Introduction to population genetics

Frédéric Guillaume

University of Zürich – IEU

SystemsX 2015

Population Genetics 1

Outline

Introduction to Population Genetics:

- Hardy-Weinberg law
- Wright-Fisher model of genetic drift
- Effective population size
- The coefficient of inbreeding and loss of genetic diversity

Population Genetics

What is Population Genetics?

What does it predict?

Population Genetics

Population Genetics predicts:

- the distribution of allele frequencies and
- the changes in allele frequencies

in a population.

Population Genetics

What are the major processes affecting allele frequencies in a population?

Population Genetics

Major processes are:

- selection,
- mutation,
- drift,
- migration,
- and the mating system.

The Hardy-Weinberg law

Hardy-Weinberg law

The Hardy-Weinberg law predicts the genotype proportions in a population under the assumption of:

- no mutation,
- no selection,
- no migration,
- no genetic drift (infinitely large population),
- random mating,

that is, under the assumption of **no changes in allele frequencies**.

Hardy-Weinberg genotype frequencies

Consider a population of N individuals mating at random and producing N offspring for the next generation.

Consider one locus with two alleles A and a that segregate in that population.

Before reproduction, each adult produces a infinite number of gametes following Mendelian segregation.

Allele frequencies in the gamete pool:

$$\text{A: } p = \frac{N_A}{N_A+N_a}; \qquad \text{a: } q = \frac{N_a}{N_A+N_a} = 1 - p; \qquad p + q = 1$$

p = proportion of A alleles in the gamete pool

q = proportion of a alleles in the gamete pool

Hardy-Weinberg genotype frequencies

$$\text{A: } p = \frac{N_A}{N_A+N_a}; \qquad \text{a: } q = \frac{N_a}{N_A+N_a} = 1 - p; \qquad p + q = 1$$

Offspring are produced by randomly drawing **2** gametes from the pool.

What is the probability that an offspring is:
AA:
Aa:
aa:

Hardy-Weinberg genotype frequencies

Allele frequencies from genotype frequencies:

Genotype	AA	Aa	aa
Frequency	x_{11}	x_{12}	x_{22}

$$x_{11} + x_{12} + x_{22} = 1$$

Hardy-Weinberg genotype frequencies

Allele frequencies from genotype frequencies:

Genotype	AA	Aa	aa
Frequency	x_{11}	x_{12}	x_{22}

$$x_{11} + x_{12} + x_{22} = 1$$

$$p = x_{11} + x_{12} \times \frac{1}{2} \quad \left(= \frac{2 \times N_{AA}}{2 \times N} + \frac{N_{Aa}}{2 \times N} \right)$$

$$q = x_{22} + x_{12} \times \frac{1}{2} \quad \left(= \frac{2 \times N_{aa}}{2 \times N} + \frac{N_{Aa}}{2 \times N} \right)$$

Hardy-Weinberg genotype frequencies

Allele frequencies from genotype frequencies:

Genotype	AA	Aa	aa
Frequency	x_{11}	x_{12}	x_{22}
	p^2	$2pq$	q^2

$$x_{11} + x_{12} + x_{22} = p^2 + 2pq + q^2 = 1$$

Hardy-Weinberg genotype frequencies

Allele frequencies from genotype frequencies:

Genotype	AA	Aa	aa
Frequency	x_{11}	x_{12}	x_{22}
	p^2	$2pq$	q^2

$$x_{11} + x_{12} + x_{22} = p^2 + 2pq + q^2 = 1$$

$$p = x_{11} + x_{12} \times \frac{1}{2} = p^2 + pq$$

$$q = x_{22} + x_{12} \times \frac{1}{2} = q^2 + pq$$

(remember that $q = 1 - p$)

Hardy-Weinberg law

Assumptions:

1. allele frequencies do not differ between males and females (or individuals are hermaphrodites)
2. individuals mate at random
3. generations are non-overlapping
4. meiosis is *fair* or equilibrated, there is no segregation distortion
5. population size is infinite, and frequency of matings is as expected from allele frequencies
6. individuals produce the same number of offspring, on average
7. offspring have the same probability of survival (no selection)
8. there is no new genetic material (no mutation, no migration)

Hardy-Weinberg law

Under these assumptions:

- H-W Law insures **constancy** of allele and genotype frequencies and thus **preservation** of genetic variation,
- genotype frequencies will reach H-W equilibrium **in one generation** (or two if unequal allele frequencies between males and females),
- any departure from H-W equilibrium means that at least one assumption **is violated**.

Hardy-Weinberg law

Under these assumptions:

- H-W Law insures **constancy** of allele and genotype frequencies and thus **preservation** of genetic variation,
- genotype frequencies will reach H-W equilibrium **in one generation** (or two if unequal allele frequencies between males and females),
- any departure from H-W equilibrium means that at least one assumption **is violated**.

H-W law is a "**zero force law**" of population genetics.

H-W law is a "**zero force law**" of population genetics.

H-W law describes what happens in absence of any *evolutionary* force. It can be considered as the **First Law of Population Genetics**.

Departure from H-W genotypic proportions is **sufficient** to infer that some force is acting on the population.

It is, however, *not necessary*, as the genotypic proportions and allele frequencies may be conserved when several forces are opposing each other (e.g. under mutation-selection balance).

Hardy-Weinberg law

We next consider two causes of departure from H-W equilibrium:

- finite population sizes (random genetic drift),
- mutation.

Genetic Drift

Reproduction, and the union of gametes, is a **random process**.

To better grasp the concept of genetic drift, it is very useful to interpret allele and genotype frequencies as *probabilities*.

Genetic Drift

Reproduction, and the union of gametes, is a **random process**.

To better grasp the concept of genetic drift, it is very useful to interpret allele and genotype frequencies as *probabilities*.

Zygotes are created, in a randomly mating population, by drawing gametes at random from the gamete pool. The allele frequency p in the gamete pool is then equivalent to the probability $P(x = A)$ of drawing an A allele at random (with replacement):

$$P(x = A) = p$$

$$P(x = a) = q$$

$$P(x = a) = 1 - P(x = A)$$

Genetic Drift

Because reproduction is a **random process**:

- the number of alleles in the next generation cannot be predicted exactly. However, we know the probability of drawing N'_A A alleles out of $2N$ alleles with probability p. This is given by the **Binomial distribution**,
- if, say, $N_A = 8$ and $2N = 40$, the Binomial distribution gives the probabilities for each $N'_A \in \{0, 40\}$ in the next generation,
- each repetition of the process will give slightly different results for N'_A.

Genetic Drift

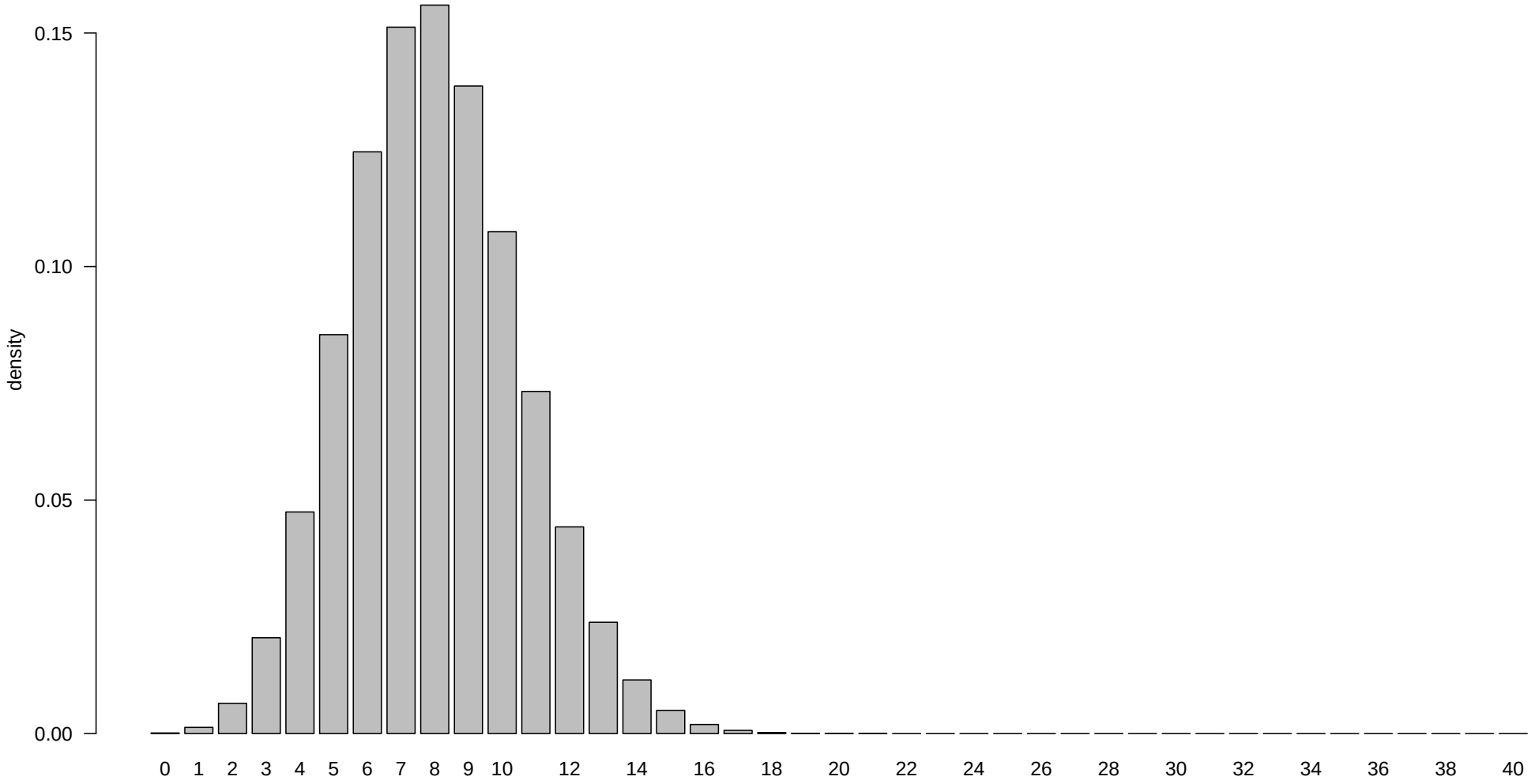
Because reproduction is a **random process**:

- the number of alleles in the next generation cannot be predicted exactly. However, we know the probability of drawing N'_A A alleles out of $2N$ alleles with probability p. This is given by the **Binomial distribution**,
- if, say, $N_A = 8$ and $2N = 40$, the Binomial distribution gives the probabilities for each $N'_A \in \{0, 40\}$ in the next generation,
- each repetition of the process will give slightly different results for N'_A.

