THE ORGANIZATION OF GENETIC DIVERSITY

Constancy of allele frequencies:

-HARDY WEINBERG EQUILIBRIUM

Changes in allele frequencies:

- MUTATION and RECOMBINATION

- GENETIC DRIFT and POPULATION STRUCTURE

- MIGRATION and GENE FLOW

- NATURAL SELECTION

WHY MODELS?

- Understanding the mechanisms by which the evolutionary forces act on allele frequencies allows producing mathematical models that approximate reality.

- It allows understanding the subtle inteplay between these forces.

- The derived equations allow estimating parameters of interest.

- Allow testing different hypotheses.

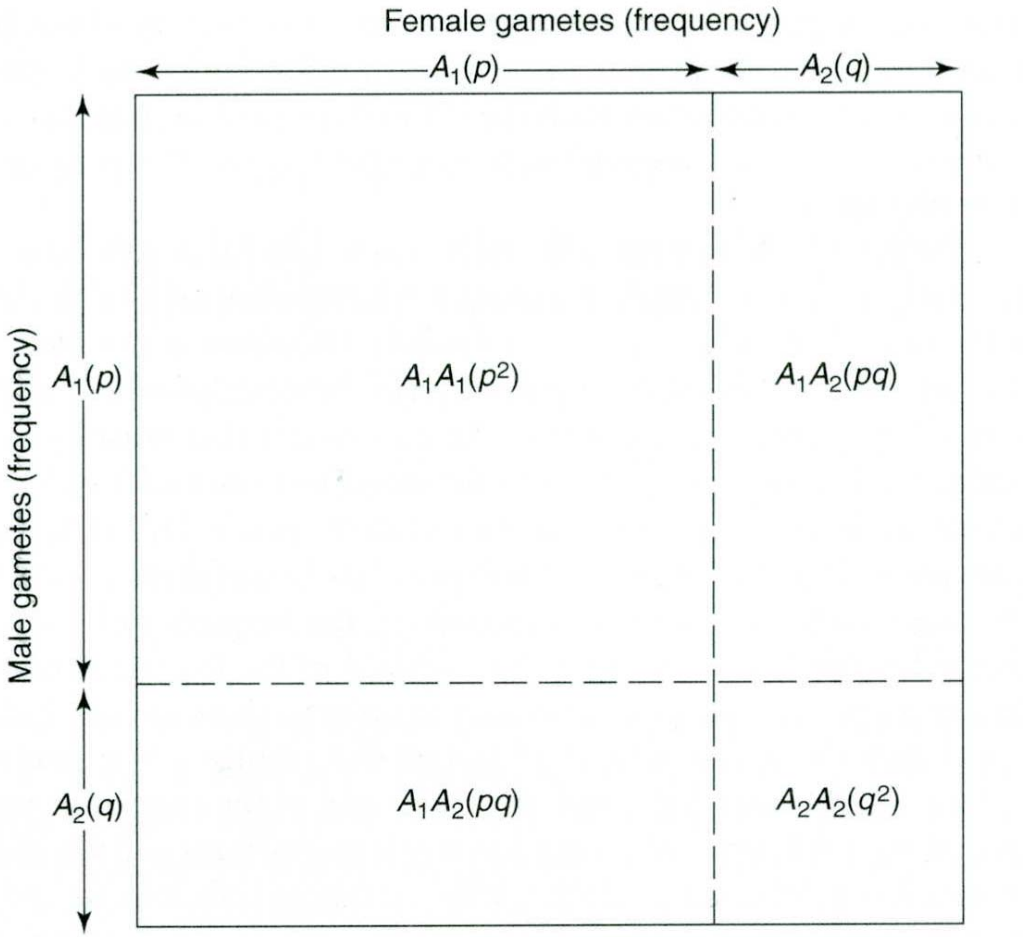
- Alternative models can be compared to determine which provides the best fit to the data.

HARDY-WEINBERG EQUILIBRIUM

- diploid organisms;
- sexual reproduction;
- random mating;
- non-overlapping generations;
- equal allele frequencies in both sexes;
- mutation/selection/migration is absent;
- population size is infinite (no genetic drift).

The genotype frequencies can be predicted/deduced.

HARDY-WEINBERG EQUILIBRIUM



Allele frequencies:

A_1 – p

A_2 – q

Genotype frequencies:

A_1A_1 – p^2

A_1A_2 – pq + pq = 2pq

A_2A_2 – q^2

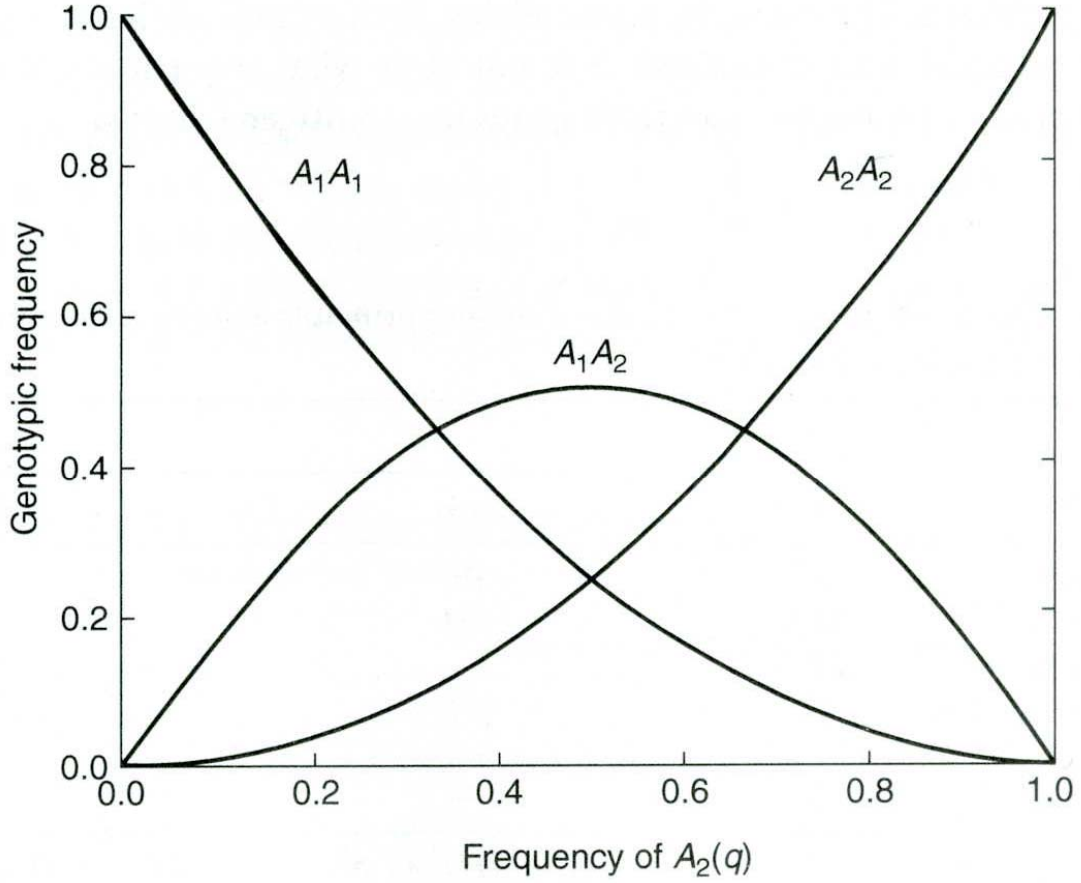
HARDY-WEINBERG EQUILIBRIUM

Male genotypes (frequencies)	*Female genotypes (frequencies)*		
	$A_1A_1(P)$	$A_1A_2(H)$	$A_2A_2(Q)$
$A_1A_1(P)$	P^2	PH	PQ
$A_1A_2(H)$	PH	H^2	HQ
$A_2A_2(Q)$	PQ	HQ	Q^2

		Progeny		
Mating type	*Frequency*	A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	P^2	P^2	—	—
$A_1A_1 \times A_1A_2$	$2PH$	PH	PH	—
$A_1A_1 \times A_2A_2$	$2PQ$	—	$2PQ$	—
$A_1A_2 \times A_1A_2$	H^2	$\frac{1}{4}H^2$	$\frac{1}{2}H^2$	$\frac{1}{4}H^2$
$A_1A_2 \times A_2A_2$	$2HQ$	—	HQ	HQ
$A_2A_2 \times A_2A_2$	Q^2	—	—	Q^2
Total	1	$(P+\frac{1}{2}H)^2 = p^2$	$2(P+\frac{1}{2}H)(Q+\frac{1}{2}H) = 2pq$	$(Q+\frac{1}{2}H)^2 = q^2$

The random cross of genotypes can be interpreted as the random union of gametes. This is particularly useful because it allows following the changes in frequency of a single allele instead of two genotypes.

HARDY-WEINBERG EQUILIBRIUM



HARDY-WEINBERG EQUILIBRIUM

Estimating allele and genotype frequencies:

$P=N_{A_1A_1}/N$ $H=N_{A_1A_2}/N$ $Q=N_{A_2A_2}/N$

$$p = (N_{A_1A_1} + 1/2\ N_{A_1A_2}) / N$$

$$q = (1/2\ N_{A_1A_2} + N_{A_2A_2}) / N$$

HARDY-WEINBERG EQUILIBRIUM

The example of Cystic Fibrosis:

- Causes difficulty in breeding, sinus infections, poor growth, diarrhea, etc.

