinian evolution is driving the maintenance of diversity. This pro-diversity force is called "**balancing selection**". It can happen when multiple types of a gene offer different advantages.



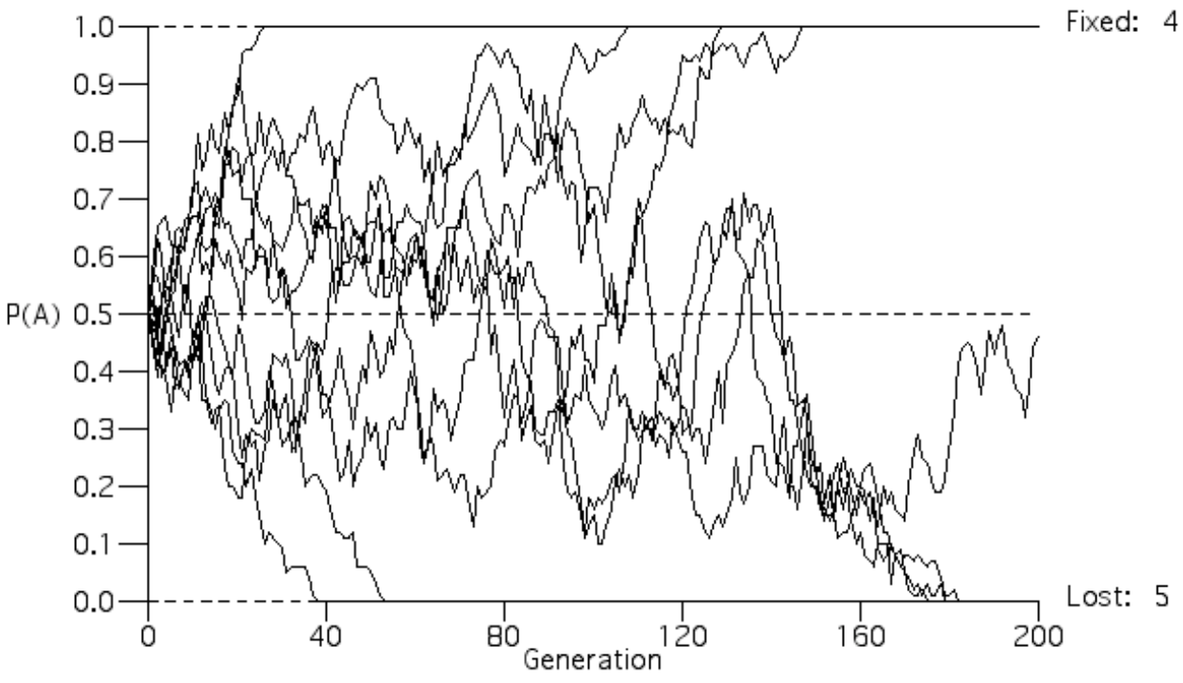
Note: Selection > Positive Selection > Balancing Selection

Balancing selection refers to a number of selective processes by which multiple alleles are actively maintained in the gene pool of a population at frequencies above that of gene mutation. **There are several mechanisms by which balancing selection works to maintain polymorphism.** The two major and most studied are **heterozygote advantage** and **frequency-dependent selection**.

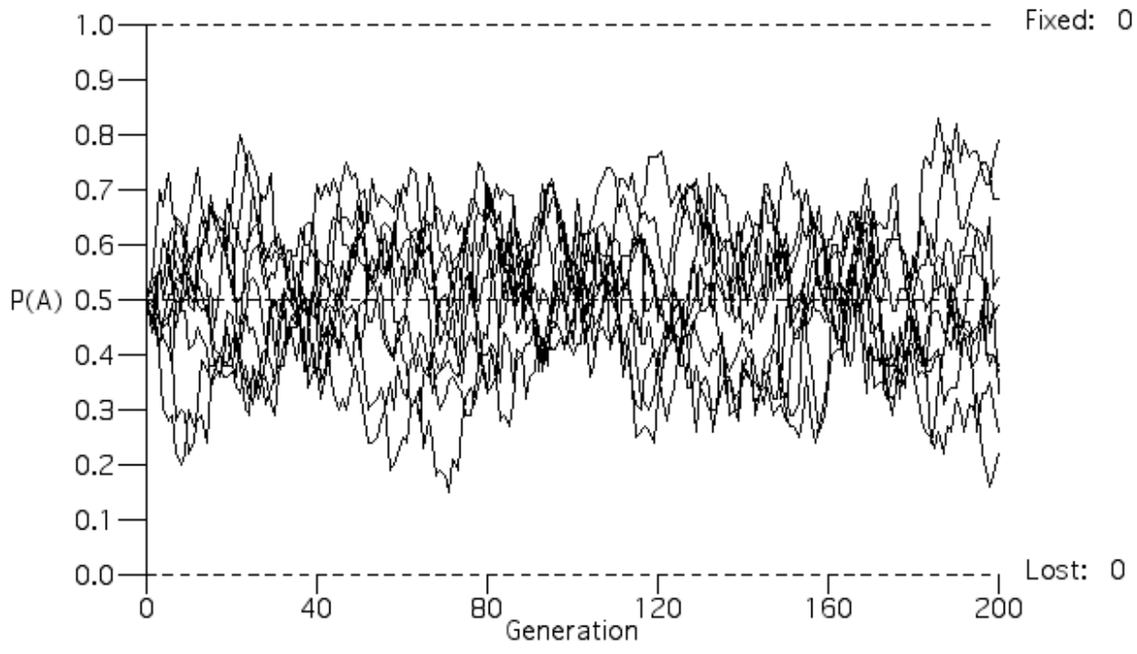
Heterozygote advantage - A situation in which a single disadvantageous allele is not selected out of a population, because, when a individual is heterozygous for that allele, it gains some sort of local advantage by having the disadvantageous allele. For example, the allele for **sickle-cell anemia** offers "resistance" to malaria. If a person in an area high in malaria is heterozygous for sickle-cell anemia, the "resistance" they gain to malaria outweighs the disadvantage of having heterozygous sickle-cell anemia. A person homozygote for the mutant allele will not have a greater advantage, even if they are completely resistant to malaria. What's the use of being resistant to malaria if you're blood can't carry oxygen?

Frequency-dependent selection (e.g. mimicry) occurs when the fitness of a trait depends on its frequency in a population. It is possible for the fitness of a genotype to increase (**positive frequency-dependent**) or decrease (**negative frequency-dependent**) as the genotype frequency in the population increases. Positive frequency dependent selection occurs when a trait has higher fitness when it is common than when it is rare. Negative frequency dependent selection occurs when a trait has higher fitness when it is rare than when it is common. Negative frequency dependent selection will turn out to have some important implications for evolution of other traits.

Heterozygote Advantage - Selection for the heterozygote and against the homozygotes. Allele frequencies move toward an equilibrium defined by the fitness values of the two homozygotes.



With genetic drift alone polymorphisms is decreasing over time.



With balancing selection, polymorphisms would be maintained for a longer time than under the pure drift model.

Balancing selection maintains polymorphism.

• Natural selection may favor non-poisonous butterflies that have the same color pattern as poisonous butterflies. This system is called **Batesian mimicry**. When they are rare, birds will tend to avoid the mimics, because they will have already have encountered a poisonous butterfly of the same appearance. But when the non-poisonous type is common, the previous encounters of birds with butterflies of their appearance are more likely to have been rewarding; the birds will not avoid eating them, and their fitness will be lower. The fitness of the mimics is **negatively frequency-dependent**.

