Genetic drift

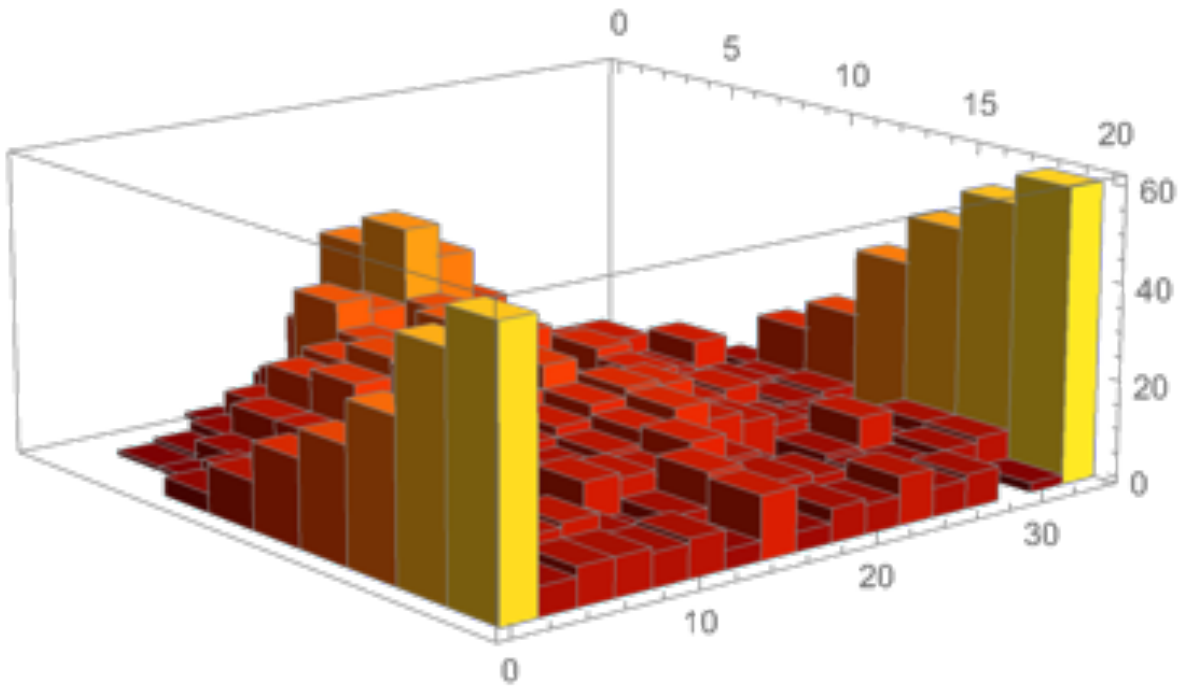
BIOL 434/509

Genetic drift

Genetic drift is the change in allele frequency from one generation caused by chance sampling in finite populations.

What happens in an individual population is unpredictable, but we can describe the distribution of allele frequencies among replicate populations.

Fly eye color (red [normal] vs. brown): Buri 1956:



Tribolium body color (black vs. reddish brown (normal)): Rich, Bell and Wilson 1979

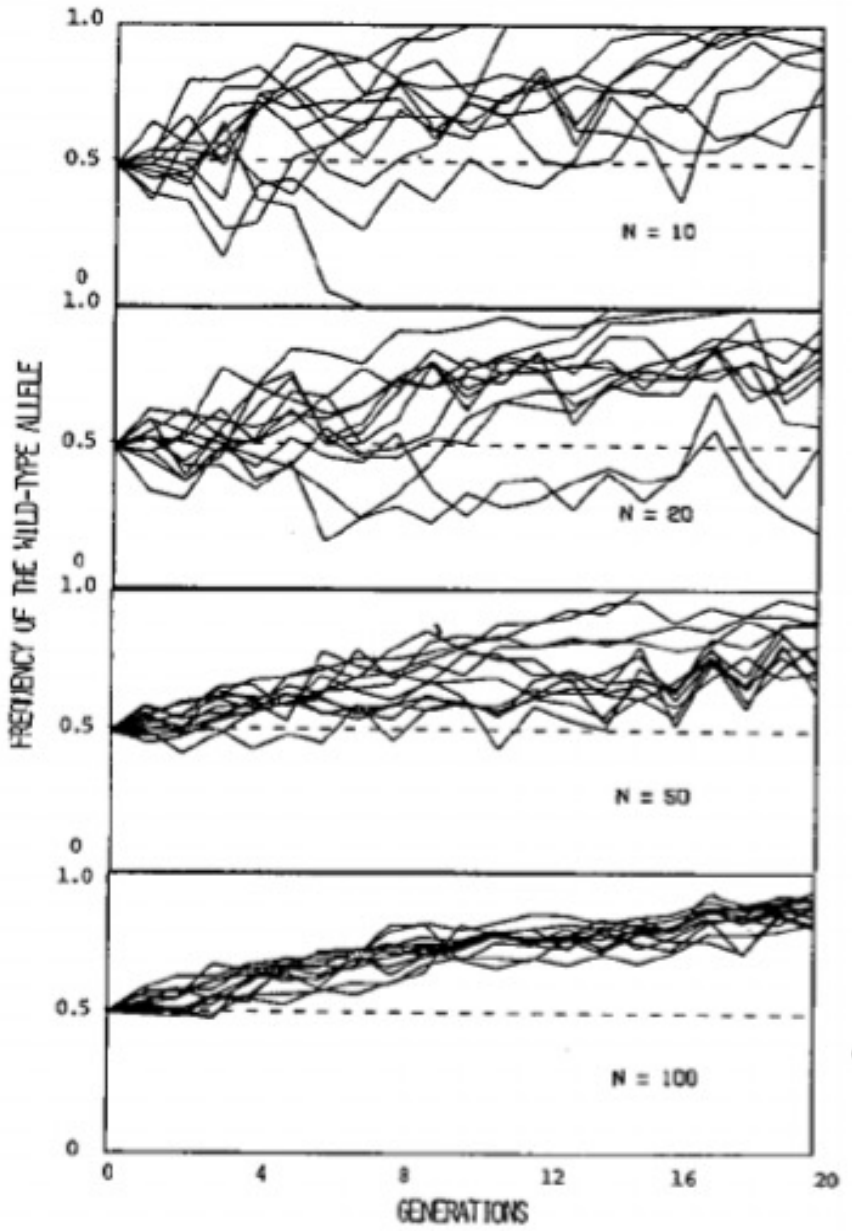


FIG. 1. Frequency of the wild-type allele (b^+) over generations for each population within the different population size groups.

