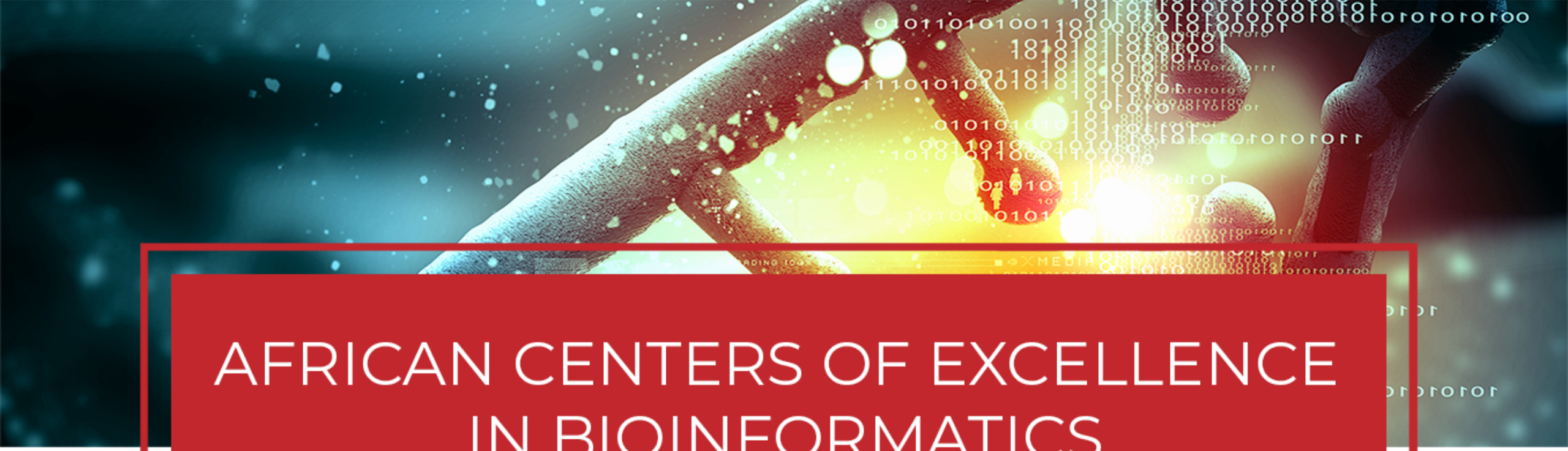
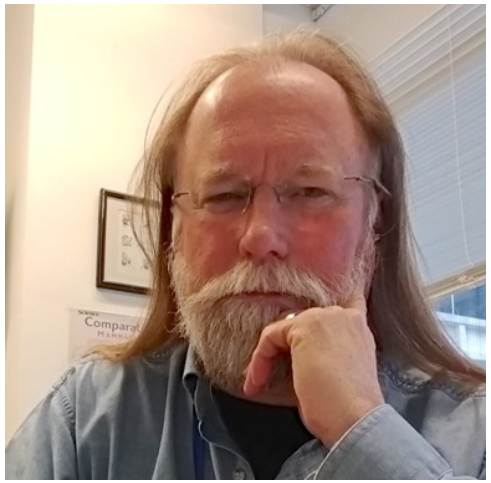


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

HARDY-WEINBERG EQUILIBRIUM, BASIC POPULATION GENETICS, AND COALESCENT MODELS

Today's Instructor



Dr. Kurt Wollenberg, Ph.D. in Genetics

Ongoing Computational Biology projects:

- Hepatitis B molecular evolution
- CLAG protein family evolution

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

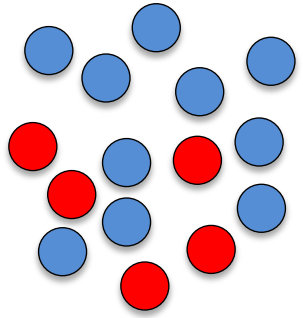
Class Materials

- Directory on Uganda ACE server:
  - File directory: user@kla-ac-bio-03:/home/bcbb_teaching_files
  - Large data files
- NIAID github repository:
  - https://github.com/niaid/ACE-2020
  - Code
  - Data files
  - Copies of lecture slides

ALLELE FREQUENCY

15 *haploid* organisms

5
10



● = …ATGAACCGATACAGG…

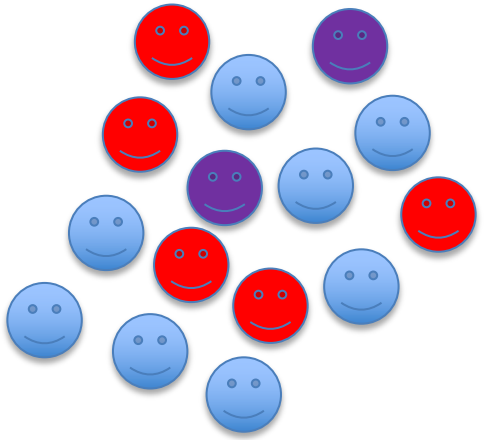
● = …ATGAACGGATACAGG…

Population size (N) = 15
Blue allele (n_1) = 10
Red allele (n_2) = 5
Blue allele frequency (p) = 10/15 = 0.67
Red allele frequency (q) = 5/15 = 0.33

ALLELE FREQUENCY

15 *diploid* organisms

8
2
5



…ATGAACCGATACAGG…
…ATGAACCGATACAGG…

…ATGAACCGATACAGG…
…ATGAACGGATACAGG…

…ATGAACGGATACAGG…
…ATGAACGGATACAGG…

Population size N = 15
Blue genotype (A_1A_1) = 8
Purple genotype (A_1A_2) = 2
Red genotype (A_2A_2) = 5

Number of gene copies ($2N$) = 30
Blue allele ($2\ A_1A_1 + A_1A_2$) = 16 + 2 = 18
Red allele ($2A_2A_2 + A_1A_2$) = 10 + 2 = 12
Blue allele frequency (p) = 18/30 = 0.6
Red allele frequency (q) = 12/30 = 0.4

GENETIC DIVERSITY

HETEROZYGOSITY

Consider a diploid individual

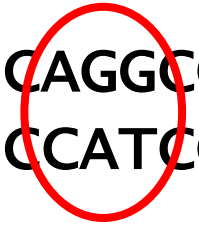
Locus 1	Locus 2	Locus 3
...ATGACCGATCAGG...	...CATATAAGGCTAGTCT...	...GGCCTATCGTAAACT...
...ATGACGGATCAGG...	...CATATAAGGCAACTCT...	...GTCCTAGGGTAACCT...

GENETIC DIVERSITY

Heterozygous loci within a diploid individual

Locus 1	Locus 2	Locus 3	Locus 4
…ATGACCGAT…	…ATAAGGCTAGTCT…	…TATCGTAAACT…	…GACGTAGT…
…ATGACGGAT…	…ATAAGGCAACTCT…	…TAGGGTAACCT…	…GACGTAGT…

