Genetic Drift

COMP 571 - Fall 2010

Luay Nakhleh, Rice University

Outline

(1) The effects of sampling

(2) Models of genetic drift

(3) Effective population size

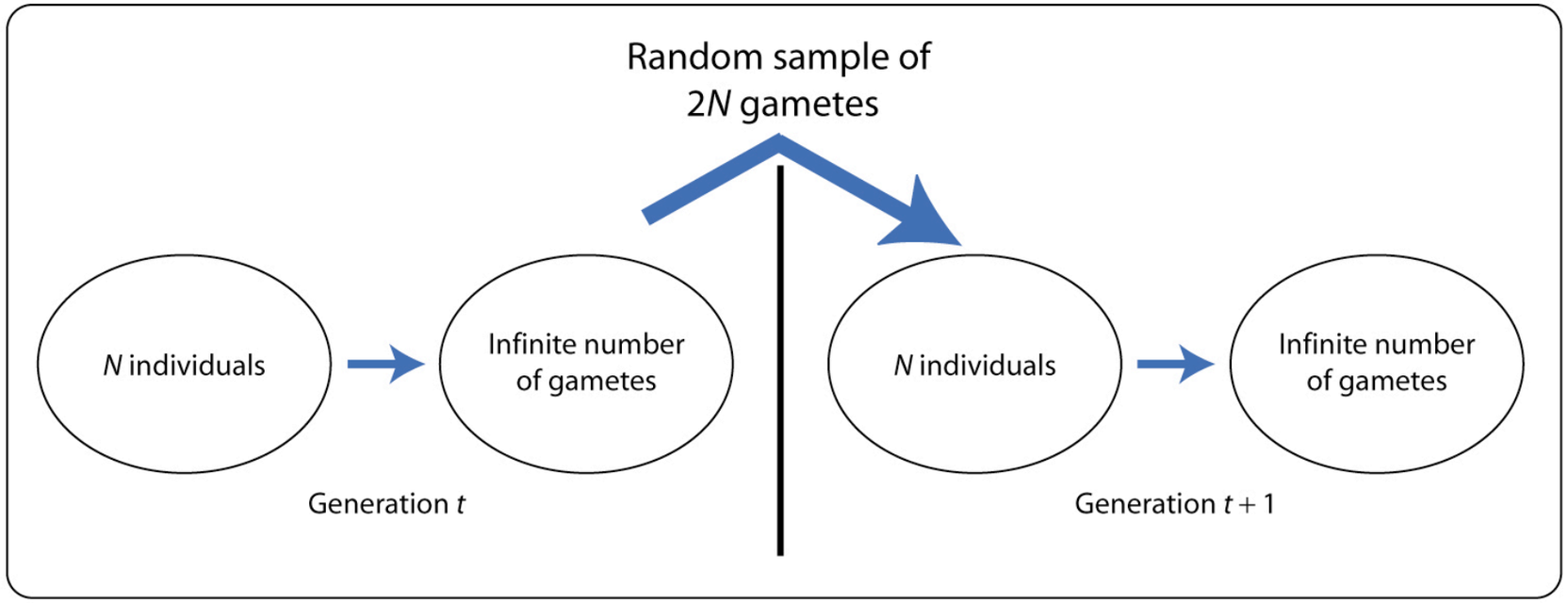
(4) Drift and inbreeding

(5) The coalescent model

(1) The Effects of Sampling

* One of the assumption of HW is that population size is very large, effectively infinite
* However, all biological populations, without exception, are finite
* How does the population size affect allele and genotype frequencies?

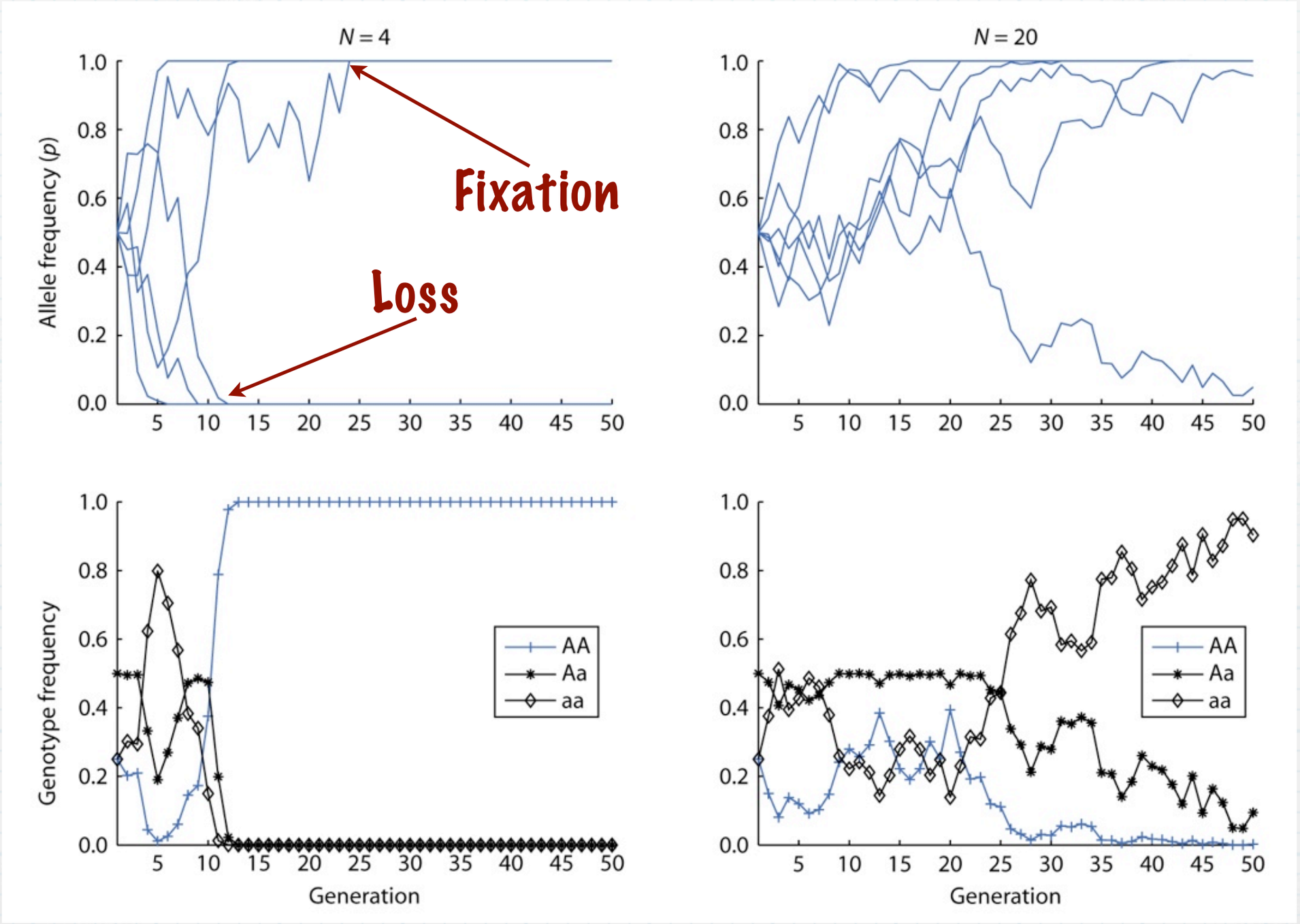
(1) The Effects of Sampling



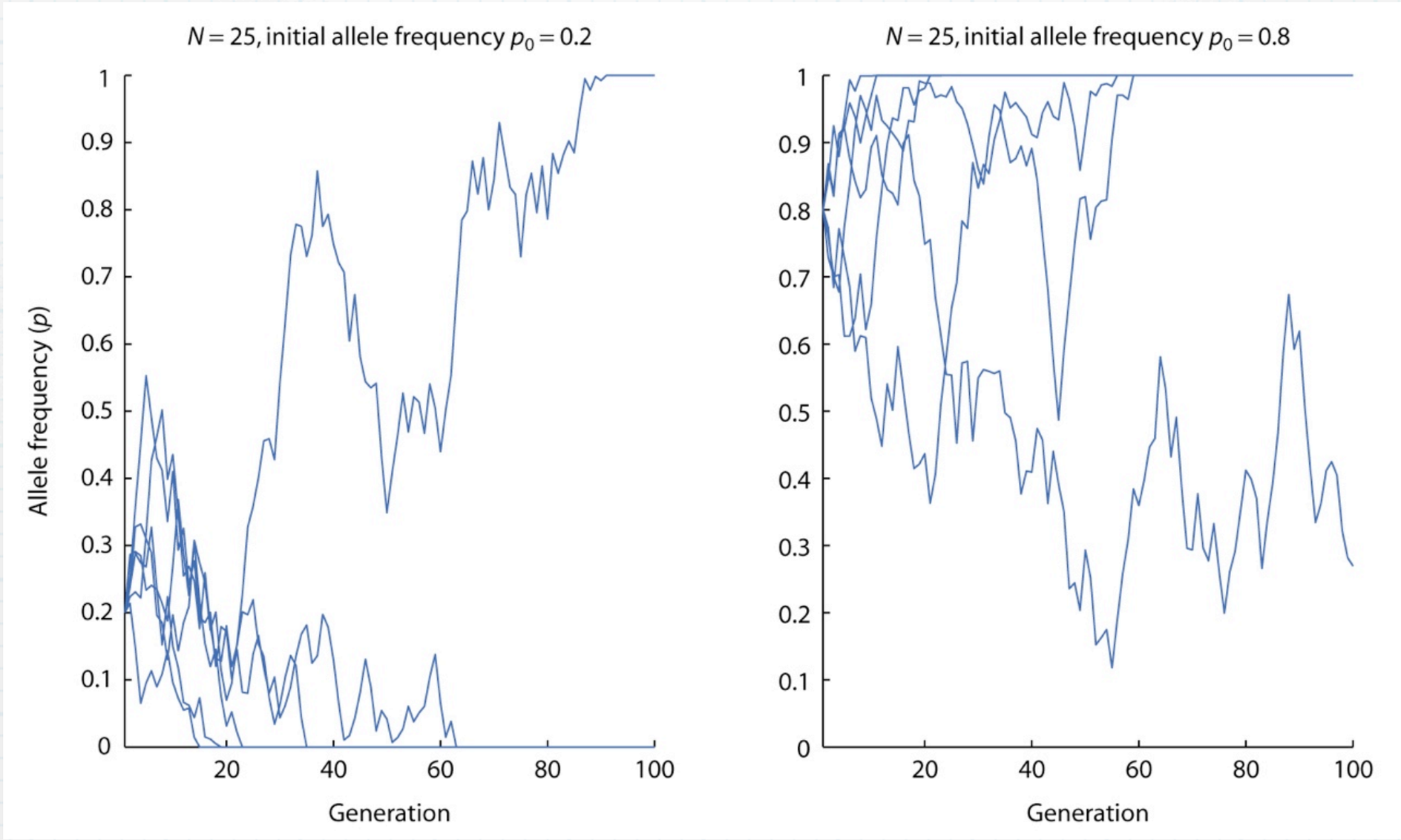
The **Wright-Fisher model** of **genetic drift** uses a simplified view of biological reproduction where all sampling occurs at one point: sampling 2N gametes from an infinite gamete pool. Genetic drift takes place only in the random sample of 2N gametes to form the next generation.

The Wright-Fisher model makes assumptions identical to those of HW, with the exception that the population is finite

(1) The Effects of Sampling



(1) The Effects of Sampling



(1) The Effects of Sampling

* Under the Wright-Fisher model:
  * the direction of changes in allele frequency is random
  * the magnitude of random fluctuations in allele frequencies from generation to generation increases as the population size decreases
  * fixation or loss is the equilibrium state
  * genetic drift changes allele frequencies and thereby genotype frequencies
  * the probability of eventual fixation of an allele is equal to its initial frequency

(2) Models of Genetic Drift

* Three probability models can be used to illustrate properties of the process of genetic drift:
  * the binomial probability distribution
  * Markov chains
  * diffusion approximation (we will not cover it)

(2) Models of Genetic Drift

The Binomial Probability Distribution

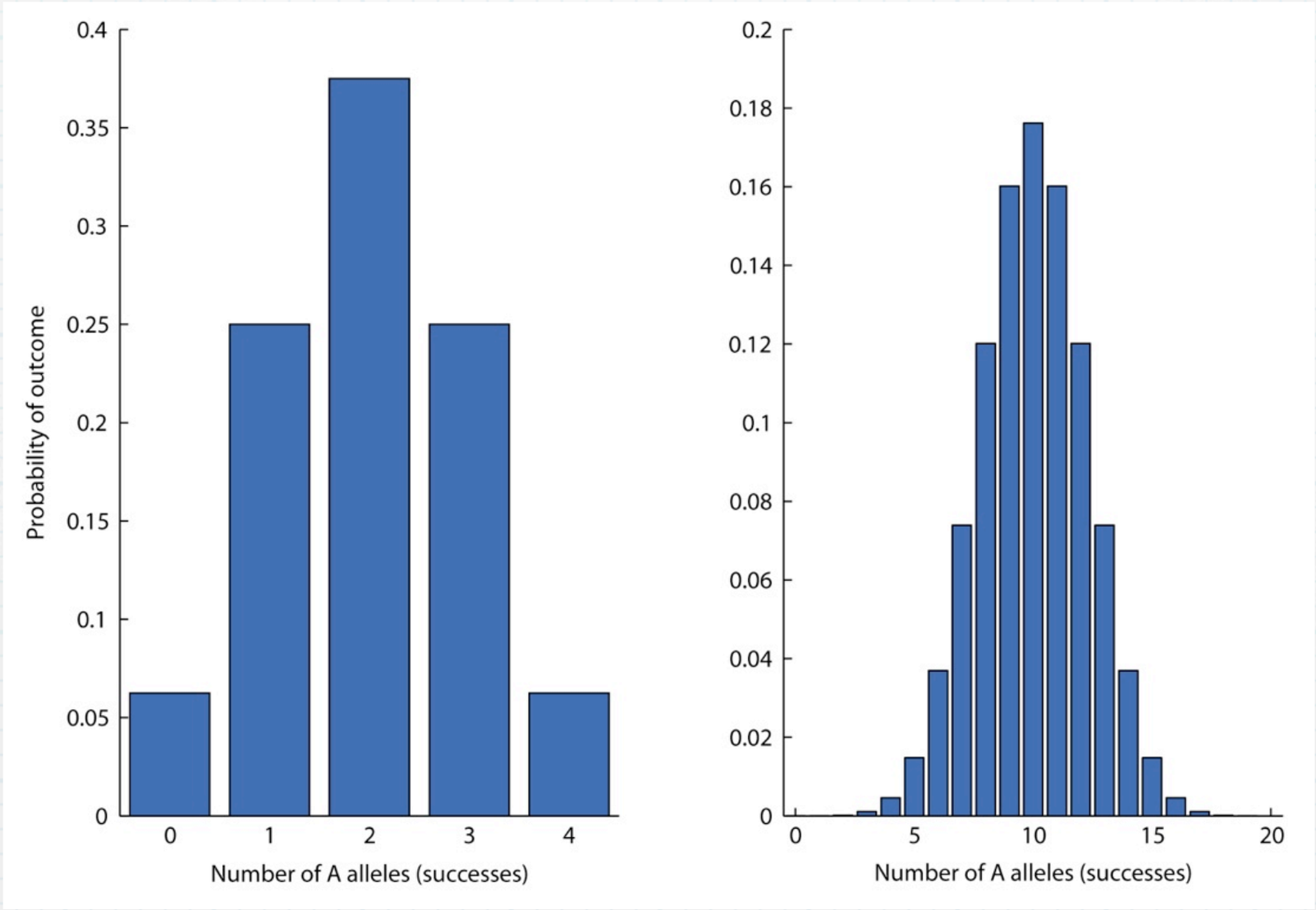
* Assuming a single locus with alleles A and a at frequencies p and q at a certain generation t (p+q=1)