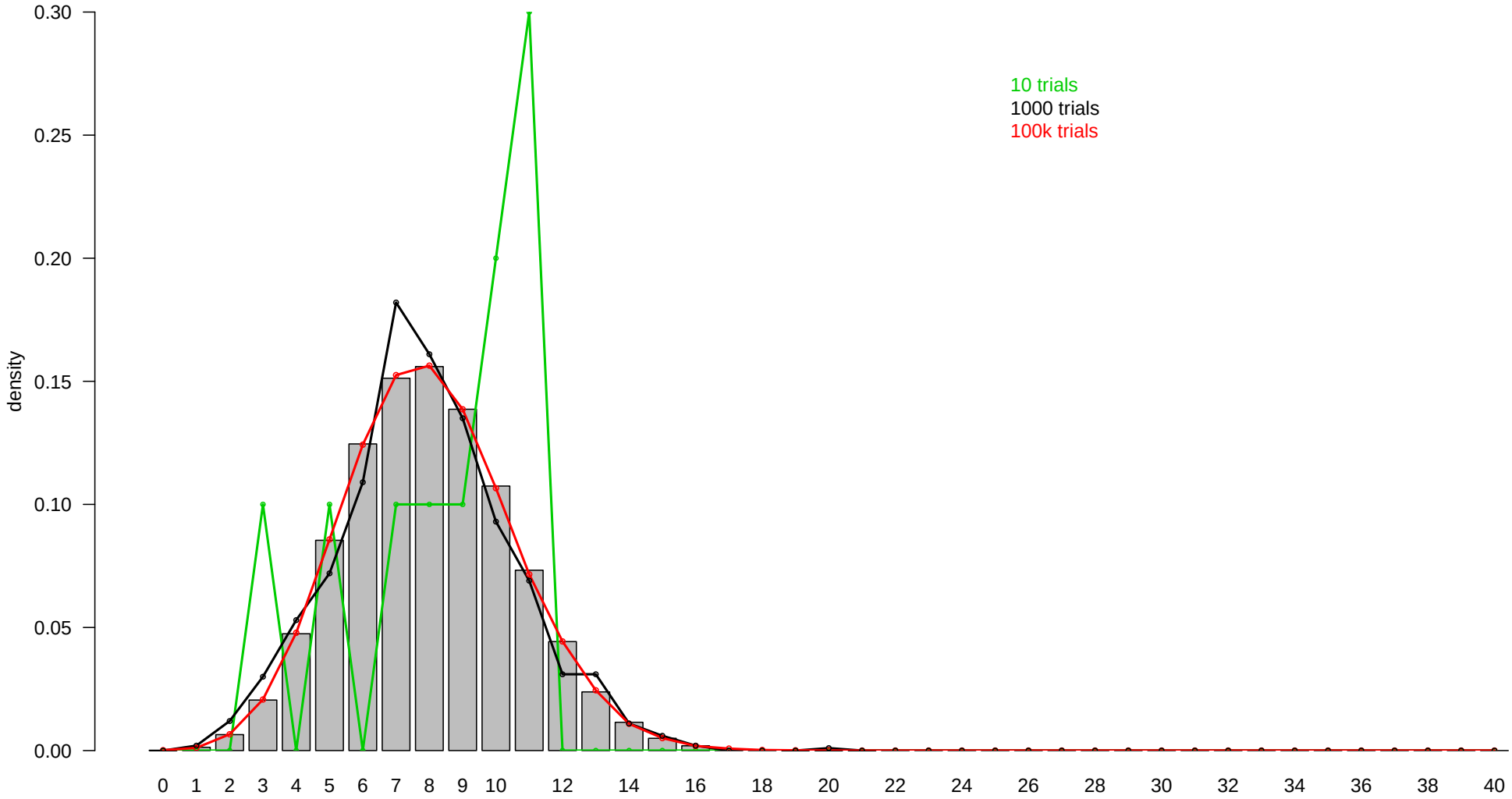
Binomial Distribution

$$\mathrm{P}\{j \text{ alleles A}\} = \binom{2N}{j} p^j q^{2N-j} = \frac{(2N)!}{j!(2N-j)!} p^j q^{2N-j}$$

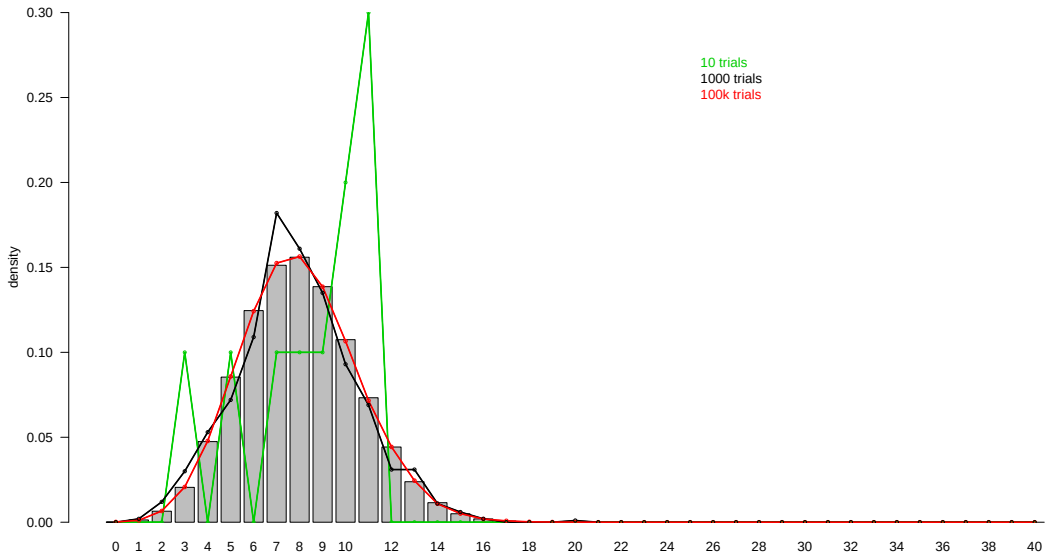
For the example above with $p = 8/40$ and $2N = 40$, the *density distribution* of the Binomial distribution is:



The Binomial density distribution is our *expectation* over many trials:

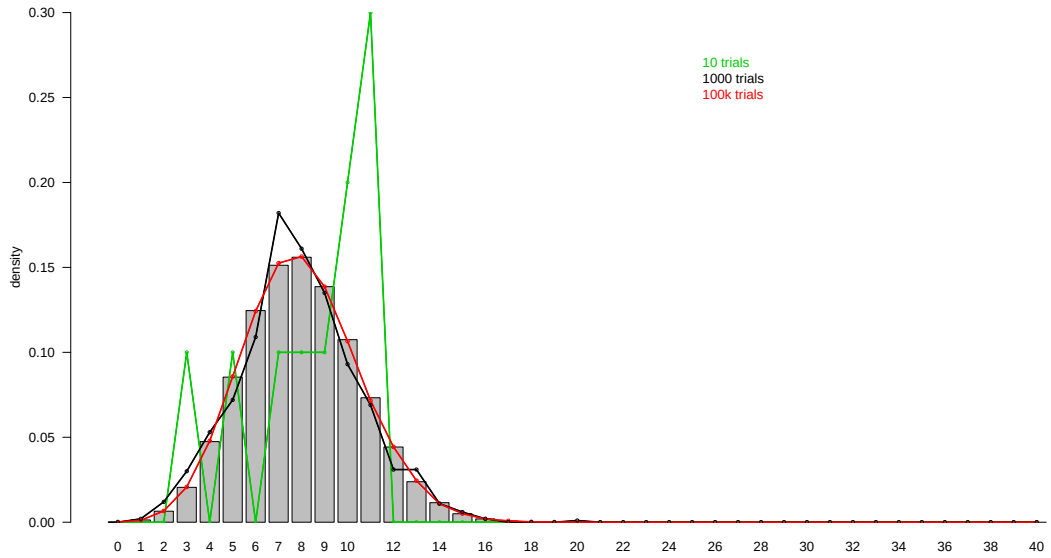


One trial: draw an allele $2N$ times with $\mathrm{P}(x = A) = p$, and count the number of A's you get. This is similar to doing one generation of random mating in a diploid population of size N.