https://keholsinger.shinyapps.io/Genetic-Drift/

An **ideal population** is one in which all individuals have an equal and independent probability of being the parent of each individual in the next generation.

Some of the unreasonable assumptions of the ideal population will be fixed later with the idea of the effective population size.

Drift in diploids

N individuals = 2*N*

With an ideal population and two different alleles, the number of copies of a particular allele in the *next* generation follows a binomial distribution

$$P(x) = \binom{2N}{x} p^x (1-p)^{2N-x}$$

General binomial distribution	Population Genetics for diploids
n = # of trials	$2N$ = # of alleles
X = number of "successes"	X = # of A alleles
p = probability of success	p = allele frequency
$E[X] = np$	$E[X] = 2Np$
$V[X]=np(1-p)$	$V[X]=2Np(1-p)$

$$E\left[p'\right] = p \qquad V\left[p'\right] = \frac{pq}{2N}$$

$$E[p'] = p \qquad V[p'] = \frac{pq}{2N}$$

So with pure drift…

- The expected value of the allele frequency doesn't change.
- The amount of drift is inversely proportional to population size.

For haploids, there are *N* alleles in the population, so

$$E[p'] = p \qquad\qquad V[p'] = \frac{pq}{N}$$

Fixation

Random genetic drift can continue until one allele is *fixed* (i.e. reaches a frequency of 1) or *lost* (reaches a frequency of 0).

Fixation of neutral alleles

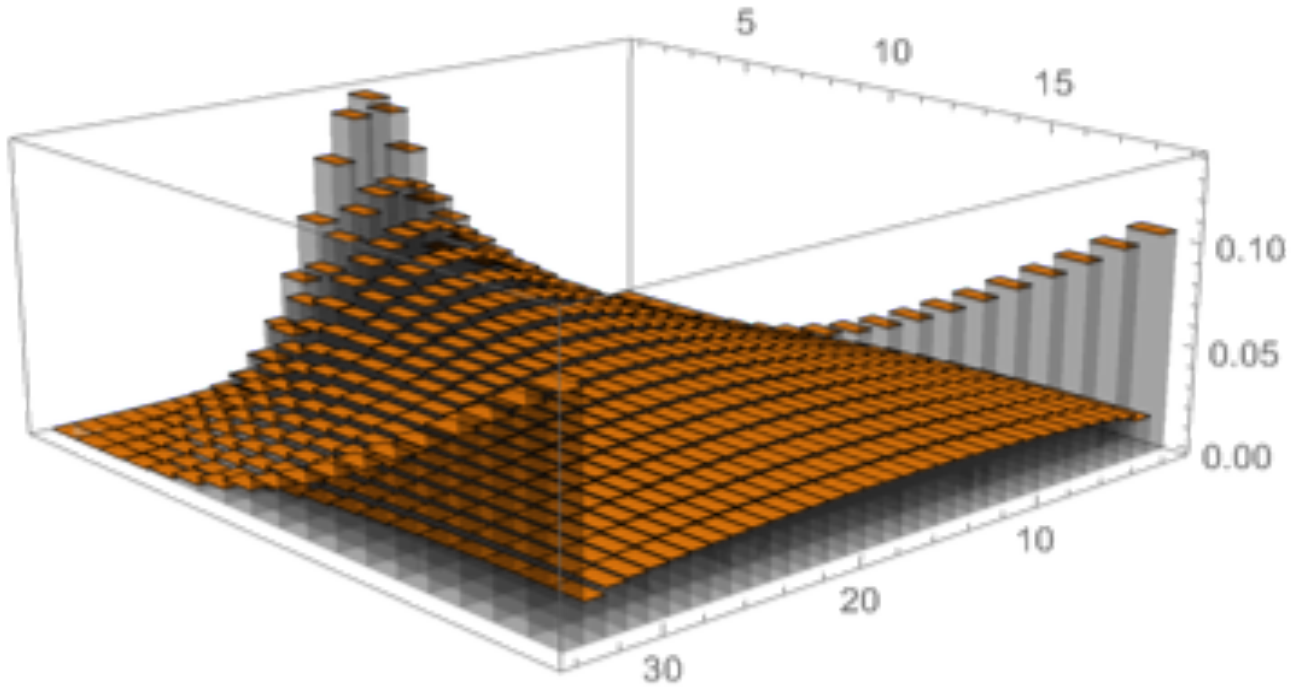
Without selection, mutation or migration, eventually every allele will be either fixed or lost.