* If 2N gametes are drawn at random to produce the zygotes of the next generation (t+1), the probability that the sample contains exactly i alleles of type A is the binomial probability

$$\binom{2N}{i} p^i q^{2N-i} = \frac{(2N)!}{i!(2N-i)!} p^i q^{2N-i}$$

(2) Models of Genetic Drift

The Binomial Probability Distribution



2N=4

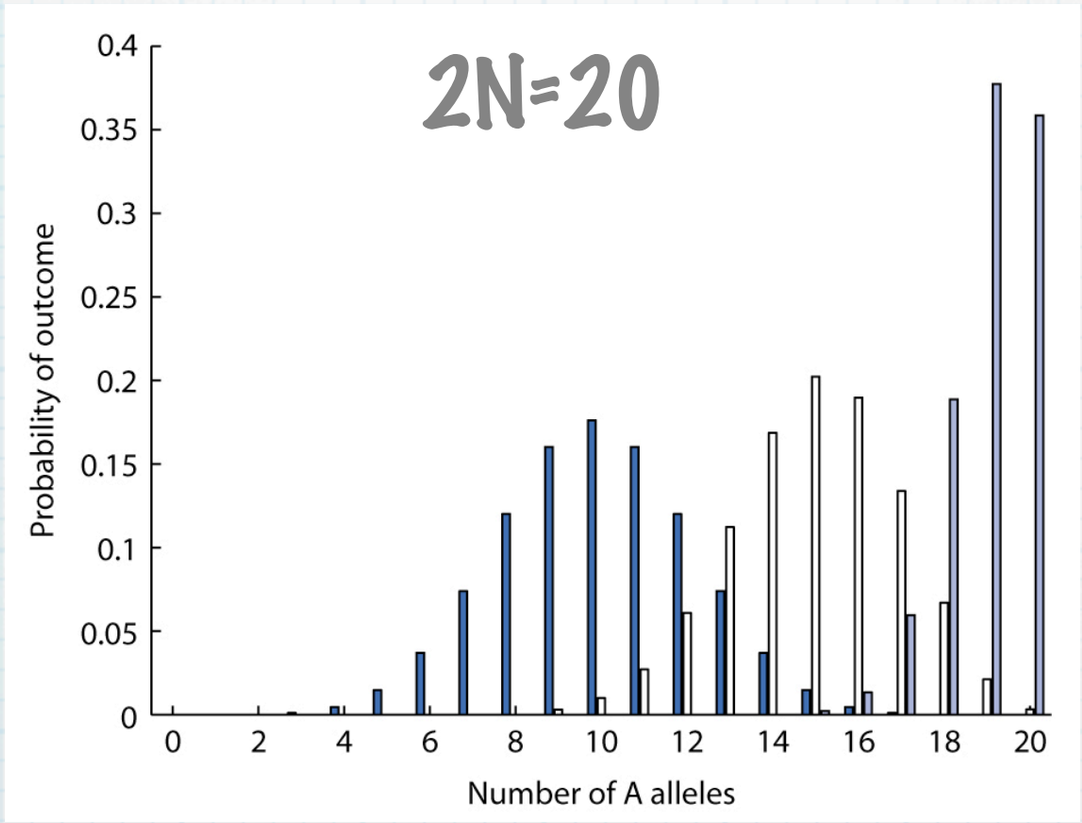
2N=20

p=q=0.5

(2) Models of Genetic Drift

The Binomial Probability Distribution

* The variance of a binomial random variable is $\sigma^2=pq$



dark blue: allele frequency = 0.5

while: allele frequency = 0.75

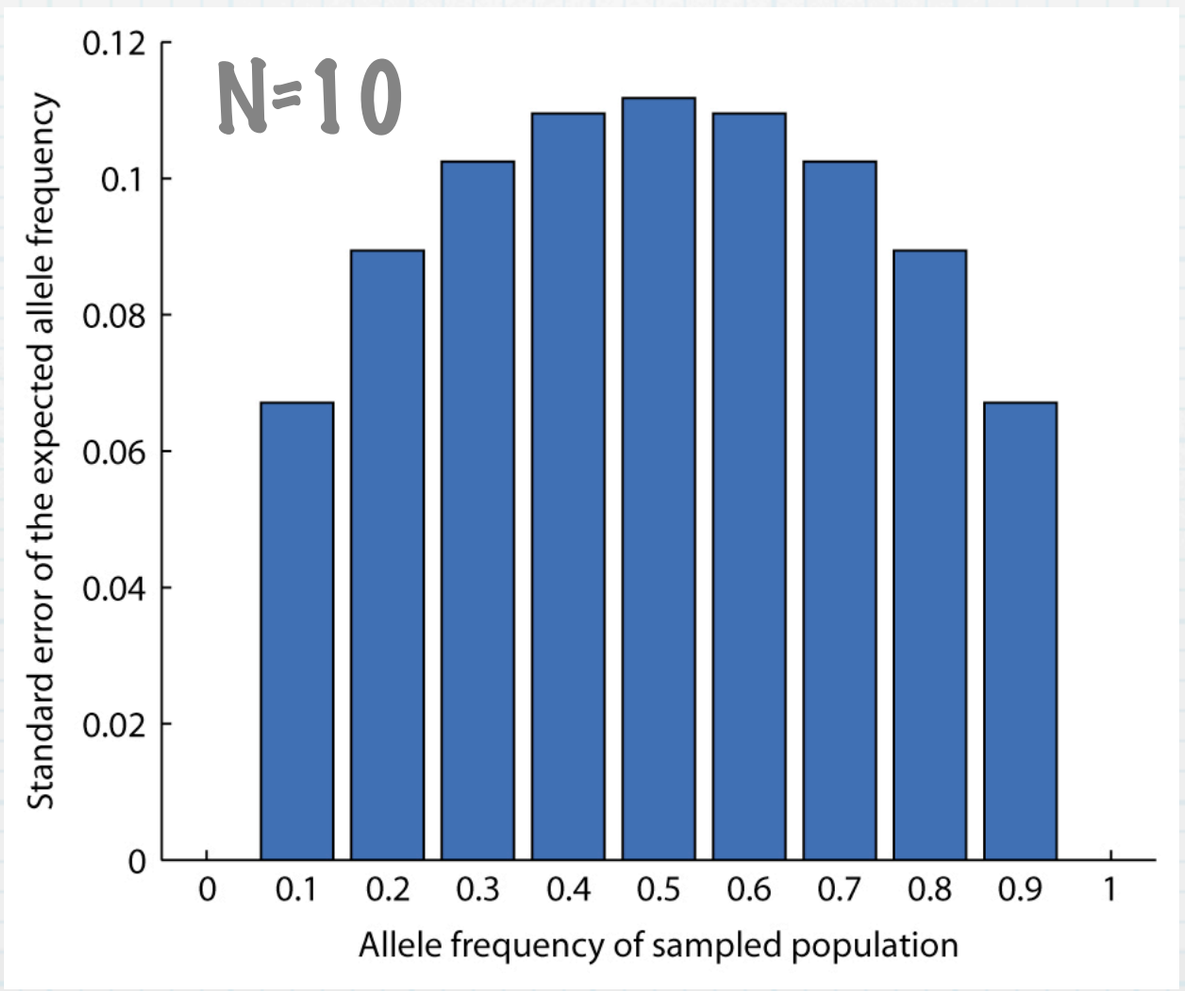
light blue: allele frequency = 0.95

(2) Models of Genetic Drift

The Binomial Probability Distribution

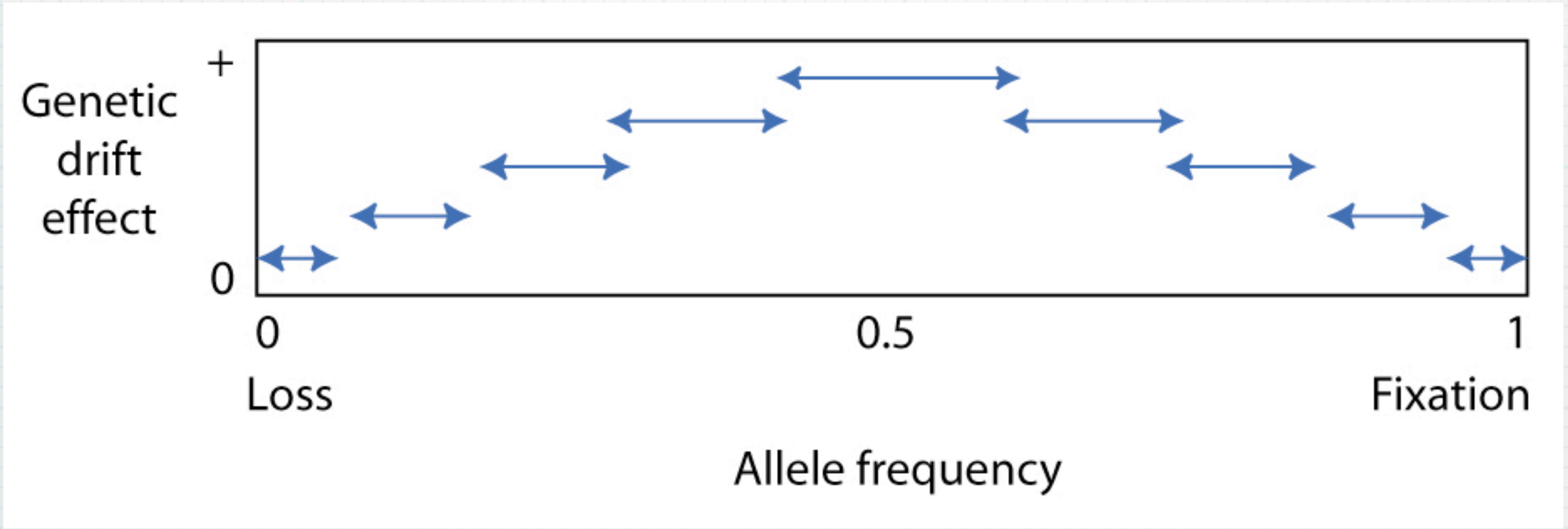
* The standard deviation and standard error are, respectively:

$$\sigma = \sqrt{pq} \qquad SE = \sqrt{\frac{pq}{2N}}$$



(2) Models of Genetic Drift

The Binomial Probability Distribution



Genetic drift is less effective at spreading out the distribution of allele frequencies as alleles approach fixation or loss

(2) Models of Genetic Drift

The Binomial Probability Distribution

* Exercise: Two independent laboratory populations of the fruit fly D. melanogaster were observed for two generations. The populations each had a size of N=24 individuals with an equal number of males and females. In the first generation, both populations were found with fA=p=0.5. In the second generation, one population showed fA=p=0.458 and the other fA=p=0.521. What are the chances of observing these allele frequencies after one generation of genetic drift?

(2) Models of Genetic Drift

Markov Chains

* Consider populations composed of a diallelic locus in a single diploid individual
* The number of A alleles are denoted by P(0), P(1), P(2)
* Question: What are chances that a population starting out in one of these three states ends up in one of these three states due to sampling error?
* This can be answered with the help of the transition probability from i alleles to j alleles in a generation

$$P_{i \to j} = \binom{2N}{j} p^j q^{2N-j}$$

(2) Models of Genetic Drift

Markov Chains

One generation later (t=1)		Initial state: # of A alleles (t=0)		
A alleles	Expected frequency	2	1	0
2	$P_{t=1}(2)$	$= (P_{2\to2})\,P_{t=0}(2) + (P_{1\to2})\,P_{t=0}(1) + (P_{0\to2})\,P_{t=0}(0)$		
1	$P_{t=1}(1)$	$= (P_{2\to1})\,P_{t=0}(2) + (P_{1\to1})\,P_{t=0}(1) + (P_{0\to1})\,P_{t=0}(0)$		
0	$P_{t=1}(0)$	$= (P_{2\to0})\,P_{t=0}(2) + (P_{1\to0})\,P_{t=0}(1) + (P_{0\to0})\,P_{t=0}(0)$		

The expected frequency of populations with zero, one, or two A alleles in generation one (t=1) based on the previous generation (t=0)

(2) Models of Genetic Drift

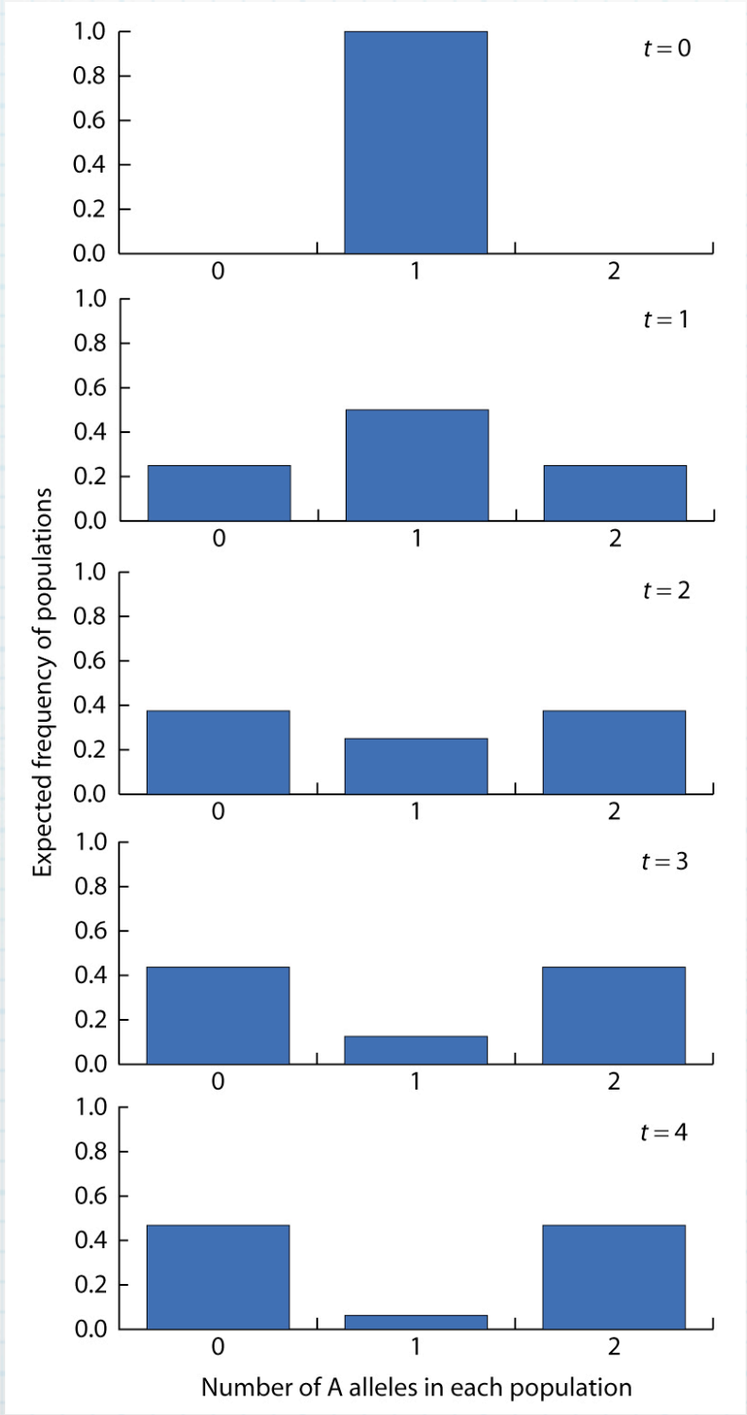
Markov Chains

One generation later (t=1)		Initial state: # of A alleles (t=0)		
A alleles	Expected frequency	2	1	0
2	$P_{t=1}(2) = (P_{2\to2})\,P_{t=0}(2) + (P_{1\to2})\,P_{t=0}(1) + (P_{0\to2})\,P_{t=0}(0)$			
1	$P_{t=1}(1) = (P_{2\to1})\,P_{t=0}(2) + (P_{1\to1})\,P_{t=0}(1) + (P_{0\to1})\,P_{t=0}(0)$			
0	$P_{t=1}(0) = (P_{2\to0})\,P_{t=0}(2) + (P_{1\to0})\,P_{t=0}(1) + (P_{0\to0})\,P_{t=0}(0)$			