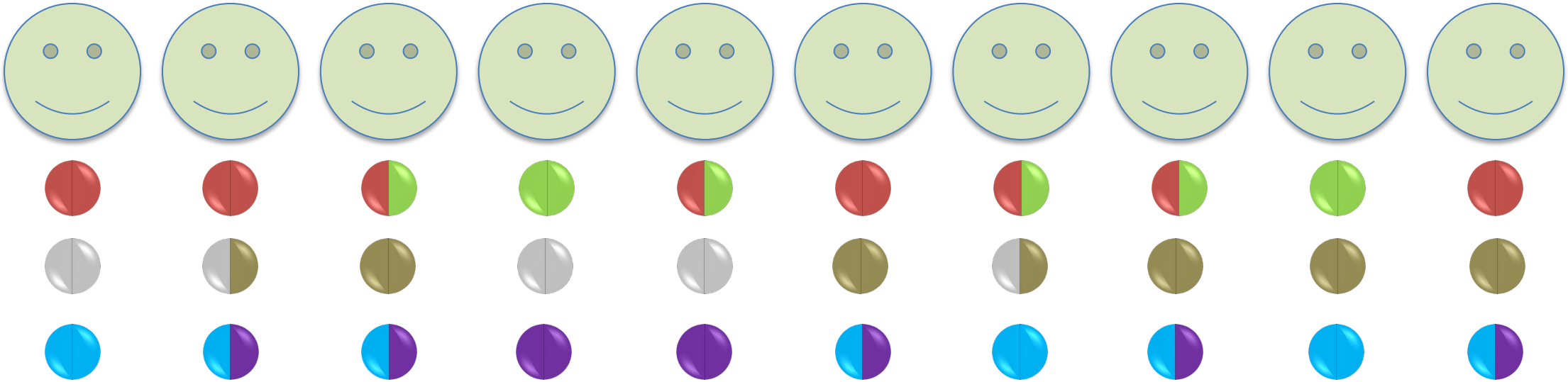
Locus 5	Locus 6	Locus 7	Locus 8
…AGGACGTTAT…	…GGCAGGCG…	…GCAAGAG…	…AGCCTATAG…
…AGGACGTTAT…	…GGCCATCG…	…GTAATCG…	…AGCCTATAG…



Counting loci: 5 heterozygous loci out of 8, frequency = 0.625

GENETIC DIVERSITY

Heterozygous loci within a population of diploid individuals



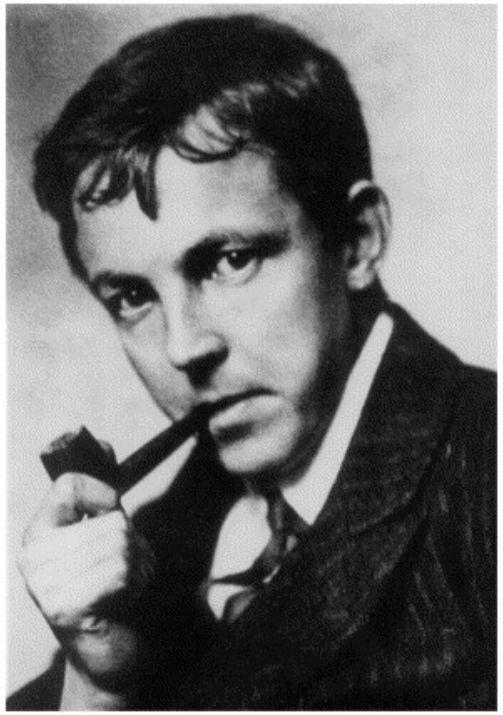
Population size = 10

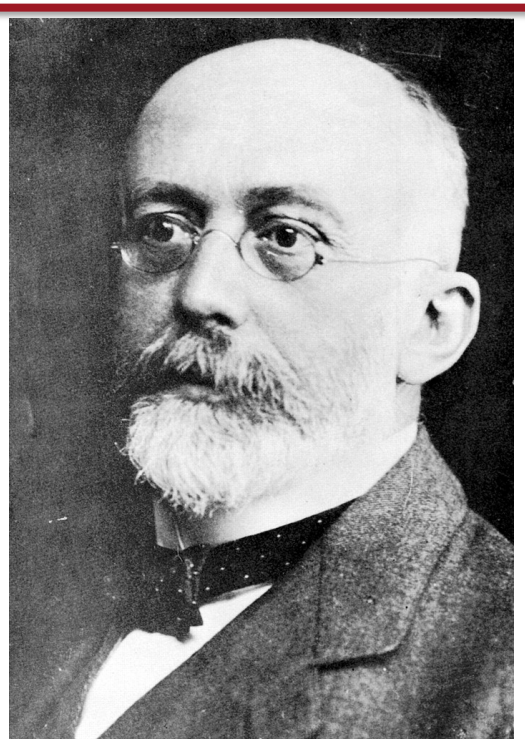
Locus 1: 6 homozygous, 4 heterozygous

Locus 2: 8 homozygous, 2 heterozygous

Locus 3: 5 homozygous, 5 heterozygous

HARDY-WEINBERG EQUILIBRIUM

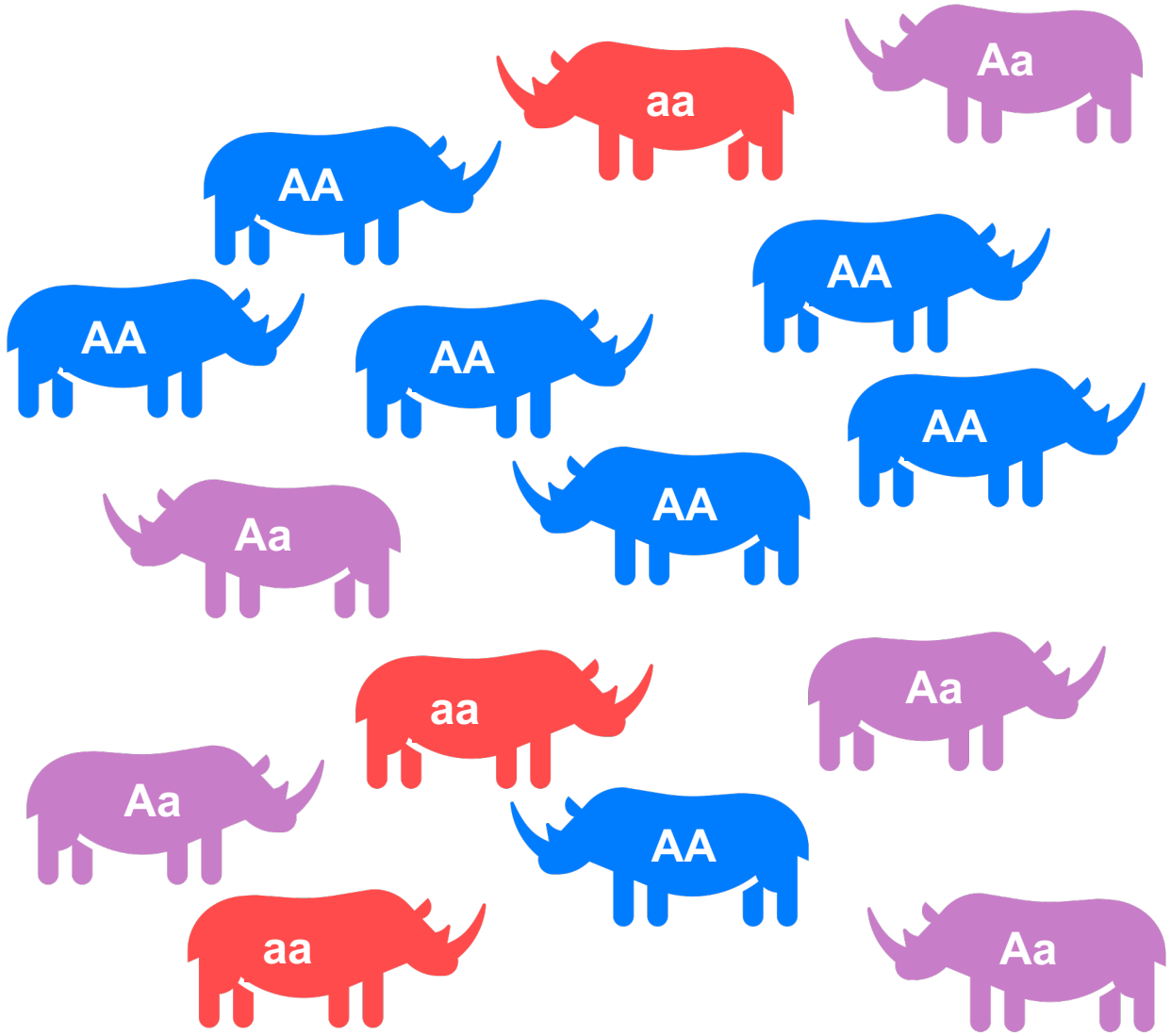




In 1908 G. H. Hardy and W. Weinberg independently derived the result that, for ideal, randomly mating populations of infinite size with no mutation, migration (gene flow), selection, or miotic drive, allele frequencies do not change across generations and for diploid organisms follow a multinomial (binomial for two alleles) distribution.