The probability of eventual fixation of an allele affected only by drift = p (its allele frequency)

Modelling drift

1. Wright-Fisher model
2. Diffusion models
3. Identity in state
4. Coalescence

Wright-Fisher model



More on this later in term…

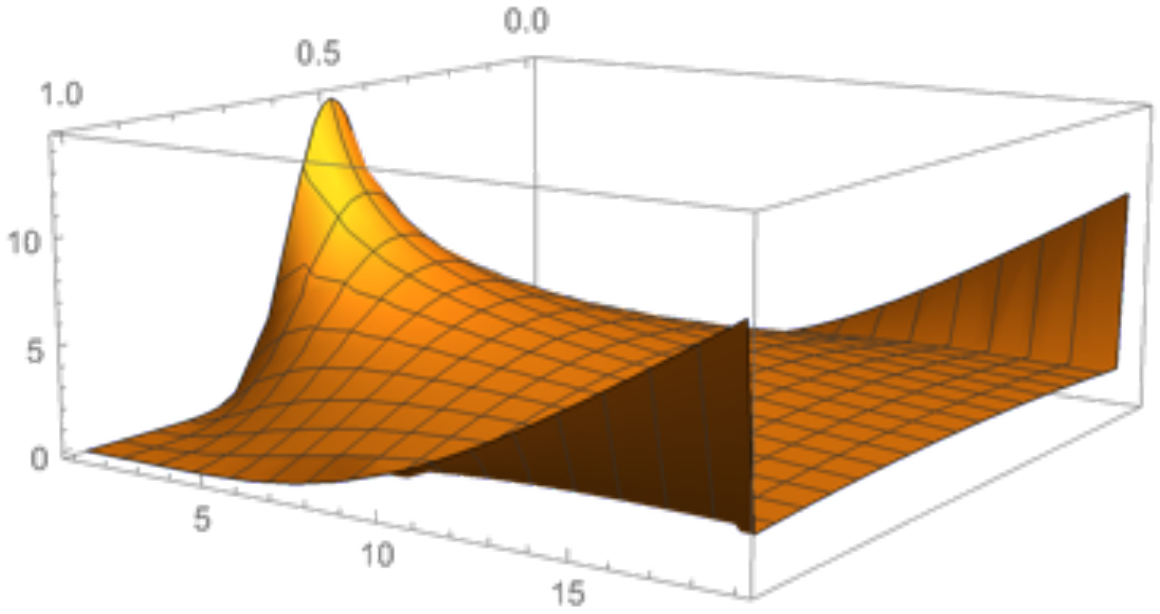
Diffusion models

A continuous approximation to the Wright-Fisher.

Source of many useful results

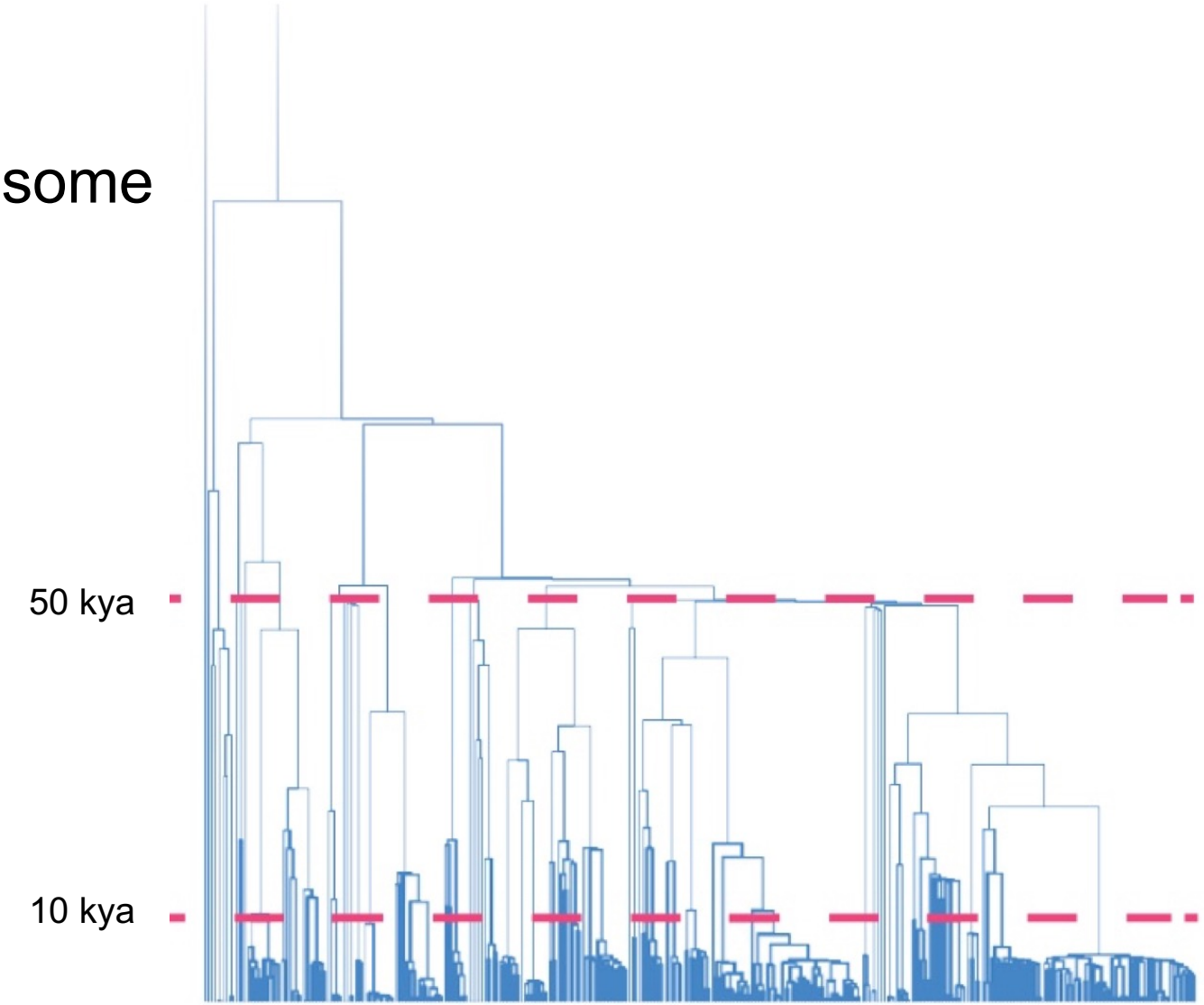
$$\frac{\partial \phi(p,x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\frac{\phi}{2N_e} - \frac{\partial}{\partial x}M_{\partial x}\phi$$

Diffusion models



Coalescence – understanding gene trees

Human Y chromosome



Karmin et al. (2015, Genome Res 25:1)

Coalescence

The **coalescent** considers the genealogy of the alleles in a population *backwards in time*, starting from the present.