The expected frequency of populations with zero, one, or two A alleles in generation one (t=1) based on the previous generation (t=0)

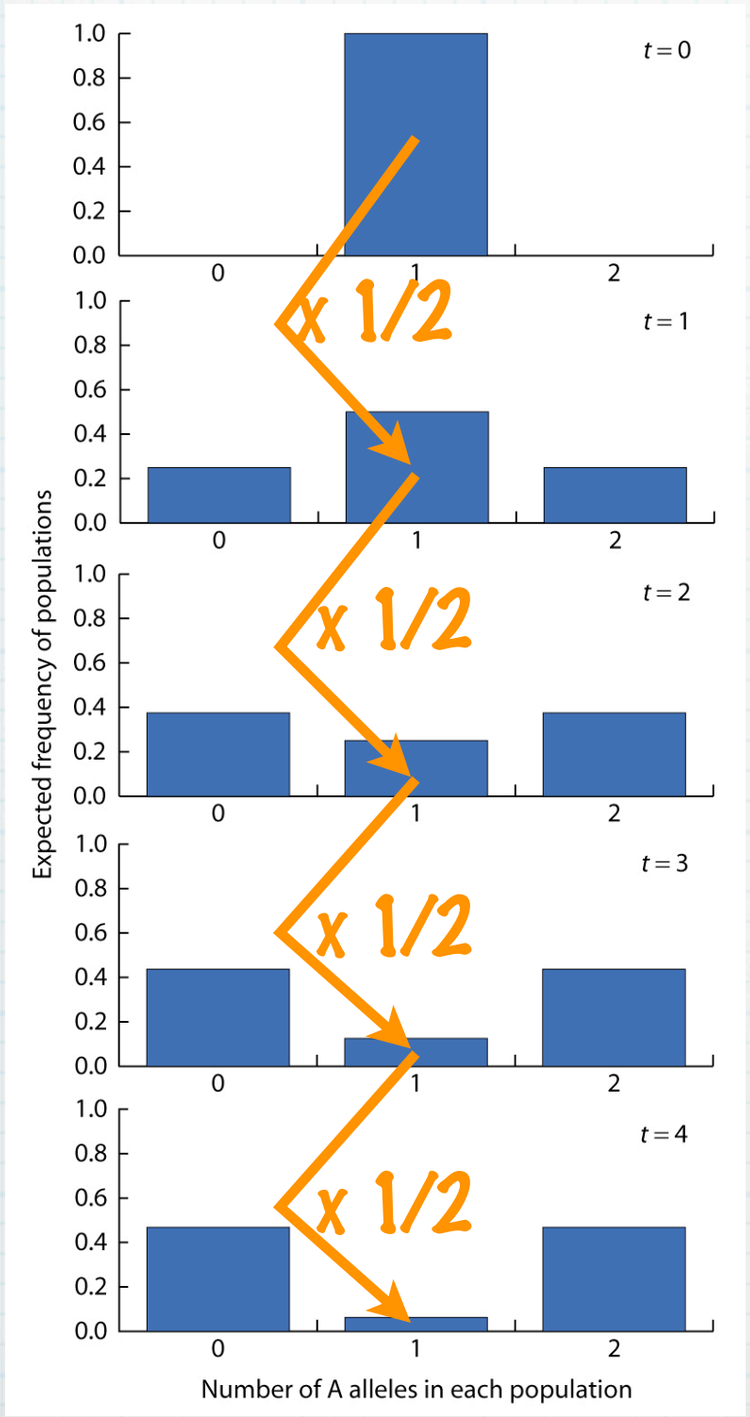
(2) Models of Genetic Drift

Markov Chains



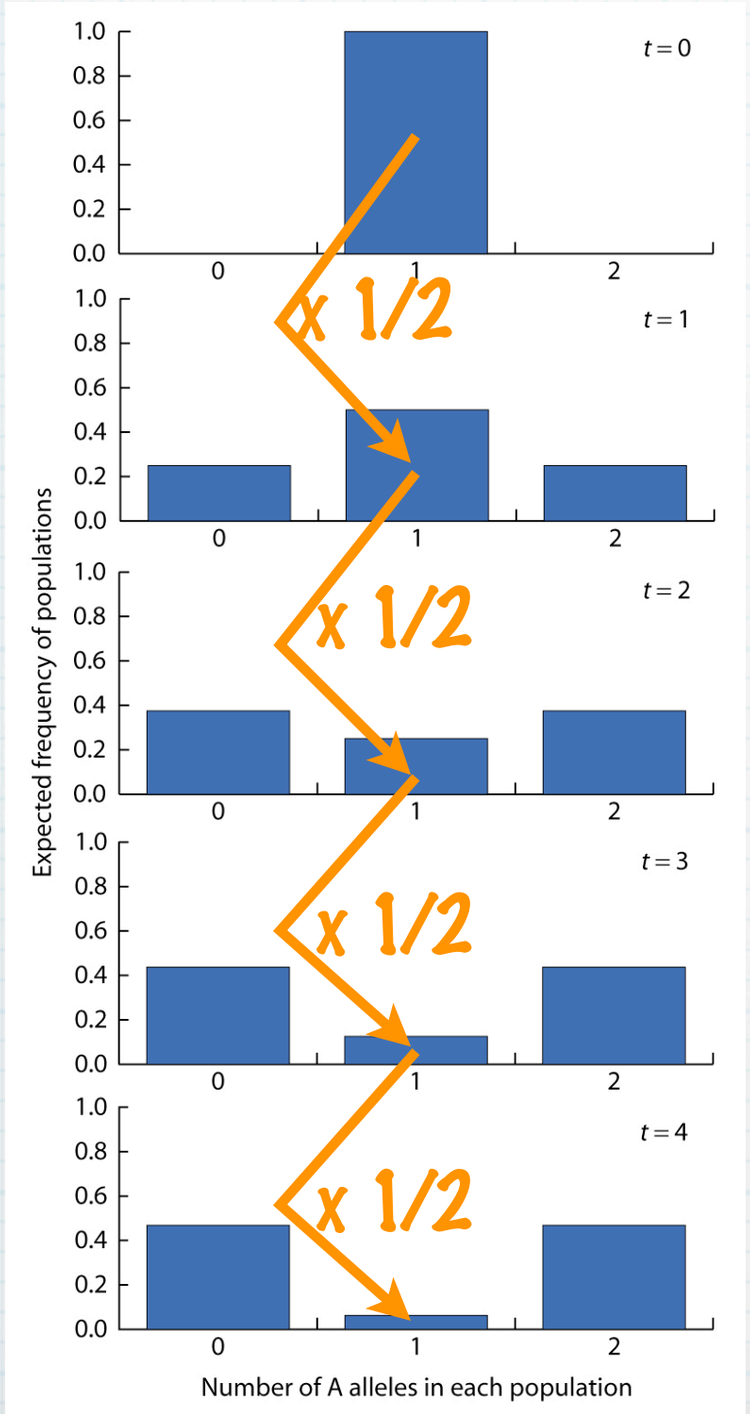
(2) Models of Genetic Drift

Markov Chains



(2) Models of Genetic Drift

Markov Chains

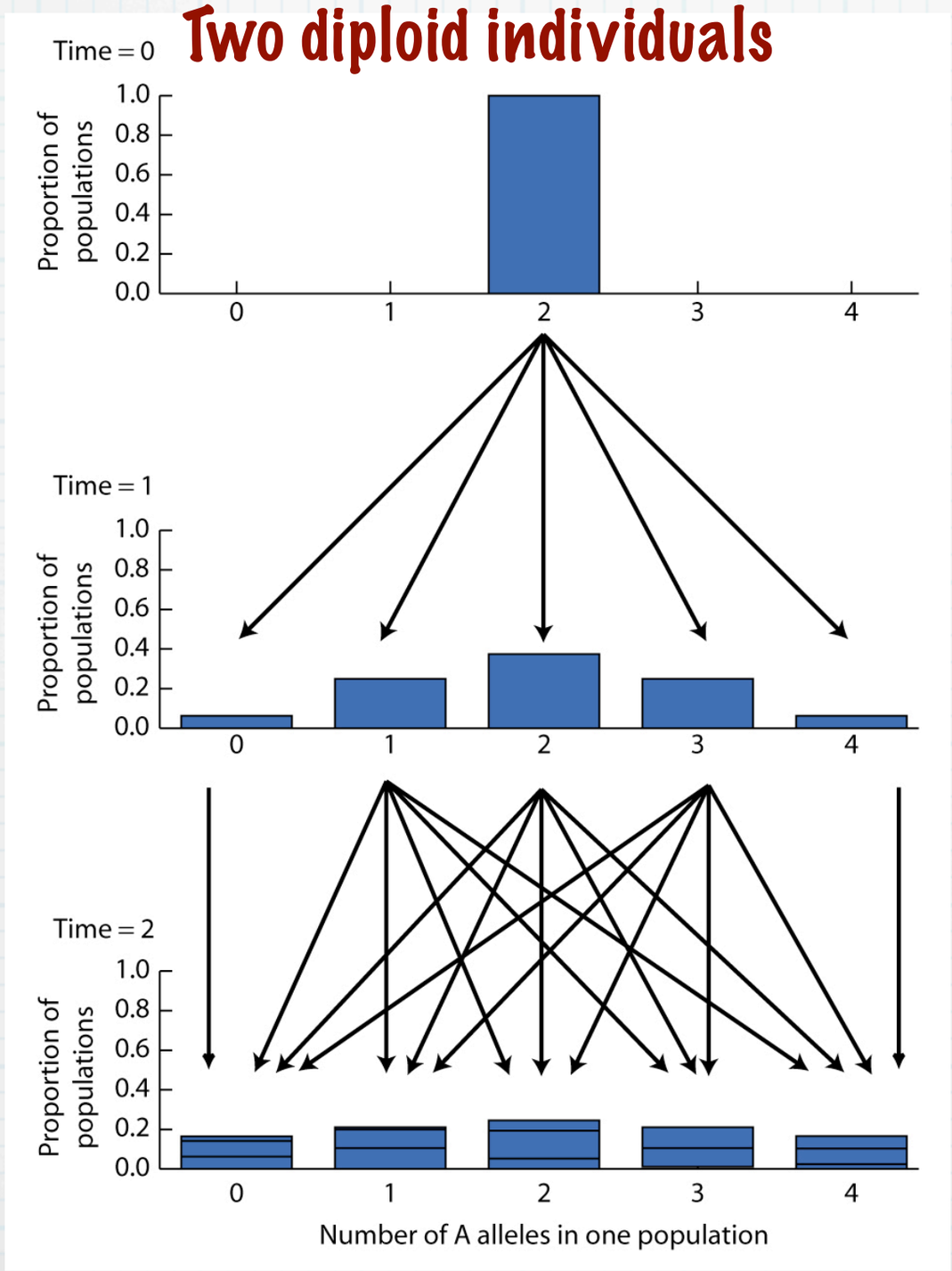


The rate at which genetic variation is lost from the collection of many populations is

1 - (1/2N)