HARDY-WEINBERG EQUILIBRIUM

- Consider a diploid, randomly mating population
- Phenotypes are blue, red, and purple
- Color controlled by a single locus
- There are two alleles at this locus
- Genotypes are AA, Aa, and aa
- Frequency of allele **A** is *p*.
- Frequency of allele **a** is *q* = 1-*p*.



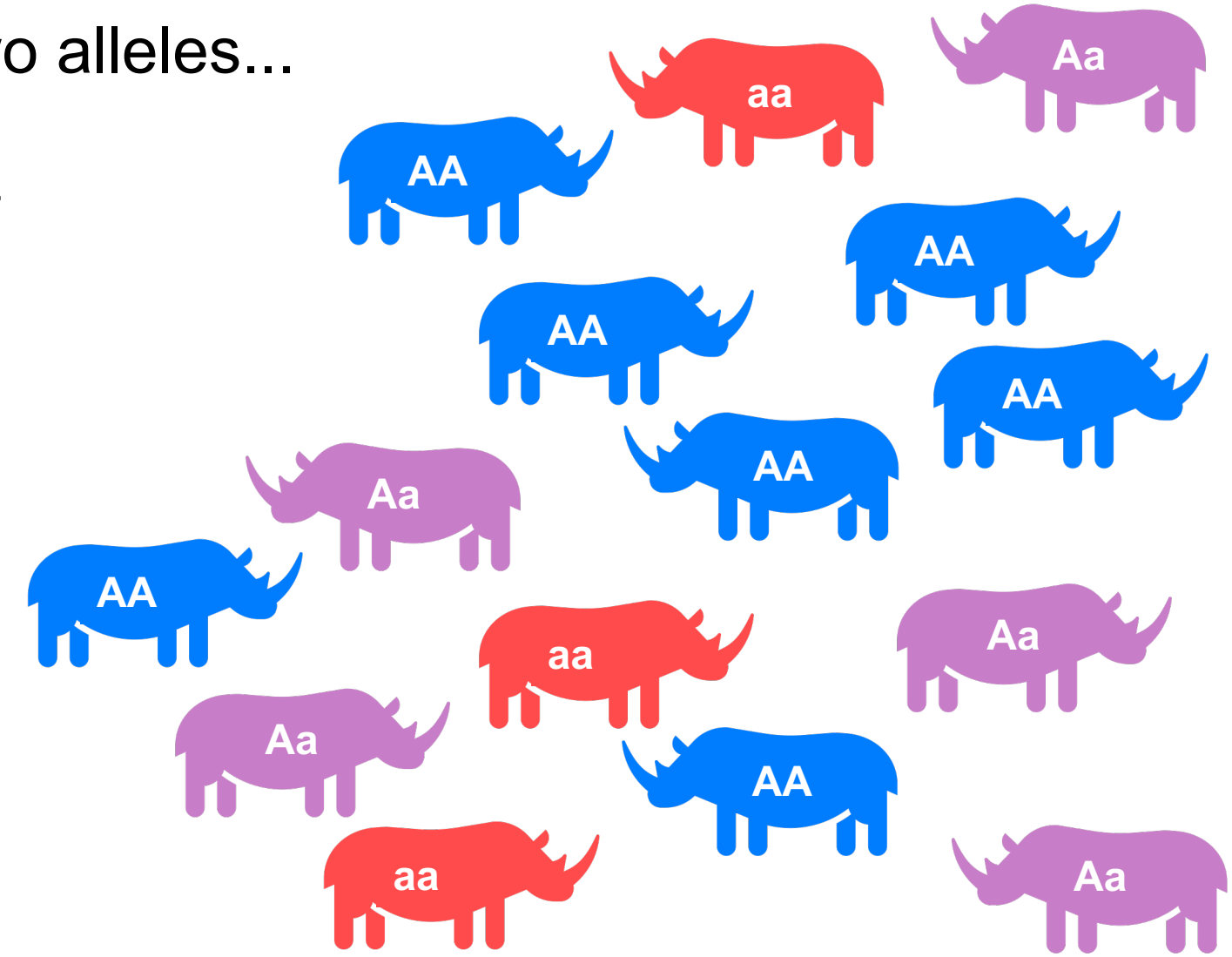
HARDY-WEINBERG EQUILIBRIUM

For our diploid locus with two alleles...

AA count = 7, f(AA) = 7/15 = 0.47
Aa count = 5, f(Aa) = 5/15 = 0.33
aa count = 3, f(aa) = 3/15 = 0.20

A = (2x7)+5 = 19
a = (2x3)+5 = 11

2*N* genes = 2x15 = 30
p = 19/30 = 0.63
q = 11/30 = 0.37



HARDY-WEINBERG EQUILIBRIUM

For large gamete pools in randomly mating populations all gamete fusion events will be independent.

Therefore, the probability of two gametes combining is the product of the frequency of each type of gamete.

$P(AA) = p \times p = p^2$ $P(Aa) = p \times q$ $P(aa) = q \times q = q^2$

HARDY-WEINBERG EQUILIBRIUM

For our ideal population, the offspring in the following generation will have the genotypes AA, Aa, aA, aa

In terms of gamete probabilities, offspring frequencies are

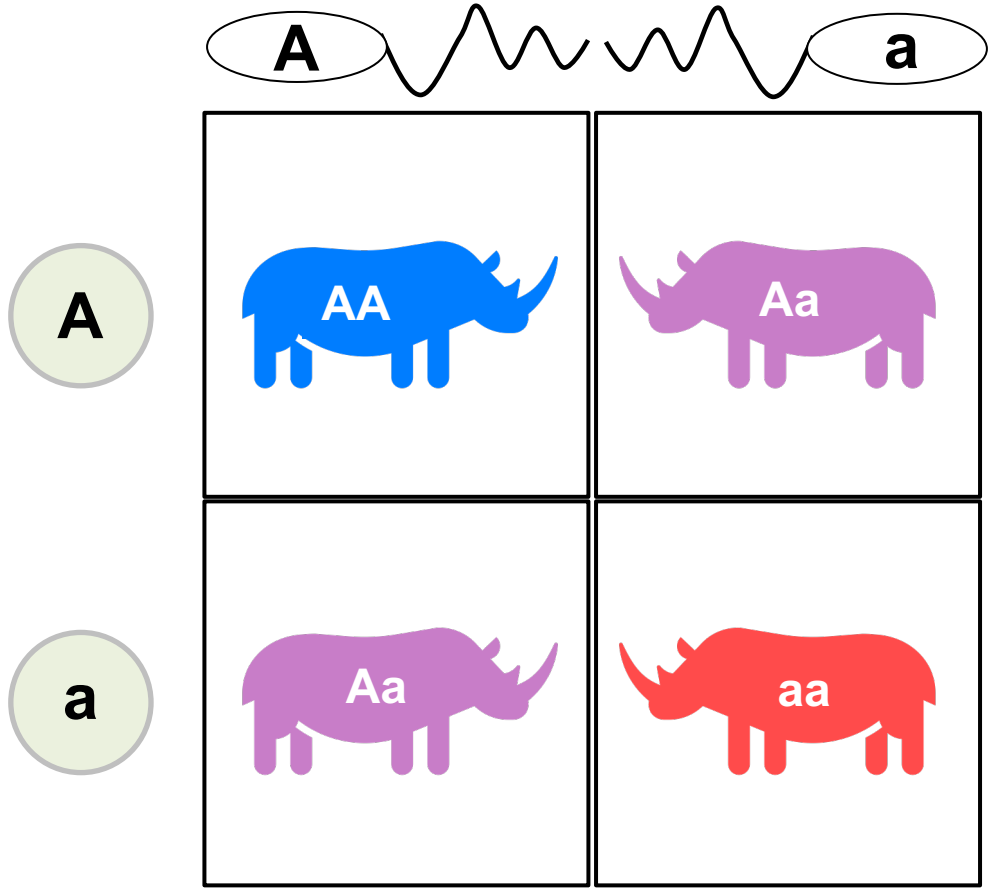
$f(AA) = p^2$

$f(Aa) = 2pq$

$f(aa) = q^2$

Total population frequency is

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



HARDY-WEINBERG EQUILIBRIUM

What if there are more than two alleles?