With negatively frequency-dependent fitnesses (as in Batesian mimicry), it is possible for natural selection to maintain a polymorphism. When a genotype is rare, it is relatively favored by selection and it will increase in frequency; as it becomes more common, its fitness decreases and there may come a point at which it is no longer favored. At that point, the fitnesses of the different genotypes are equal and natural selection will not alter their frequencies: they are at **equilibrium**.

• In other butterflies, such as in central and south American Heliconius, there are several morphs within a species, each morph having a different color pattern. All the morphs are poisonous. When a morph is common, it will be more likely that birds will have already learned to avoid them, whereas birds will not yet have learned to avoid a rare morph. An individual of a rare morph is therefore more likely to be the unlucky prey that educates the bird, and gets killed in the process. The fitness of each morph is **positively frequency-dependent**.

A different example of **negative frequency dependent selection** occurs in fruit flies and is called the "**rare male advantage**". Female fruit flies in a population prefer to mate with a male with an unusual phenotype. Suppose for example that most individuals in the population have red eyes, but a few have white eyes. White eyed males will attract more females. This is not just because females like white eyes. If most males in the population have white eyes but a few have red eyes, females will mate preferentially with the red-eyed males.

Proceedings of the National Academy of Sciences
Vol. 65, No. 2, pp. 345–348, February 1970

The Mating Advantage of Rare Males in *Drosophila*

Lee Ehrman*

THE ROCKEFELLER UNIVERSITY, NEW YORK CITY

Communicated by Theodosius Dobzhansky, November 20, 1969

Abstract. The mating advantage of rare *Drosophila* males is tested using two eye color mutants. In one experiment, the flies remained for three hours in observation chambers containing 25 pairs; in another experiment they stayed for 24 hours in mass cultures of 200 individuals. The outcome of this latter experiment was followed for ten generations, with all competition other than that for mates eliminated. For initial frequencies of 80 per cent for the common and 20 per cent for the rare type, the frequencies converged to approximate equality because the rare males were favored as mates. When the formerly rare type increases in frequency, it loses its mating advantage, and a balanced equilibrium is eventually attained.

A_1A_1 A_1A_2 A_2A_2

$$w_{11} = w_{12} > w_{22} = 0$$

What is the expected fate of allele A_2?

A_1A_1 A_1A_2 A_2A_2

$$w_{11} = w_{12} > w_{22} = 0$$

	A_1	A_1
A_1	A_1A_1	A_1A_1
A_1	A_1A_1	A_1A_1

	A_1	A_1
A_1	A_1A_1	A_1A_1
A_2	A_1A_2	A_1A_2

	A_1	A_2
A_1	A_1A_1	A_1A_2
A_2	A_1A_2	A_2A_2

A_1A_1 A_1A_2 A_2A_2

$$w_{12} > w_{11} = w_{22} = 0$$

What is the expected genotype frequency after 10 generations?

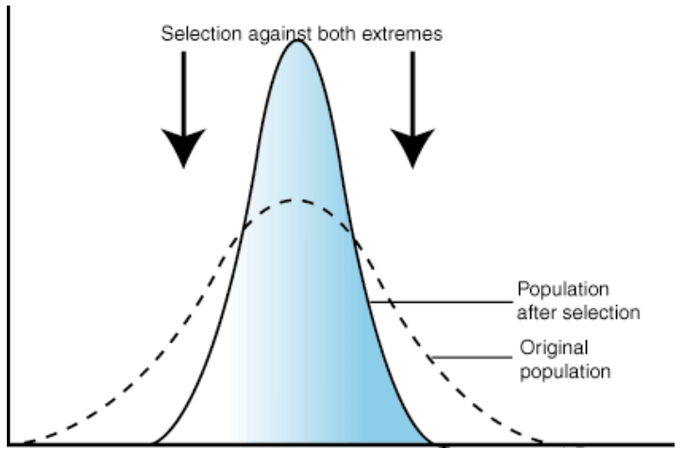
A_1A_1 A_1A_2 A_2A_2

$$w_{12} > w_{11} = w_{22} = 0$$

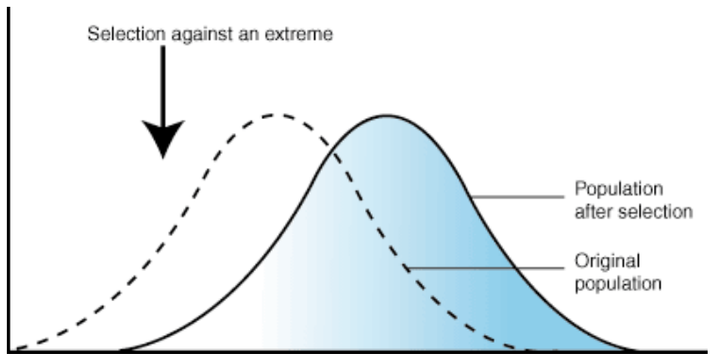
$$f(A_1A_1) = f(A_2A_2) : 25\,\%$$

$$f(A_1A_2) : 50\,\%$$

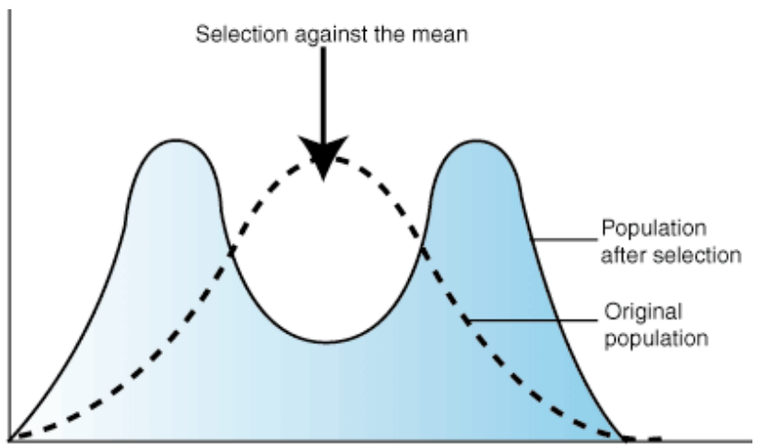
Stabilising selection



Directional selection



Disruptive selection



The relative contribution of the three genotypes to the next generation is determined by the product of the relative fitness and the frequency before selection of that genotype.

$$f(A_1A_1) = \frac{p^2\omega_{11}}{\bar{\omega}}$$

$$f(A_1A_2) = \frac{2pq\omega_{12}}{\bar{\omega}}$$

$$f(A_2A_2) = \frac{q^2\omega_{22}}{\bar{\omega}}$$

$$p' = \frac{p^2\omega_{11} + pq\omega_{12}}{\bar{\omega}}$$

$$q' = \frac{pq\omega_{12} + q^2\omega_{22}}{\bar{\omega}}$$

The **mean fitness** of the population is the sum of the relative contribution of the different genotypes:

$$\bar{\omega} = p^2\omega_{11} + 2pq\omega_{12} + q^2\omega_{22}$$

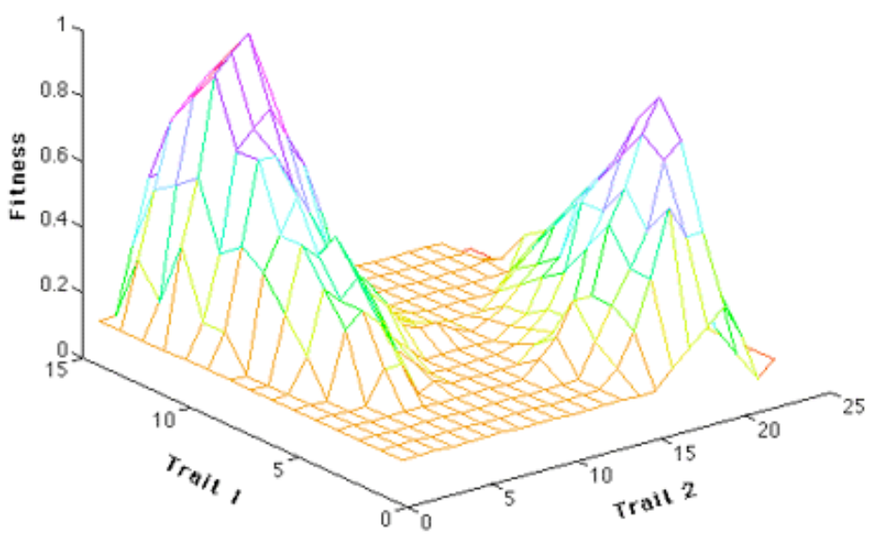
Genotype	A_1A_1	A_1A_2	A_2A_2	A_1A_3	A_2A_3	A_3A_3
Fitness	w_{11}	w_{12}	w_{22}	w_{13}	w_{23}	w_{33}
Frequency	p^2	$2pq$	q^2	$2pr$	$2qr$	r^2

$$p' = \frac{(p^2\omega_{11} + pq\omega_{12} + pr\omega_{13})}{\bar{\omega}}$$

$$q' = \frac{(q^2\omega_{22} + pq\omega_{12} + qr\omega_{23})}{\bar{\omega}}$$

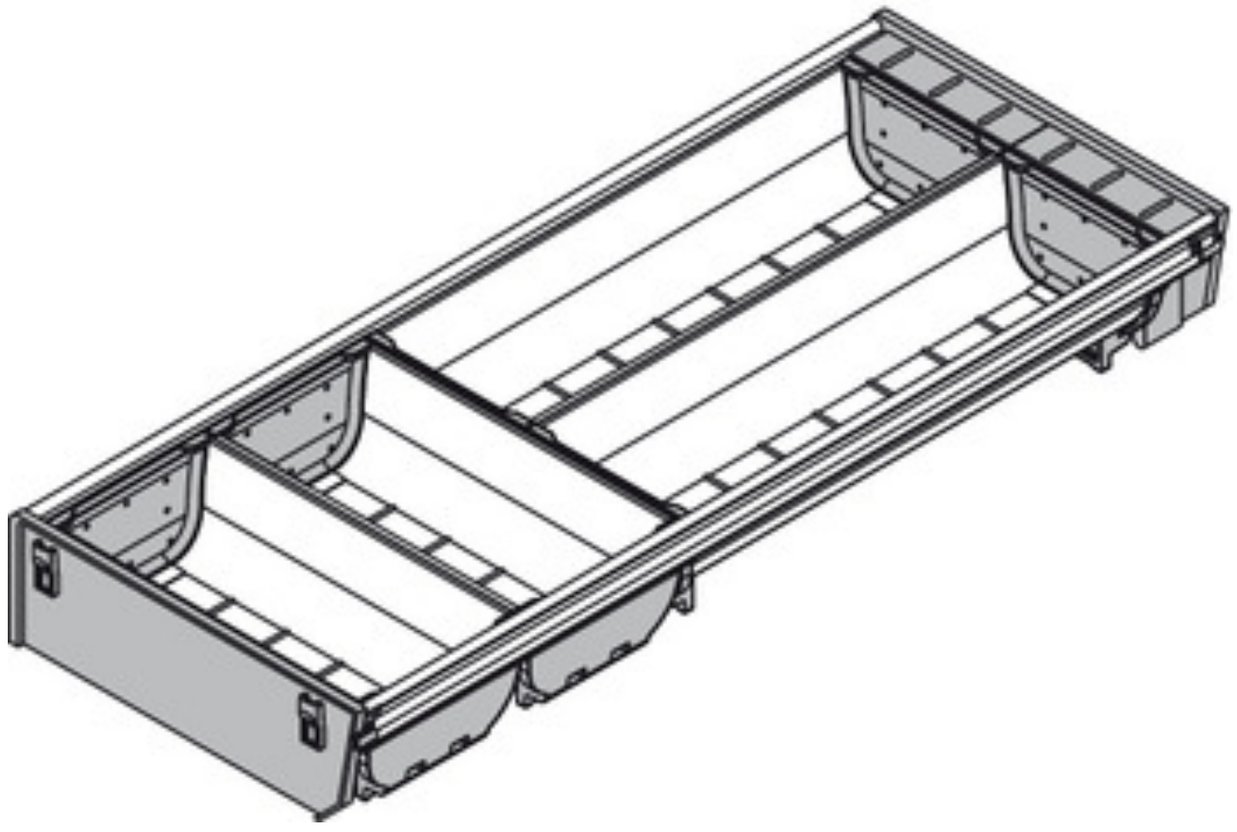
$$r' = \frac{(r^2\omega_{33} + pr\omega_{13} + qr\omega_{23})}{\bar{\omega}}$$

In a population, there may be just one coadapted gene complex, or there might be several different combinations of traits, each of which could have high fitness. This latter possibility gives rise to another concept: that of an **adaptive landscape**. An adaptive landscape is the description of the fitnesses of all possible combinations of different traits in a population. Adaptive landscapes are frequently represented graphically; fitness is plotted on a vertical axis and trait values for different genes are plotted on other axes. Combinations of traits that have high fitness thus appears as peaks, and combinations that have low fitness appear as valleys. Here is an example of an adaptive landscape:



This example shows two traits, and a situation in which there are two combinations of traits, the peaks in the graph, shown in purple, that have high fitness, while other combinations of the traits have low fitness. These peaks in the adaptive landscape can be called **adaptive peaks**; note that they are also combinations of different genetic traits that, together, have high fitness, so they are coadapted gene complexes. An adaptive peak and a coadaptive gene complex are thus basically the same thing.

PopGen ▷ Subdivision



The **fixation index** (F_{ST}) is a measure of population differentiation due to genetic structure.