- It is a genetically transmitted disorder, caused by a mutation in the gene for the protein *cystic fibrosis transmembrane conductance regulator* .

- This gene is required to regulate the components of sweat, digestive juices, and mucus.

HARDY-WEINBERG EQUILIBRIUM

The example of Cystic Fibrosis:

The frequency of the recessive homozygous genotype (A_2A_2) is 1:1700;

The frequency of "A_2" is then $q = \sqrt{1/1700} = 0.024$;

The frequency of "A_1A_2" is $2pq = 2 \times 0.976 \times 0.024 = 0.047 =$ 1:21.

Although the disease occurs only in 1 out of 1700 people, 1 in 21 is carrier of the allele causing the disease.

HARDY-WEINBERG EQUILIBRIUM

X-linked locus (XX females and XY males)

P_f, H_f e Q_f: frequencies of diploid genotypes "A_1A_1", "A_1A_2" and "A_2A_2" in females.

P_m e Q_m: frequencies of haploid genotypes "A_1" and "A_2" in males.

The frequencies of "A_2" in both sexes are:

$q_f = Q_f + 1/2H_f$

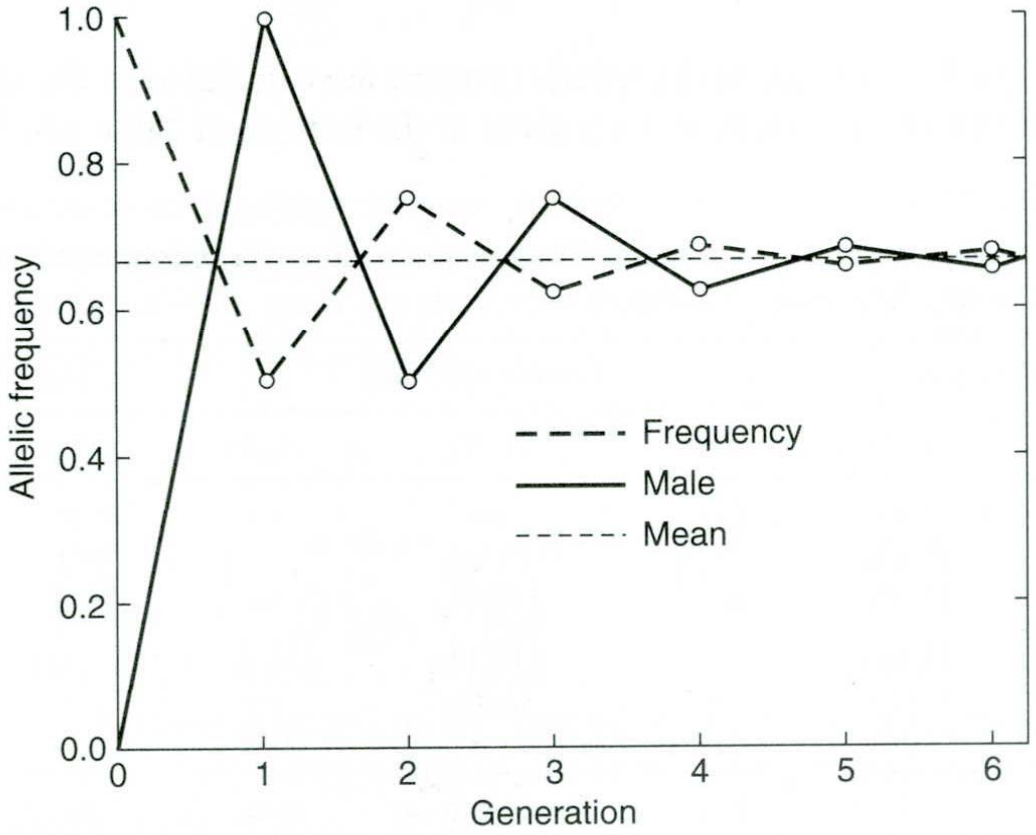
$q_m = Q_m$

the mean allelic frequency of "A_2" is

$$q = 2/3q_f + 1/3q_m$$

HARDY-WEINBERG EQUILIBRIUM

If the frequency of one allele at na X-linked locus are $q_f = 1.0$ and $q_m = 0.0$ on the first generation (q = 2/3)…



HARDY-WEINBERG EQUILIBRIUM

There are no evolutionary forces acting other than what is imposed by the mechanism of reproduction: it serves as basis to compare with more complex models.

Once the equilibrium is reached, it predicts the constancy of allelic frequencies through time (if genetic drift, mutation, migration and selection are **absent**).

TESTING HWE

χ^2

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

Df = # genotypes - # estimated parameters - 1

Fisher's exact test

MEASURING GENETIC DIVERSITY AND DISTANCE

PROPORTION OF POLYMORPHIC LOCI

$$P = x/m$$

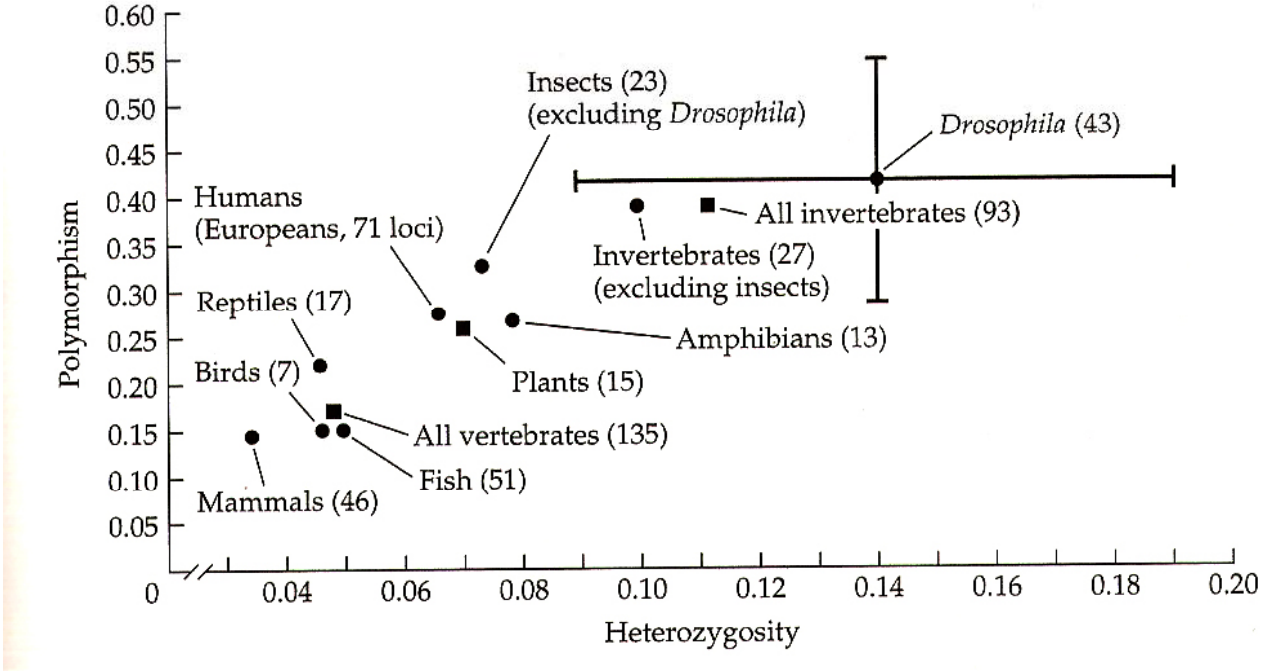
x – nr. of polimorphic *loci*

m – nr. of analysed *loci*

HETEROZYGOSITY

$$He = 1 - \sum_{i=1}^{n} p_i^2 \qquad Ho = \sum_{i<j}^{n} P_{ij}$$

n – Nr. of alleles



MEAN NUMBER OF ALLELES

n_a – mean number of alleles/locus

n_e – effective number of alleles $\quad 1/ \Sigma x_i^2$

$n_e \leq n_a$

n_e minimizes the importance of rare alleles as source of variation.