For diploid organisms ($2N$ genes for population size N)

$$(p + q + r + s + t + \ldots)^2 = 1$$

For example, with 4 alleles with frequencies p, q, r, s genotype frequencies are

$$p^2 + 2pq + 2pr + 2ps + 2qr + 2qs + 2rs + q^2 + r^2 + s^2 = 1$$

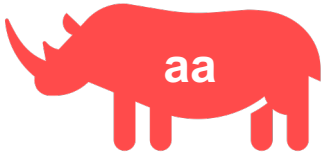
HARDY-WEINBERG EQUILIBRIUM

Hardy-Weinberg Equilibrium describes a population in terms of allele frequencies *p*, *q*, etc.

For this ideal population the genotype frequencies (P_{AA}, P_{Aa}, P_{aa}, etc.) can change but, if equilibrium conditions hold, they will return to the binomial frequencies in the next generation.

HARDY-WEINBERG EQUILIBRIUM

10,000 *diploid* organisms



Genotype counts: *AA* = 6000 *aa* = 4000

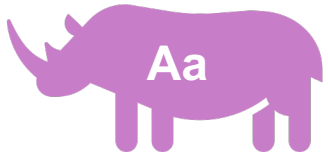
Number of gene copies (2*N*) = 20,000
Blue allele: 2 *AA* = 12,000
Red allele: 2 *aa* = 8,000
Blue allele frequency (*p*) = 12,000/20,000 = 0.6
Red allele frequency (*q*) = 8,000/20,000 = 0.4

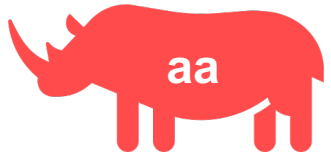
HARDY-WEINBERG EQUILIBRIUM

10,000 *diploid* organisms – the next generation

Blue allele frequency (p) = 0.6 Red allele frequency (q) = 0.4

Under Hardy-Weinberg Equilibrium in the next generation the genotype frequencies will be: $P_{AA} = p^2 = 0.36$, $P_{Aa} = 2pq = 0.48$, $P_{aa} = q^2 = 0.16$





Genotype counts: 3600 4800 1600

HARDY-WEINBERG EQUILIBRIUM

Is a population in Hardy-Weinberg Equilibrium?

Test for agreement with Hardy-Weinberg using the chi-squared test

Observed genotype frequencies = the data we've collected = *O*

Expected genotype frequencies = binomial distribution based on allele frequencies = *E*

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of Freedom = # of genotypes - # of alleles

HARDY-WEINBERG EQUILIBRIUM

Is this population in Hardy-Weinberg Equilibrium?

Genotype counts: *AA* = 6000 *aa* = 4000

Allele frequency (*p*) = 12,000/20,000 = 0.6

Genotype counts, expected: *AA* = 10000 x $(0.6)^2$ = 3600

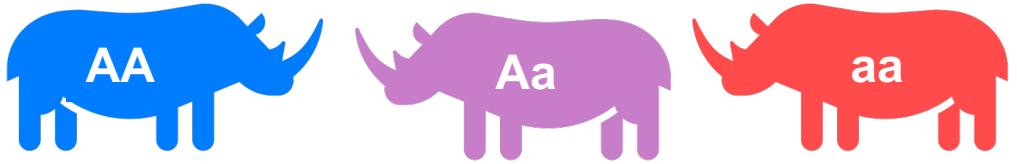
Aa = 10000 x 2(0.24) = 4800

aa = 10000 x $(0.4)^2$ = 1600

$$\chi^2 = \frac{(6000-3600)^2}{3600} + \frac{(0-4800)^2}{4800} + \frac{(4000-1600)^2}{1600} = 1600 + 4800 + 3600 = 10000 > 3.841$$

HARDY-WEINBERG EQUILIBRIUM

Is this population in Hardy-Weinberg Equilibrium?



Genotype counts: *AA* = 5000 *Aa* = 2000 *aa* = 3000

Allele frequency (p) = 12,000/20,000 = 0.6

Genotype counts, expected: *AA* = 10000 x $(0.6)^2$ = 3600

Aa = 10000 x 2(0.24) = 4800

aa = 10000 x $(0.4)^2$ = 1600

$$\chi^2 = \frac{(5000-3600)^2}{3600} + \frac{(2000-4800)^2}{4800} + \frac{(3000-1600)^2}{1600} = 544.4 + 1633.3 + 1225 = 3402.7$$

HARDY-WEINBERG EQUILIBRIUM

Is this population in Hardy-Weinberg Equilibrium?

Allele frequency (p) = 0.6

Genotype counts, expected (N = 10,000) : AA = 3600 Aa = 4800 aa = 1600

N	AA obs	Aa obs	aa obs	p	χ^2	Pr (χ^2 df=1)
10,000	6000	0	4000	0.6	10,000	0
10,000	5000	2000	3000	0.6	3403	0
10,000	3625	4750	1625	0.6	1.08	0.30
1,000	500	200	300	0.6	340.3	5.55×10^{-76}
100	50	20	30	0.6	34.03	5.43×10^{-9}
100	40	40	20	0.6	2.78	0.096