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

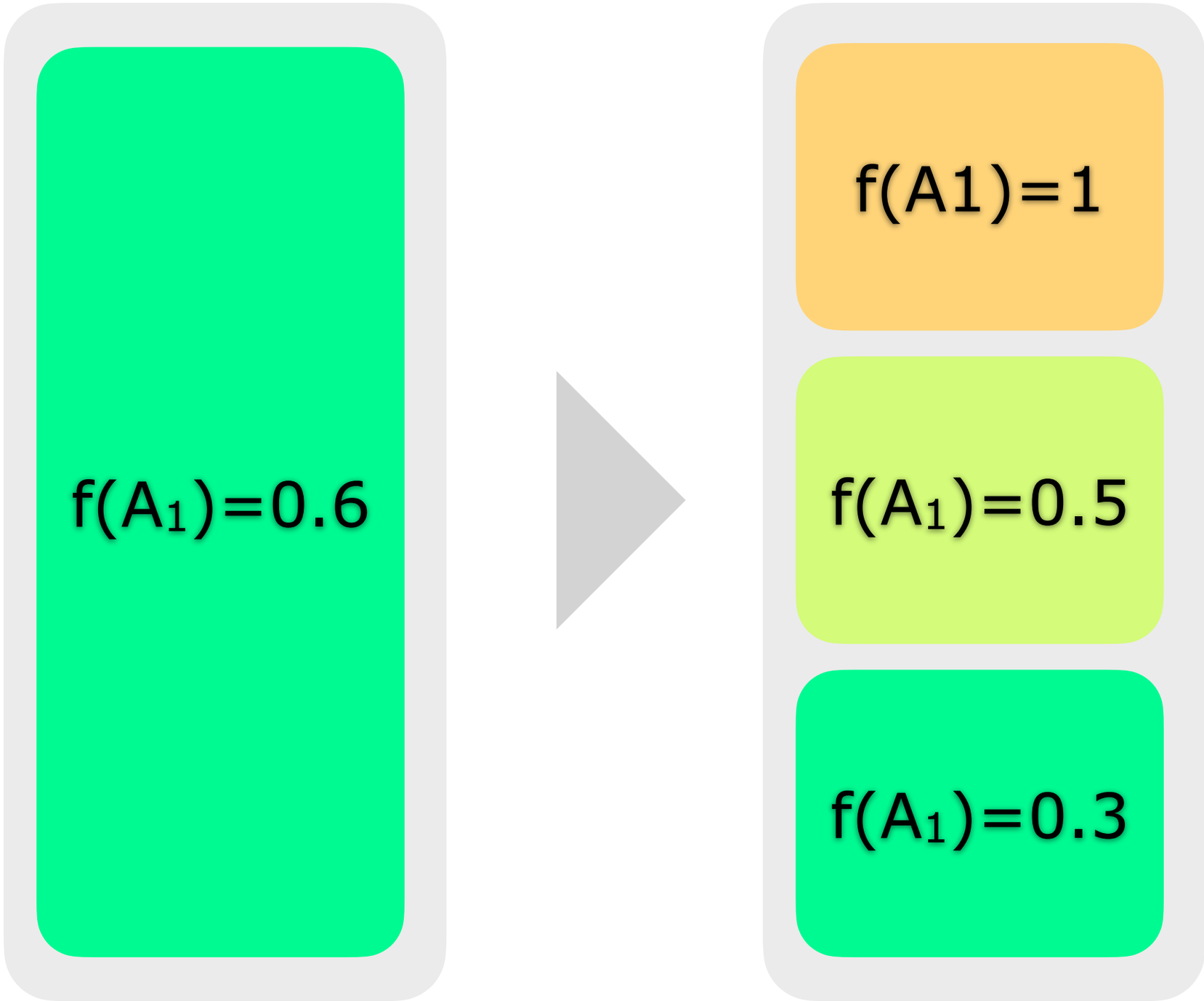
Expected Heterozygosity

Expected average heterozygosity assuming HWE

$$H_S = \sum \frac{2 p_i q_i}{n}$$

Heterozygosity expected **without** subdivision

$$H_T = 2\overline{p}\overline{q}$$


f(A1)=1
f(A1)=0.6
f(A1)=0.5
f(A1)=0.3

f(A1)=1

f(A1)=0.5

f(A1)=0.3

Expected average heterozygosity assuming HWE

$$H_S = \sum \frac{2p_i q_i}{n} = \frac{0 + 2*0.25 + 2*0.21}{3} = 0.307$$

Heterozygosity expected without subdivision

$$\bar{p} = \frac{1 + 0.5 + 0.3}{3} = 0.6$$

$$\bar{q} = 1 - \bar{p} = 0.4$$

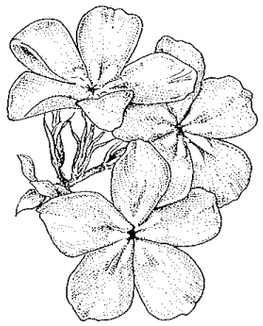
$$H_T = 2\bar{p}\bar{q} = 2*0.6*0.4 = 0.48$$

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{0.307}{0.48} = 0.36$$

BOX 3.1 *Calculation of F-statistics*

Levin (1978) scored allele frequencies at the *Pgm-2* locus in 43 Texas subpopulations of *Phlox cuspidata*. Forty of these subpopulations were fixed for the *b* allele (listed together in the first row of the table below). In the other three subpopulations the frequencies of *b* were 0.49, 0.83, and 0.91, with observed heterozygote frequencies of 0.17, 0.06, and 0.06, respectively:

Subpopulation	p_i	H_I
1–40	1	0
41	0.49	0.17
42	0.83	0.06
43	0.91	0.06



Phlox cuspidata

From these data we can calculate the three hierarchical *F*-statistics and their components as follows:

$\overline{H}_I$ = the observed proportion (frequency, not numbers) of heterozygotes within subpopulations, averaged over all subpopulations:

$$\overline{H}_I = \frac{(40\times 0)+0.17+0.06+0.06}{43} = 0.0067$$

$\overline{H}_S$ = the expected proportion of heterozygotes within subpopulations, assuming random mating (= $2p_iq_i$), averaged over all *n* subpopulations:

$$\overline{H}_S = \frac{\sum_{i=1}^{n} 2p_iq_i}{n} =$$

$$\frac{(40\times 0)+2(0.49\times 0.51)+2(0.83\times 0.17)+2(0.91\times 0.09)}{43} = 0.0220$$

H_T = the expected proportion of heterozygotes over the entire metapopulation (=$2\bar{p}\,\bar{q}$). H_T is not itself an average because there is only one metapopulation, but since there is a unique value for *p* and *q* for each subpopulation, the average allele frequencies across all subpopulations are used:

$$\bar{p} = \frac{(40\times 1)+0.49+0.83+0.91}{43} = 0.9821$$

$$1-\bar{p} = \bar{q} = 0.0179$$

$$H_T = 2\bar{p}\bar{q} = 2\times 0.9821\times 0.0179 = 0.0352$$

Box 3.1 continued

We can now calculate the three *F*-statistics using the definitions given in the text:

$$F_{IS} = \frac{\overline{H}_S - \overline{H}_I}{\overline{H}_S} = \frac{0.0220-0.0067}{0.0220} = 0.70$$

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} = \frac{0.0352-0.0220}{0.0352} = 0.38$$

$$F_{IT} = \frac{H_T - \overline{H}_I}{H_T} = \frac{0.0352-0.0067}{0.0352} = 0.81$$

Check these using equation 3.9:

$$(1-F_{IS})(1-F_{ST}) = (1-F_{IT})$$

$$(1-0.7)(1-0.38) = (1-0.81)$$

$$(0.3)(0.62) = 0.19$$

$$0.19 = 0.19$$

F_{ST} or F_{IT} are metapopulation level measures of population structure, quantifying the degree of subpopulation differentiation within the total population, and the overall amount of reduction in heterozygosity. Therefore, it doesn't make sense to calculate separate F_{ST} or F_{IT} for each subpopulation. Each subpopulation can have its own value for F_{IS}, however, because this is just the inbreeding coefficient we calculated before. To compare the reduction in heterozygosity across the three hierarchical levels, we use the average F_{IS} calculated previously.