PROPORTION OF VARIABLE SITES

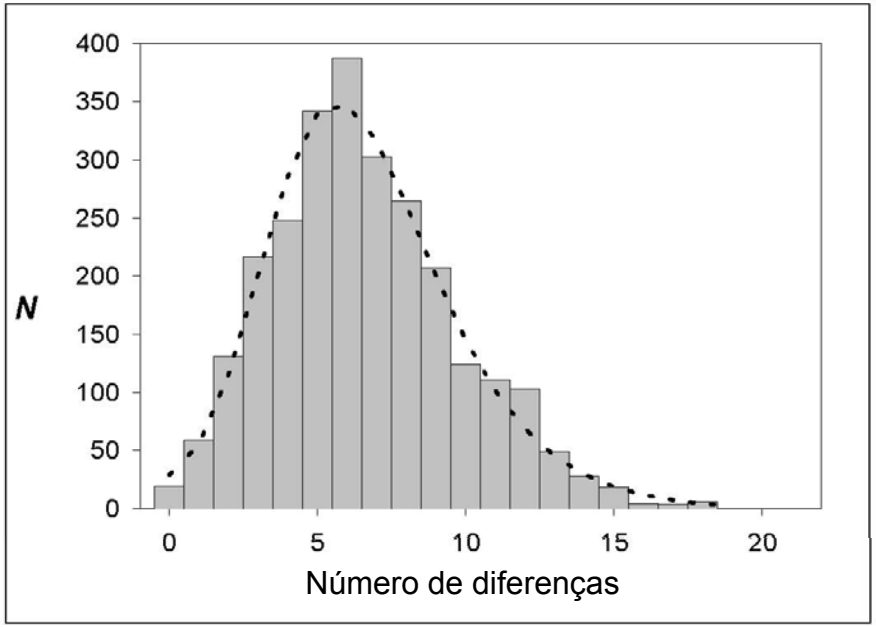
$$p_n = S/N$$

S – nr. of variable sites

N – nr. of analysed sites

NUCLEOTIDE DIVERSITY

$$\pi = \sum_{ij} p_i p_j \pi_{ij}$$

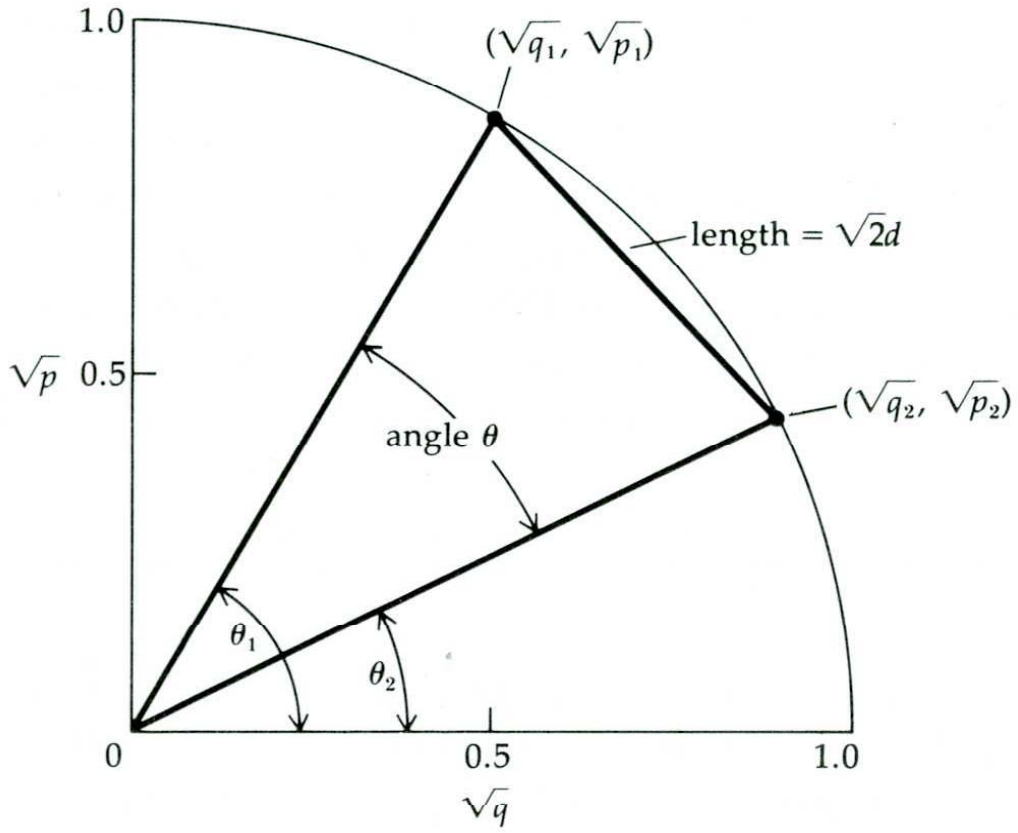


p_i – *frequency of sequence i*

p_j - *frequency of sequence j*

π_{ij} – *proportion of nucleotides that differ between sequences i and j*

GENETIC DISTANCES



GENETIC DISTANCES

Nei's distance

$$I = J_{XY} / (J_X J_Y)^{1/2}$$

$$J_{XY} = \Sigma p_{ix} p_{iy}$$

$$J_X = \Sigma p_{ix}^2$$

$$J_Y = \Sigma p_{iy}^2$$

$$D = - \ln (I)$$

GENETIC DISTANCES

Measure	Reference
$\delta\mu^2$	(Goldstein *et al.*, 1995)
R_{ST}	(Slatkin, 1995)
D_{SW}	(Shriver *et al.*, 1995)

GENETIC DISTANCES

$$d_{XY} = \sum_{ij}^{n} p_i p_j d_{ij}$$

Mean number of substitutions per position between sequences of two populations.

$$d_A = d_{XY} - (d_X + d_Y)/2$$

Mean number of substitutions per position between sequences of two populations, corrected for the distances observed within each population.

“FORCES” SHAPING GENETIC DIVERSITY

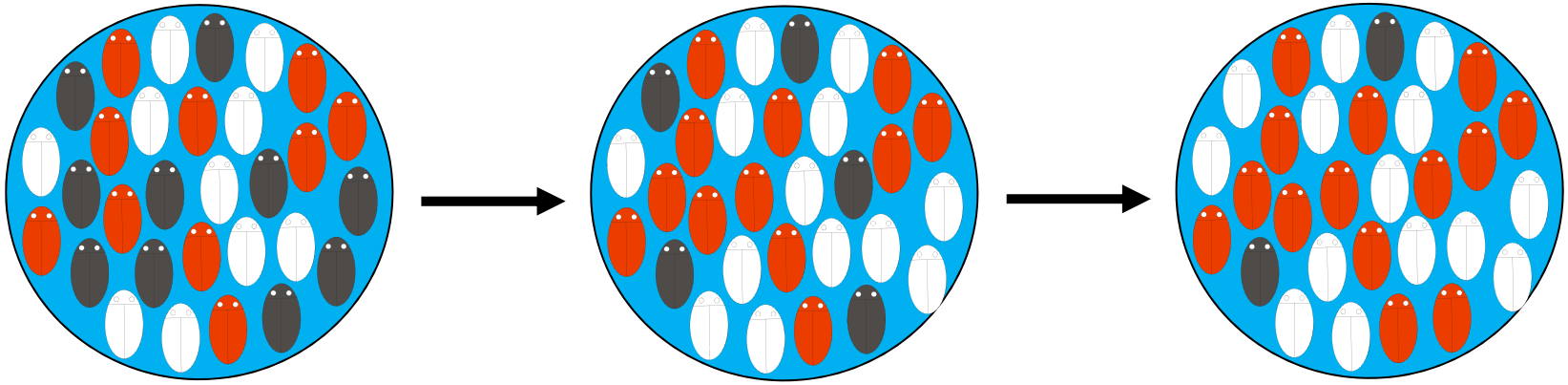
GENETIC DRIFT

GENETIC DRIFT

No population is infinitely large as is assumed in the HWE theorem: each generation is a finite sample of the genetic composition of the previous one.

Therefore, variation in allele frequency between generations can occur simply due to this stochastic process of sampling: GENETIC DRIFT.

GENETIC DRIFT



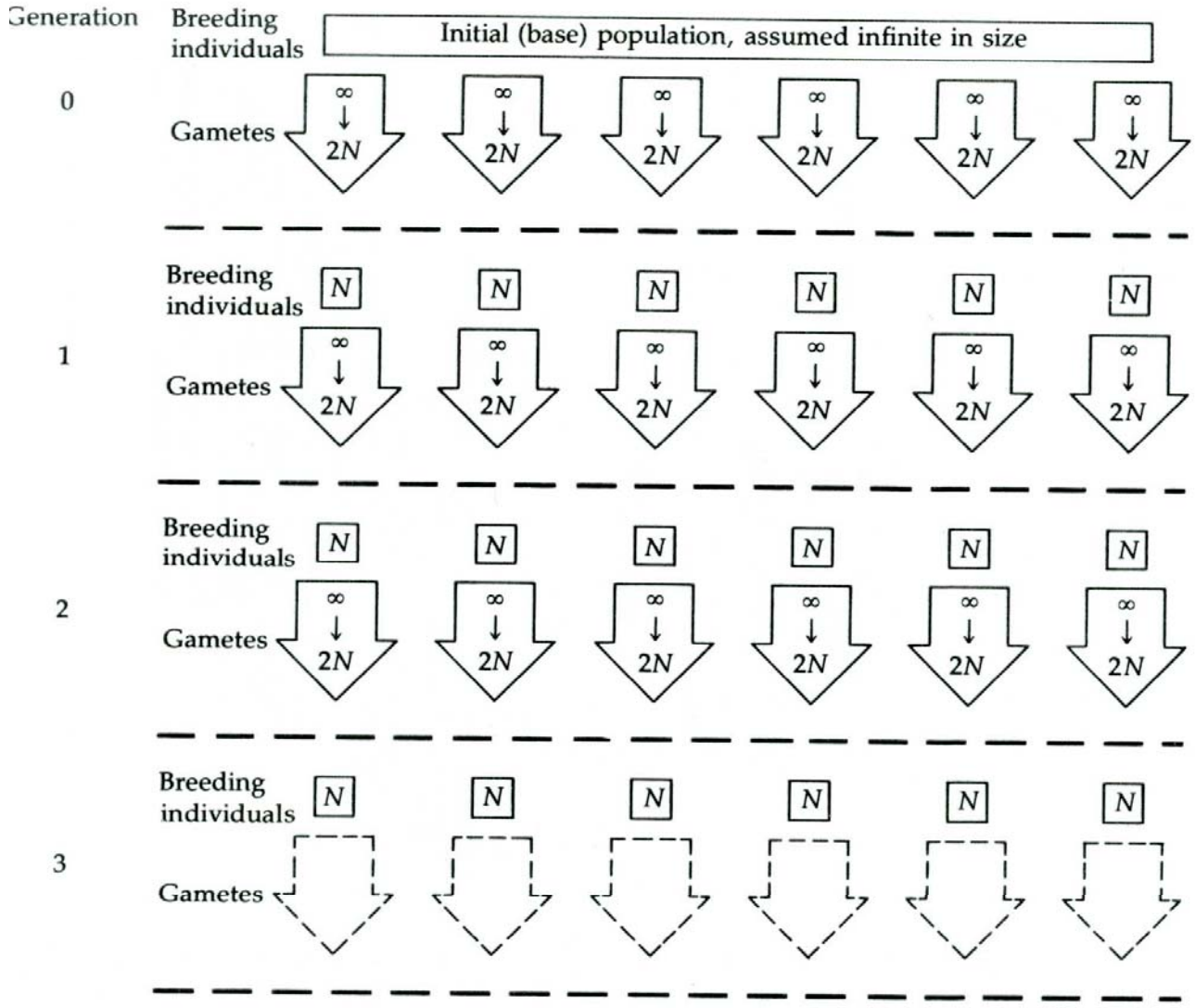
- From the process of random sampling alone populations cannot gain new alleles.