The reason why the green and black lines do not match well with the underlying distribution is because of **sampling error**. The sampling error of the Binomial is equal to its **variance**:

$$\mathrm{Var}\{j\} = 2Npq$$



The reason why the green and black lines do not match well with the underlying distribution is because of **sampling error**. The sampling error of the Binomial is equal to its **variance**:

$$\mathrm{Var}\{j\} = 2Npq$$

from this variance, we can derive the **sampling variance** of the allele frequency:

$$\mathrm{Var}\{p'\} = \mathrm{Var}\left\{\frac{j}{2N}\right\} = \frac{\mathrm{Var}\{j\}}{4N^2} = \frac{pq}{2N}$$

Wright-Fisher model

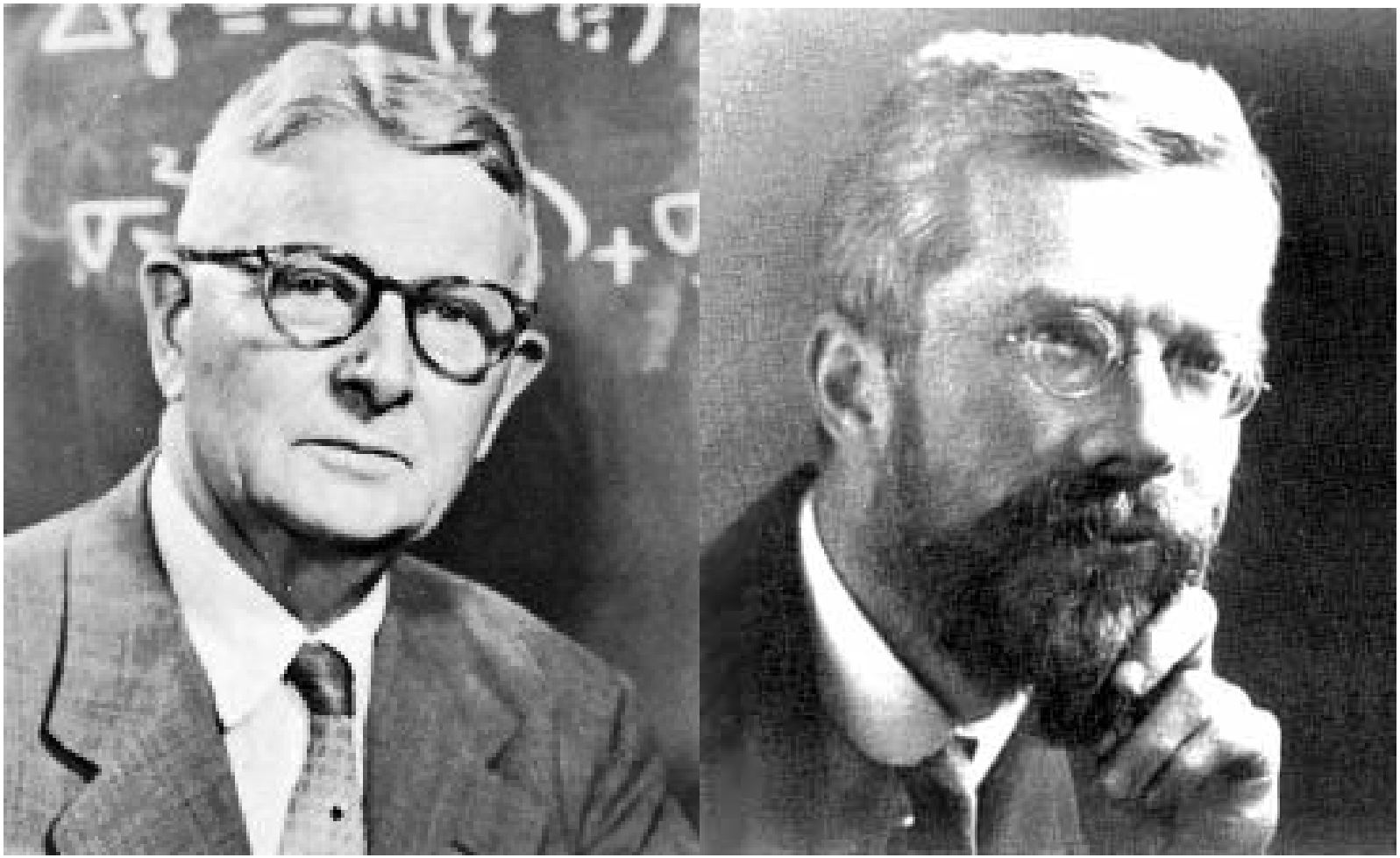


Figure 1: Sewall Wright (1889–1988) and Ronald A. Fisher (1890–1962)

Wright-Fisher model

Random sampling of N offspring (N zygotes from 2N gametes) from N parents, with replacement, in a population with random mating.

Assumptions:

- size N is constant, and **not infinite**
- random-mating
- non-overlapping generations
- one locus – or free recombination

Wright-Fisher model

Under Wright-Fisher model assumptions, the **sampling variance** of the allele frequencies in the *next generation* is that of the Binomial sampling process:

$$\mathrm{Var}\{p'\} = \sigma_p^2 = \frac{pq}{2N}.$$

Wright-Fisher model

Under Wright-Fisher model assumptions, the **sampling variance** of the allele frequencies in the *next generation* is that of the Binomial sampling process:

$$\mathrm{Var}\{p'\} = \sigma_p^2 = \frac{pq}{2N}.$$

Thus, we expect more stochasticity in allele frequencies over time in small populations than in large populations.

Wright-Fisher model

Under Wright-Fisher model assumptions, the **sampling variance** of the allele frequencies in the *next generation* is that of the Binomial sampling process:

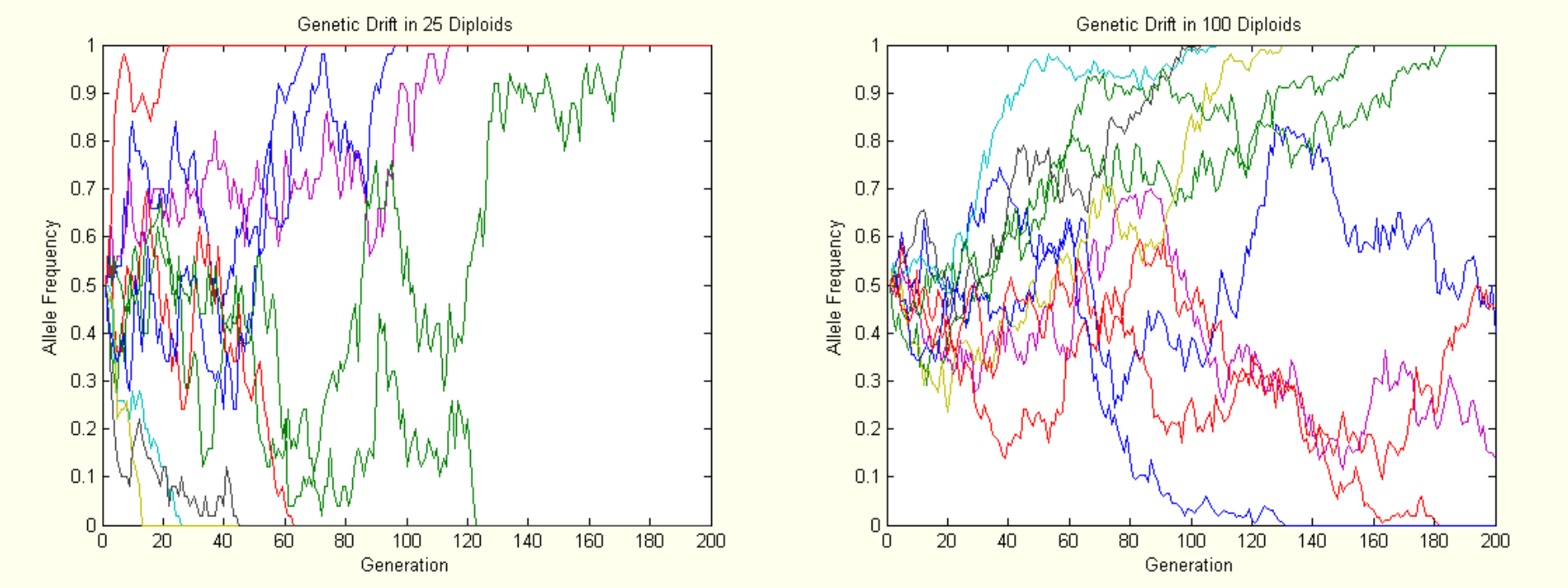
$$\mathrm{Var}\{p'\} = \sigma_p^2 = \frac{pq}{2N}.$$

Thus, we expect more stochasticity in allele frequencies over time in small populations than in large populations.

The **intensity** of the process is **stronger** in **small** populations.