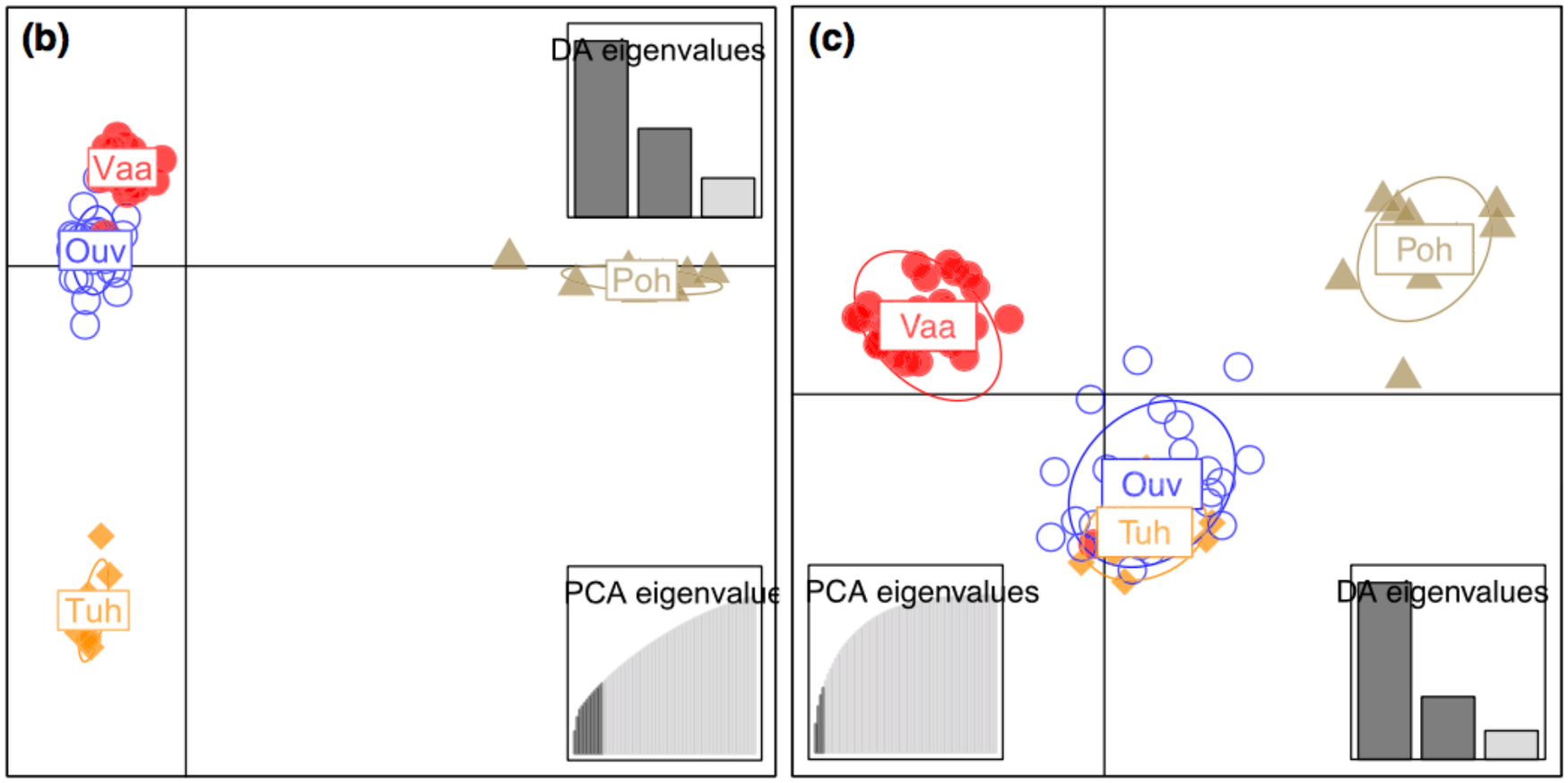
In this example F_{IS} is very large, demonstrating a high level of inbreeding in these self-fertilizing plants. This high F_{IS} comes entirely from the three unfixed subpopulations. The forty fixed subpopulations do not contribute to this estimate of inbreeding, because there is no genetic variation and thus no heterozygotes to be reduced in frequency. Mathematically, these 40 subpopulations are represented by zeros in both H_I and H_S, so they do not affect F_{IS}. F_{ST} is also quite large, which is also likely due to the high level of self-fertilization. Much of gene flow in plants is through movement of pollen by wind or animal pollinators, so when most pollen stays on the same plant, as it does in highly selfing species, it greatly reduces gene flow from pollen movement among subpopulations. Note that this high differentiation is mainly due to subpopulation #41, which is the only one with a lower frequency of the *b* allele.

The **conservation and management of endangered species** requires information on their genetic diversity, relatedness and population structure. The main genetic markers applied for these questions are **microsatellites** and **single nucleotide polymorphisms (SNPs)**, the latter of which remain the more resource demanding approach in most cases.

Source: Lemopoulos et al. (2018) Comparing RADseq and microsatellites for estimating genetic diversity and relatedness - Implications for brown trout conservation.

SNP-Data

Microsatellite-Data



Poh: Pohjajoki
Tuh: Tuhkajoki
Vaa: Vaarainjoki
Ouv: Hatchery

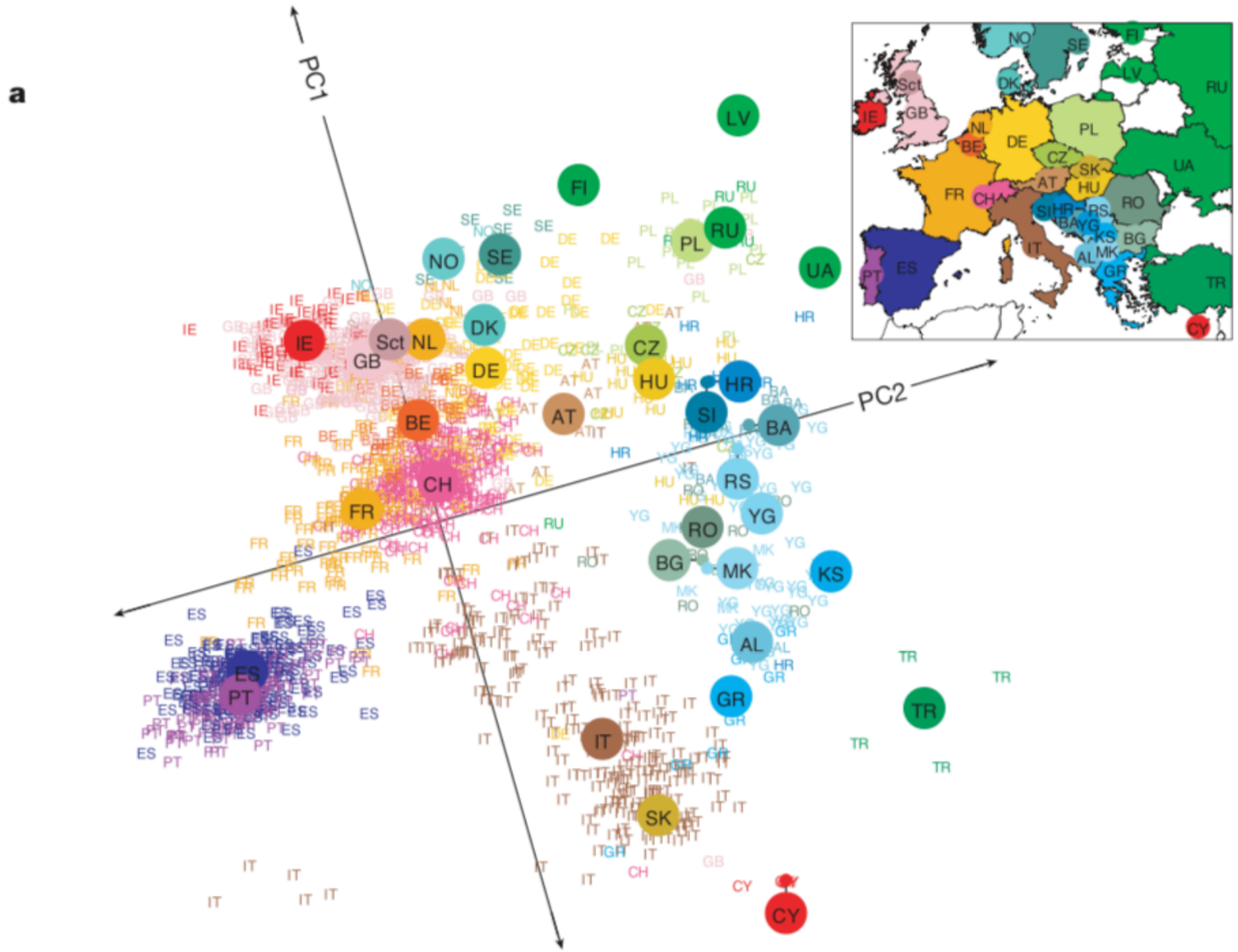
Grouping of *Salmo trutta* samples on **D**iscriminant **a**nalysis of **p**rincipal **c**omponents (DAPC) based on RADseq data (b) and microsatellite data (c) from the same individuals.