MRCA = "**M**ost **r**ecent **c**ommon **a**ncestor" : most recent shared ancestor of two (or more) alleles.

https://phytools.shinyapps.io/coalescent-plot/

Haploids

Probability of coalescence in previous generation of two alleles is $1/N_e$.

Distribution of time to MRCA is a geometric distribution. Therefore:

Mean time to MRCA of two alleles is

$1/ (1/N_e) = N_e$.

The variance of the time to MRCA is approximately N_e^2.

Diploids

Twice as many alleles as haploids, so

Probability of coalescence between two alleles in a given generation is $1/(2N_e)$.

Therefore the mean time to MRCA of two alleles is $2N_e$, with variance $4N_e^2$.

Multiple alleles

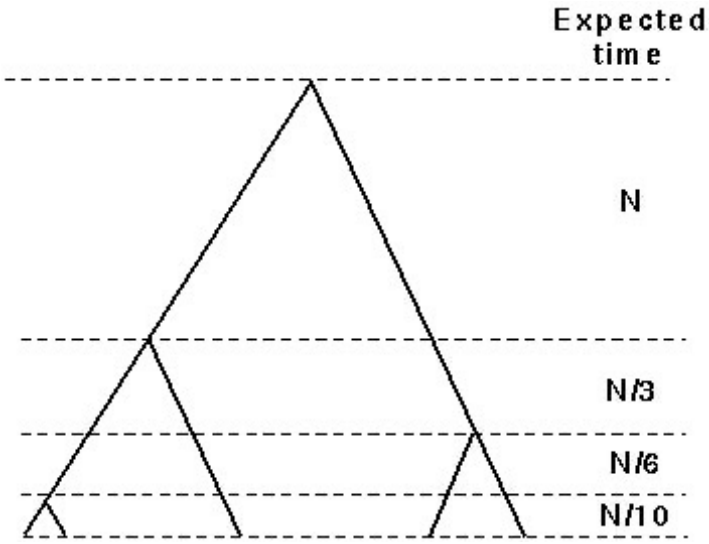
Out of *n* alleles, there are $\binom{n}{2}$ pairs of alleles.

Therefore the probability that any coalescent event happens among those *n* alleles approximately $\binom{n}{2}/N_e$ for haploids (or $\binom{n}{2}/(2N_e)$ for diploids).

Multiple alleles

Distribution of time until any coalescent event happens is a geometric distribution with mean $N_e / \binom{n}{2}$ for haploids. (For diploids, $2N_e / \binom{n}{2}$)

Multiple alleles

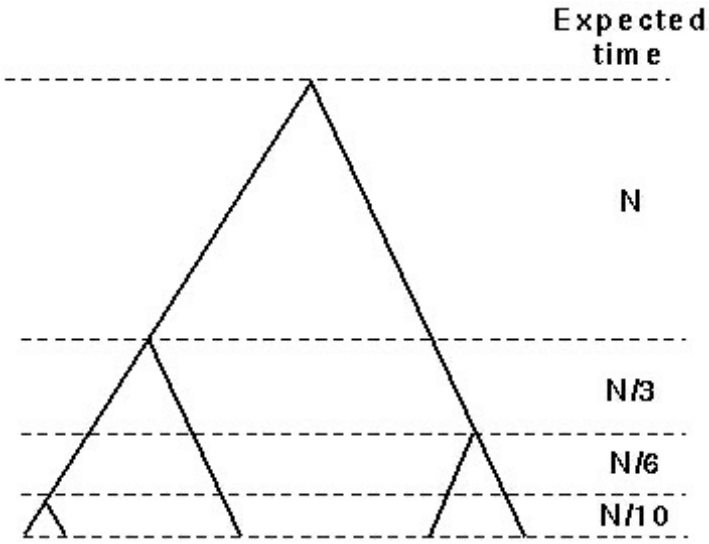


After a coalescent event has happened, there are now n – 1 remaining lineages.

The mean time until the next (i.e. previous) coalescent event is now $N_e / \binom{n}{2}$.

(For diploids, $2N_e / \binom{n}{2}$)

Multiple alleles

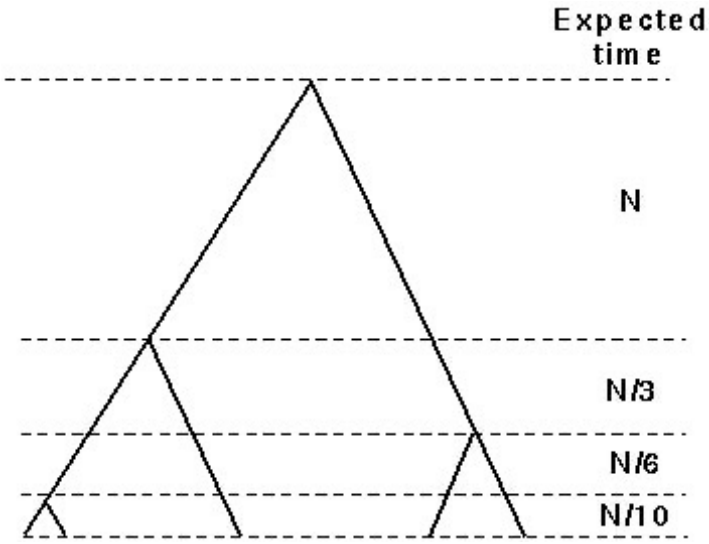


Expected time to coalescence of all *n* alleles = $\sum_{i=2}^{n} \frac{N}{\binom{i}{2}} = 2N_e \frac{n-1}{n}$.

(Note that this reduces to N_e for n=2.)