HARDY-WEINBERG EQUILIBRIUM

Deviations from Hardy-Weinberg: Changes in allele frequency over time

Deterministic causes

Mutation

Migration

Meiotoic Drive

Selection

Non-random mating

Stochastic Causes

Finite Population Size

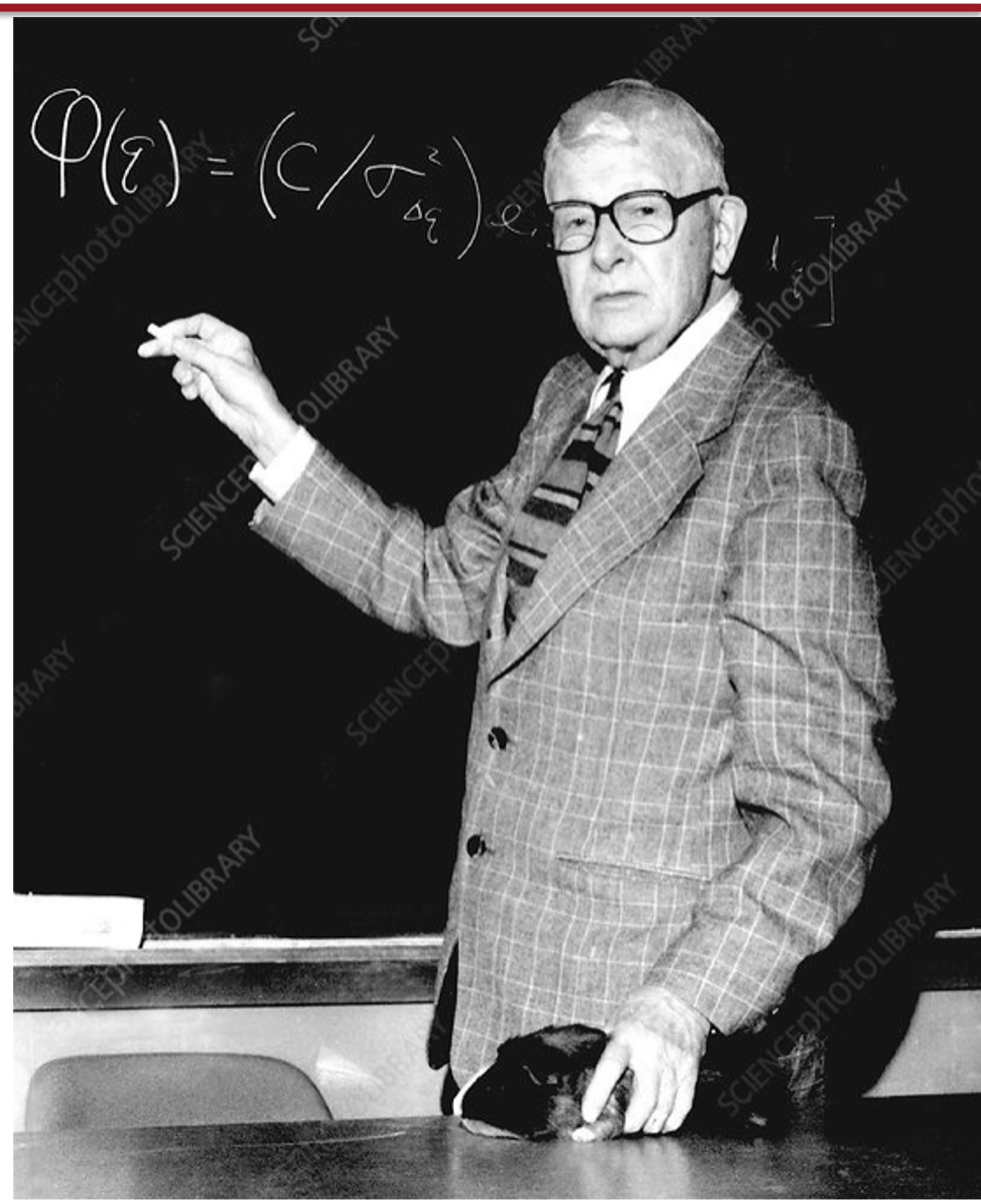
Sampling of Gametes

Genotyping Errors

POPULATION GENETICS

Effective Population Size – N_e

Defined by Sewall Wright in 1931 to be the size of an idealized population (one with no mutation, migration, selection, or other factors biasing allele transmission) that would have the same effect of random sampling on gene frequency as in the actual population being studied.



EFFECTIVE POPULATION SIZE

Variance in allele frequency from one generation to the next is proportional to the inverse of the population size.

For an ideal population: $V(p_{t+1}) = p_t(1-p_t)/(2N)$

Phenomena that violate the assumptions of Hardy-Weinberg equilibrium will generally increase allele frequency variance in the next generation. The effective population size N_e is the size of a population that leads to the next-generation allele frequency variance expected under HWE.

EFFECTIVE POPULATION SIZE

What can lead to $N \neq N_e$?

Overlapping generations: leads to some individuals being pre- and postreproductive, giving an effective population size less than the census population size.

Skewed mating systems: Polygamy/polyandry such that not every individual of one gender is able to reproduce. $N_e = 4N_m N_f / N_m + N_f$

Population bottlenecks: long-term effective population size is the harmonic mean of historical population sizes. $N_e = n/(1/N_1 + 1/N_2 + ... + 1/N_n)$

POPULATION GENETICS

Other effect of finite population size: Inbreeding

Inbreeding: an increase in the probability that two copies of an allele are identical due to common ancestry.

This probability is described by *f*, the inbreeding coefficient.

INBREEDING

Assume a large pool of alleles from which we will draw two alleles to make an offspring.

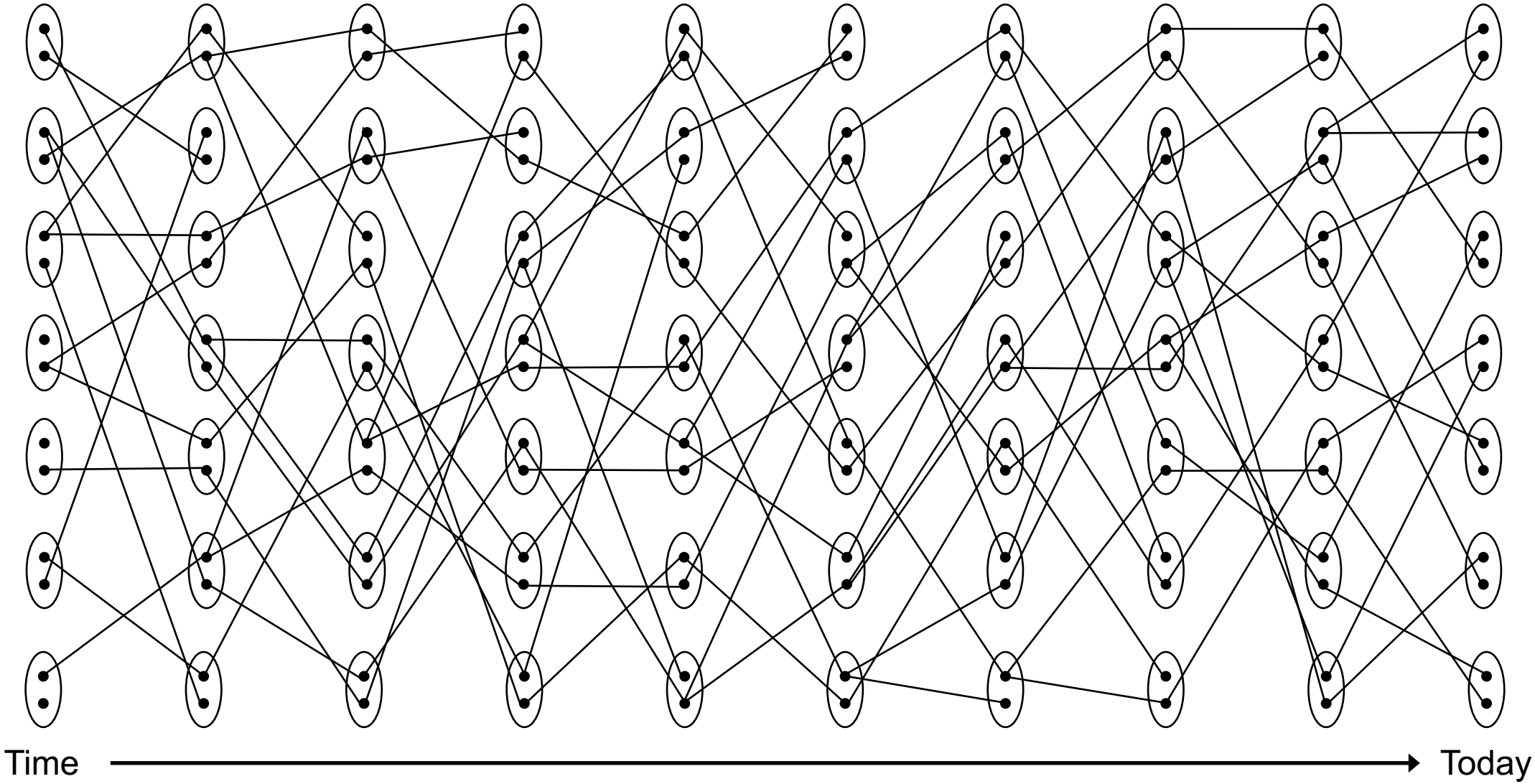
- The first allele is drawn with probability *p*, its frequency in the pool.
- The inbreeding coefficient *f* is the probability that the next allele drawn is identical by descent (ibd).
- By this definition, a second identical allele not identical by descent will be drawn with probability (1-*f*).
- This gives two possible draws for homozygotes with the probabilities *pf* (ibd) and $p^2(1-f)$ (not ibd).