Source: Lemopoulos et al. (2018) Comparing RADseq and microsatellites for estimating genetic diversity and relatedness - Implications for brown trout conservation.

Heterozygosity has been associated with components of **fitness** in numerous studies across a wide range of taxa. Because heterozygosity is associated with **individual performance** it is also expected to be associated with **population dynamics**. However, investigations into the association between heterozygosity and population dynamics have been rare because of difficulties in linking evolutionary and ecological processes. The choice of heterozygosity measure is a further issue confounding such studies as it can be biased by individual differences in the frequencies of the alleles studied, the number of alleles at each locus as well as the total number of loci typed.

Source: Di Fonzo et al. (2011) The Population Growth Consequences of Variation in Individual Heterozygosity.

Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations.



Population structure within Europe. a, A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe.

Source: Novembre et al. (2008) Genes mirror geography within Europe.

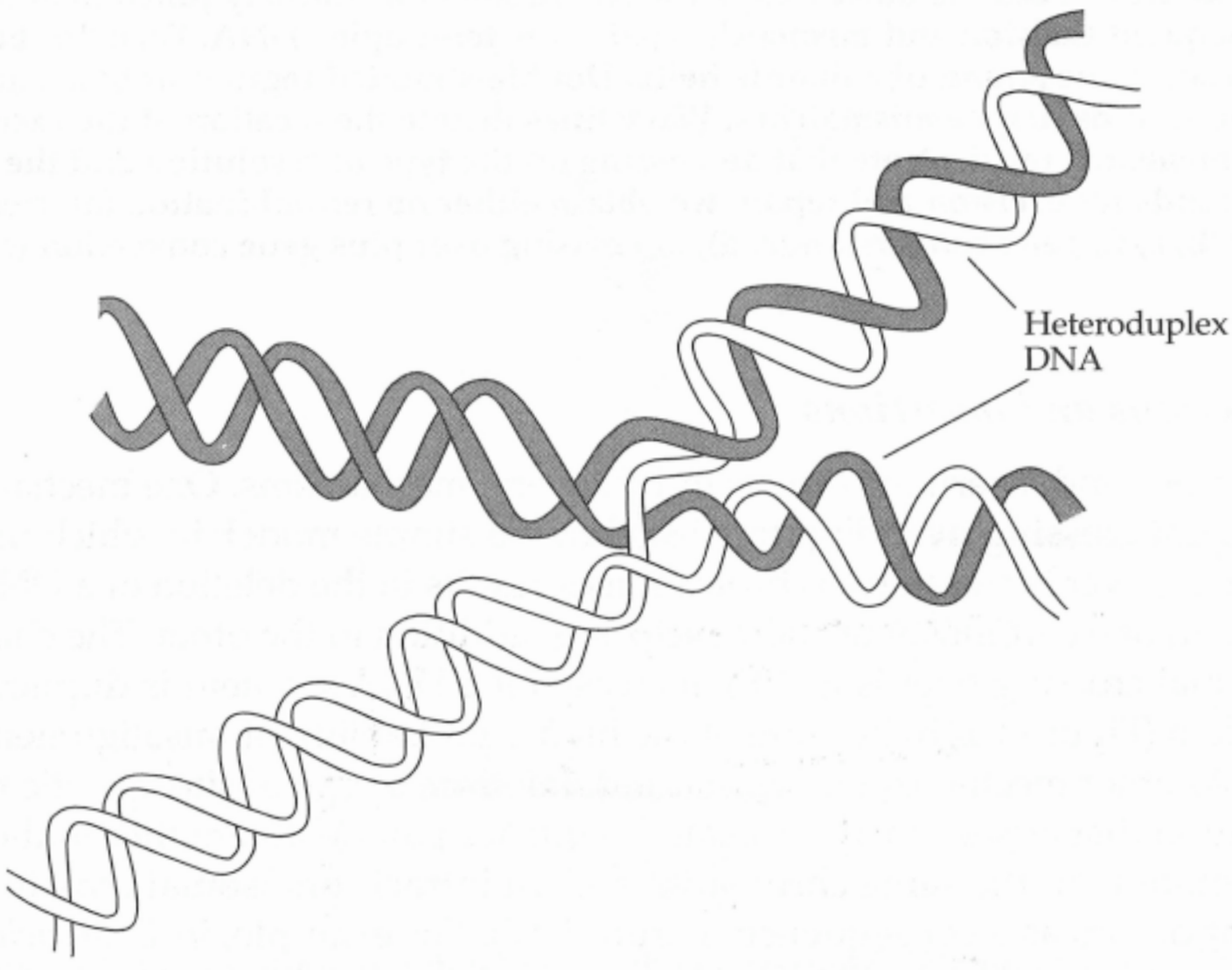


FIGURE 1.14 The Holliday structure. Note the heteroduplex DNA composed of mismatched chromatid strands (gray and white strands).

We consider two loci (A and B), each segregating for two alleles (A_1 / A_2 and B_1 / B_2)

```
--A1-?-B1--
--A2-?-B2--
```

Allele frequencies:
freq(A_1) = p_1 / freq(A_2) = p_2 ($p_1+p_2=1$)
freq(B_1) = q_1 / freq(B_2) = q_2 ($q_1+q_2=1$)

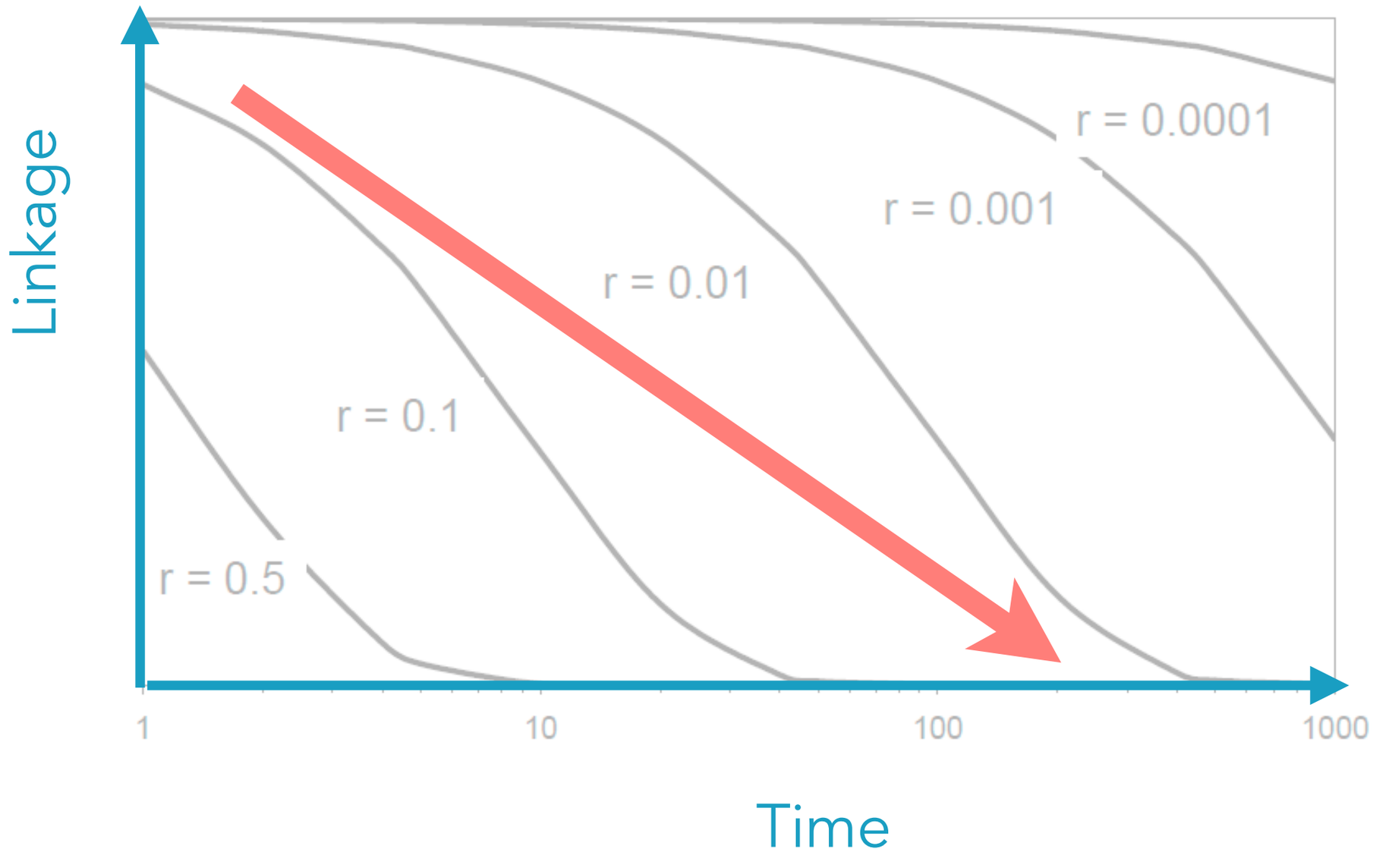
gametes (haplotypes)	frequency
A_1B_1	$x_{11}=p_1q_1+D$
A_1B_2	$x_{12}=p_1q_2+D$
A_2B_1	$x_{21}=p_2q_1$
A_2B_2	$x_{22}=p_2q_2$