- Genetic drift thus leads to a loss of diversity either by loss or fixation of alleles.

THE WRIGHT-FISHER MODEL

- diploid organisms;
- sexual reproduction;
- random mating;
- non-overlapping generations;
- equal allele frequencies in both sexes;
- mutation/selection/migration is absent;
- POPULATION SIZE IS CONSTANT - N

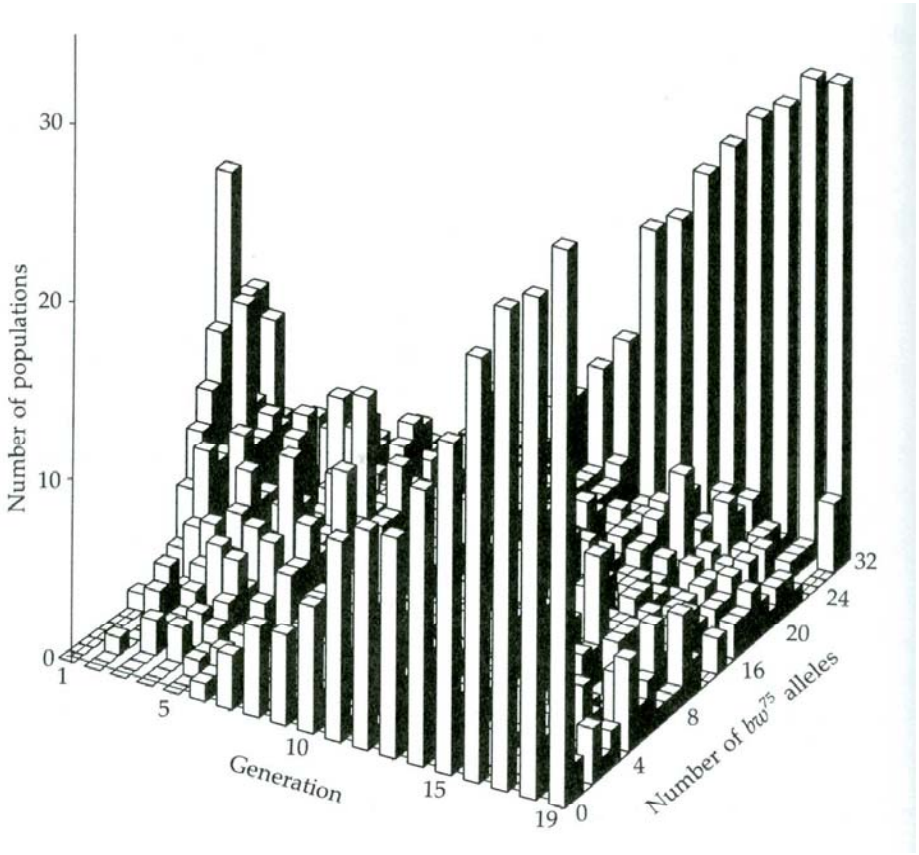
THE WRIGHT-FISHER MODEL



AN EXPERIMENT

Buri, 1956

- *Drosophila*
- Frequencies of eye colour:
  - $bw^{75}bw^{75}$: light brown
  - $bw^{75}bw$: red
  - bwbw: brown
- 107 subpopulations:
  - 8 females
  - 8 males
  - All $bw^{75}bw$



- The probability of fixation of an allele is its initial frequency.

- The probability of fixation of a new allele is thus 1/2N.

THE WRIGHT-FISHER MODEL

- Example: Biallelic locus A_1/A_2

Population N=4 diploid individuals.

Sample of 2N gametes to form N individuals.

Possibilities: take 0, 1, 2, … , 8 alleles "A_1" (the remaining being "A_2").

- The probability of each of these possibilities is given by the binomial distribution:

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

THE WRIGHT-FISHER MODEL

State

Absorbing state

Probability of state transition*

Matrices

*(directly obtained from the binomial distribution)