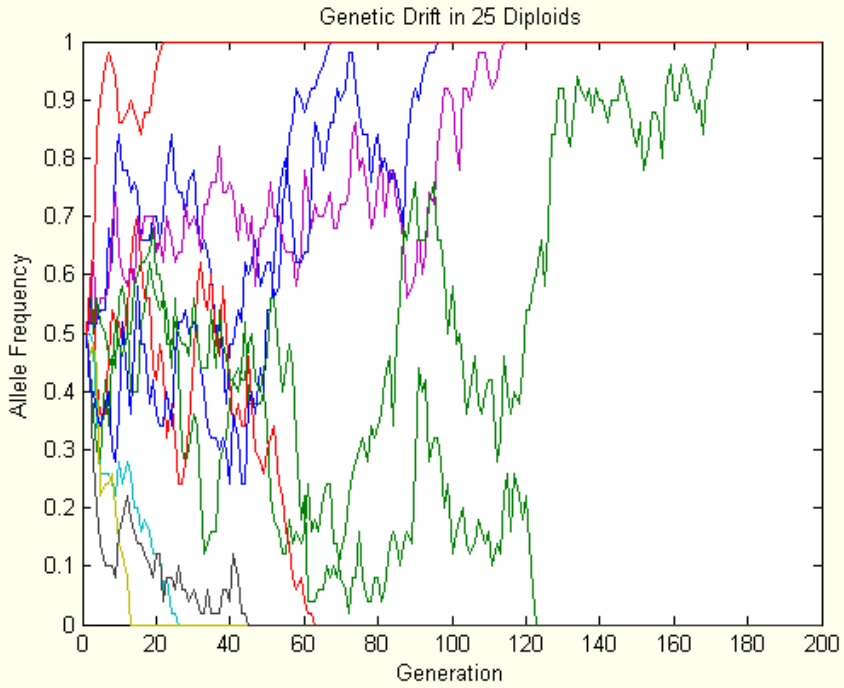
Genetic Drift

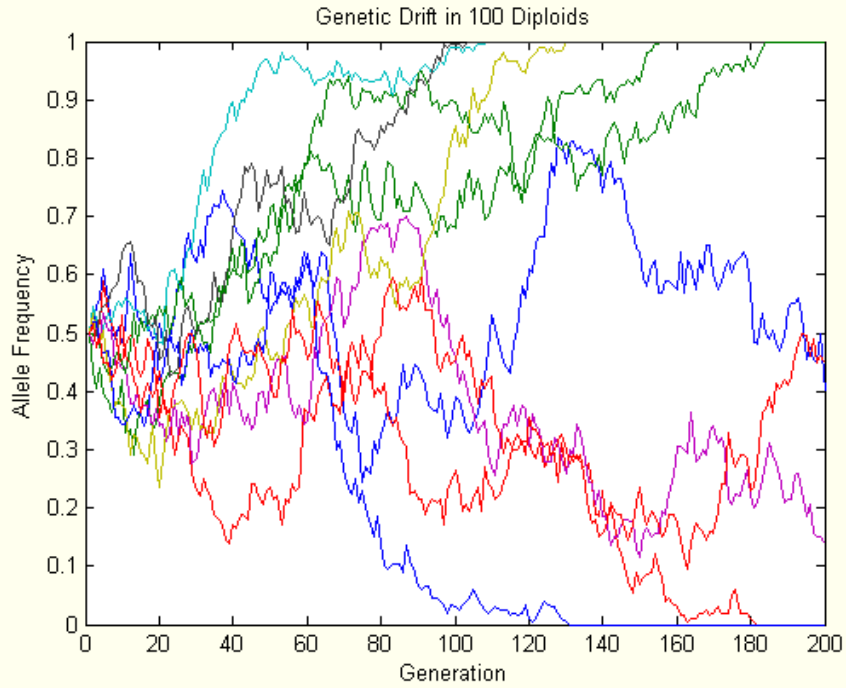
Simulating the evolution of allele frequencies over time in two small populations (N = 25 and N = 100):



each line on the graph is a separate experiment (e.g., a separate population or experimental line kept in the lab and reared for 200 generations)

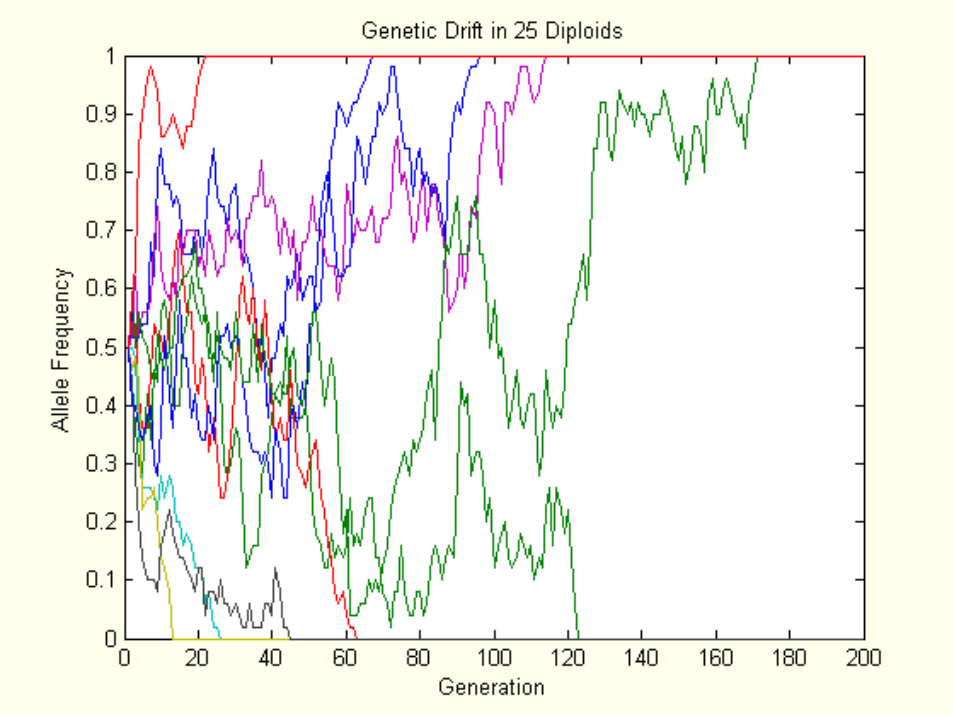
Genetic Drift





Allele frequencies are more stochastic when **N** is small, and eventually reach **loss** ($p = 0$) or **fixation** ($p = 1$) of one of the alleles.
This random process is called **genetic drift**.
Its magnitude depends on population size (i.e. sampling variance).

Genetic Drift in the Wright-Fisher model



Properties in one population

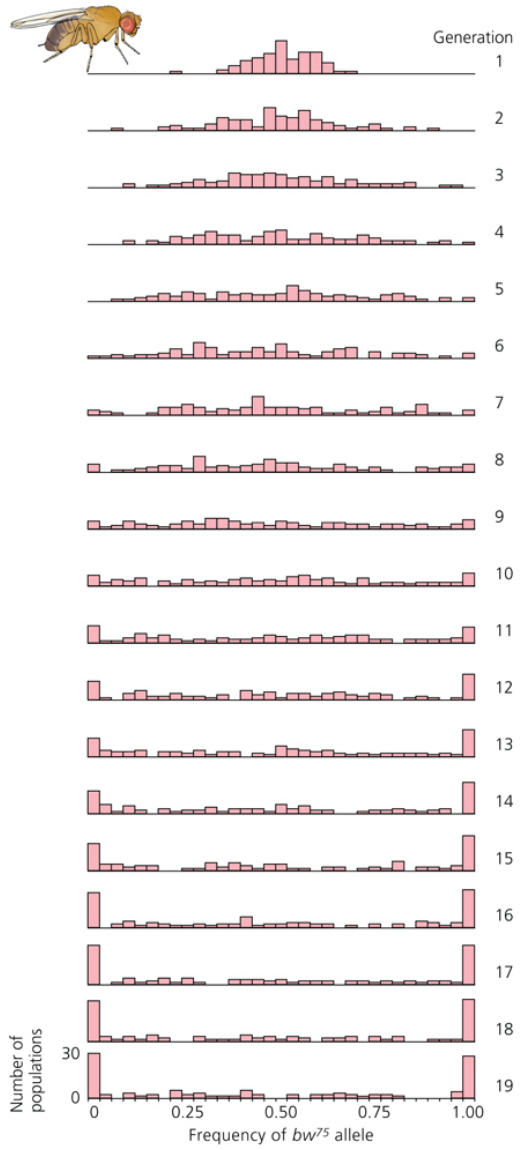
- The exact allele frequency at each time step t cannot be predicted.

- There are two **equilibria**: $(p = 0)$ and $(p = 1)$.

- The **probability of fixation** of an allele is its frequency p $(= \frac{1}{2N}$ for a new mutation).

- The **expected time to fixation** $(p = 1)$ is $\propto N$ and $\propto \frac{1}{p}$, and is $\approx 4N$ for a new mutation (with $p = \frac{1}{2N}$).