(2) Models of Genetic Drift

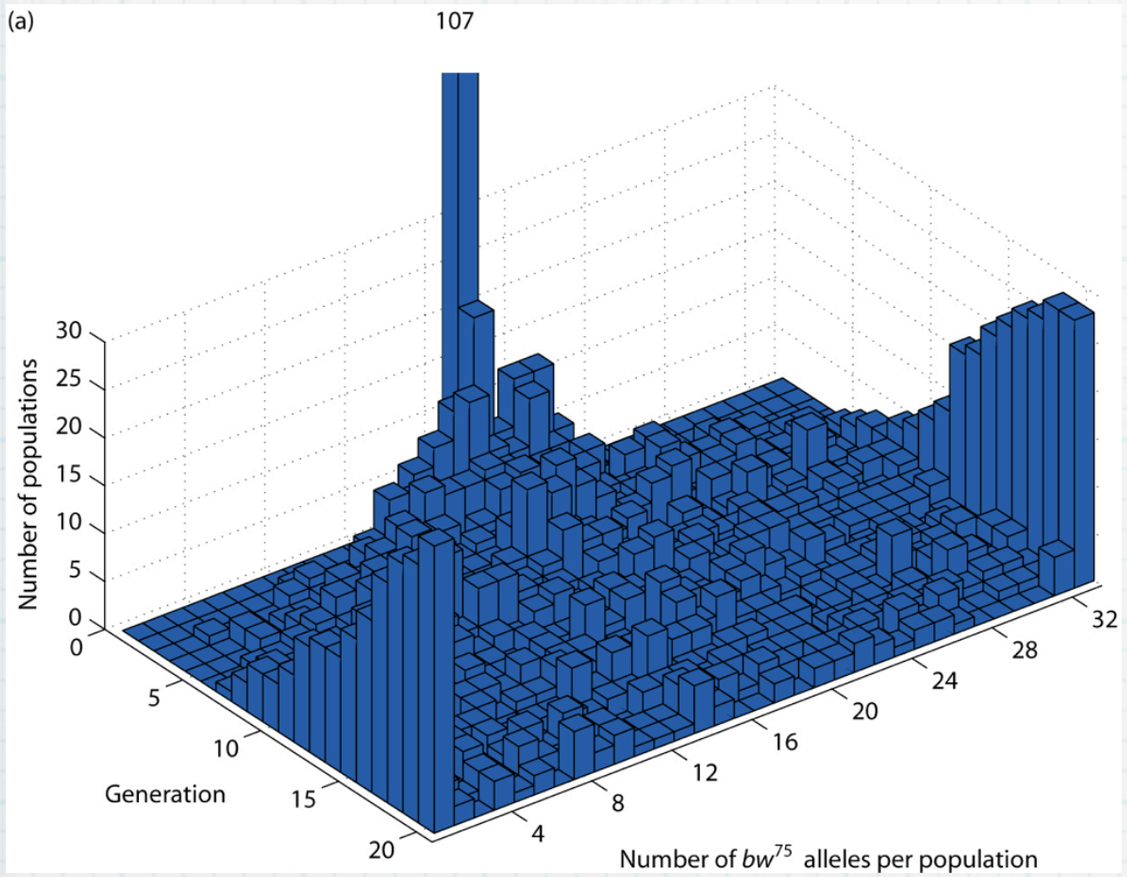
Markov Chains



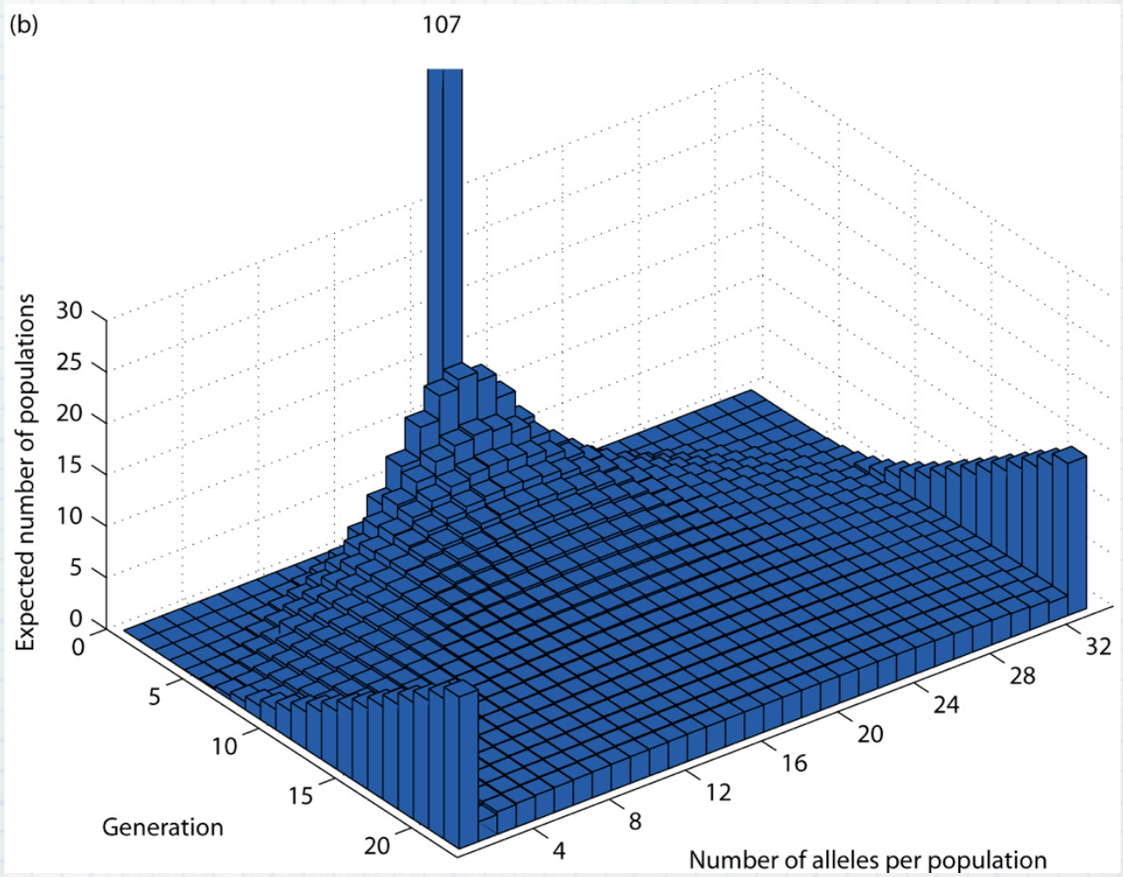
Genetic drift modeled by a Markov chain

(2) Models of Genetic Drift

Markov Chains



Allelic states for 107 D. melanogaster populations where 16 individuals (8 of each sex) were randomly chosen to start each new generation. Initially, all 107 populations had equal numbers of the wild-type and bw alleles.



The expected frequency of populations in each allelic state determined with a Markov chain model for a population size of 16 with 107 populations that initially have equal frequencies of two alleles.

(2) Models of Genetic Drift

Markov Chains

* The average time to fixation for alleles that eventually fix in a population and the average time to loss for alleles that eventually are lost from a population are, respectively:

$$\overline{T}_{fix} = -4N\frac{(1-p)\ln(1-p)}{p} \qquad \overline{T}_{loss} = -4N\frac{p\ln p}{1-p}$$

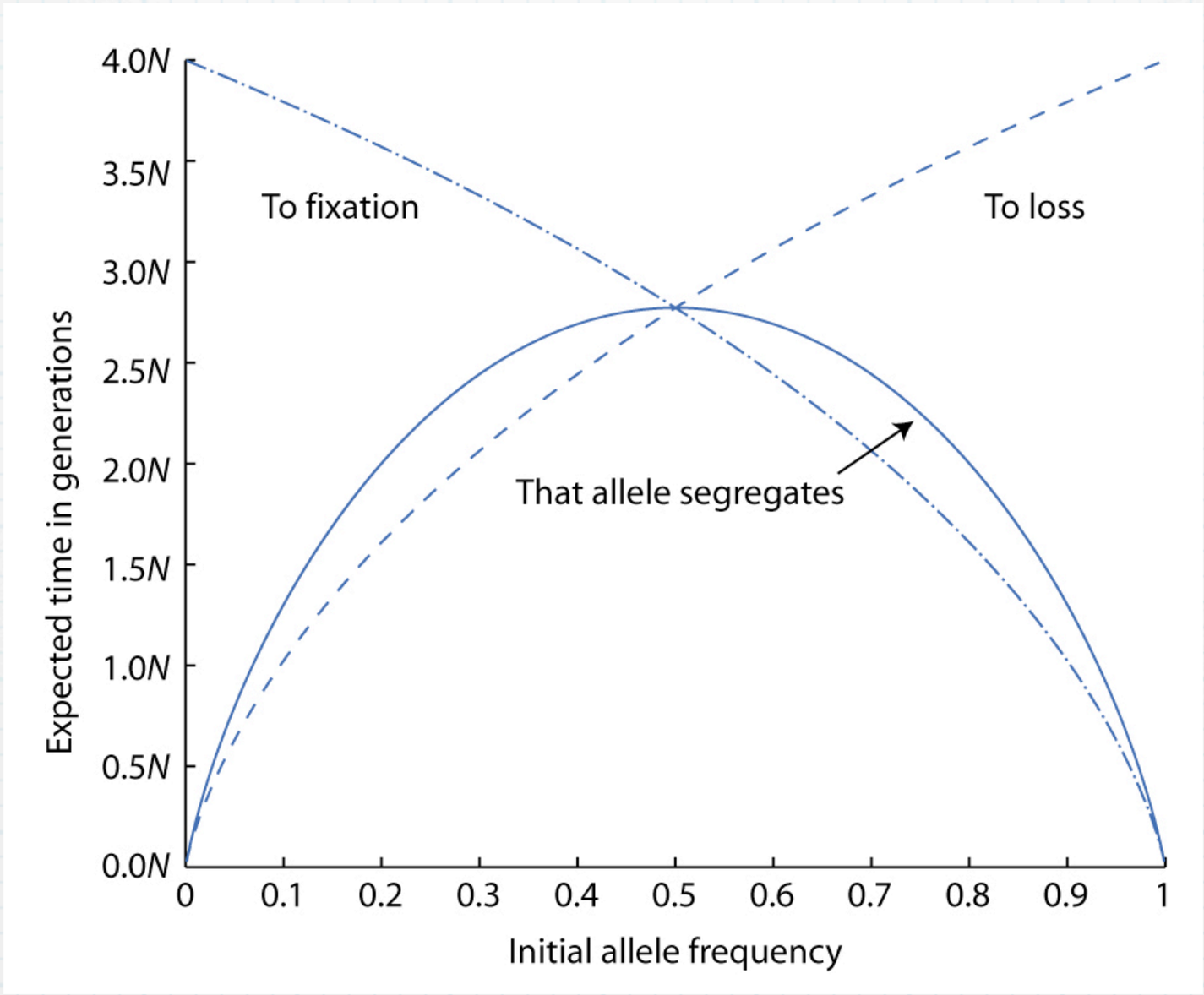
where p is the initial allele frequency.

* The mean persistence time of an allele is:

$$\overline{T}_{segregate} = p\overline{T}_{fix} + (1-p)\overline{T}_{loss} = -4N[(1-p)\ln(1-p) + p\ln p]$$

(2) Models of Genetic Drift

Markov Chains



(3) Effective Population Size

* In most biological populations it is difficult, or impossible, to determine the number of gametes that contribute to the next generation

* Therefore, we need to define the size of populations in a way that's different from "the number of individuals in a population"

* The definition of the population size in population genetics relies on the dynamics of genetic variation in the population (i.e., **the size of a population is defined by the way genetic variation in the population behaves**)

(3) Effective Population Size

* **Census population size** (**N**): the number of individuals in a population

* **Effective population size** (**N_e**): the size of an ideal Wright-Fisher population that maintains as much genetic variation or experiences as much genetic drift as an actual population regardless of census size

(3) Effective Population Size

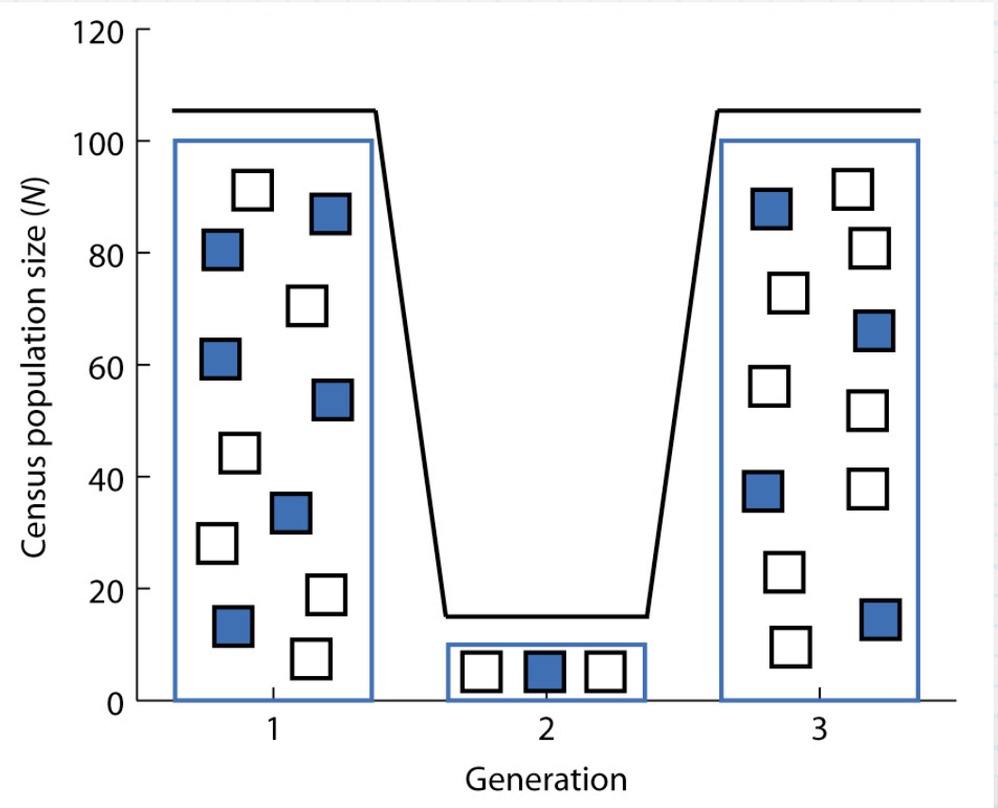
* Cases where a population may deviate from the ideal of the Wright-Fisher model:
  * Fluctuating population size
  * Breeding sex ratio
  * Variation in family size
  * Subdivided population

(3) Effective Population Size

* Fluctuation in population size over time (e.g., the cases of genetic bottleneck and founder events):

$$\frac{1}{N_e} = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{N_i}$$

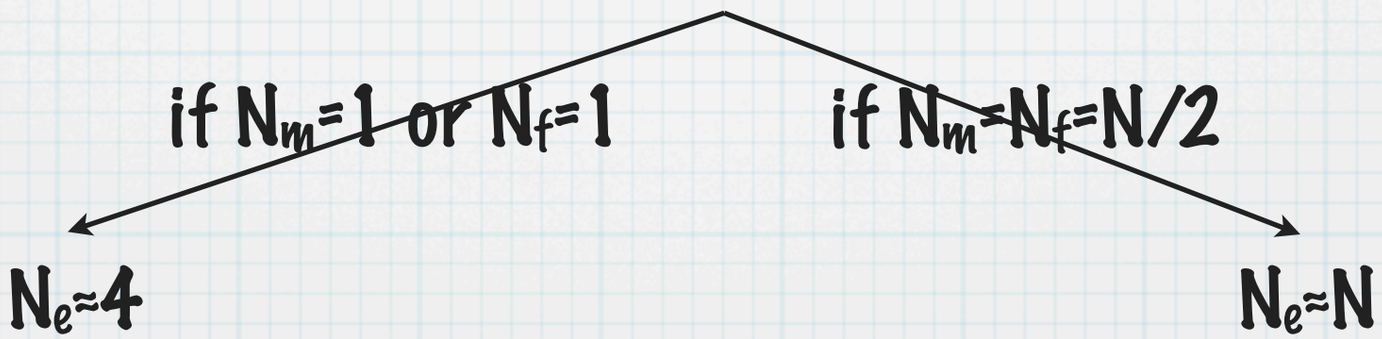
← Census population size at generation i



(3) Effective Population Size

* When N_m and N_f are the numbers of females and males, respectively, breeding in the population, and all other assumptions of Wright-Fisher populations are met, we have:

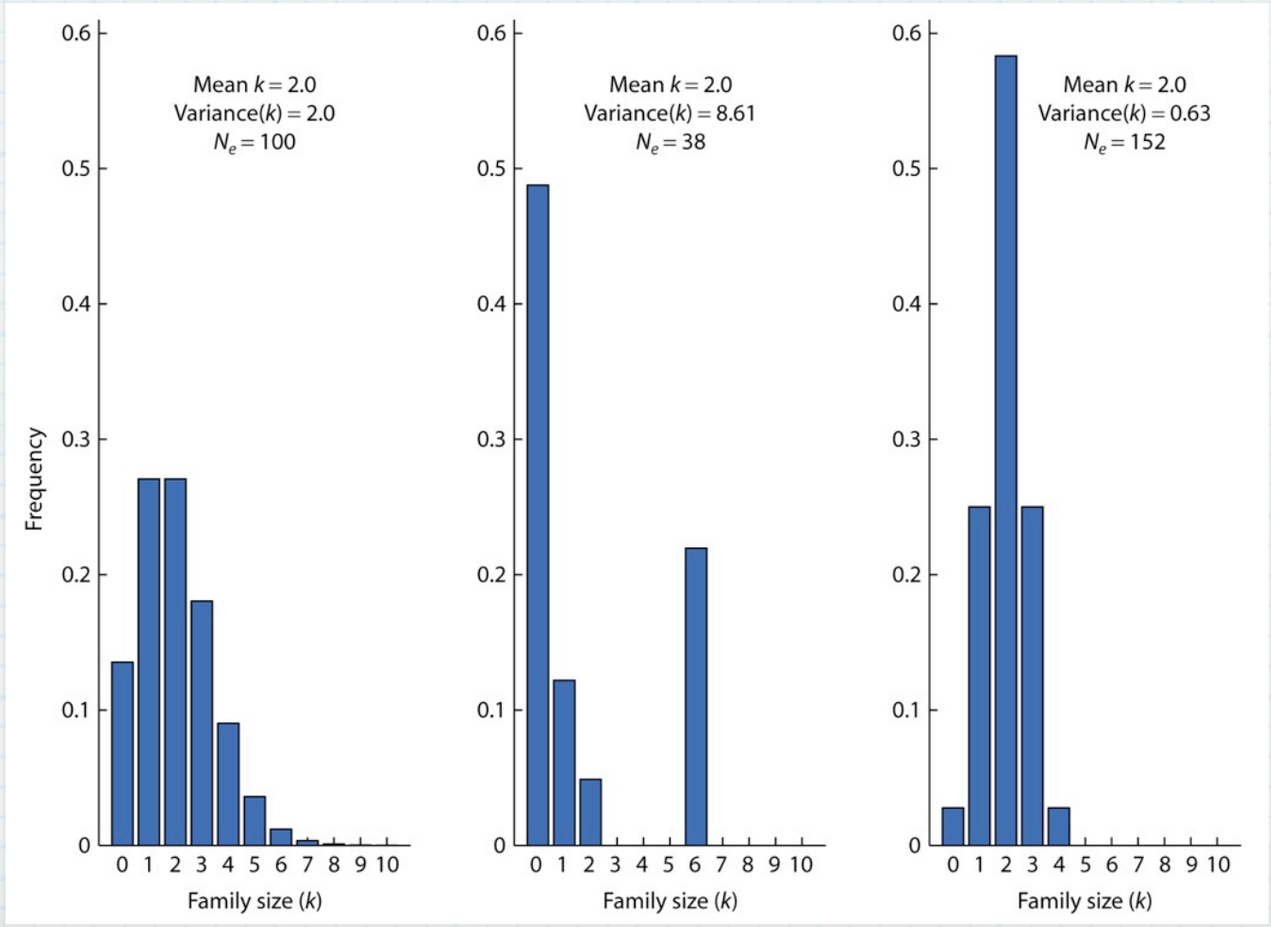
$$N_e = 4\frac{N_m N_f}{N_m + N_f}$$



(3) Effective Population Size

* When there is variation in family size, where N_{t-1} is the size of the parental population and k is the number of gametes that result in progeny, we have:

$$N_e = \frac{4N_{t-1}}{var(k) + \overline{k}^2 - \overline{k}}$$



(3) Effective Population Size

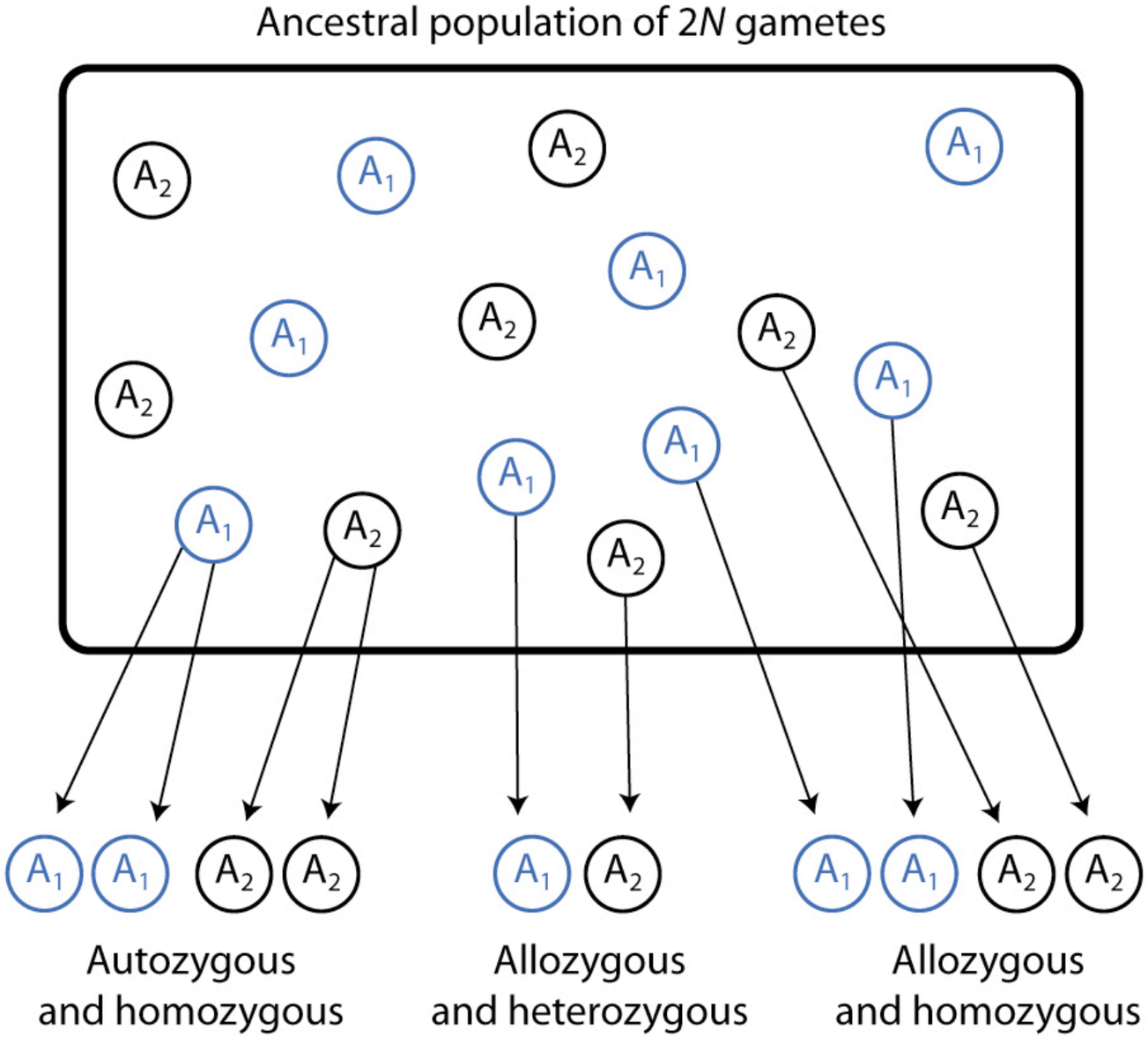
* When the population is divided into d demes, each of size N, and migration between demes occurs at rate m, we have:

$$N_e = N + \frac{(d-1)^2}{4md}$$

(4) Parallelism Between Drift and Inbreeding

* So far: genetic drift and population size
* We now demonstrate that finite population size can be thought of as a form of inbreeding
* Genetic drift occurs due to finite population size
* As populations get smaller, the probability of inbreeding increases
* Therefore, genetic drift and the tendency for inbreeding are interrelated phenomena, connected to the size of the population

(4) Parallelism Between Drift and Inbreeding



Autozygosity: given that one allele has been sampled, what is the probability of sampling the same allele on the next draw? $1/(2N_e)$

Allozygosity: given that one allele has been sampled, what is the probability of sampling a different allele on the next draw? $1-(1/(2N_e))$

(4) Parallelism Between Drift and Inbreeding

* We can use the probability of autozygosity in a finite population to define the fixation index as

$$F_t = \frac{1}{2N_e}$$

for generation t under the assumption that none of the alleles in the gamete pool in generation t-1 are identical by descent

* To make it more general, we have

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

sampling between generations

the proportion of apparently allozygous alleles that are actually autozygous due to past sampling or inbreeding

(4) Parallelism Between Drift and Inbreeding

* By definition, F is the reduction in heterozygosity as well as the increase in homozygosity compared to HW expected frequencies

* If F is proportional to the homozygosity and amount of inbreeding, then 1-F is proportional to the amount of heterozygosity and random mating

(4) Parallelism Between Drift and Inbreeding

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

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

Eq. 2.20: $H_t = 2pq(1-F_t)$

$$\frac{H_t}{2pq} = \left(1 - \frac{1}{2N_e}\right)\left(\frac{H_{t-1}}{2pq}\right)$$

$$H_t = \left(1 - \frac{1}{2N_e}\right) H_{t-1}$$

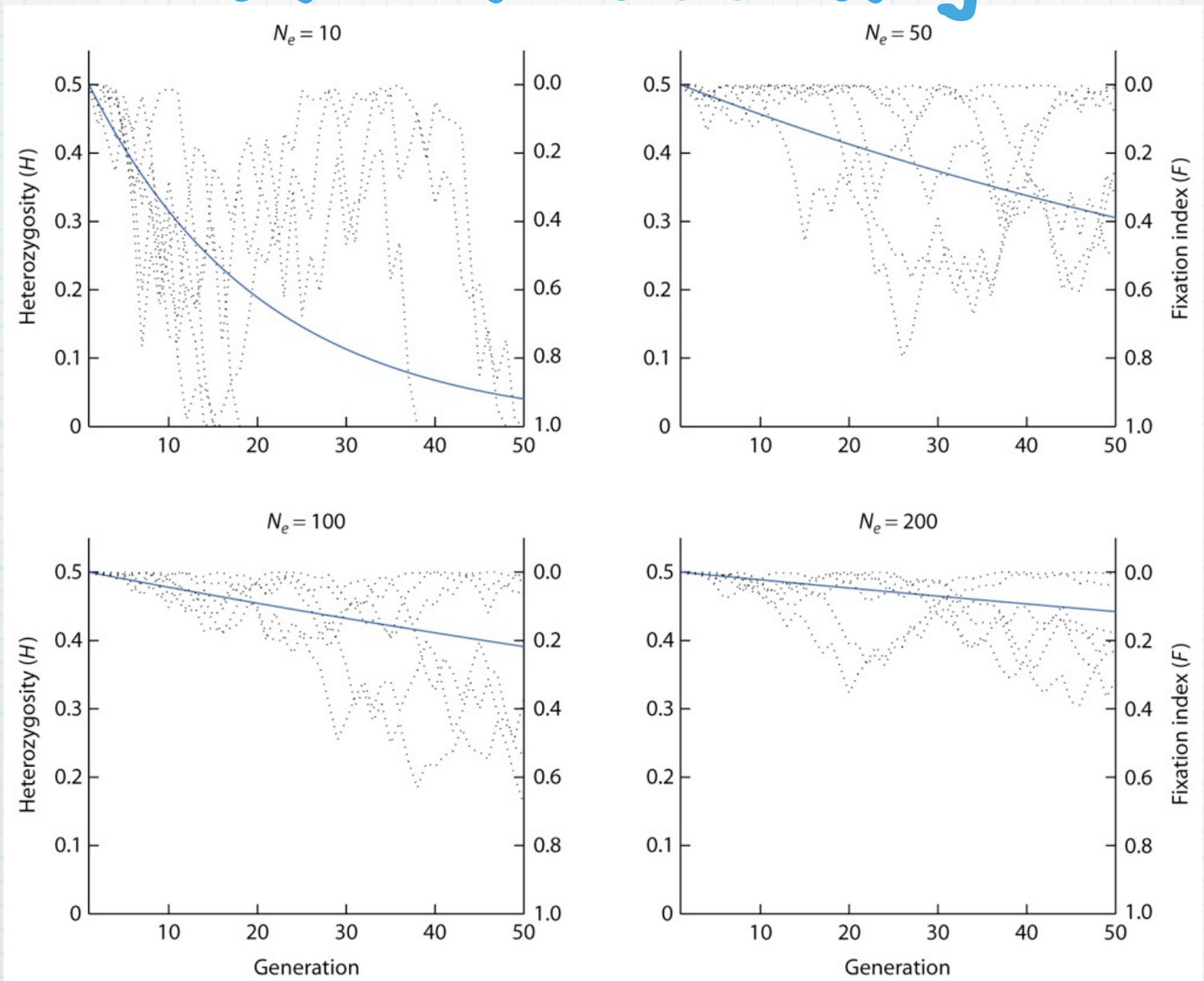
$$H_t = \left(1 - \frac{1}{2N_e}\right)^t H_0$$

H_t: heterozygosity after t generations

H_0: initial heterozygosity

$$H_t \approx H_0 e^{-t/2N}$$

(4) Parallelism Between Drift and Inbreeding



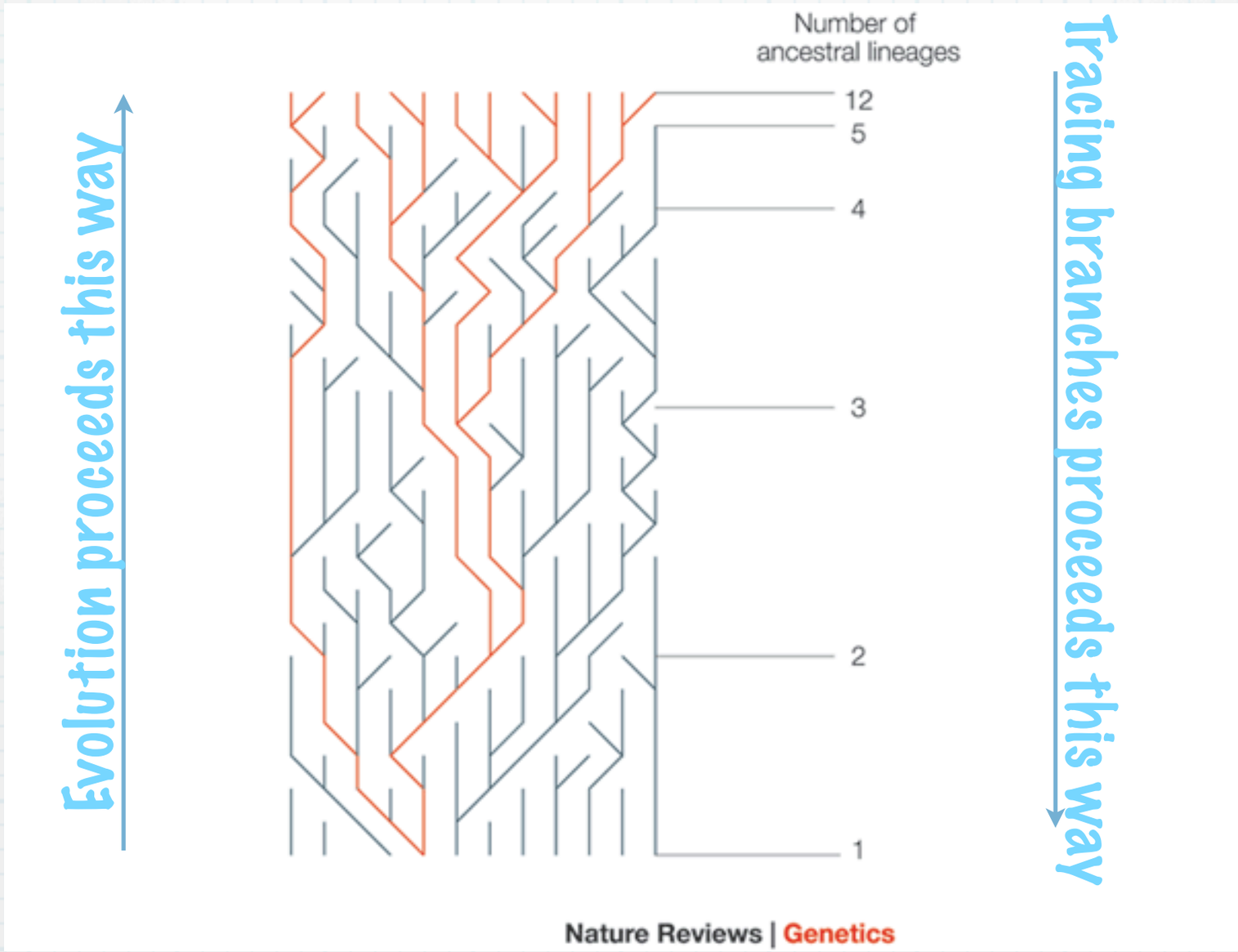
The decline in heterozygosity as a consequence of genetic drift in finite populations. The solid lines show the expected heterozygosity H_t. The dotted lines are levels of heterozygosity (2pq) in six replicate finite populations experiencing genetic drift.