$\sum x_{ij} = 1.0$

allele	frequency
A_1	$p_1=x_{11}+x_{12}$
A_2	$p_2=x_{21}+x_{22}$
B_1	$q_1=x_{11}+x_{21}$
B_2	$q_2=x_{12}+x_{22}$

↘ $x_{11}+x_{12} = p_1q_1+p_1q_2 = p_1(q_1+q_2) = p_1$

→ In the presented situation, there is no linkage disequilibrium and gamete frequencies can be accurately followed using allele frequencies.


Linkage
r = 0.0001
r = 0.001
r = 0.01
r = 0.1
r = 0.5
1
10
100
1000
Time

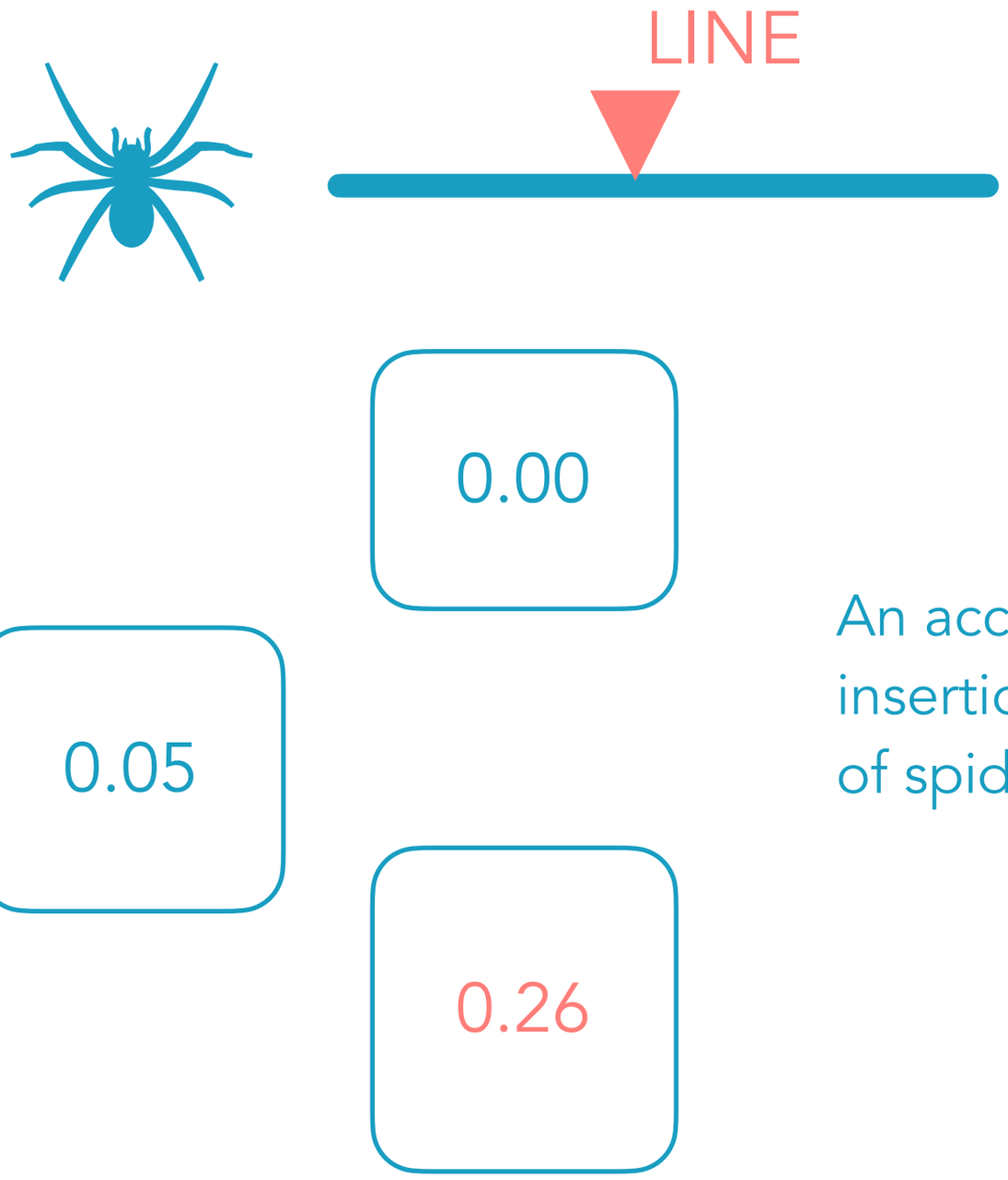
The human major histocompatibility complex (MHC) is a cluster of linked loci involved in immune response that is found in all vertebrates. These loci have a major role in fighting pathogens. There are consistent non-random associates of alleles at different loci (linkage disequilibrium) in the MHC.

With the advent of genome wide linkage maps of polymorphic DNA markers it has been possible to examine the influence of genes on the susceptibility to disease in a comprehensive way. Studies of this type have shown that within the human population there are multiple genetic loci which are involved in susceptibility to common autoimmune conditions such as diabetes and rheumatoid arthritis. It is likely that most autoimmune conditions have some genetic component. It is important to distinguish these kind of genetic susceptibilities from traditional inherited disease. No single predisposing allele has to be present for the disease to occur, rather the presence of combinations of susceptibility alleles significantly increases the probability of that individual developing the specific disease.