Genetic Drift in the Wright-Fisher model



Properties in many populations

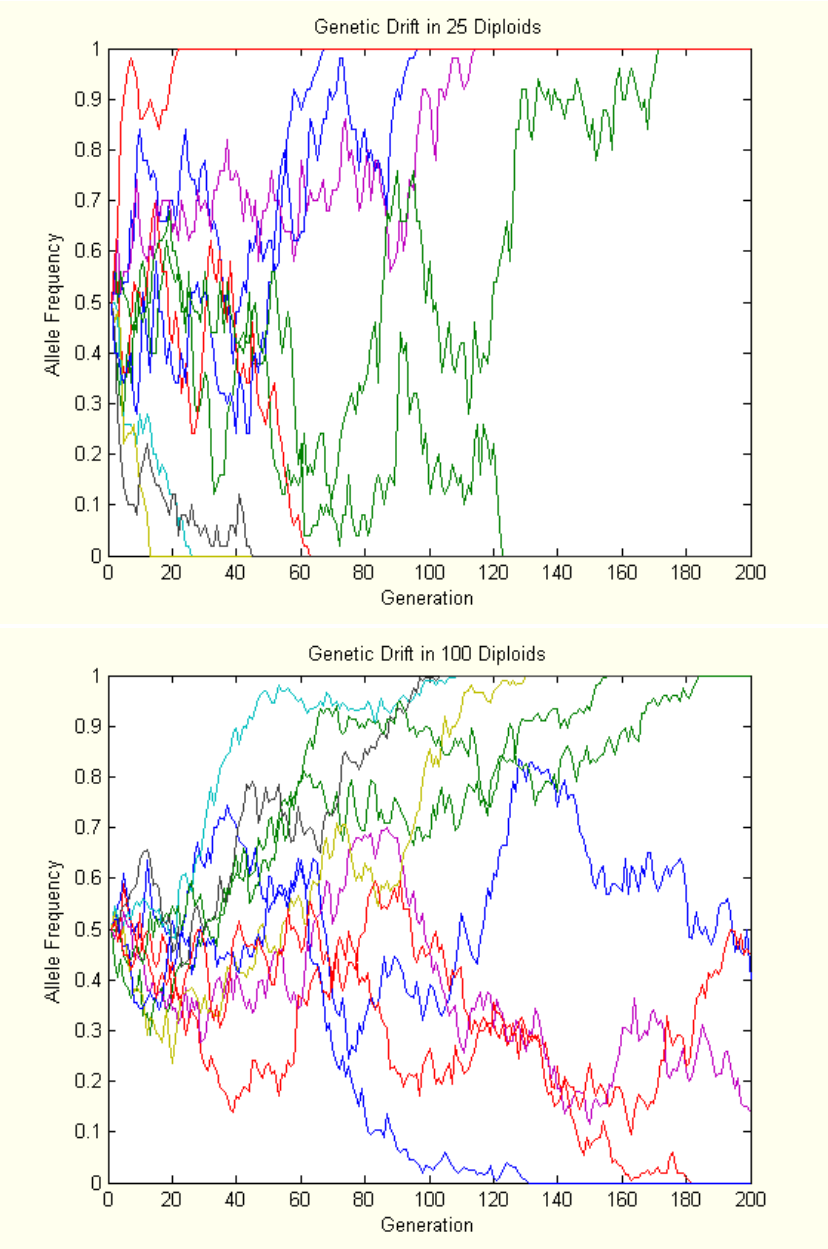
• The mean allele frequency **among replicates** (lines) is equal to the initial allele frequency:

$$\overline{p}_{equ} = p_0.$$

• The variance in allele frequencies among lines increases with time (by the addition of the Binomial sampling variance), and is:

$$\sigma^2_{p_t} = p_0 q_0 \left[1 - \left(1 - \frac{1}{2N}\right)^t\right]$$

Genetic Drift in the Wright-Fisher model



Consequences:

- Within-population genetic variance will erode with time.

- Among-population genetic variance will increase with time.

$\longrightarrow$ Drift is a dispersive process, it increases genetic differentiation of populations.

$\longrightarrow$ The magnitude of drift is the sampling variance $\sigma^2 = \frac{pq}{2N}$.

$\longrightarrow$ Drift causes a decay of **heterozygosity**.

Effective population size N_e

No real population meet the assumptions of the Wright-Fisher model. Is the theory then still valid outside of this abstract model?

Fortunately, the answer is YES.

The relationship between population size and strength of drift allows us to define the concept of an **effective population size,** N_e.

Effective population size N_e

No real population meet the assumptions of the Wright-Fisher model. Is the theory then still valid outside of this abstract model?

Fortunately, the answer is YES.

The relationship between population size and strength of drift allows us to define the concept of an **effective population size,** N_e.

N_e: Size of an **ideal** population having the same **amount of drift** as the observed population of size N.

Effective population size N_e

No real population meet the assumptions of the Wright-Fisher model. Is the theory then still valid outside of this abstract model?

Fortunately, the answer is YES.

The relationship between population size and strength of drift allows us to define the concept of an **effective population size,** N_e.

N_e: Size of an **ideal** population having the same **amount of drift** as the observed population of size N.

The **amount of drift** is given by the **sampling variance** in the Wright-Fisher model:

$$\sigma_p^2 = \frac{pq}{2N}.$$

Effective population size N_e

N_e: Size of an **ideal** population having the same **amount of drift** as the observed population of size N.

The **amount of drift** is given by the **sampling variance** in the Wright-Fisher model:

$$\sigma_p^2 = \frac{pq}{2N}.$$

The **effective population size** is then defined as:

$$N_e = \frac{pq}{2\sigma_{obs}^2}$$

Effective population size N_e

N_e helps us convert our real population into an ideal (panmictic) population where we can precisely describe the genetic process of drift. It suffices, for that, to substitute N_e for N in all above formulas.

Effective population size N_e

N_e helps us convert our real population into an ideal (panmictic) population where we can precisely describe the genetic process of drift. It suffices, for that, to substitute N_e for N in all above formulas.

In most cases, real populations have $N_e < N$, because:

Effective population size N_e

N_e helps us convert our real population into an ideal (panmictic) population where we can precisely describe the genetic process of drift. It suffices, for that, to substitute N_e for N in all above formulas.

In most cases, real populations have $N_e < N$, because:

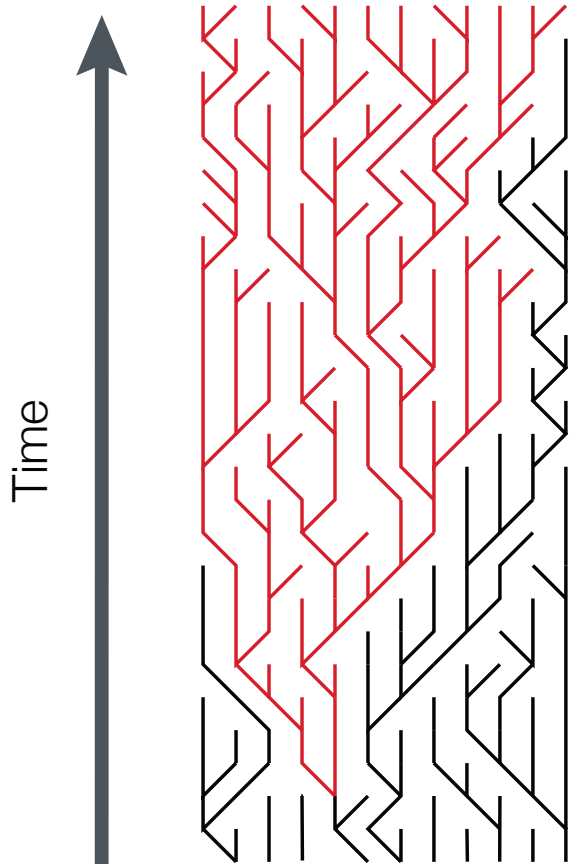
- N varies in time
- $N_m \neq N_f$, e.g. fewer males contribute to reproduction than females
- some individuals have a higher reproductive success than others
- the variance in reproductive number is greater than expected by chance (i.e. not Binomial)

Identity By Descent

To understand the effects of drift on genotype frequencies, we need to introduce the concept of **Identity By Descent (IBD)** and the **Coefficient of Inbreeding (F)**.

Inbreeding coefficient

The loss of genetic variation is caused by **inbreeding**, or the build-up of **Identity-by-Descent (IBD)** in the population.



Identity By Descent

Identity by descent is an inevitable consequence of finite population size.

Consider this:

• each individual in a population has 2^t ancestors at time t. A population of $N = 10000$ will be descending from $10000 \times 2^{15} \approx 3.3 \times 10^8$ individuals 15 generations ago.

Identity By Descent

Identity by descent is an inevitable consequence of finite population size.

Consider this:
• each individual in a population has 2^t ancestors at time t. A population of $N = 10000$ will be descending from $10000 \times 2^{15} \approx 3.3 \times 10^8$ individuals 15 generations ago.

• in the Wright-Fisher model, in each generation, about 37% of individuals don't leave any descendent ($P(x = 0) \approx e^{-1} \approx 0.37$). Therefore, at generation $t = 15$ there will be $N \times (0.63)^t \approx 10$ ancestors for our population.

Identity By Descent

Identity by descent is an inevitable consequence of finite population size.

Consider this:
• each individual in a population has 2^t ancestors at time t. A population of $N = 10000$ will be descending from $10000 \times 2^{15} \approx 3.3 \times 10^8$ individuals 15 generations ago.

• in the Wright-Fisher model, in each generation, about 37% of individuals don't leave any descendent ($P(x = 0) \approx e^{-1} \approx 0.37$). Therefore, at generation $t = 15$ there will be $N \times (0.63)^t \approx 10$ ancestors for our population.

$\Rightarrow$ Therefore, individuals in a population of constant size will inevitably share a large part of their ancestors and thus carry copies of the same ancestral gene. This is called **Indentity by descent**.

Inbreeding coefficient

The probability that the two alleles carried by a diploid individual at a given locus are identical by descent **(IBD)** is given by the **coefficient of inbreeding F**.

Inbreeding coefficient

The probability that the two alleles carried by a diploid individual at a given locus are identical by descent **(IBD)** is given by the **coefficient of inbreeding F**.

In a randomly mating population with 2N alleles, the probability to draw twice a copy of the same allele is $\frac{1}{2N}$ (self-fertilization is allowed).

Inbreeding coefficient

The probability that the two alleles carried by a diploid individual at a given locus are identical by descent **(IBD)** is given by the **coefficient of inbreeding F**.

In a randomly mating population with 2N alleles, the probability to draw twice a copy of the same allele is $\frac{1}{2N}$ (self-fertilization is allowed).

Therefore, with finite population sizes (1/2N not too small), random sampling of gametes will also cause changes in **genotype frequencies**.

Inbreeding coefficient

The probability that the two alleles carried by a diploid individual at a given locus are identical by descent **(IBD)** is given by the **coefficient of inbreeding F**.

In a randomly mating population with 2N alleles, the probability to draw twice a copy of the same allele is $\frac{1}{2N}$ (self-fertilization is allowed).

Therefore, with finite population sizes (1/2N not too small), random sampling of gametes will also cause changes in **genotype frequencies**.

$\Rightarrow$ **Heterozygosity** will change over time.