(For diploids, $4N_e \frac{n-1}{n}$)

Multiple alleles



Expected time to coalescence of all *n* alleles

$$= \sum_{i=2}^{n} \frac{N}{\binom{i}{2}} = 2N_e \frac{n-1}{n}.$$

For a large number of alleles *n*, this time to MRCA of all alleles is approximately $2N_e$ for haploids (or $4N_e$ for diploids).

Approximately half of the time waiting for the MRCA is waiting for the last two alleles to coalesce.

Identity in state

Let *F* be the probability that two alleles drawn at random are the same. (This is called the probability of identity in state.)

$$F' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F$$

$$F' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F$$

example of using conditional probability:

Pr[2 alleles are the same] =

Pr[2 alleles are the same | they came from the same parent allele] x

Pr[they came from the same parent allele] +

Pr[2 alleles are the same | they came from different parent alleles] x

Pr[they came from different parent alleles]

Multiple generations

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t (1 - F_0)$$

Haploids

$$F_t = 1 - \left(1 - \frac{1}{N}\right)^t (1 - F_0)$$

Relationship to Identity by descent

Similar equations for probability of identity by descent, f

$$f_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1}$$

$$f_t = 1 - \left(1 - \frac{1}{2N}\right)^t (1 - f_0)$$

Normally we define the period of time we are looking for changes in identity by descent such that $f_0 = 0$.

Loss of heterozygosity

Heterozygosity is simply :

$H = 1 - \Pr[\text{identity in state}] = 1 - F$

In terms of identity by descent:

$$H_t = (1 - f_t)H_0$$

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

Loss of genetic variance

Genetic variance within a population reduces on average over time through drift:

$$V_{G,t} = (1 - f_t)V_{G,0}$$

$$V_{G,t} = \left(1 - \frac{1}{2N}\right)^t V_{G,0}$$

Probability of fixation

Fixation occurs when an allele reaches a frequency of 1 in a population.

The probability of fixation of an allele not subject to selection is its current allele frequency.

An allele that has just appeared by mutation is at frequency $1/(2N)$. If it is neutral this is also its probability of fixation.

Effective population size, N_e

Predicts amount of drift in non-ideal population

$$F_t = 1 - \left(1 - \frac{1}{2N_e}\right)^t (1 - F_0)$$

The effective population size of a population is the size of an ideal population which acts the same as the real population in question.

Effective population size, N_e

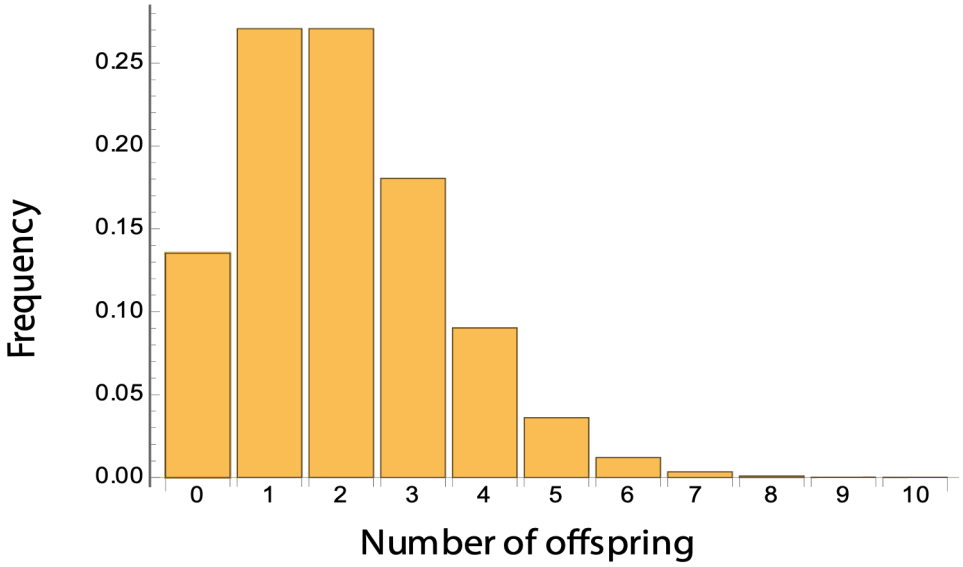
$$F_t = 1 - \left(1 - \frac{1}{2N_e}\right)^t (1 - F_0)$$

$$f_t = 1 - \left(1 - \frac{1}{2N}\right)^t (1 - f_0)$$

$$V_{G,t} = \left(1 - \frac{1}{2N}\right)^t V_{G,0}$$

Real populations have greater variance in reproductive success than assumed by ideal population.