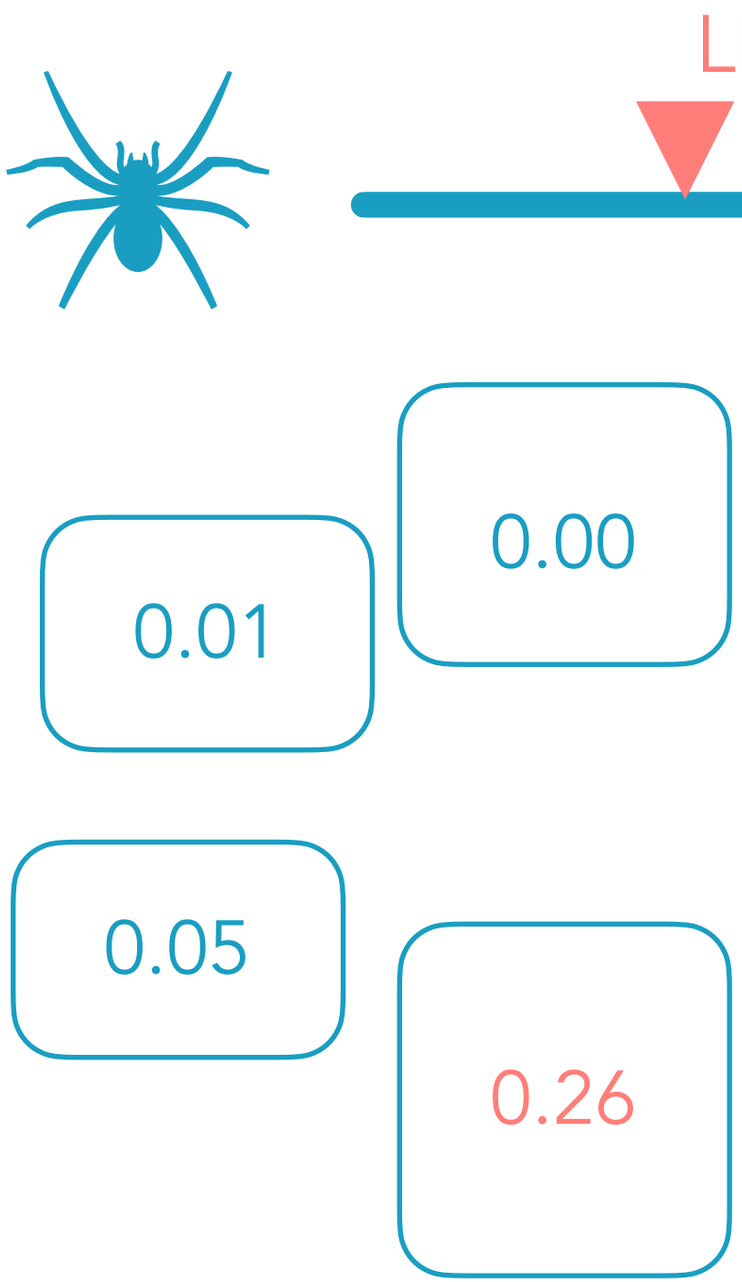
Haplotype frequencies for HLA-A and HLA-B loci. The color indicates a deficiency (red), or excess (blue) of the haplotype.

		HLA-A alleles			
		A1	A2	A3	overall HLA-B allele frequencies
HLA-B alleles	B7	**0.0074**	0.026	0.0477	0.1143
	B8	0.0672	0.011	0.0019	0.0971
	B35	0.0029	0.0178	0.0257	0.1052
	B44	0.0089	0.0503	0.0068	0.1242
overall HLA-A allele frequencies		0.1439	0.2855	0.1335	

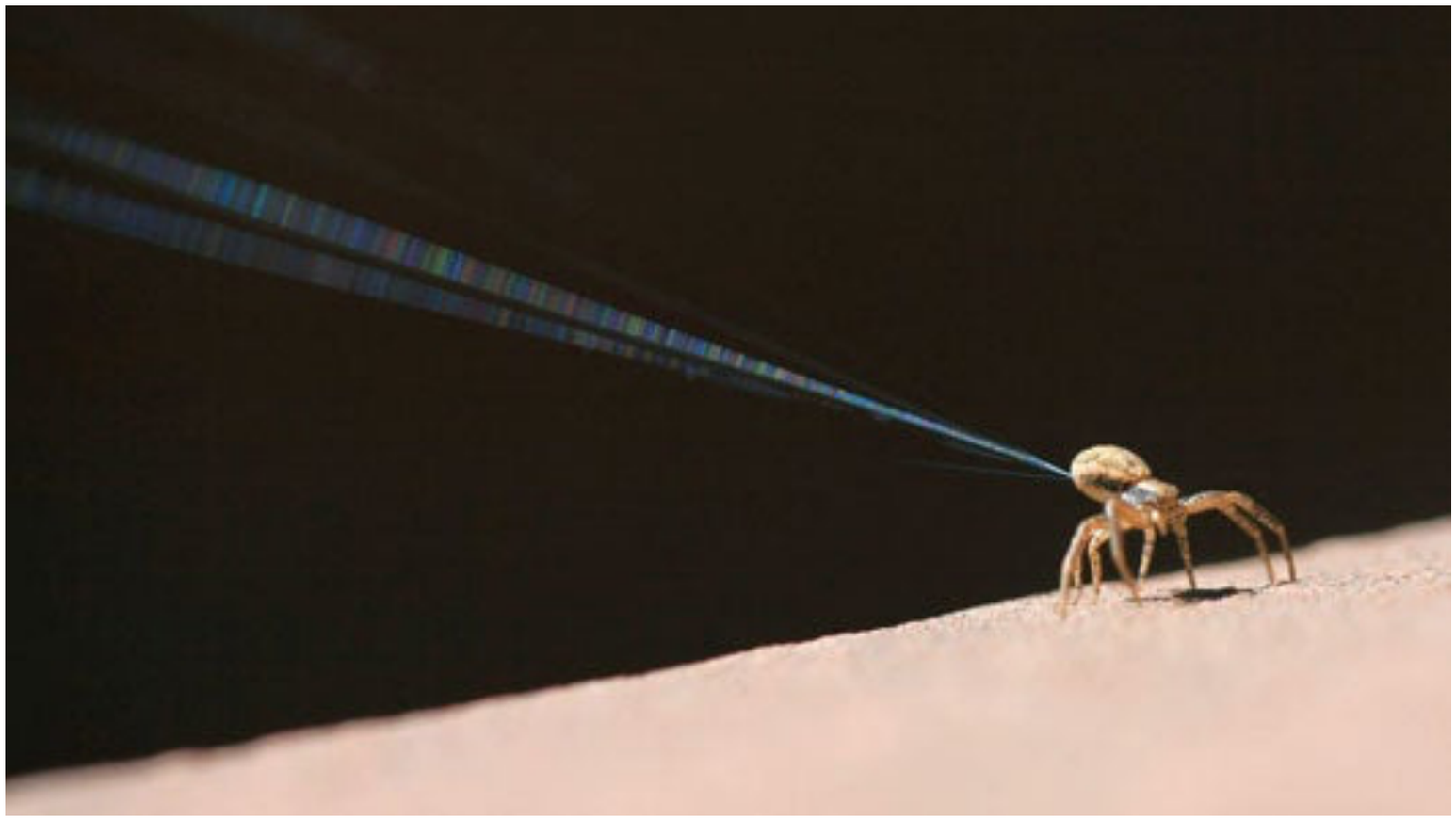
The frequency of A1-B7 haplotype is 0.0074. At linkage equilibrium it would be expected to have a frequency of 0.1439*0.1143=0.0164, so that it shows a deficiency of -0.0090. This deficiency is the likage disequilibrium associated with this halotype and is 55% of the maximum value that D could have for alleles with these frequencies,



An accumulation of a TE (e.g. LINE) insertion was detected in a subpopulation of spiders. How can this be explained?



- random observation
- genetic hitchhiking / selective sweeps
- positive selection



Not all animals need wings to move through the air. Some spiders have evolved an ability called ballooning. They spin silk threads that remain attached to their bodies. This technique allows the spiders to passively travel across extensive distances, even on relatively calm days.