INBREEDING

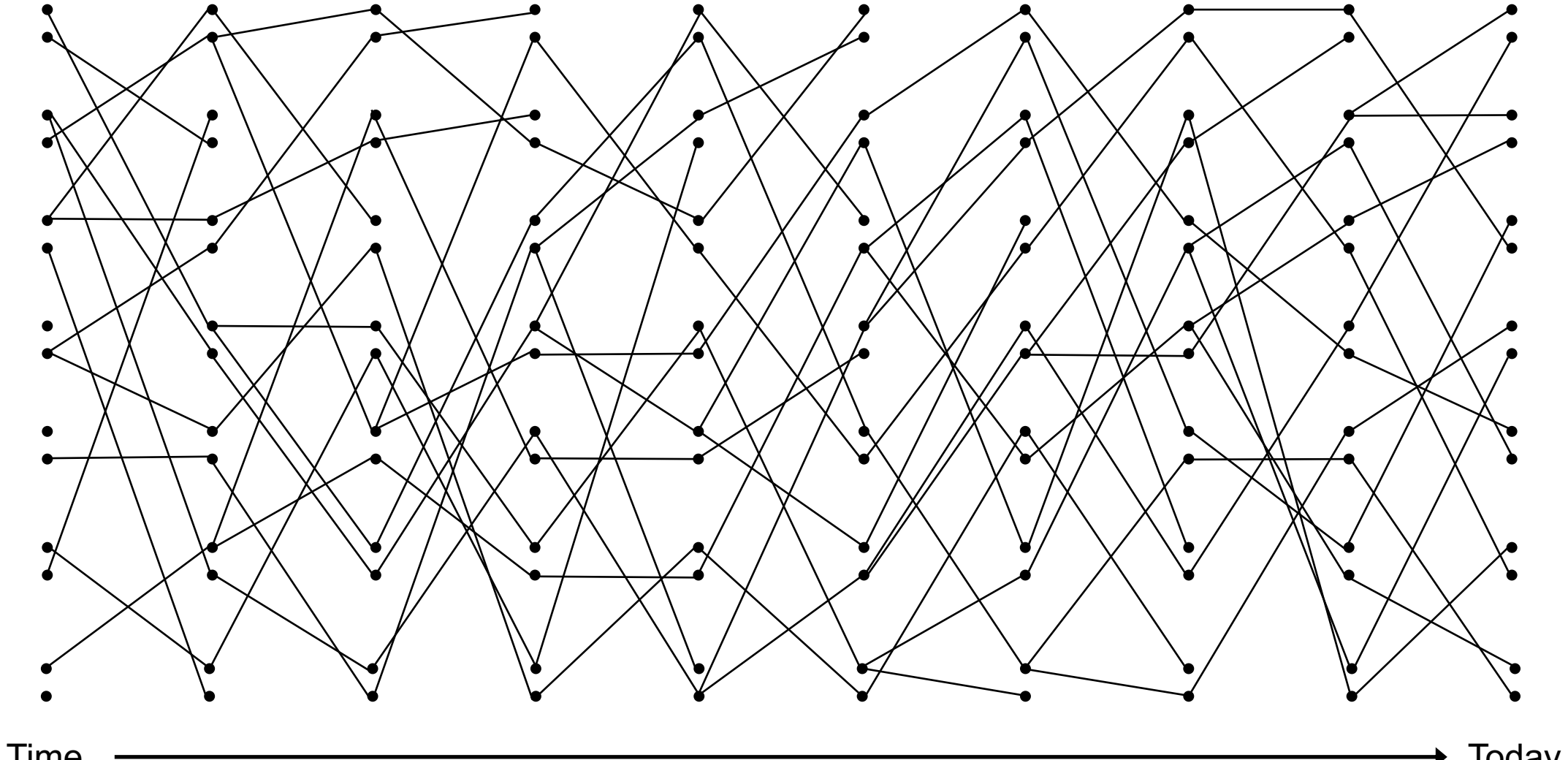
Assume a large pool of alleles from which we will draw two alleles to make an offspring.

- Genotype AA frequency is $P_{AA} = pf + p^2(1-f) = p^2 + fpq$
- Genotype Aa frequency is $P_{Aa} = 0f + 2pq(1-f) = 2pq(1-f)$
- Genotype aa frequency is $P_{aa} = q^2 + fpq$
- When inbreeding coefficient f is positive there is a deficit of heterozygotes proportional to $(1-f)$.
- Inbreeding affects only genotype frequencies, not allele frequencies.

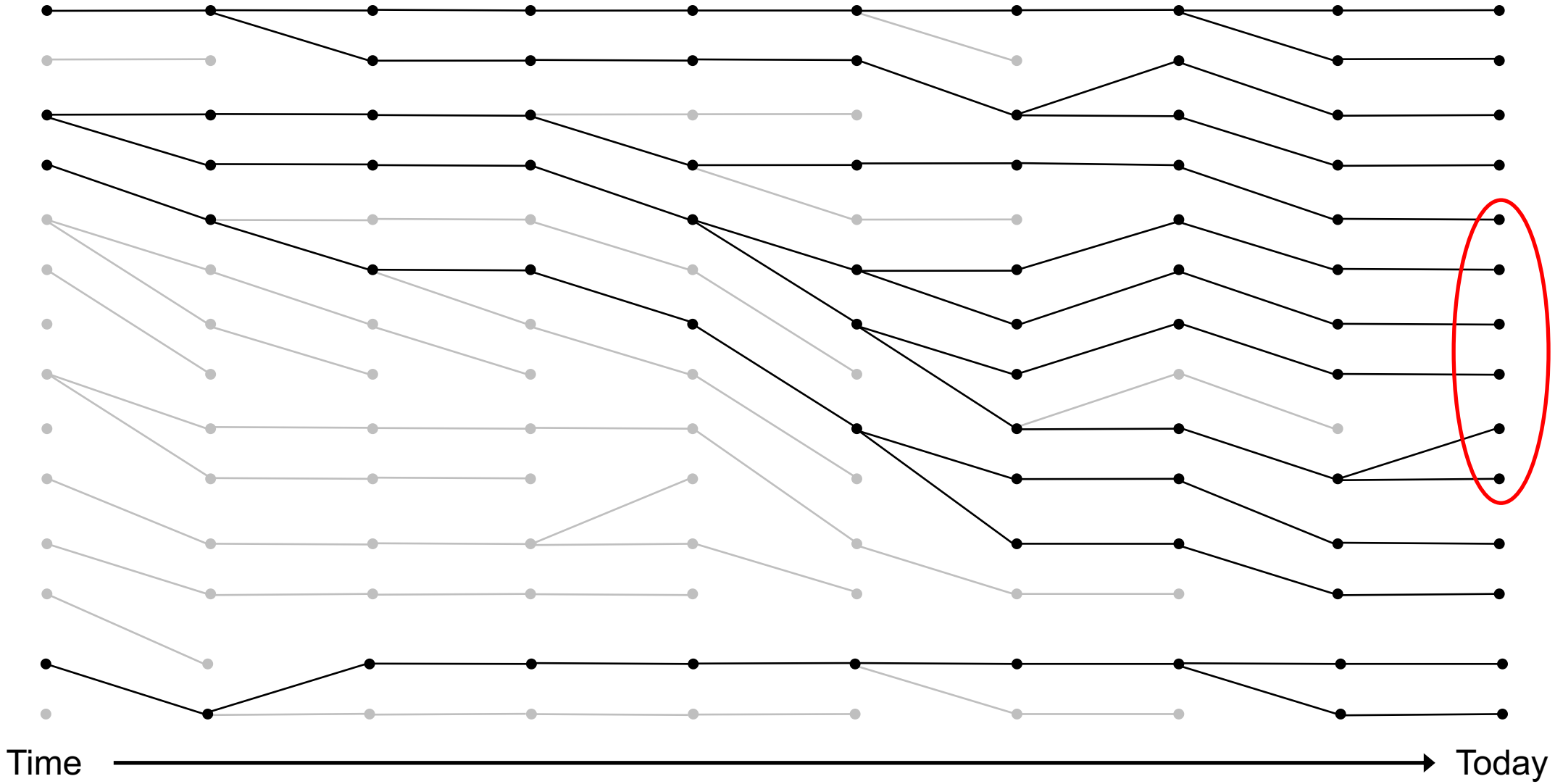
COALESCENT MODELS



COALESCENT MODELS



COALESCENT MODELS



COALESCENT MODELS

- Gene genealogy: the relationships among members of a set of nonrecombining genetic elements not subject to selection on genotype.
- Kingman (1982)/Tajima (1983): a population consists of 2N gene copies each with distinct genealogical histories.
- A coalescent approach examines data at the present time and models its behavior in the past.
- At generation *t* the state of the data in the previous generation (*t*+1) only depends on the states at generation *t*. This is, by definition, a Markov process.