(4) Parallelism Between Drift and Inbreeding

* Conclusions:
  * Genetic drift causes populations to become more inbred in the sense that autozygosity and homozygosity increase even though mating is random
  * Mating systems where there is consanguineous mating cause genetic variation in populations to behave as if the effective population size were smaller than it would be under complete random mating

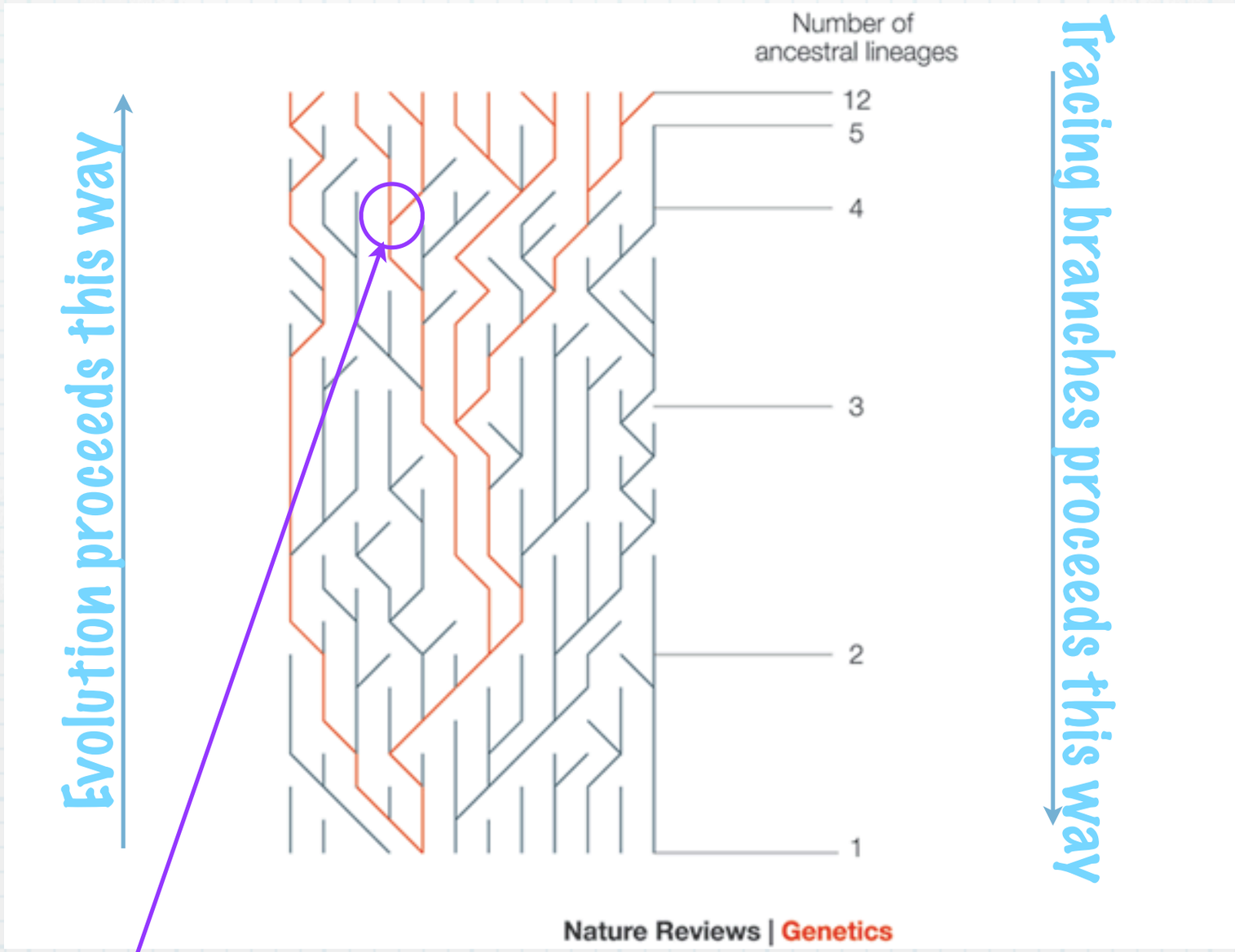
(5) The Coalescent Model

Modeling the branching of lineages



(5) The Coalescent Model

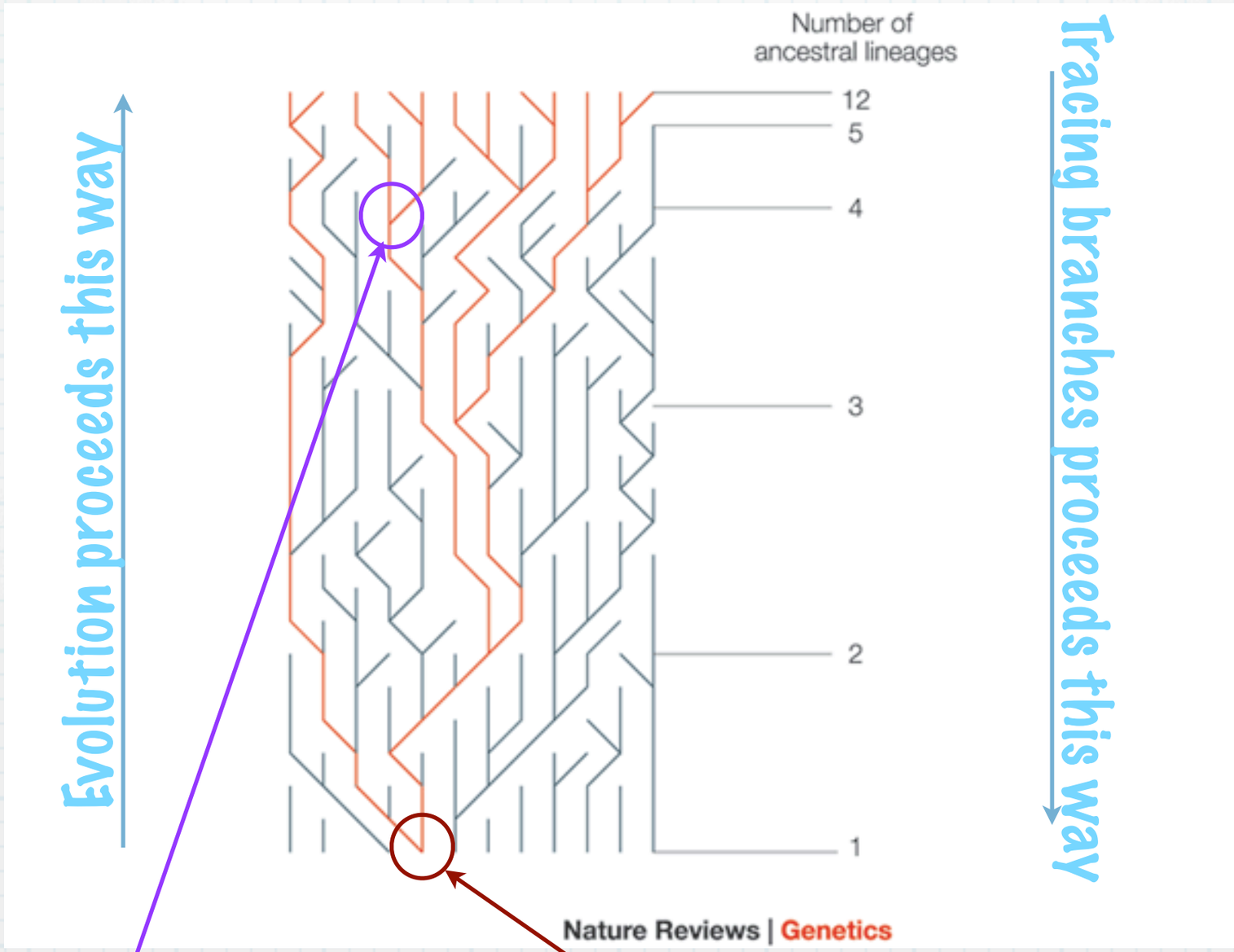
Modeling the branching of lineages



coalescent event

(5) The Coalescent Model

Modeling the branching of lineages



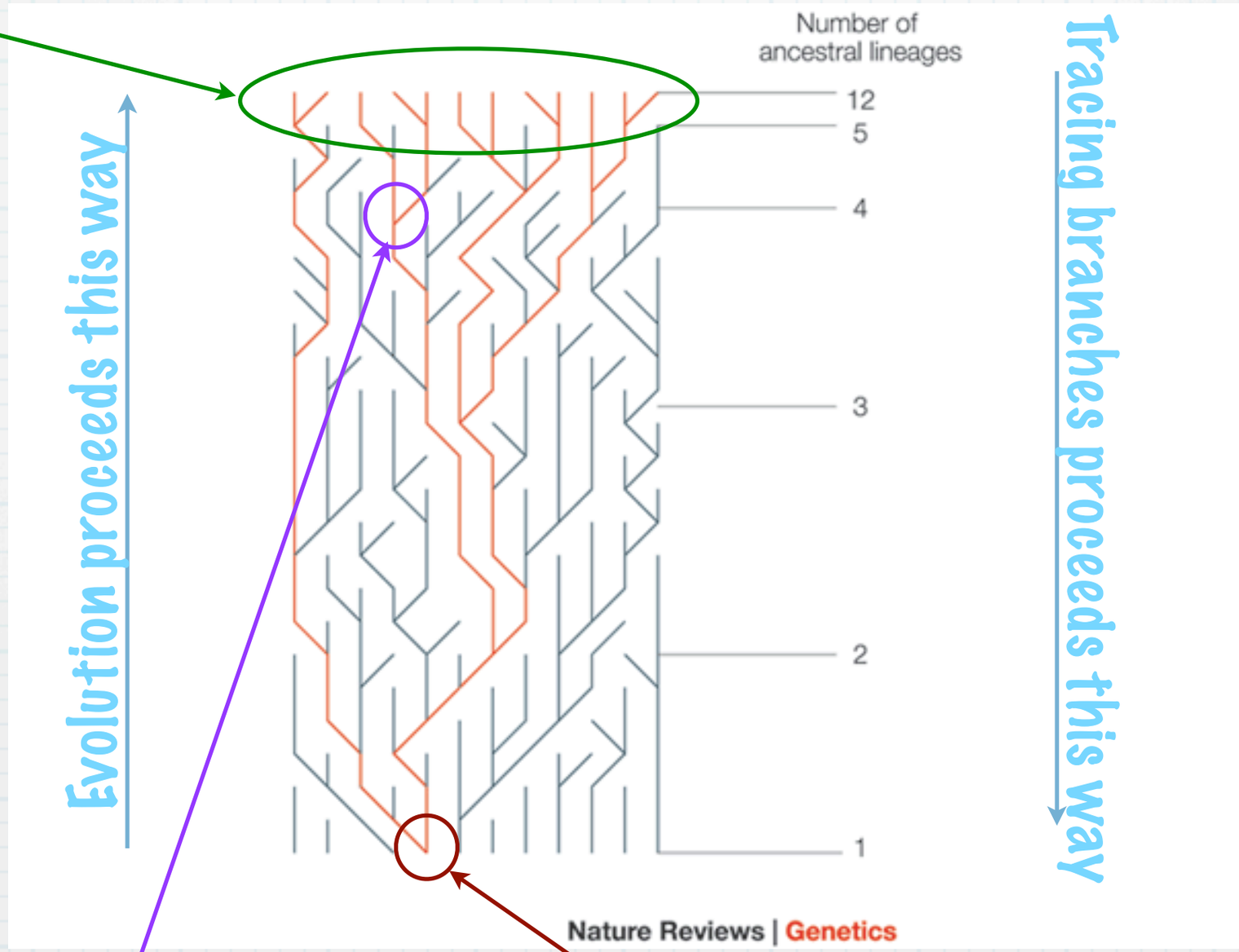
coalescent event

most recent common ancestor (MRCA)

(5) The Coalescent Model

Modeling the branching of lineages

fixation



coalescent event

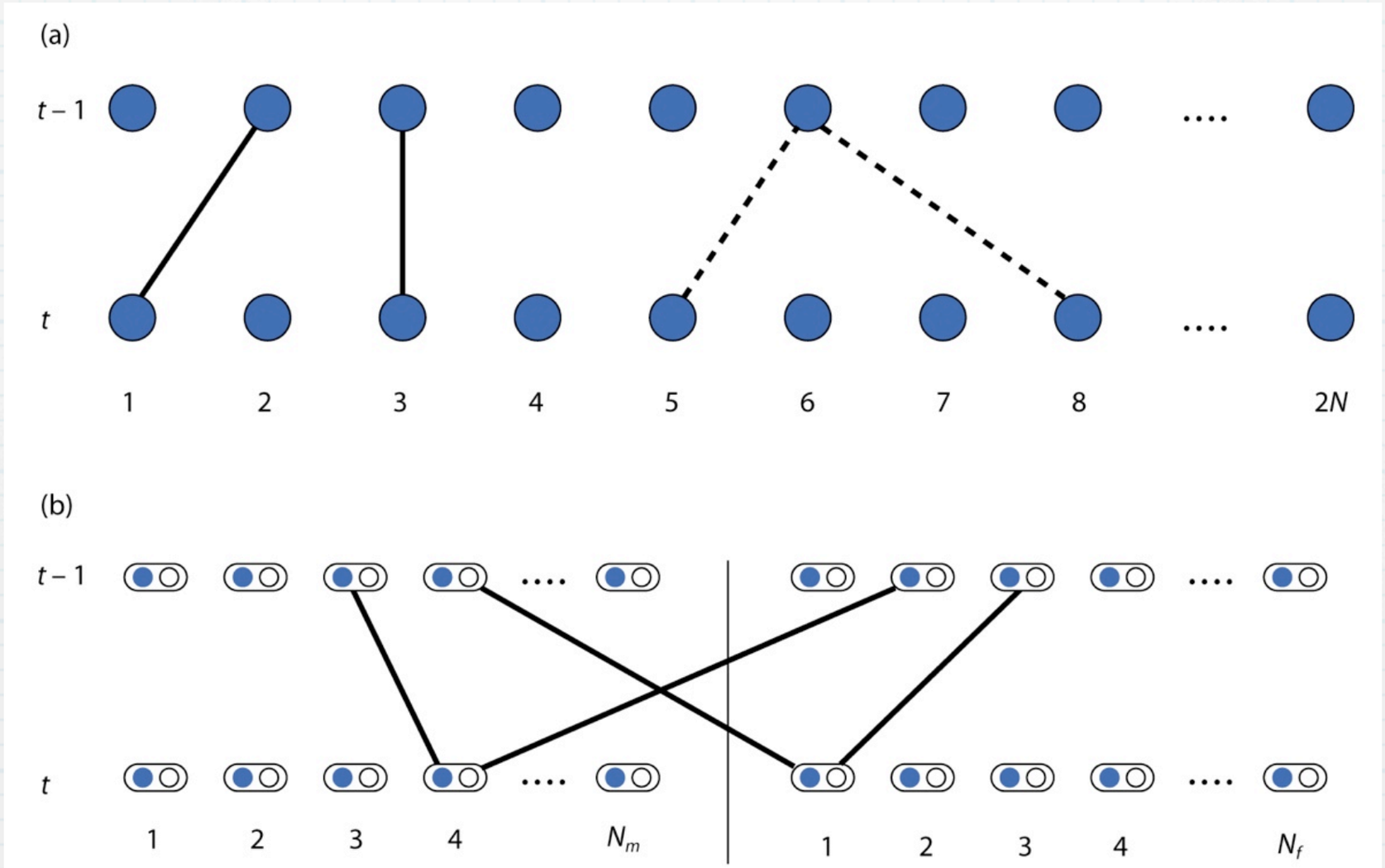
most recent common ancestor (MRCA)