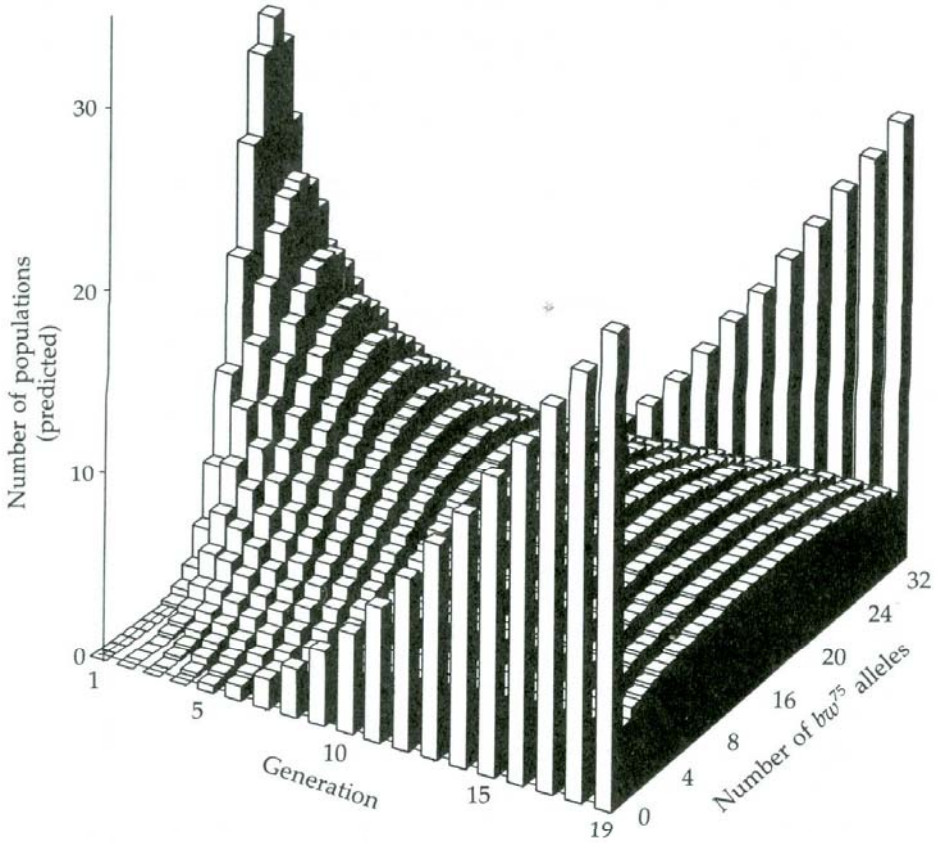
$$T_{ij} = \binom{2N}{j} p^j q^{2N-j}$$

THE WRIGHT-FISHER MODEL: MATRICES OF STATE TRANSITION

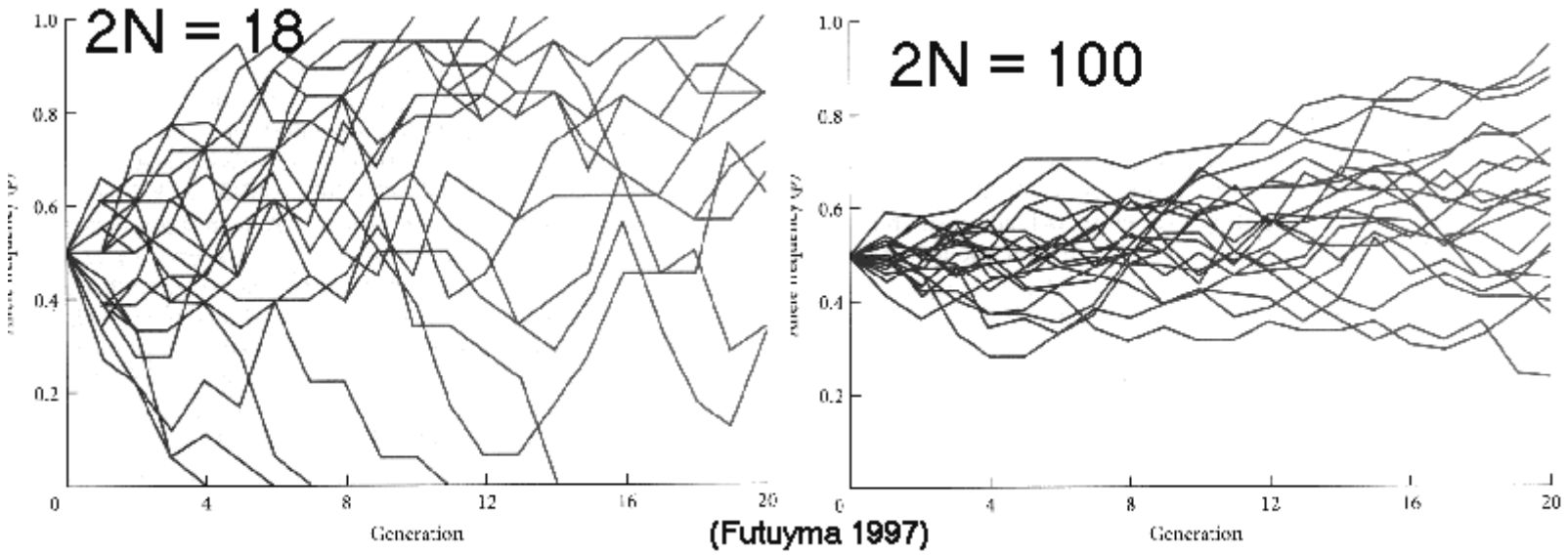
Matrix of probabilities of state transition in a population with 2N=4:

		Number of *A* alleles in generation $t + 1$				
		0	**1**	**2**	**3**	**4**
Number of *A* alleles in generation *t*	**0**	1	0	0	0	0
	1	0.316	0.422	0.211	0.047	0.004
	2	0.062	0.25	0.375	0.25	0.062
	3	0.004	0.047	0.211	0.422	0.316
	4	0	0	0	0	1

THE WRIGHT-FISHER MODEL



GENETIC DRIFT AND POPULATION SIZE



The magnitude of genetic drift is related to the size of the population being sampled.

$t = 4N_e$

EFFECTIVE POPULATION SIZE (N_e)

Is the size of a Wright-Fisher population that displays the same amount of genetic drift as the population under study.

Census (N) *vs.* effective population size (N_e):

- Fluctuations of N_e through time.
- Variance of reproductive success.
- Unequal sex-ratios.
- Overlapping generations.
- Population structure.
- Etc.

EFFECTIVE POPULATION SIZE (N_e)

$$1/N_e = (1/t)\,(1/N_1 + 1/N_2 + \ldots + 1/N_t)$$

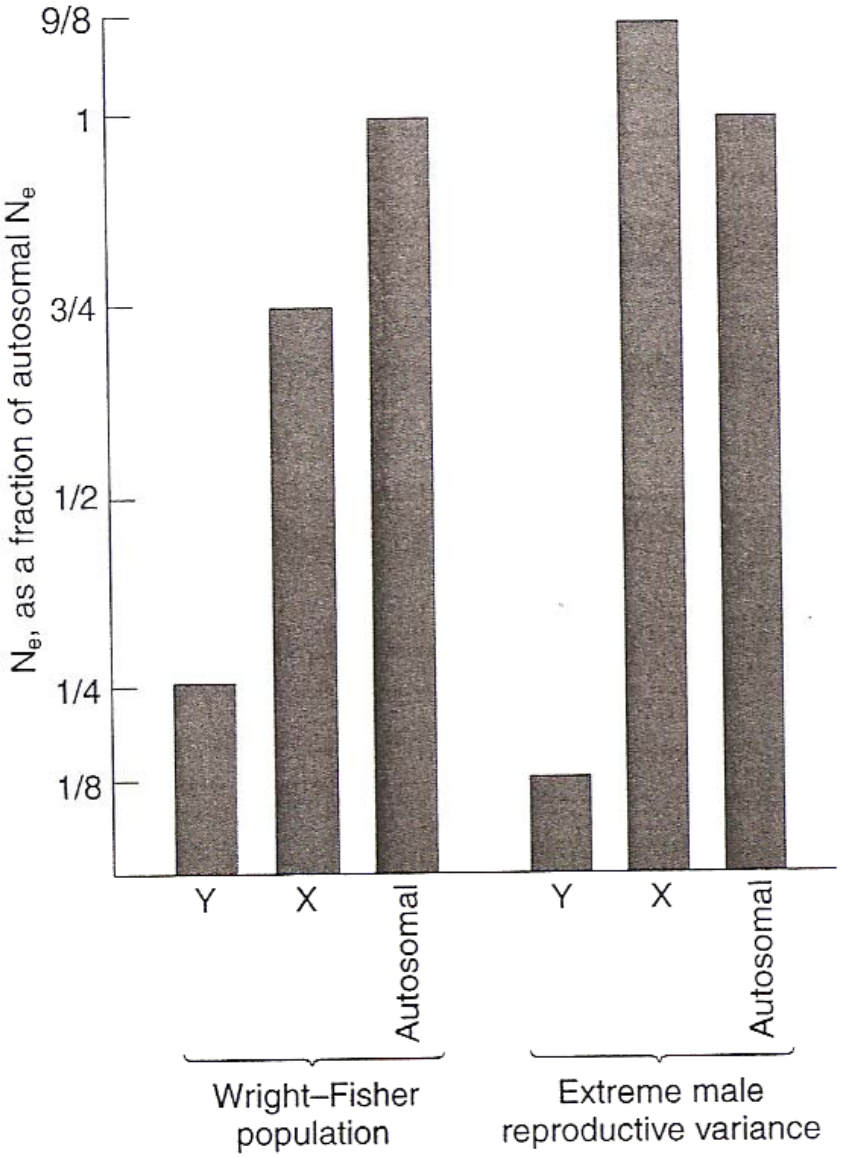
Example: a population has suffered a “bottleneck”:

$N_0 = 1000$, $N_1 = 10$, $N_2 = 1000$

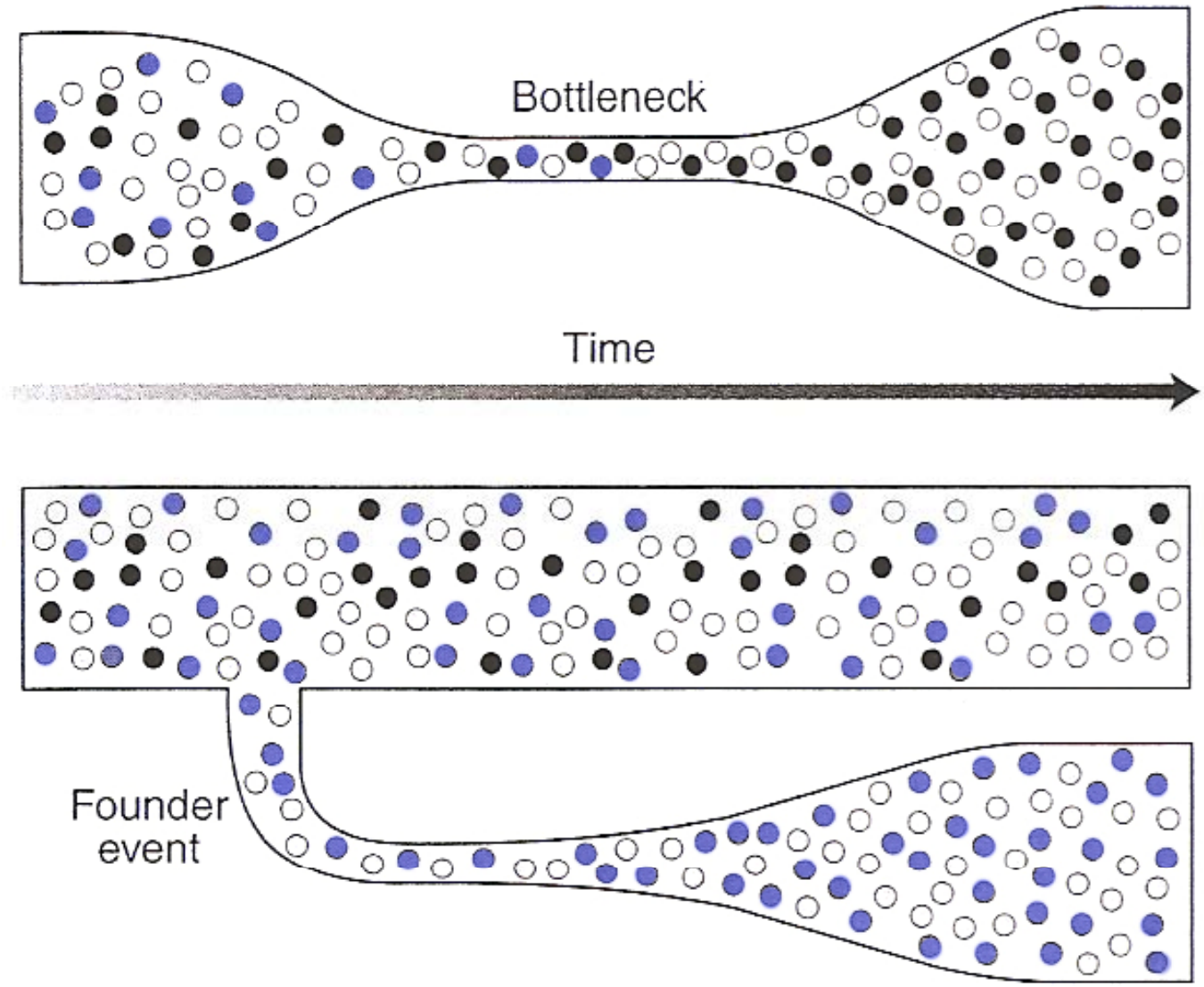
$1/N_e = (1/3)(1/1000 + 1/10 + 1/1000)$

$N_e = 29.4$ (the arithmetic mean would be 670)