COALESCENT MODELS

- A coalescent event: two lineages in generation *t* have a common ancestor (coalesce) in generation *t*+1.
- For a diploid population with 2N alleles present in each generation the probability that two alleles in generation *t* have the same ancestor in generation *t*+1 is $P_C = \frac{1}{2N}$
- The probability these two alleles do not coalesce is $P_{NC} = 1 - \frac{1}{2N}$
- The probability that two alleles have not coalesced over *t* generations and then do at generation *t*+1 is $P_{C,t+1} = (1 - \frac{1}{2N})^t \frac{1}{2N}$

COALESCENT MODELS

- The probability that two alleles have not coalesced over *t* generations and then do at generation *t*+1 is $P_{C,t+1} = (1 - \frac{1}{2N})^t \frac{1}{2N}$
- For reasonably large values of 2N (>100) this can be approximated as $P_{C,t+1} = \frac{1}{2N} e^{-\frac{t}{2N}}$
- For large values of t this approximates an exponential distribution, giving E(Time to coalescence) = 2*N* generations and Var(T) = $4N^2$

COALESCENT MODELS

Probability each of *k* gene copies comes from a different ancestor is

$$P_{NC,t+1} = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right)\left(1 - \frac{3}{2N}\right)\ldots\left(1 - \frac{k-1}{2N}\right)$$

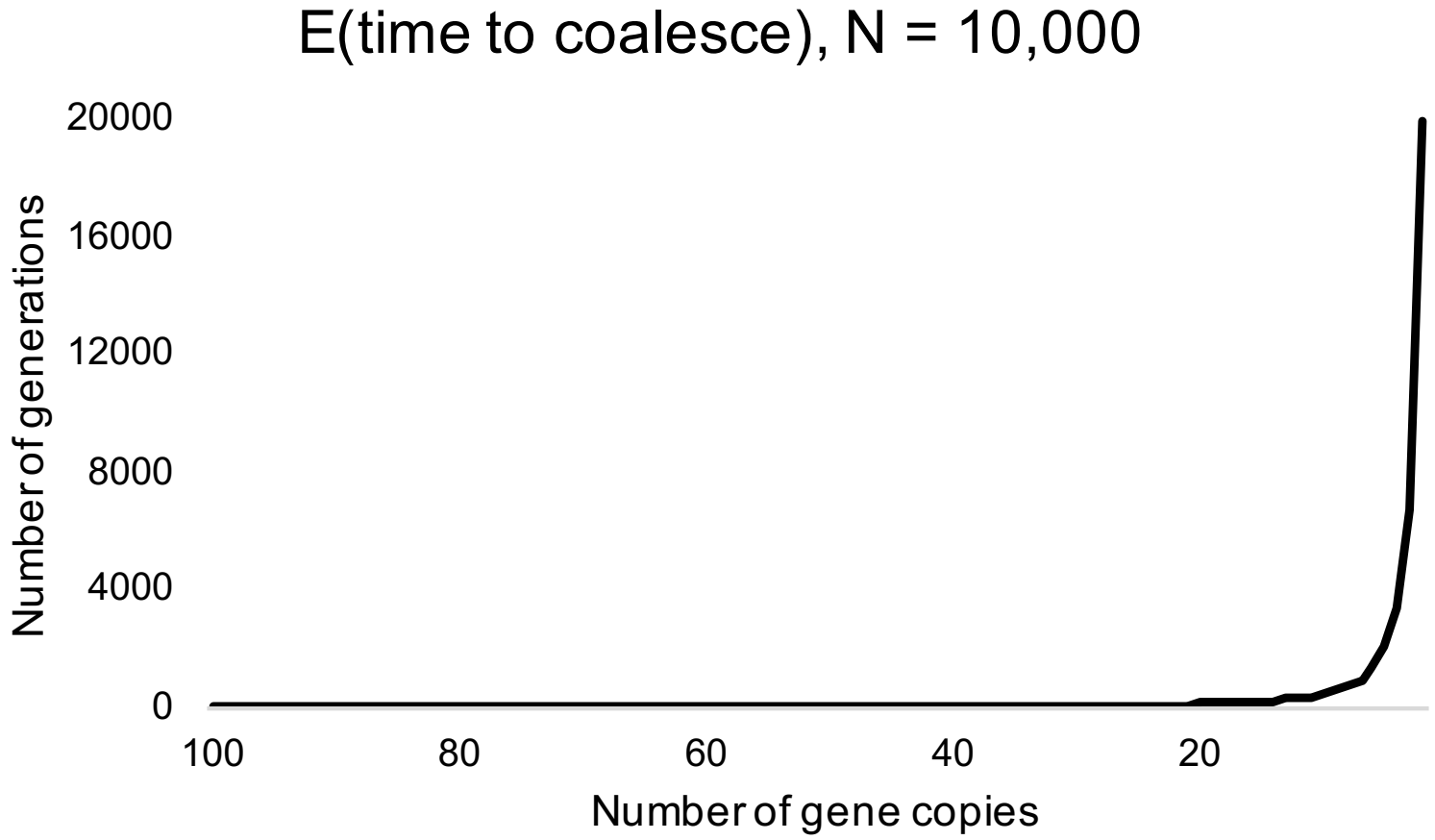
$$P_{NC,t+1} = 1 - \left(\frac{1+2+3+\cdots+(k-1)}{2N}\right) = 1 - \frac{k(k-1)}{4N}$$

The # of generations to the next coalescent event has an expected value of the reciprocal of $P_{C,t+1} = 1 - P_{NC,t+1} = \frac{k(k-1)}{4N}$ or $\frac{4N}{k(k-1)}$

Moving back through time in our coalescent the expected time to the next coalescent event gets larger, with a maximum value (2*N*) at *k*=2

COALESCENT MODELS

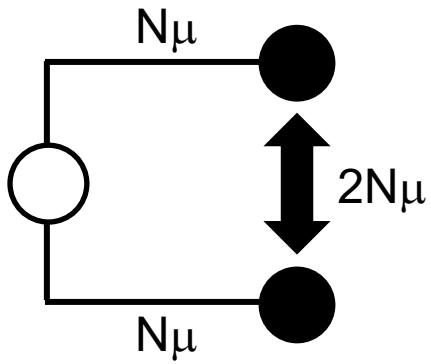
Expected time to next coalescent event = $\frac{4N}{k(k-1)}$



COALESCENT MODELS

The Coalescent (*n*-coalescent)

- For genetic data, coalescent intervals are typically expressed as the number of substitutions, with N generations expected to coalescence