(5) The Coalescent Model

Modeling the branching of lineages

haploid model

diploid model

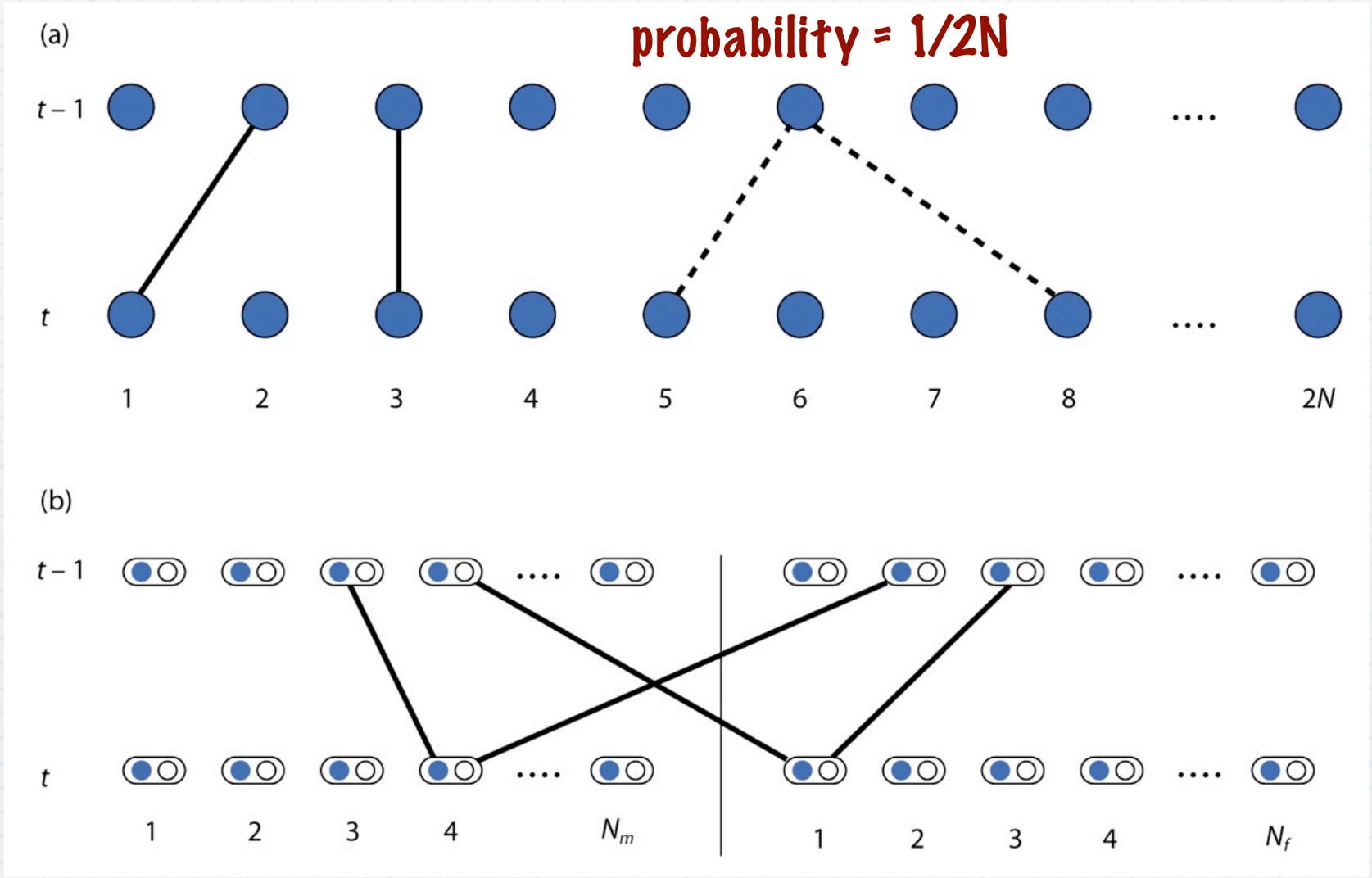


(5) The Coalescent Model

Modeling the branching of lineages

haploid model

diploid model

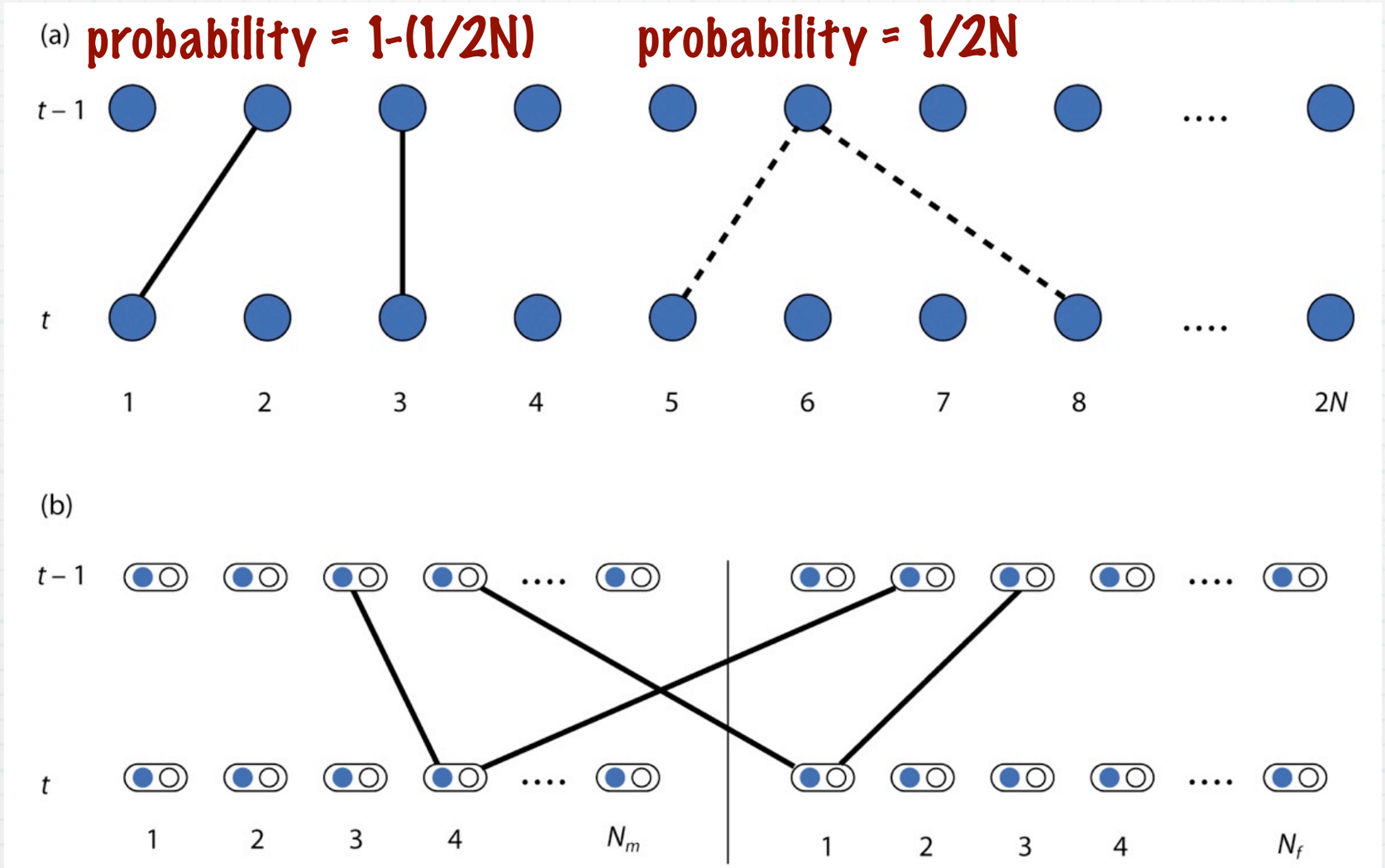


(5) The Coalescent Model

Modeling the branching of lineages

haploid model

diploid model

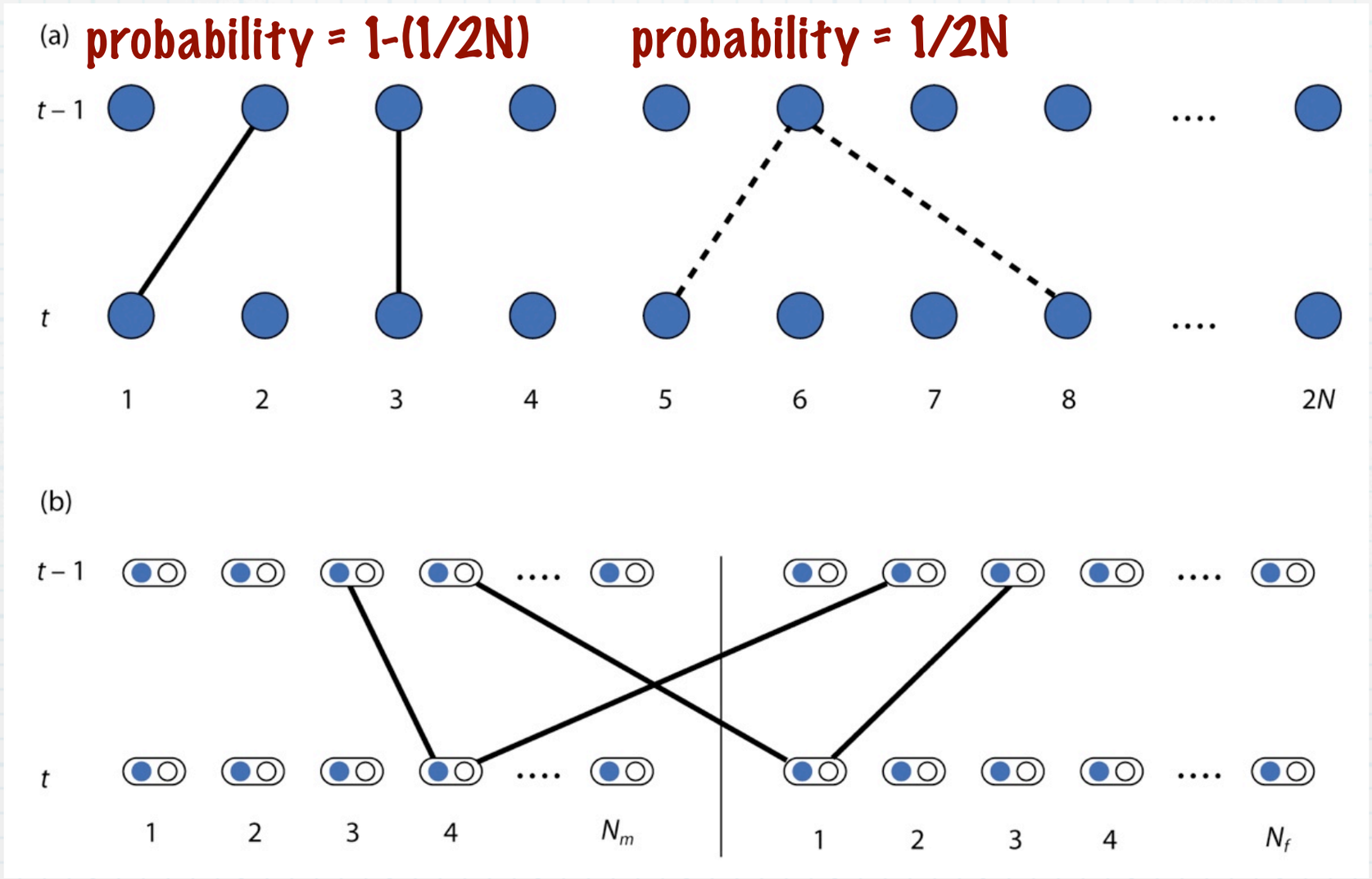


(5) The Coalescent Model

Modeling the branching of lineages

haploid model

diploid model



The haploid model with 2N lineages is routinely used to approximate the diploid model with $N=N_f+N_m$ diploid individuals

(5) The Coalescent Model

Modeling the branching of lineages

The probability that two randomly chosen lineages coalesce after exactly t generations is

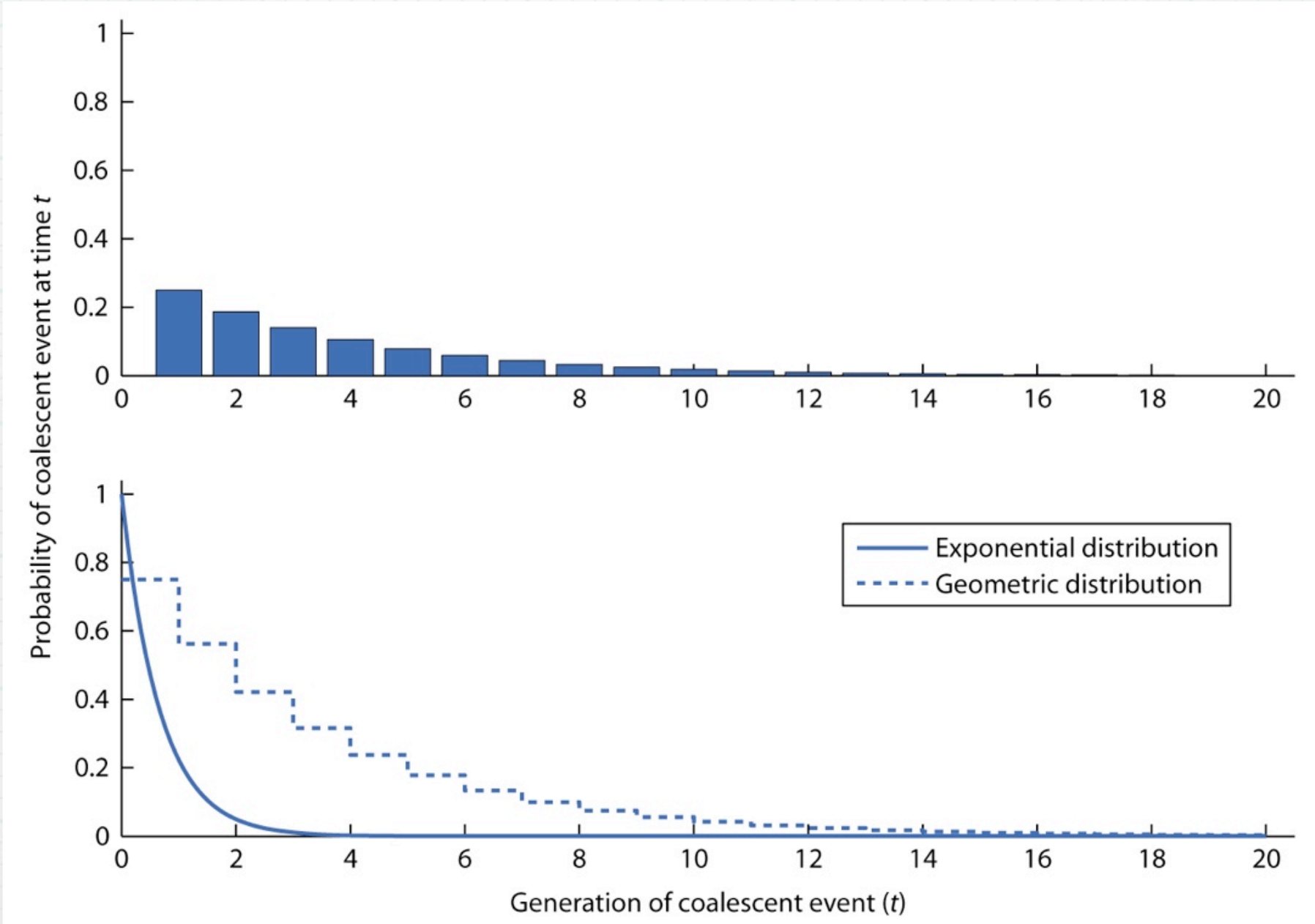
$$\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$$

The cumulative probability that two randomly chosen lineages coalesce at or before generation t is approximated by the exponential distribution

$$P(T_C \leq t) = 1 - e^{-\frac{1}{2N}t}$$

(5) The Coalescent Model

Modeling the branching of lineages



(5) The Coalescent Model

Modeling the branching of lineages

Recall that for exponential distribution, we have:

$$\text{pdf} : \lambda e^{-\lambda x} \qquad \text{cdf} : 1 - e^{-\lambda x} \qquad \mu : \frac{1}{\lambda} \qquad \sigma^2 : \frac{1}{\lambda^2}$$

It follows that the average time to coalescence and the variance in time to coalescence are, respectively

$$\frac{1}{\frac{1}{2N}} = 2N \qquad \text{and} \qquad \frac{1}{\left(\frac{1}{2N}\right)^2} = 4N^2$$