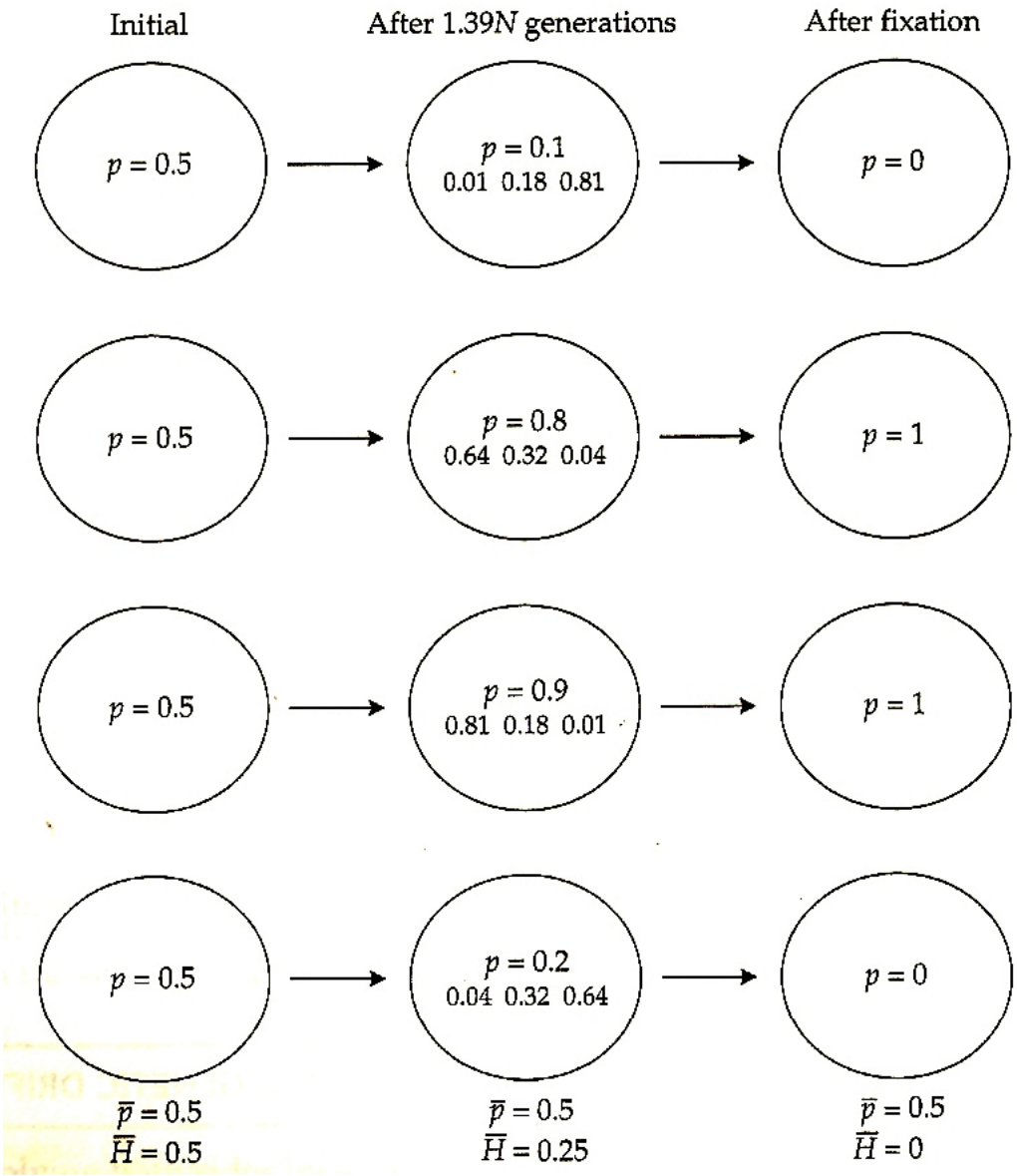
EFFECTIVE POPULATION SIZE (N_e)