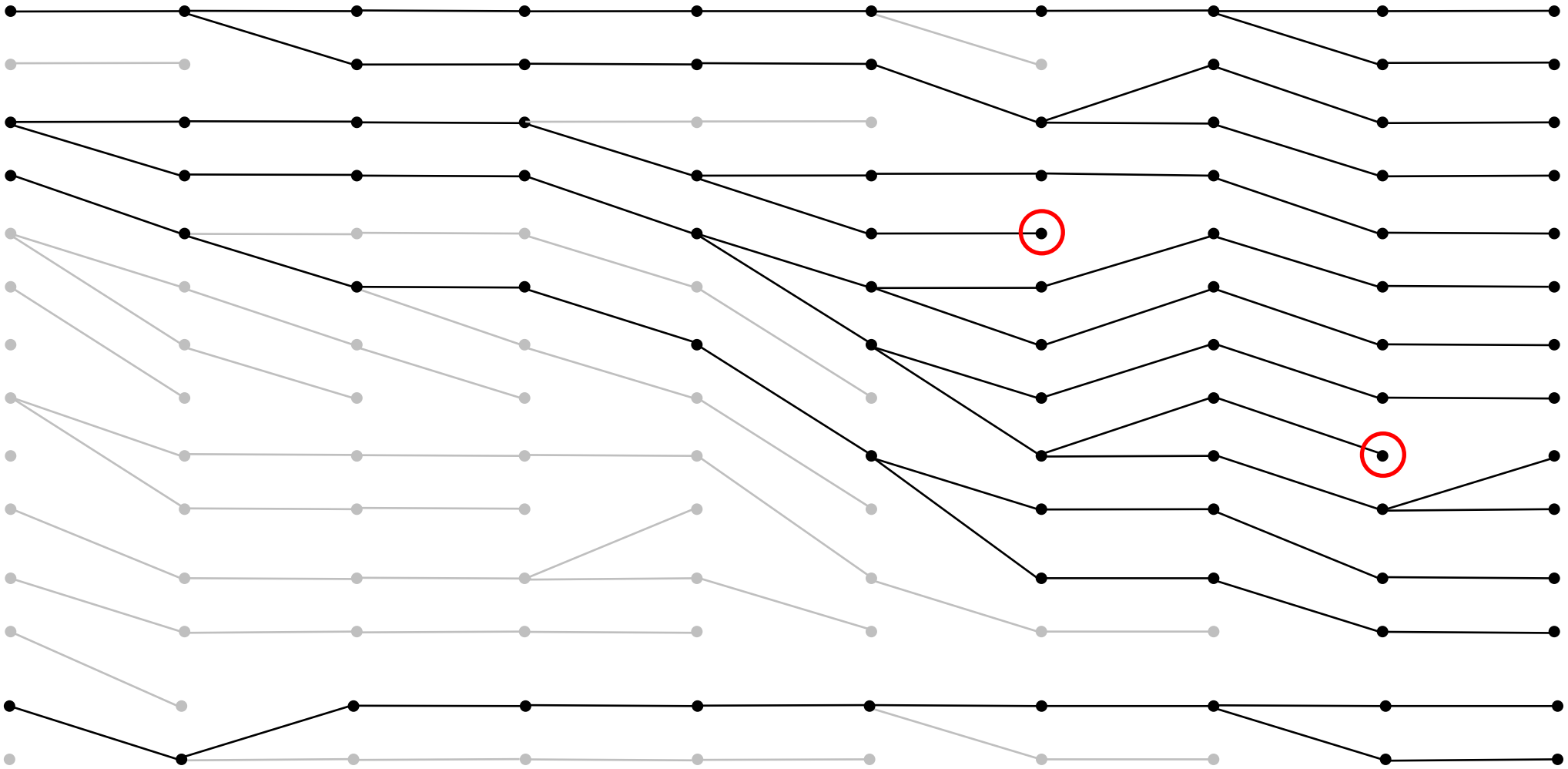


COALESCENT MODELS

The Serial Coalescent (*s*-coalescent)

- With genetic data, coalescent intervals are typically expressed as the number of substitutions
- This means that time is scaled by substitution rate μ
- For the standard coalescent, instead of N you must use $\theta = 2N\mu$ (haploid) or $4N\mu$ (diploid)
- Serial sampling (collecting sequences at different time points) allows you to separate time and substitution rate

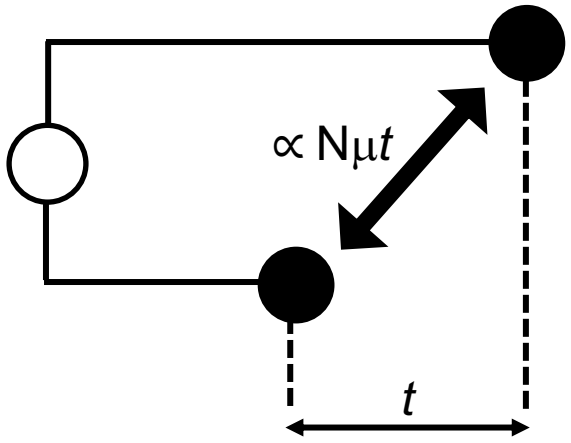
COALESCENT MODELS

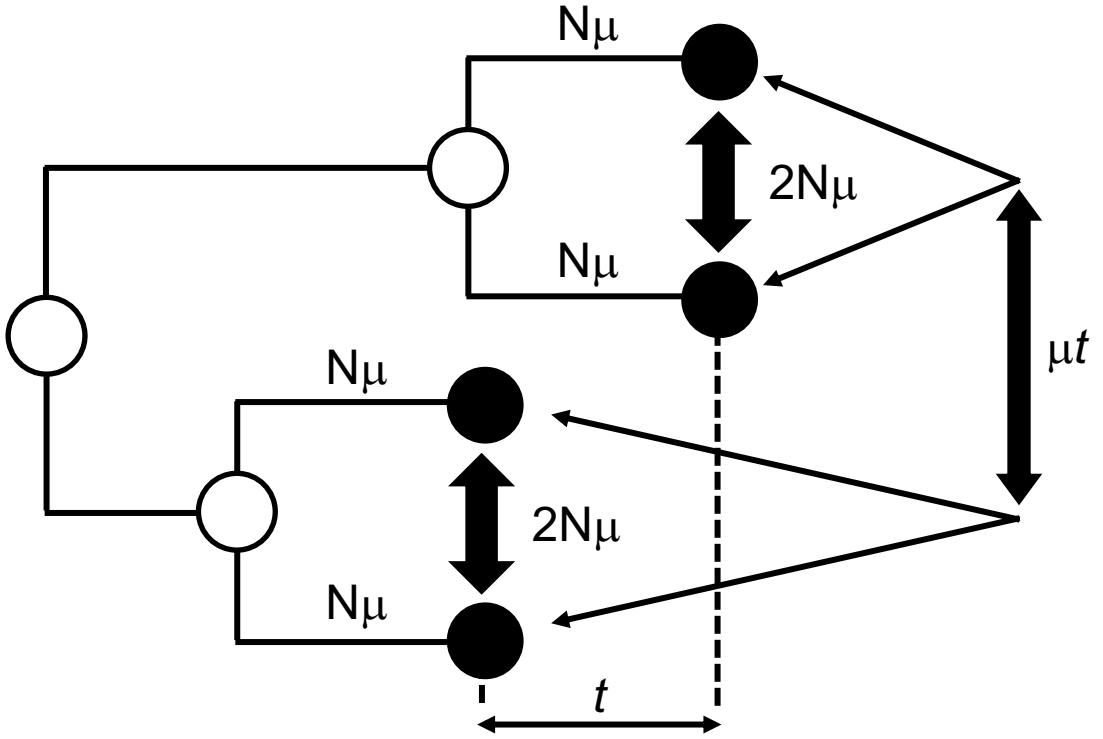


COALESCENT MODELS

The Serial Coalescent (*s*-coalescent)

- With serial sampled genetic data, coalescent intervals are now a function of time between samples and the substitution rate





COALESCENT MODELS

The *n*-Coalescent

- Number of lineages decrease monotonically over time
- Parameter estimates unreliable because the coalescence interval variance increases as number of lineages decreases

The *s*-Coalescent

- Number of lineages increases in the interval when new sequences are added
- Parameter estimates more reliable (smaller variance) as more lineages are present over past intervals
- Adding more sequences to past intervals increases reliability of parameter estimates.

HWE, Pop Gen, and Coalescence

Thank you for your time and attention