Effect of inbreeding on genotype frequencies

In a **large** population with **constant** allele frequencies:

	Allozygous		Autozygous
Homozygous, AA:	$p^2(1-F)$	$+$	pF
Heterozygous, Aa:	$2pq(1-F)$		
Homozygous, aa:	$q^2(1-F)$	$+$	qF
Total	$1-F$		F

Effect of inbreeding on genotype frequencies

In a **large** population with **constant** allele frequencies:

	Allozygous		Autozygous
Homozygous, AA:	$p^2(1-F)$	$+$	pF
Heterozygous, Aa:	$2pq(1-F)$		
Homozygous, aa:	$q^2(1-F)$	$+$	qF
Total	$1-F$		F

F can be measured from a **pedigree** (shown on the black board).

Effect of inbreeding on genotype frequencies

In a **large** population with **constant** allele frequencies:

	Allozygous		Autozygous
Homozygous, AA:	$p^2(1-F)$	$+$	pF
Heterozygous, Aa:	$2pq(1-F)$		
Homozygous, aa:	$q^2(1-F)$	$+$	qF
Total	$1-F$		F

F can be measured from a **pedigree** (shown on the black board).

$F \neq 0$ is also a consequence of random mating in small populations.

Inbreeding coefficient

In a **finite** population, with **random mating**:

Coefficient of inbreeding at time t:

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

Inbreeding coefficient

In a **finite** population, with **random mating**:

Coefficient of inbreeding at time t:

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

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

Inbreeding coefficient

In a **finite** population, with **random mating**:

Coefficient of inbreeding at time t:

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

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

Inbreeding increases by $\frac{1}{2N}$ every generation.

Inbreeding coefficient and loss of genetic diversity

Coefficient of inbreeding at time t:

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

Heterozygosity at time t (recall formula on p38):

$$H_1 = 2p_0q_0(1 - F_0)$$
$$H_1 = H_0(1 - F_0)$$
$$H_1 = H_0\left(1 - \frac{1}{2N}\right)$$

Inbreeding coefficient and loss of genetic diversity

Coefficient of inbreeding at time t:

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

Heterozygosity at time t (recall formula on p38):

$$H_1 = 2p_0q_0(1 - F_0)$$
$$H_1 = H_0(1 - F_0)$$
$$H_1 = H_0\left(1 - \frac{1}{2N}\right)$$

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

Inbreeding coefficient and loss of genetic diversity

Coefficient of inbreeding at time t:

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

Heterozygosity at time t:

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

$\Rightarrow$ Heterozygosity changes by $\Delta H = -\frac{1}{2N}$ every generation.

Formulas

- change of heterozygosity:

$$H_t = H_0 \left(1 - \frac{1}{2N_e}\right)^t = H_0 \times e^{-t/2N_e}$$

- time to halve the heterozygosity:

$$t_{1/2} = \frac{-\ln(2)}{\ln(1 - 1/2N_e)} \approx 2N_e \ln(2)$$

- estimating the inbreeding coefficient:

$$F = \frac{H_{expected} - H_{observed}}{H_{expected}}, \quad \text{with } H_{expected} = 2pq$$

Drift and Mutation

Let's turn now to the effect of mutation on genetic variation. We are here concerned with **neutral mutations**, that is, mutations unseen to selection. Such mutations are, for instance, *silent mutations* that do not change the amino-acid composition of the gene product.

Although mutations are rare events ($\approx 10^{-8}$/nucleotide in Human), they are frequent enough to maintain polymorphism in the face of genetic drift under certain conditions.

The **mutation-drift** equilibrium is reached when the loss of genetic variation by genetic drift is exactly compensated by the creation of variation by mutation.

Drift and Mutation

If μ is the mutation rate of $A \leftrightarrow a$, the probability of identity F becomes:

$$F_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}\right] (1-\mu)^2$$

Drift and Mutation

If μ is the mutation rate of $A \leftrightarrow a$, the probability of identity F becomes:

$$F_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}\right] (1-\mu)^2$$

When equilibrium is reached, $F_t = F_{t-1}$:

$$F \approx \frac{1}{4N_e\mu + 1} + O(\mu^2),$$

and the equilibrium heterozygosity ($H = 1 - F$) is:

$$H = \frac{4N_e\mu}{4N_e\mu + 1}.$$

Mutation-Drift equilibrium

Equilibrium homozygosity

$$F \approx \frac{1}{4N_e\mu + 1}$$

If $4N_e\mu \ll 1$, then drift dominates and genetic variation is eliminated from the population.

if $4N_e\mu \gg 1$, then mutation dominates and $H = 1$.

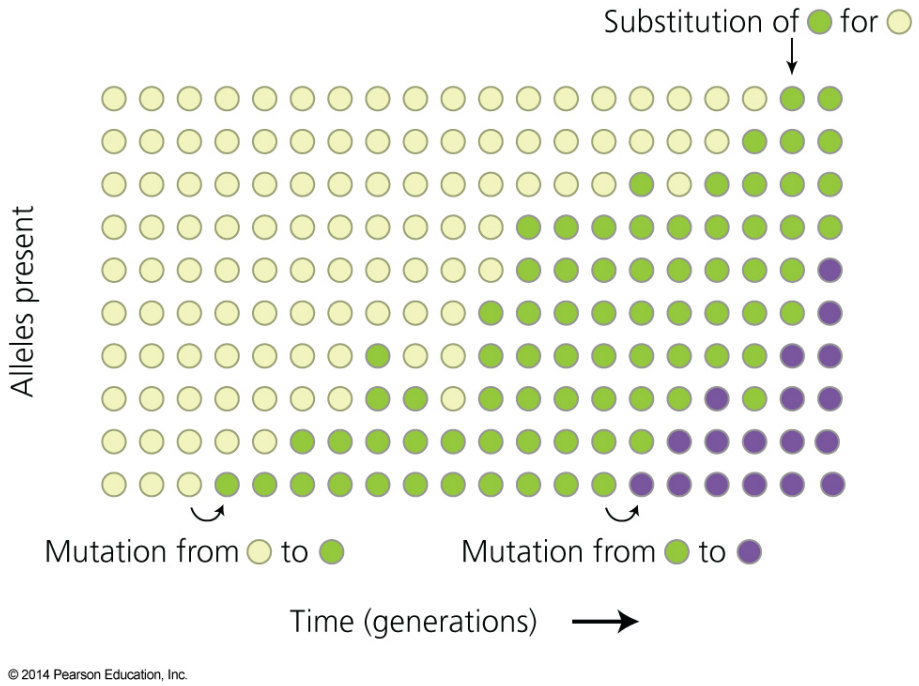
The Role of Drift in Evolution

Genetic Drift

Genetic drift matters in evolution, but how fast does it cause evolution?

Rate of evolution by genetic drift

In molecular evolution, the rate of evolution is measured by the rate of allelic substitution between lineages/species.



Substitution: replacement of an allele by **fixation** of another in a population. Substitutions are fixed differences between species at specific locations in their genome.

Rate of evolution by genetic drift

Rate of substitution of neutral mutations can be calculated as:

$$\begin{aligned}\rho &= (\#\text{ of new mutations}) \times \mathrm{P}\{\text{fixation}\} \\ &= 2N_e\mu \times \frac{1}{2N_e} \\ &= \mu\end{aligned}$$

Rate of evolution by genetic drift

Rate of substitution of neutral mutations can be calculated as:

$$\begin{aligned}\rho &= (\#\text{ of new mutations}) \times \mathrm{P}\{\text{fixation}\} \\ &= 2N_e\mu \times \frac{1}{2N_e} \\ &= \mu\end{aligned}$$

Rate of substitution of neutral alleles

$$\rho = \mu$$

The rate of substitution by genetic drift is the neutral mutation rate.

Rate of evolution by genetic drift

How important is evolution by genetic drift?

Neutral theory: (Motoo Kimura) all between-species substitutions are neutral because the fraction of advantageous mutations is so small that we can neglect them, and the majority of mutations are deleterious and do not contribute to species divergence, what remains is thus only neutral mutations.

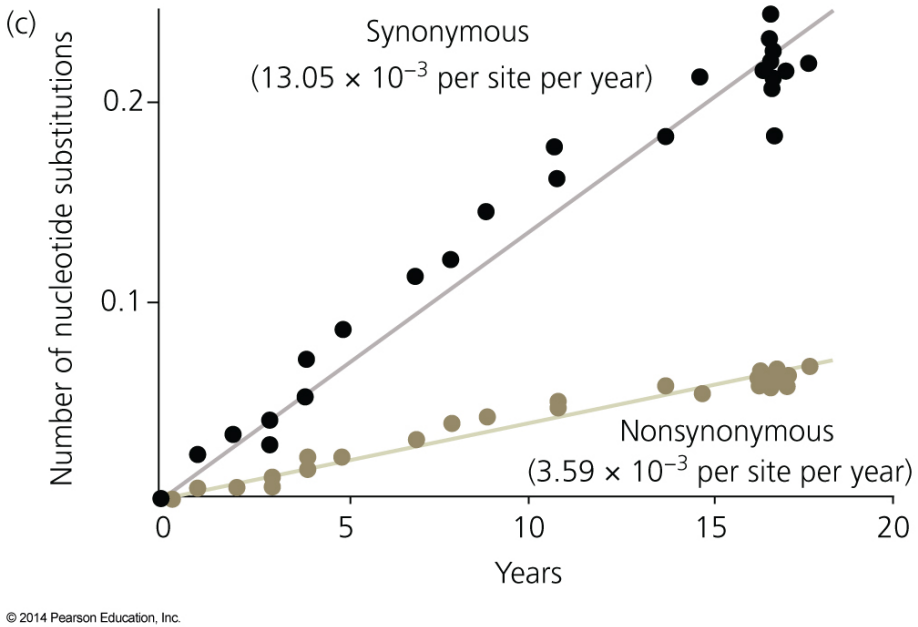
Selectionist view: (e.g. John Gillespie, Matthew Hahn) most substitutions are adaptive, and thus represent alleles fixed by natural selection instead of genetic drift.