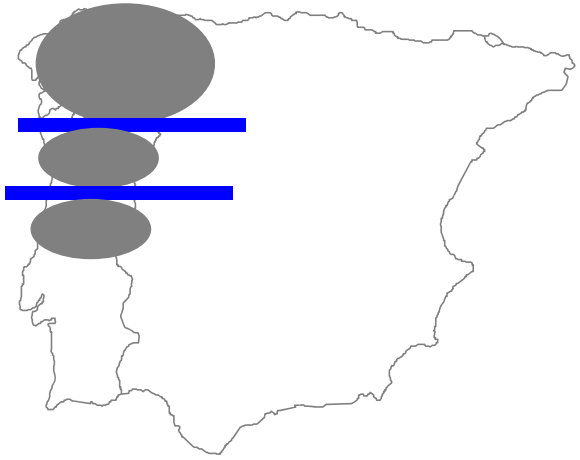


BOTTLENECKS AND FOUNDER EVENTS

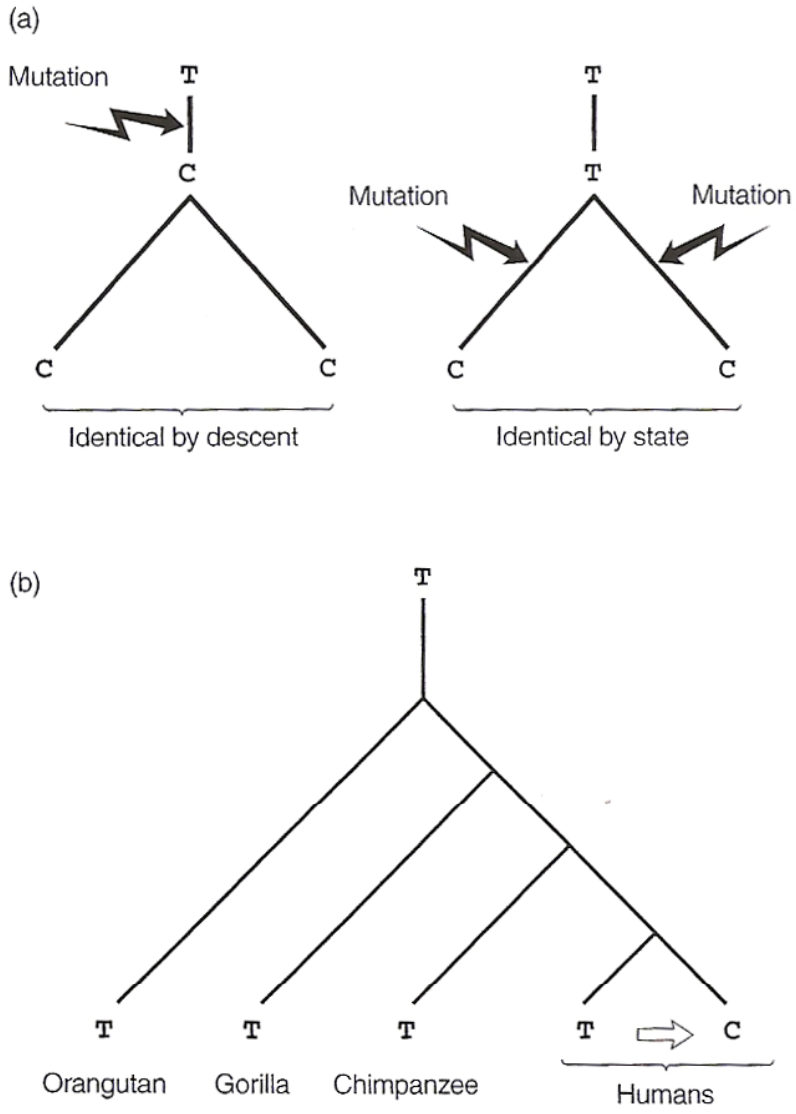


DRIFT AND INBREEDING

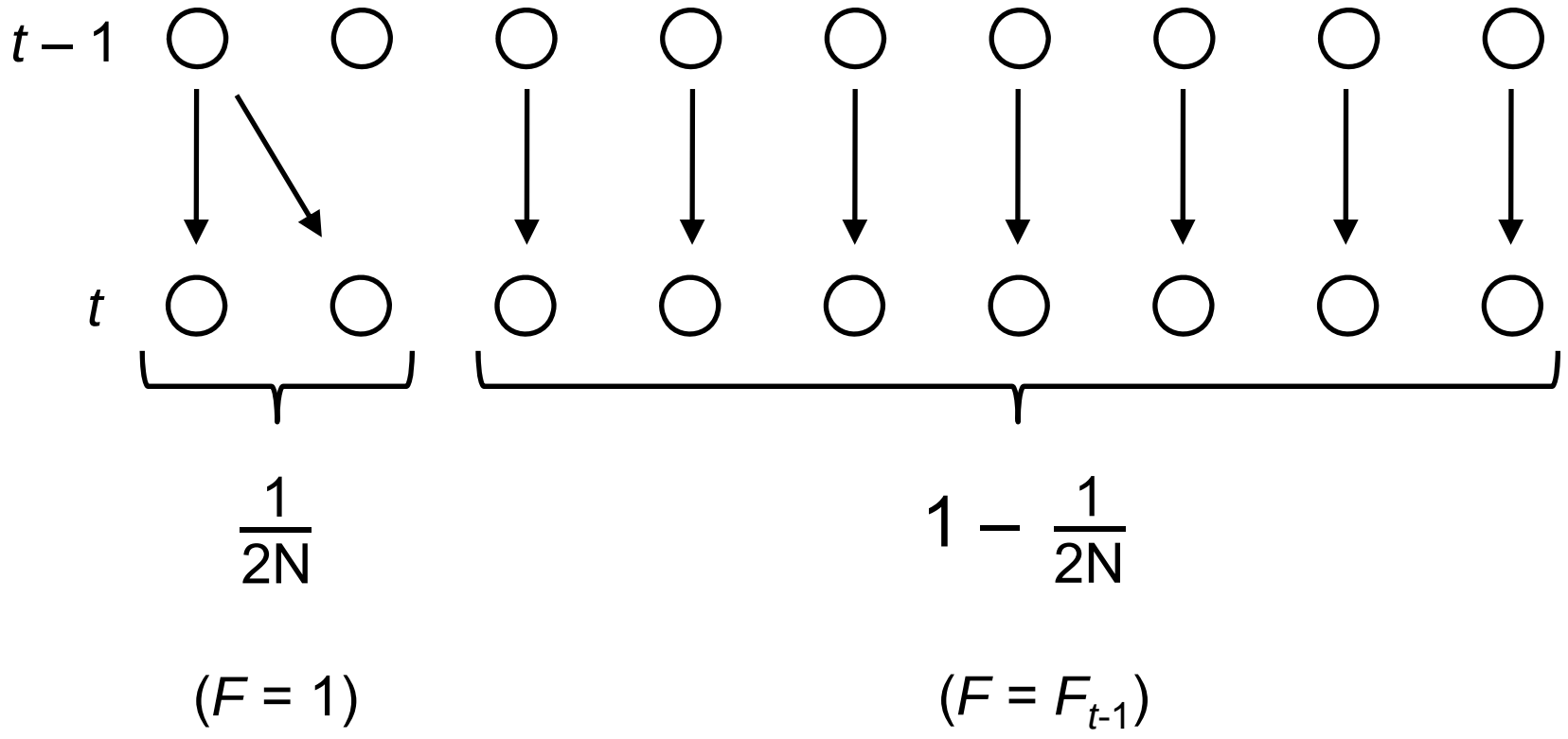




DRIFT AND INBREEDING



DRIFT AND INBREEDING



DRIFT AND INBREEDING

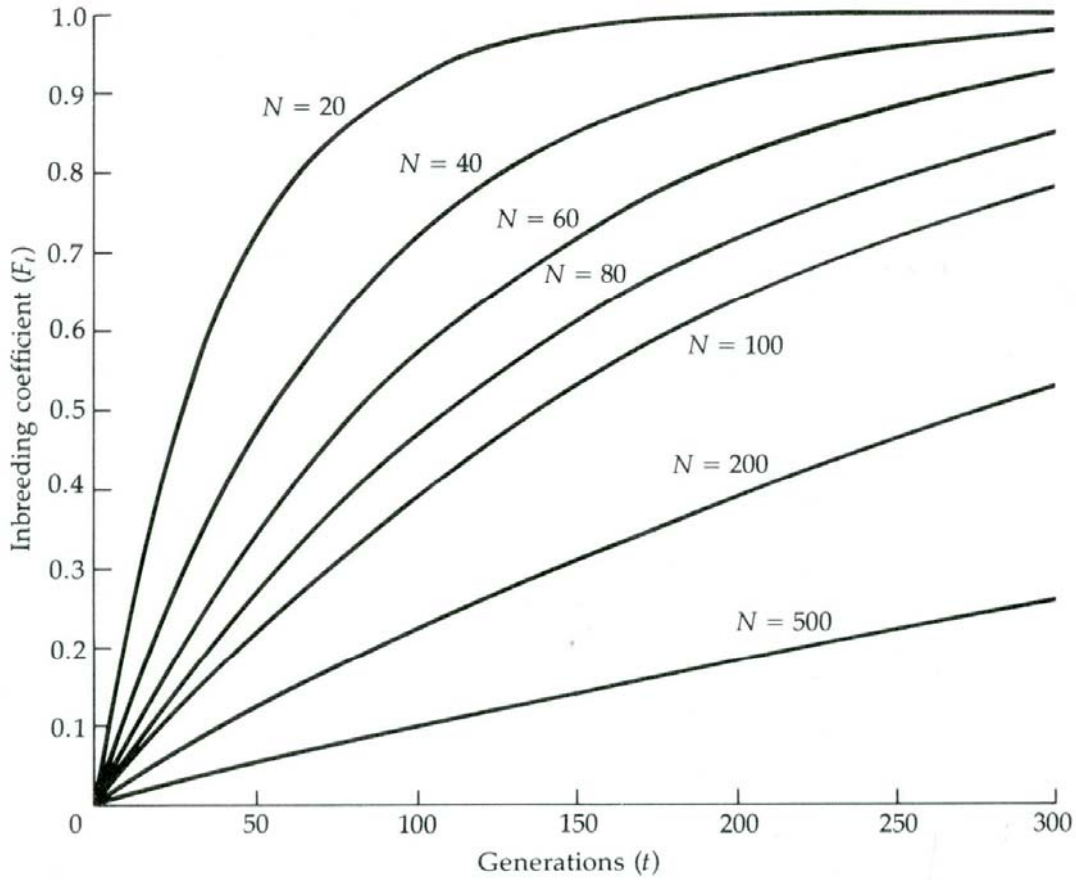
$$F_t = 1/2N + (1 – 1/2N)\, F_{t-1}$$

i.e.,

$$F_t = 1 – (1 – 1/2N)^t$$

Where F is the inbreeding coefficient represented as a probability, i.e. the probability that an individual has a pair of alleles that are identical by descent.