Assumed by ideal population



Approximately Poisson distribution

$$V = 2\left(1 - \frac{1}{N}\right)$$

European otters: Koelewijn et al. 2010

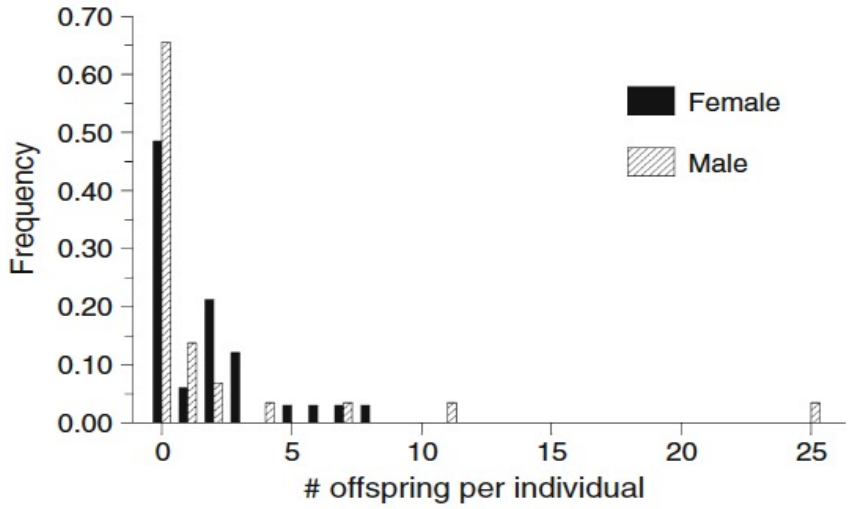
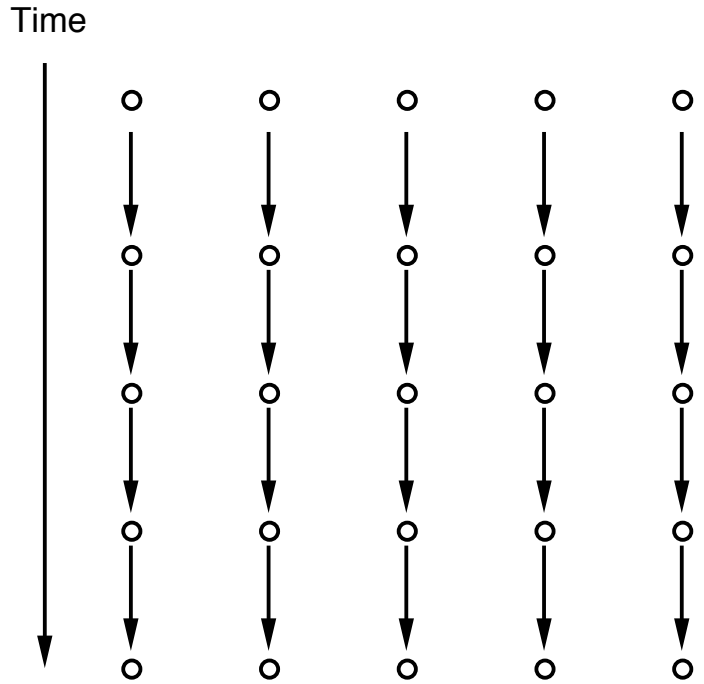


Fig. 2 Frequency distribution of the reproductive success of females and males during the period 2002–2008

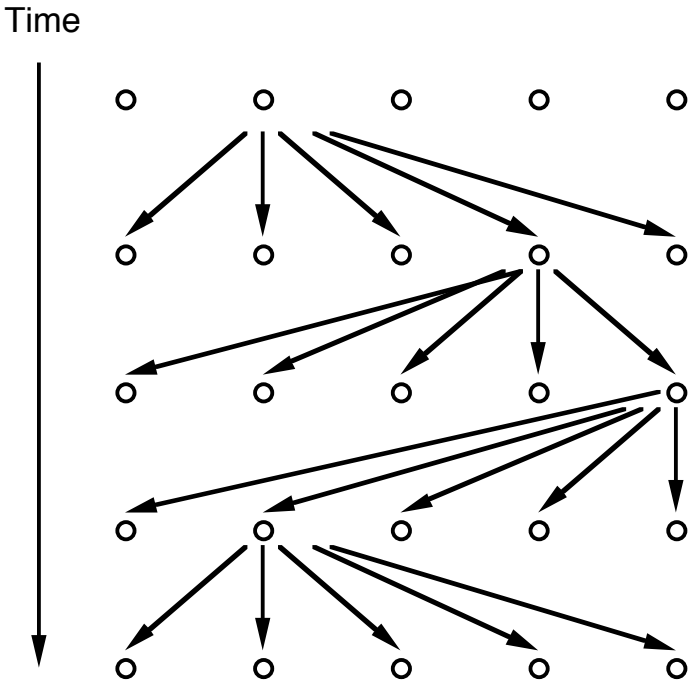
Extreme example

All alleles have exactly one offspring allele:



Other extreme

All alleles derive from one parent allele:



N_e for haploids

$$N_e = \frac{N}{V}$$

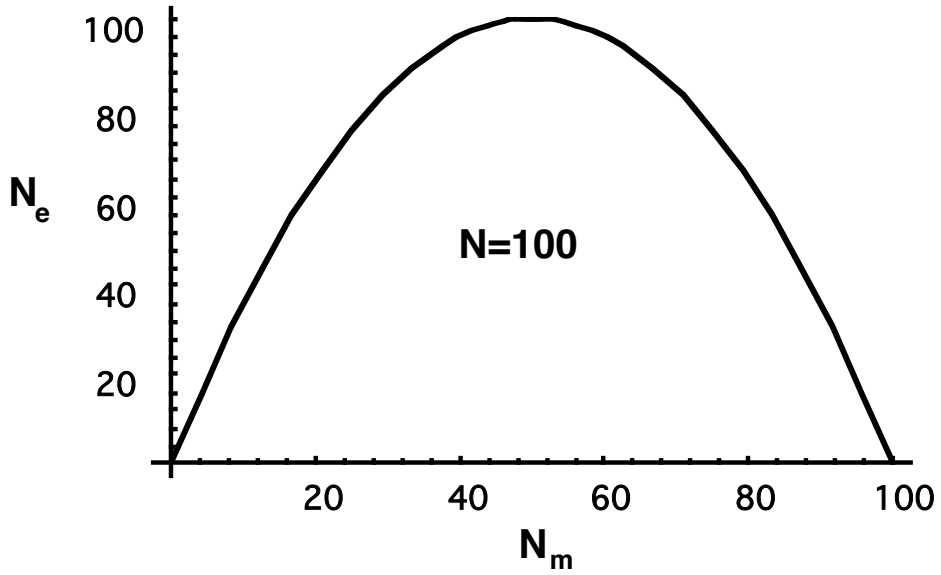
V is the variance in reproductive success

N_e for diploids

$$N_e = \frac{4N - 2}{V + 2}$$

N_e and unequal sex ratios

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



Can be derived from

$$\frac{1}{2N_e} = \frac{1}{4}\left(\frac{1}{2N_m}\right) + \frac{1}{4}\left(\frac{1}{2N_f}\right) + \frac{1}{2}(0)$$