(5) The Coalescent Model

Modeling the branching of lineages

Recall that for exponential distribution, we have:

$$\text{pdf} : \lambda e^{-\lambda x} \qquad \text{cdf} : 1 - e^{-\lambda x} \qquad \mu : \frac{1}{\lambda} \qquad \sigma^2 : \frac{1}{\lambda^2}$$

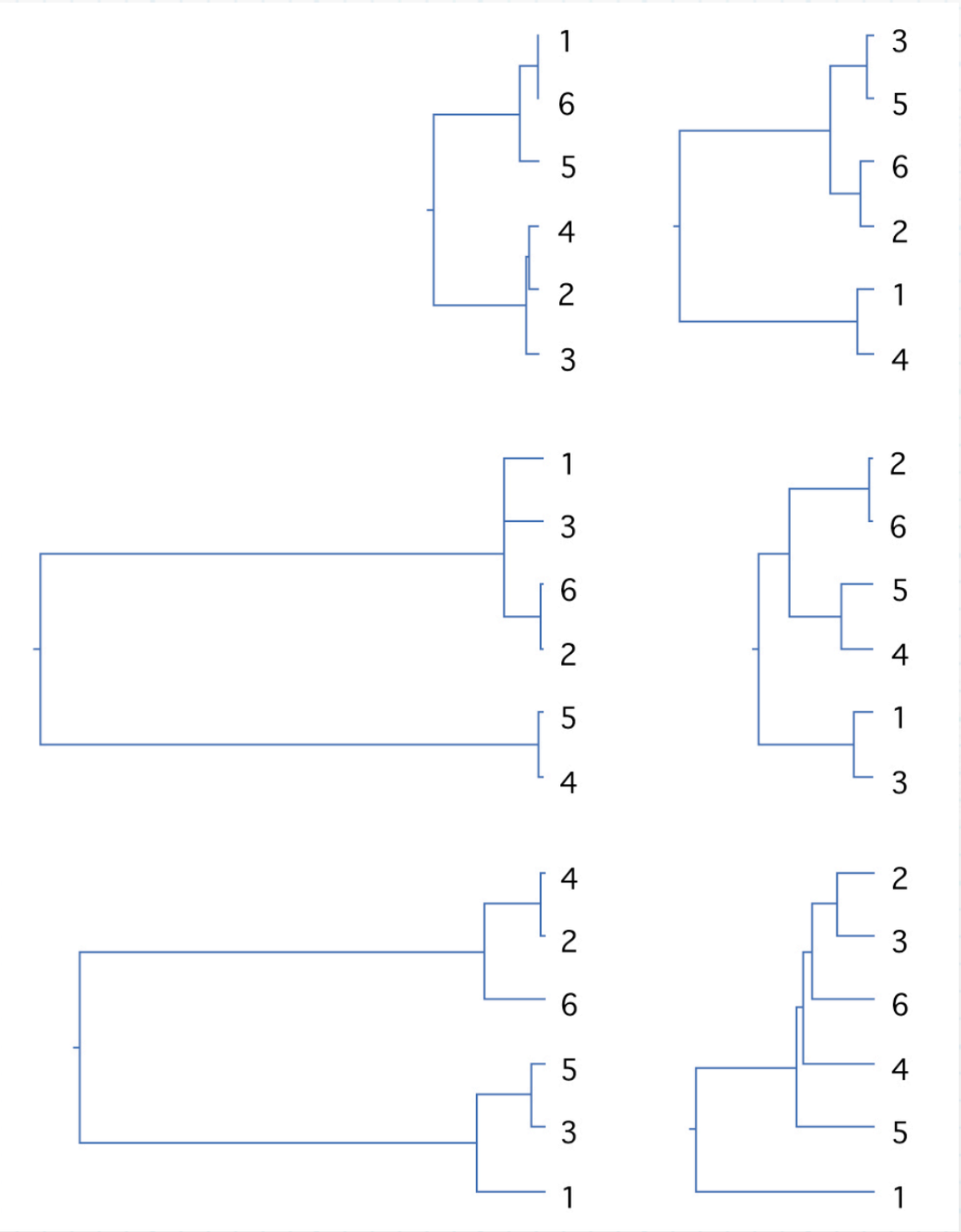
It follows that the average time to coalescence and the variance in time to coalescence are, respectively

$$\frac{1}{\frac{1}{2N}} = 2N \qquad \text{and} \qquad \frac{1}{\left(\frac{1}{2N}\right)^2} = 4N^2$$

$\Rightarrow$ the length of branches connecting lineages to their ancestors will be highly variable around their mean values

(5) The Coalescent Model

Modeling the branching of lineages



(5) The Coalescent Model

Modeling the branching of lineages

The probability of non-coalescence among k lineages is

$$\prod_{x=1}^{k-1}\left(1-\frac{x}{2N}\right)$$

If k<<2N, this can be approximated as

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

The probability of a coalescence for any one of the unique pairs of the k lineages is then

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

(5) The Coalescent Model

Modeling the branching of lineages

⇒ The probability that k lineages experience a single coalescent event t generations ago is

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

Since this probability follows an exponential distribution

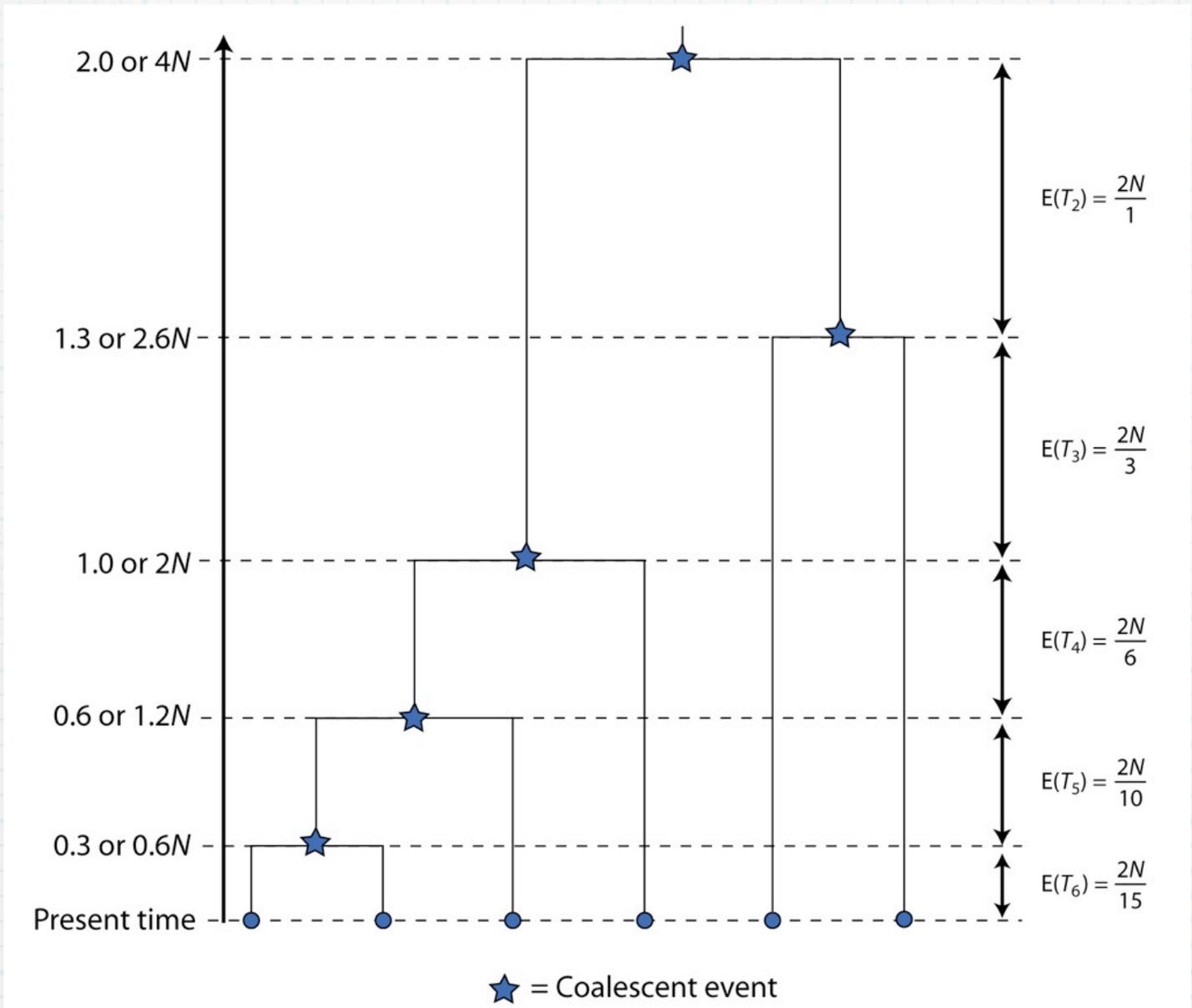
$$e^{-t\left(\frac{k(k-1)}{2}\right)\left(\frac{1}{2N}\right)}$$

the average time to coalescence for k lineages in a population of 2N is

$$\frac{2N}{\frac{k(k-1)}{2}}$$

(5) The Coalescent Model

Modeling the branching of lineages



$E(T_n)$: the expected time to coalescence for n lineages

(5) The Coalescent Model

Modeling the branching of lineages

$$E(T_i) = \frac{2}{i(i-1)} \qquad Var(T_i) = \left(\frac{2}{i(i-1)}\right)^2$$

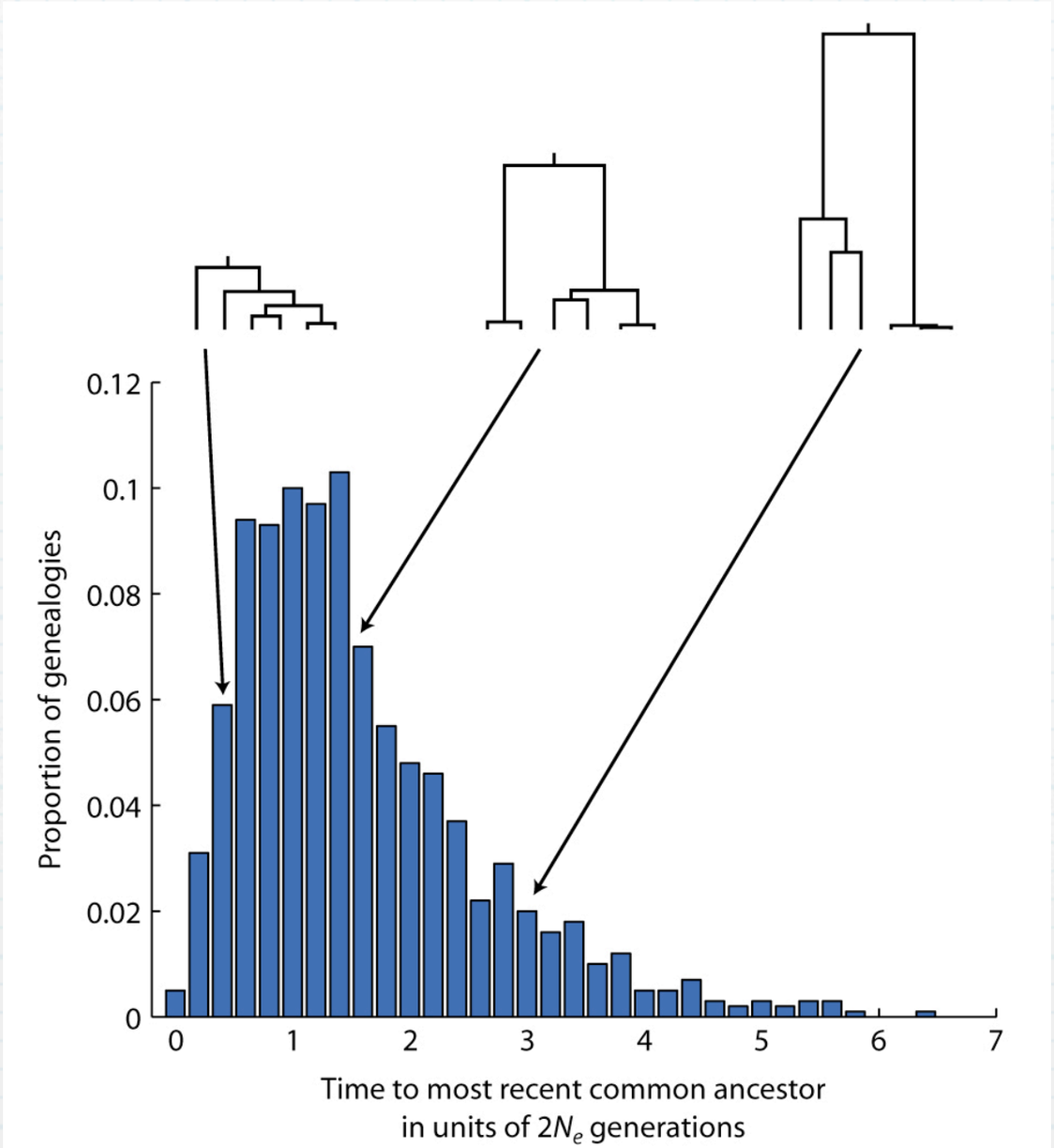
$$T_{MRCA} = E(H_k) = \sum_{i=2}^{k} E(T_i) = 2\sum_{i=2}^{k} \frac{1}{i(i-1)} = 2\left(1 - \frac{1}{k}\right)$$

$$T_{total} = \sum_{i=2}^{k} iE(T_i) = 2\sum_{i=2}^{k} \frac{i}{i(i-1)} = 2\sum_{i=1}^{k-1} \frac{1}{i}$$

$$Var(T_{MRCA}) = 4\sum_{i=2}^{k} \frac{1}{i^2(i-1)^2} \qquad Var(T_{total}) = 4\sum_{i=1}^{k-1} \frac{1}{i^2}$$

(5) The Coalescent Model

Modeling the branching of lineages



The distribution of T_{MRCA} for 1000 replicate genealogies starting with six lineages. N_e=1000.

(5) The Coalescent Model

The $g_{n,k}(t)$ function

$g_{n,k}(t)$: the probability that n lineages coalesce into k lineages within time t (i.e., that n-k coalescent events occur before time t in the past

$$g_{n,k}(t) = \begin{cases} 1 - \sum_{i=2}^{n} \frac{e^{-\binom{i}{2}t}(2i-1)(-1)^i n_{[i]}}{n_{(i)}} & \text{if } k = 1 \\ \sum_{i=k}^{n} \frac{e^{-\binom{i}{2}t}(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{i!(i-k)! n_{(i)}} & \text{if } k \geq 2 \end{cases}$$

$$n_{[i]} = n(n-1)\cdots(n-i+1)$$

$$n_{(i)} = n(n+1)\cdots(n+i-1)$$

Summary

* In finite populations, allele frequencies can change from generation to generation since the sample of gametes that found the next generation may not contain exactly the same number of each allele as the previous generation.
* Sampling error in allele frequency causes genetic drift, the random process whereby all alleles eventually reach fixation or loss.
* The Wright-Fisher model makes assumptions identical to HW in addition to assuming that each generation is founded by a sample of 2N gametes from an infinite pool of gametes.
* The action of genetic drift can be modeled by the binomial distribution, a Markov chain model, and the diffusion approximation.

Summary

* The effective population size (N_e) is the size of an ideal Wright-Fisher population that shows the same frequency behavior over time as an observed biological population regardless of its census population.

* Finite population size and consanguineous mating are analogous processes, since they both lead to increasing homozygosity and decreasing heterozygosity. The distinction is that genetic drift in finite populations causes changes in both genotype and allele frequencies while consanguineous mating changes only genotype frequencies.

* The coalescent models the branching of lineages to predict the time to the MRCA.