Neutral Theory

Why the neutral theory? Is there any evidence of it?

Neutral Theory

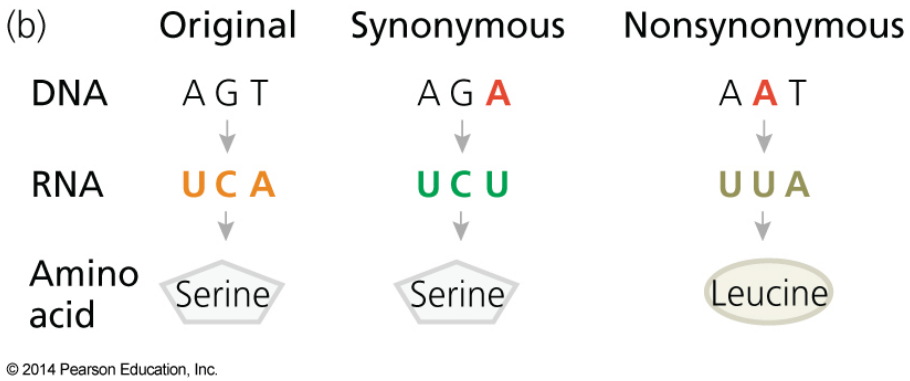
Why the neutral theory? Is there any evidence of it?



- rate of substitution is constant over time (per year)
- it is consistent across species, which have similar mutation rates
- substitutions accumulate in a clocklike fashion
- substitutions are more frequent in silent (synonymous) than non-synonymous sites of codons

Neutral Theory

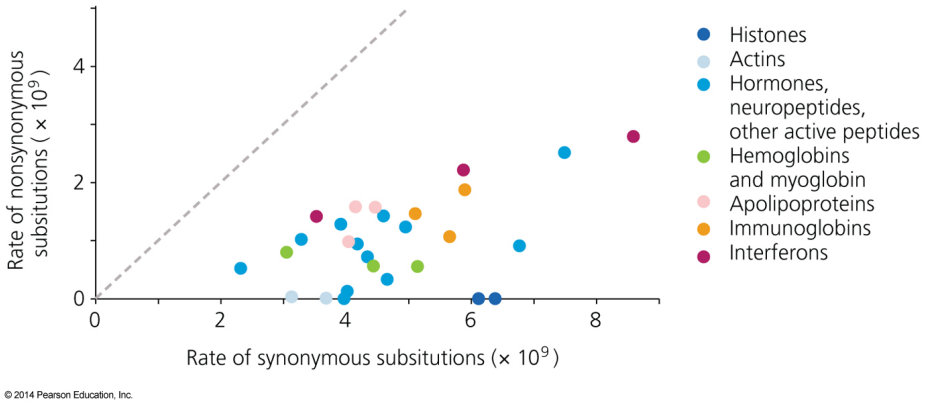
Why the neutral theory? Is there any evidence of it?

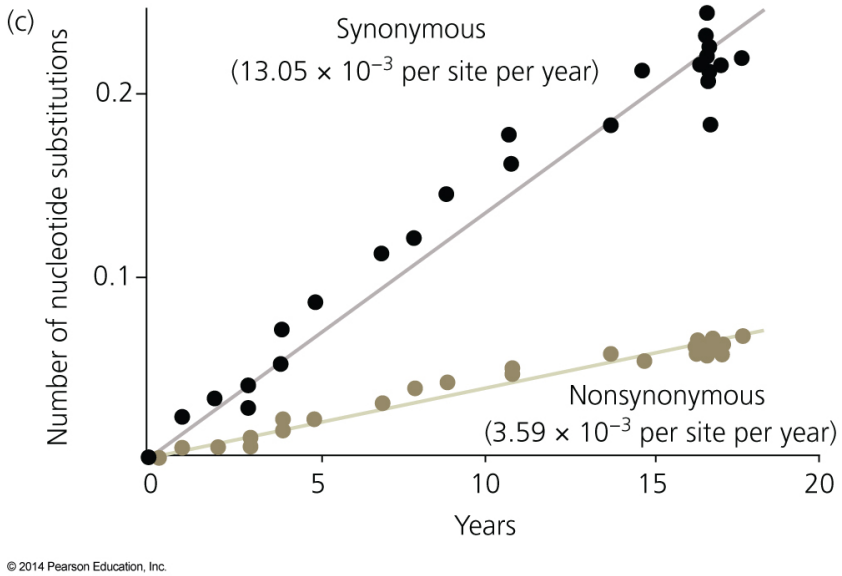


- rate of substitution is constant over time (per year)
- it is consistent across species, which have similar a.-a. mutation rates
- substitutions accumulate in a clocklike fashion
- substitutions are more frequent in silent (synonymous) than non-synonymous sites of codons

Neutral Theory

Further evidence:





- rate of synonymous substitutions always larger than non-synonymous substitutions in coding regions
- rate of substitution in pseudo-genes is highest, same order as mutation rate (e.g., $\approx 2.5 \times 10^{-8}$ between humans and chimps)
- incompatible with expectation from natural selection only (?)

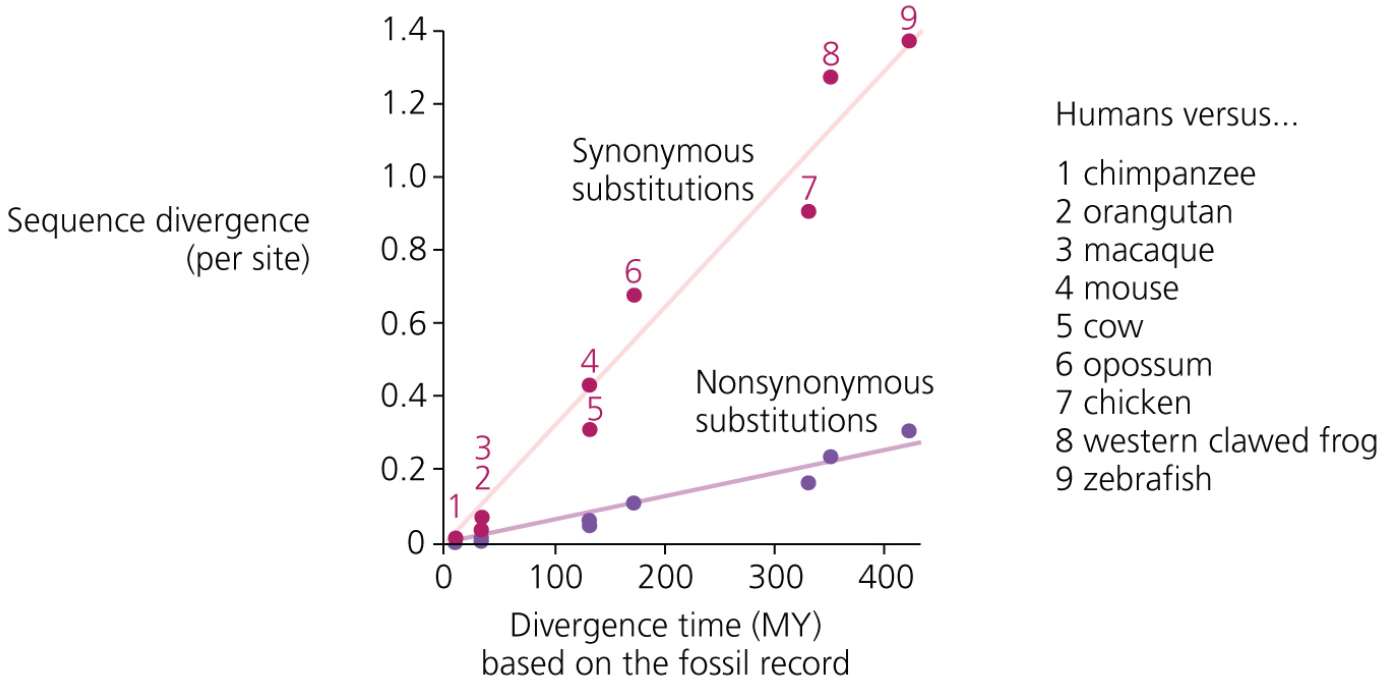
Neutral Theory

$\Rightarrow$These observations are consistent with the tenets of the Neutral Theory, that if most mutations are either deleterious or neutral, then drift dominates molecular evolution.

Neutral Theory

$\Rightarrow$These observations are consistent with the tenets of the Neutral Theory, that if most mutations are either deleterious or neutral, then drift dominates molecular evolution.

BUT, rates of substitution are calculated on a per-year basis, while the theory of genetic drift considers **generation time** as the measure of time of divergence.



Nearly-Neutral Theory

$\Rightarrow$ it is thus odd to find constant yearly rates of substitution between species that differ widely in generation time.

Nearly-Neutral Theory

Tomoko Ohta and Motoo Kimura found the answer (Otha & Kimura 1971; Otha 1972, 1977).

Nearly-Neutral Theory

Tomoko Ohta and Motoo Kimura found the answer (Otha & Kimura 1971; Otha 1972, 1977).

$\Rightarrow$ the strength of genetic drift depends on N_e, therefore, if species with small generation time have large population sizes, the two effects will compensate and rates of substitution will remain constant over time.