N_e is usually much smaller than N

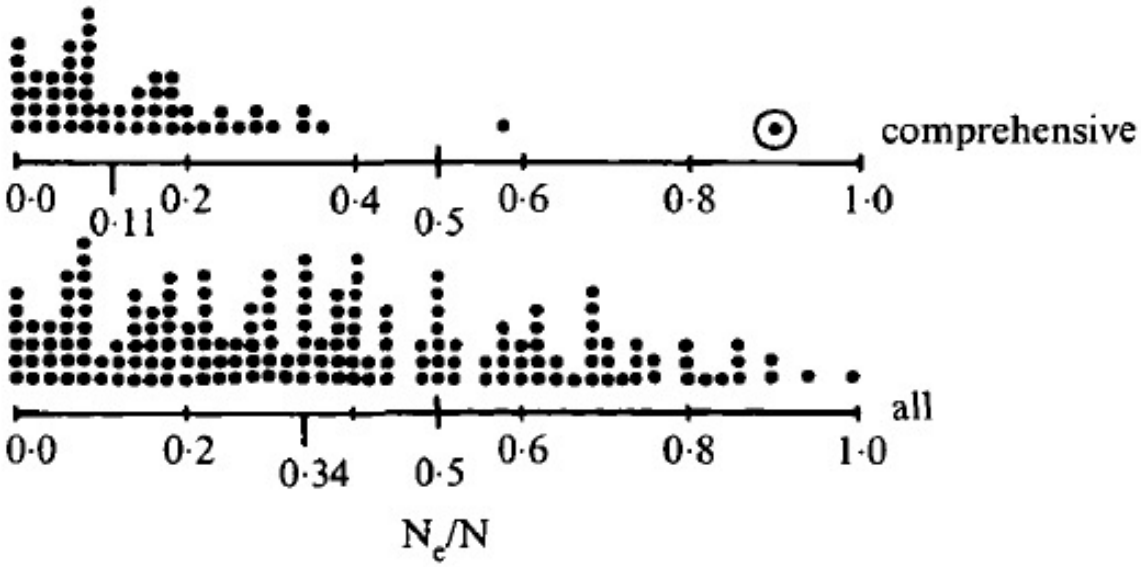
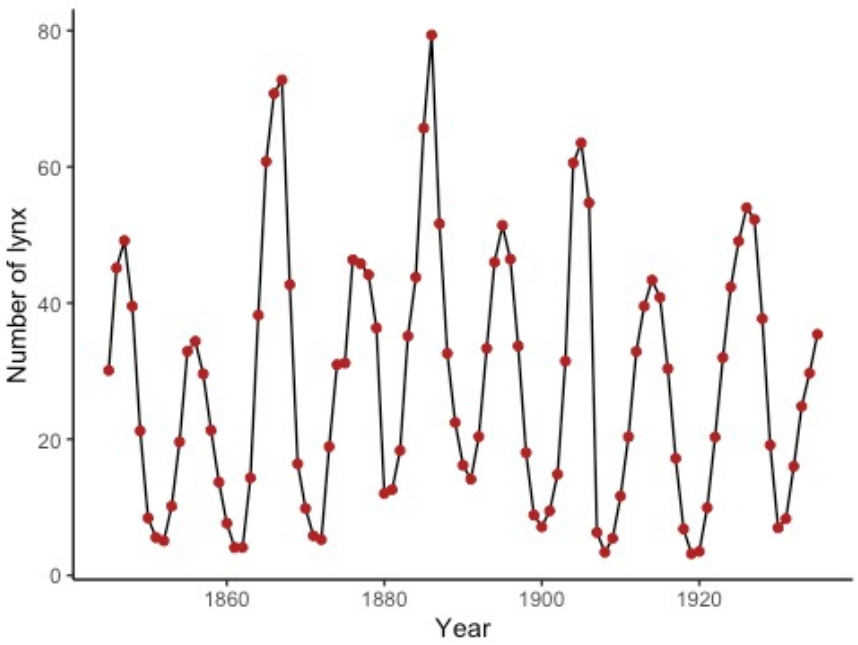


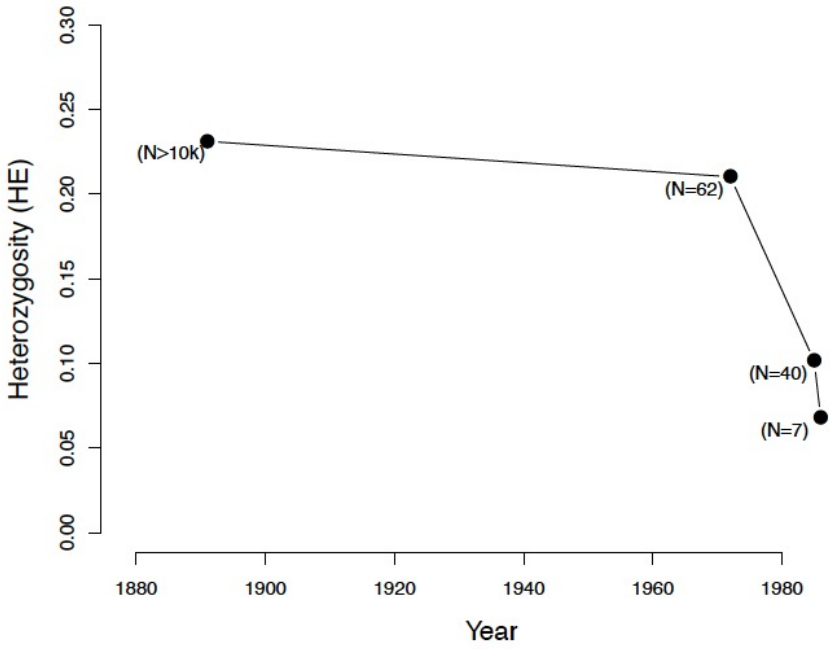
Fig. 1. Distributions of estimates of effective/actual population size (N_e/N) ratios. Comprehensive estimates (that include the effects of fluctuation in population size, variance in family size and unequal sex-ratio) are above and all estimates below. The circled outlier is for a pair mated rainbow trout (*Oncorhynchus mykiss*) population. Means of estimates are indicated below vertical lines.

Frankham 1995

If the variance in reproductive success among diploid individuals is 6, and there are 40 individuals per population per generation, what should the genetic variance within populations be after 30 generations (relative to the starting genetic variance)?

Variance in population size over time





Canadian lynx

Black-footed ferret

Variance in population size over time

$$1 - F_t = \left(1 - \frac{1}{2N_{t-1}}\right)\left(1 - \frac{1}{2N_{t-2}}\right)\left(1 - \frac{1}{2N_{t-3}}\right) \ldots (1 - F_0)$$

To find N_e so that

$$1 - F_t = \left(1 - \frac{1}{2N_e}\right)^t (1 - F_0)$$

We use: $\widetilde{N} = \frac{1}{\frac{1}{t}\sum_i \frac{1}{N_i}}$

Harmonic mean

$\tilde{N} = \frac{1}{\frac{1}{t}\sum_i \frac{1}{N_i}}$ is the harmonic mean of *N*.

Harmonic mean is very sensitive to small values.

$N_e << N$ if N is variable over time

For example:

$$N_1 = 1{,}000{,}000{,}000$$

$$N_2 = 2$$

$$N_3 = 1{,}000{,}000{,}000$$

$$\overline{N} = \frac{2{,}000{,}000{,}002}{3} = 666{,}666{,}667.3$$

$$\tilde{N} = 5.999999976 \cong 6$$

$$\tilde{N} << \overline{N}$$

Using N_e: N_e describes drift, not frequency

$$V_{G,t} = \left(1 - \frac{1}{2N_e}\right)^t V_{G,0} \qquad H_t = \left(1 - \frac{1}{2N_e}\right)^t H_0$$

$$F_t = 1 - \left(1 - \frac{1}{2N_e}\right)^t (1 - F_0)$$

$$F' = \frac{1}{2N_e} - \left(1 - \frac{1}{2N_e}\right) F$$

Expected time to coalescence of all *n* alleles = $4N_e \frac{n-1}{n}$.

Mean time to MRCA of two alleles is $2N_e$, with variance $4N_e^2$.

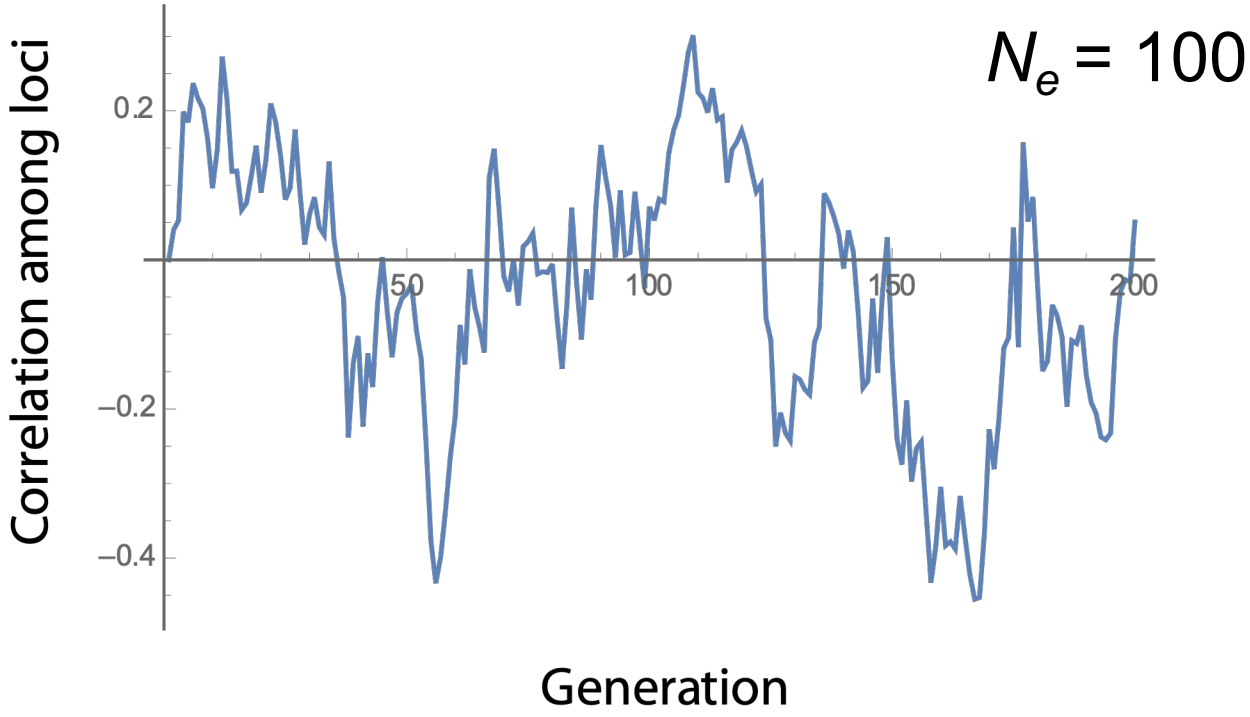
Initial frequency of new mutation: 1/2N.

Pr[fixation new neutral mutation] =1/2N.

Drift and linkage disequilibrium

Drift can generate linkage disequilibrium.

On average drift does not lead to LD, but in any given population drift may randomly create an association between alleles.



Drift and linkage disequilibrium

Drift can generate linkage disequilibrium.

$$\text{Correlation}^2 = \frac{D^2}{p_A p_a p_B p_b} = \frac{1}{1 + 4N_e r}$$

Gives effective method to estimate N_e.