Nearly-Neutral Theory

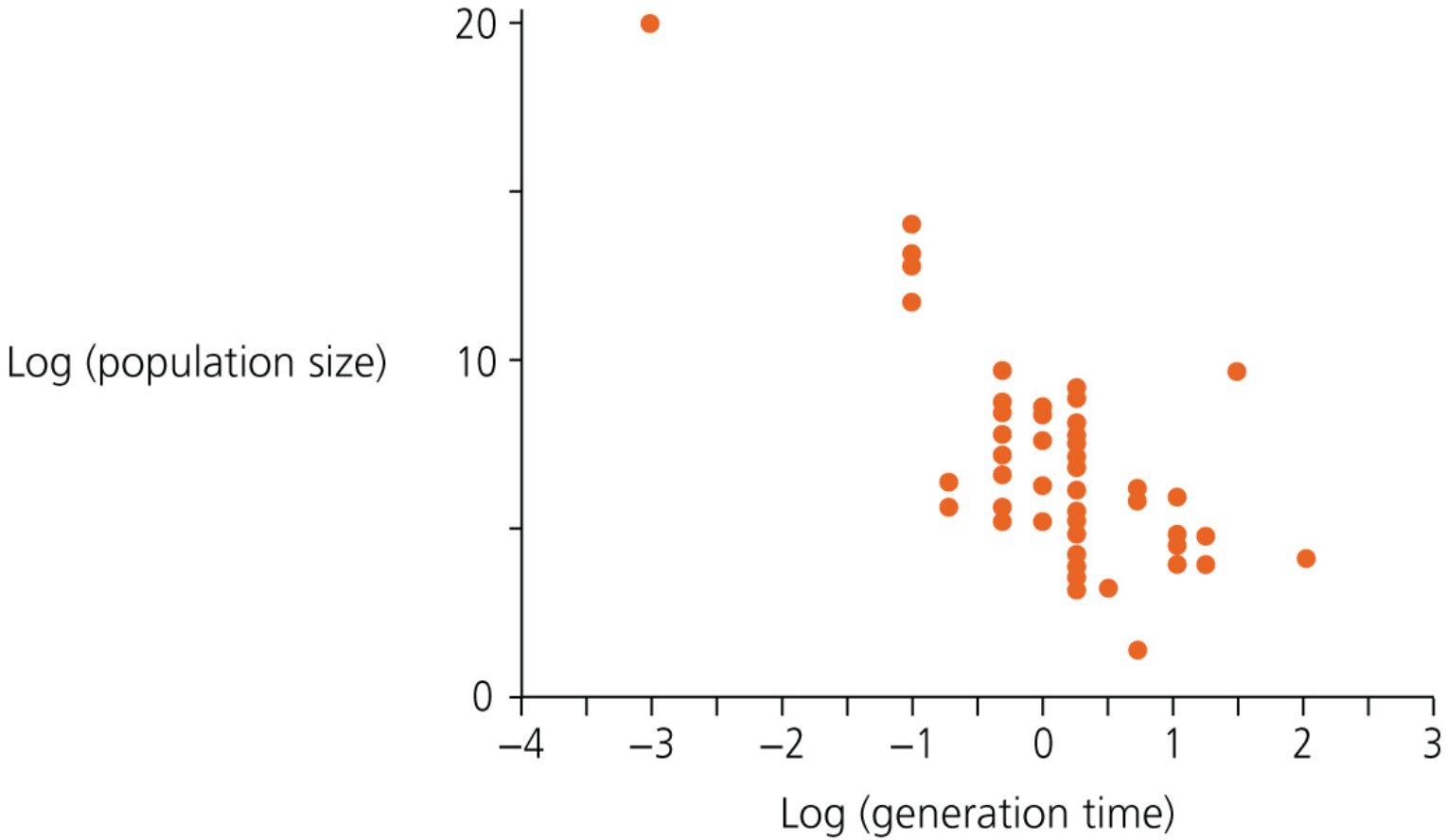
Tomoko Ohta and Motoo Kimura found the answer (Otha & Kimura 1971; Otha 1972, 1977).

$\Rightarrow$ the strength of genetic drift depends on N_e, therefore, if species with small generation time have large population sizes, the two effects will compensate and rates of substitution will remain constant over time.

$\Rightarrow$ the proportion of mutations that are **effectively neutral** depends on the effective population size N_e.

Nearly-Neutral Theory

$\Rightarrow$ there exists a negative relationship between population size and generation time in natural populations



(Chao & Carr, 1993)

Nearly-Neutral Theory

$\Rightarrow$ in **large** populations, **more** mutations appear per generation but they are **less** likely to fix by genetic drift, because drift is **weaker** ($\approx \frac{1}{2N_e}$).

$\Rightarrow$ in **small** populations, **less** mutations appear per generation but they are **more** likely to fix by genetic drift

Nearly-Neutral Theory

$\Rightarrow$ in **large** populations, **more** mutations appear per generation but they are **less** likely to fix by genetic drift, because drift is **weaker** ($\approx \frac{1}{2N_e}$).

$\Rightarrow$ in **small** populations, **less** mutations appear per generation but they are **more** likely to fix by genetic drift

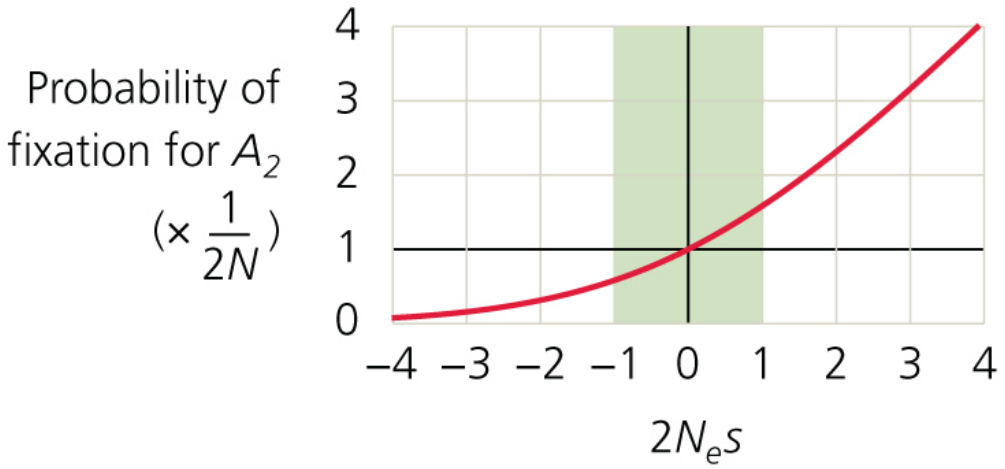
$\Rightarrow$ in addition, a **larger proportion** of mutations are **effectively neutral** in populations with a small N_e.

Nearly-Neutral Theory

$\Rightarrow$ the condition for **effective neutrality** is:

$$|s| < \frac{1}{2N_e},$$

with s the effect of the mutation on individual fitness.



N = Census population size
N_e = Effective population size

s = Selection coefficient

Genotype:	A_1A_1	A_1A_2	A_2A_2
Fitness:	1	$(1 + 0.5s)$	$(1 + s)$

Nearly-Neutral Theory

$\Rightarrow$ the condition for **effective neutrality** is:

$$|s| < \frac{1}{2N_e},$$

with s the effect of the mutation on individual fitness.

$\Rightarrow$ Therefore, the proportion of effectively neutral mutations increases as the effective population size decreases.

$\Rightarrow$ These **nearly neutral** mutations are mostly slightly deleterious mutations with $|s| \ll 10^{-2}$, because beneficial mutations are rare.

$\Rightarrow$ Selection dominates drift if $|s| > \frac{1}{2N_e}$; this condition sets a limit to selection.

Nearly-Neutral Theory

Summary: why is the molecular rate of evolution insensitive to the generation time?

Short generation time	Long generation time
• *Many* mutations on a per-year basis.	• *Few* mutations on a per-year basis.
• *Few* mutations will drift to fixation because N_e is *large* and *less* mutations are effectively neutral.	• *Many* mutations will drift to fixation because N_e is *small* and *more* mutations are effectively neutral.

Result: the difference between population-wide mutation rate and frequency of (nearly) neutral mutations cancel out

Wrap-up...

...on the role of genetic drift in evolution

- genetic drift is a potent evolutionary force that may drift (nearly) neutral alleles to fixation

Wrap-up...

...on the role of genetic drift in evolution

- genetic drift is a potent evolutionary force that may drift (nearly) neutral alleles to fixation
- the Neutral Theory seems sufficient to explain patterns of molecular evolution

Wrap-up...

...on the role of genetic drift in evolution

- genetic drift is a potent evolutionary force that may drift (nearly) neutral alleles to fixation
- the Neutral Theory seems sufficient to explain patterns of molecular evolution
- it is all about the effective population size N_e
  - probability of fixation of new mutation $= \frac{1}{2N_e}$
  - expected time to fixation of a new mutation $= 4N_e$
  - expected heterozygosity $= 4N_e\mu$ (provided that $4N_e \ll 1/\mu$)
  - expected number of nucleotide differences between two sequences is $\theta = 4N_e\mu$

Wrap-up...

...on the role of genetic drift in evolution

- genetic drift is a potent evolutionary force that may drift (nearly) neutral alleles to fixation
- the Neutral Theory seems sufficient to explain patterns of molecular evolution
- it is all about the effective population size N_e
  - probability of fixation of new mutation $= \frac{1}{2N_e}$
  - expected time to fixation of a new mutation $= 4N_e$
  - expected heterozygosity $= 4N_e\mu$ (provided that $4N_e \ll 1/\mu$)
  - expected number of nucleotide differences between two sequences is $\theta = 4N_e\mu$
- expectation under genetic drift provides a null hypothesis when inferring the action of natural selection from molecular data