DRIFT AND INBREEDING

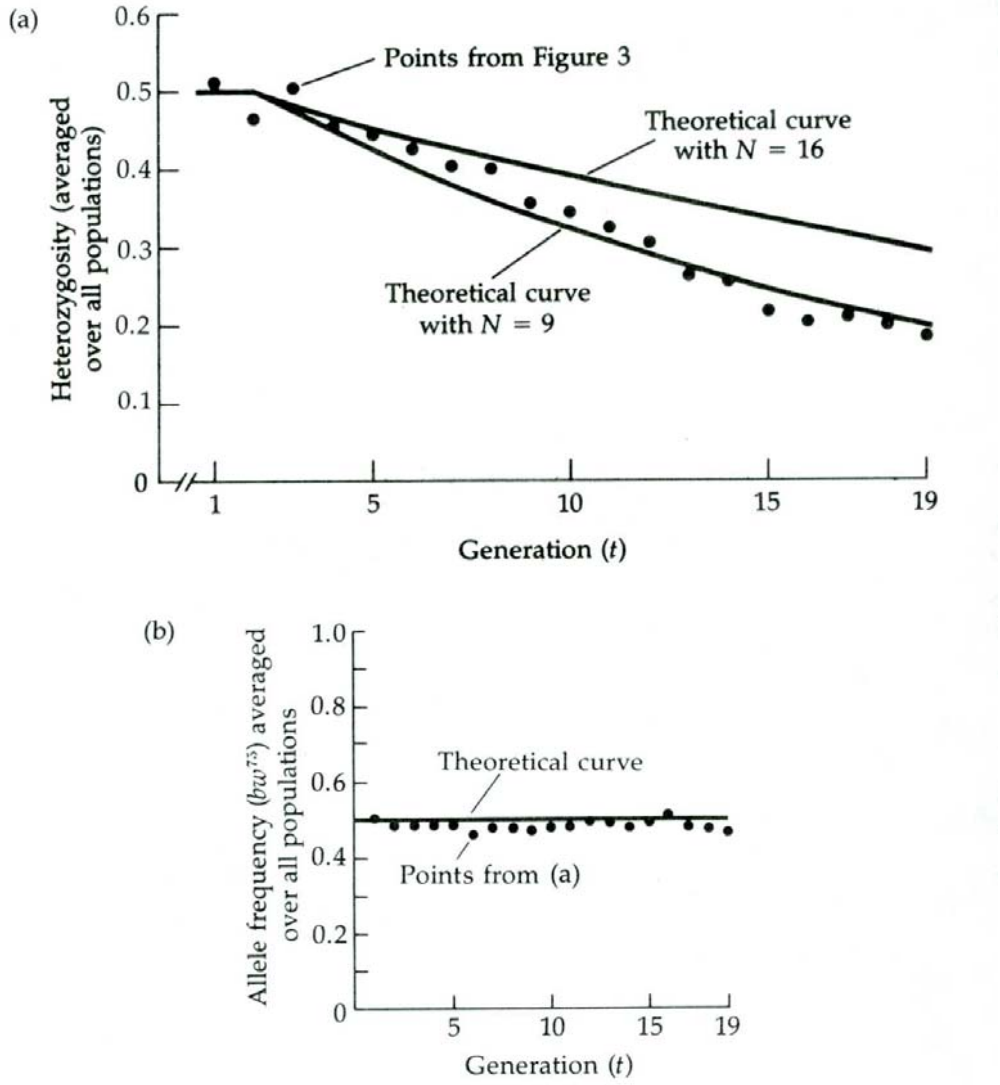


DRIFT AND INBREEDING

Thinking on heterozygosity,

$$H_t = 1 - F_t$$

DRIFT AND INBREEDING



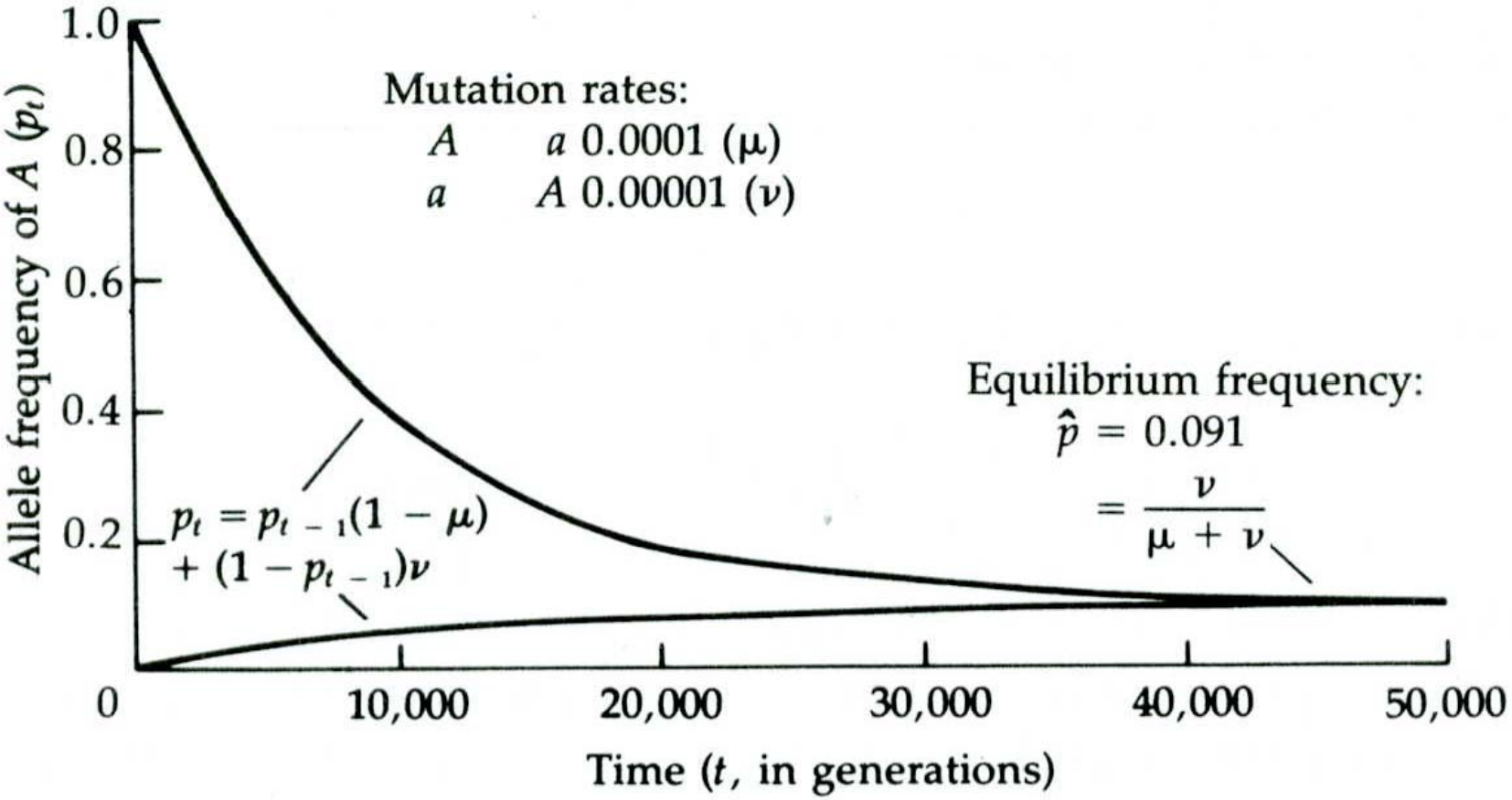
DRIFT AND INBREEDING

		GENERATION		
		0	***t***	**∞**
Inbreeding coefficient (F_t) (Average over all populations)		0	$1 - (1 - 1/2N)^t$	1
Genotype frequency (Average over all populations)	*AA*:	p_0^2	$p_0^2(1 - F_t) + p_0F_t$	p_0
	Aa:	$2p_0q_0$	$2p_0q_0(1 - F_t)$	0
	aa:	q_0^2	$q_0^2(1 - F_t) + q_0F_t$	q_0
Allele frequency (Average over all populations)	*A*:	p_0	p_0	p_0
	a:	q_0	q_0	q_0

GENERATING DIVERSITY:

MUTATION

EVOLUTION OF ALLELE FREQUENCIES UNDER MUTATION PRESSURE



MUTATION PRESSURE

Considering na average protein

300 aa – 900 nt

$4^{900} \approx 10^{542}$

One can assume that each mutation creates a new allele.

MUTATION PRESSURE

- An allele decreases in frequency as it accumulates mutations.

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

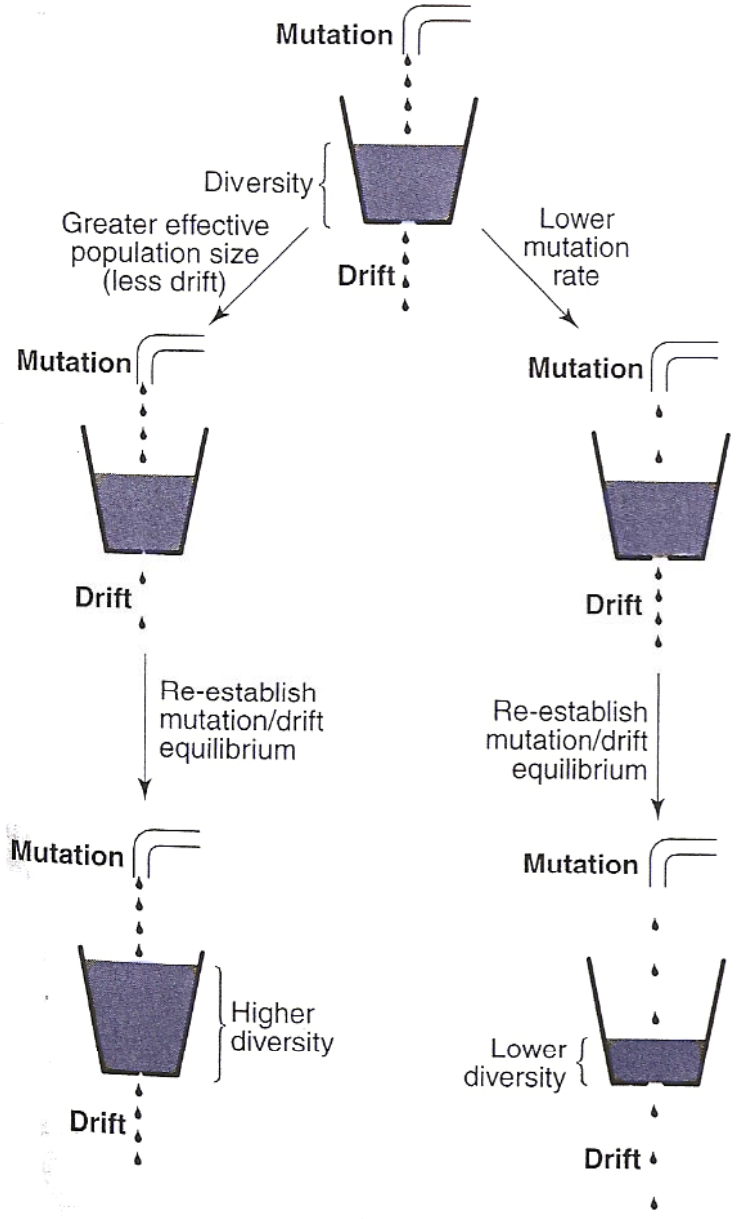
- This assumes that mutation is rare enough that there are no recurrent mutations: INFINITE ALLELES MODEL.

- This means that this model assumes that any mutation generates a new, previously absent, allele.

MUTATION-DRIFT EQUILIBRIUM

$\theta = 4N_e\mu$

How many alleles can be maintained in a population at equilibrium?



INFINITE ALLELES MODEL

According to this model, F can be calculated as before, but including the possibility of mutation:

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

Solving this expression alllows determining F at the equilibrium

$$F = 1/ 4N\mu + 1$$

INFINITE ALLELES MODEL

This model can be extended also to the presence of migration.

Since according to this model the alleles identical by state are also identical by descent, the probability of autozygosity is also the proportion of homozygous genotypes:

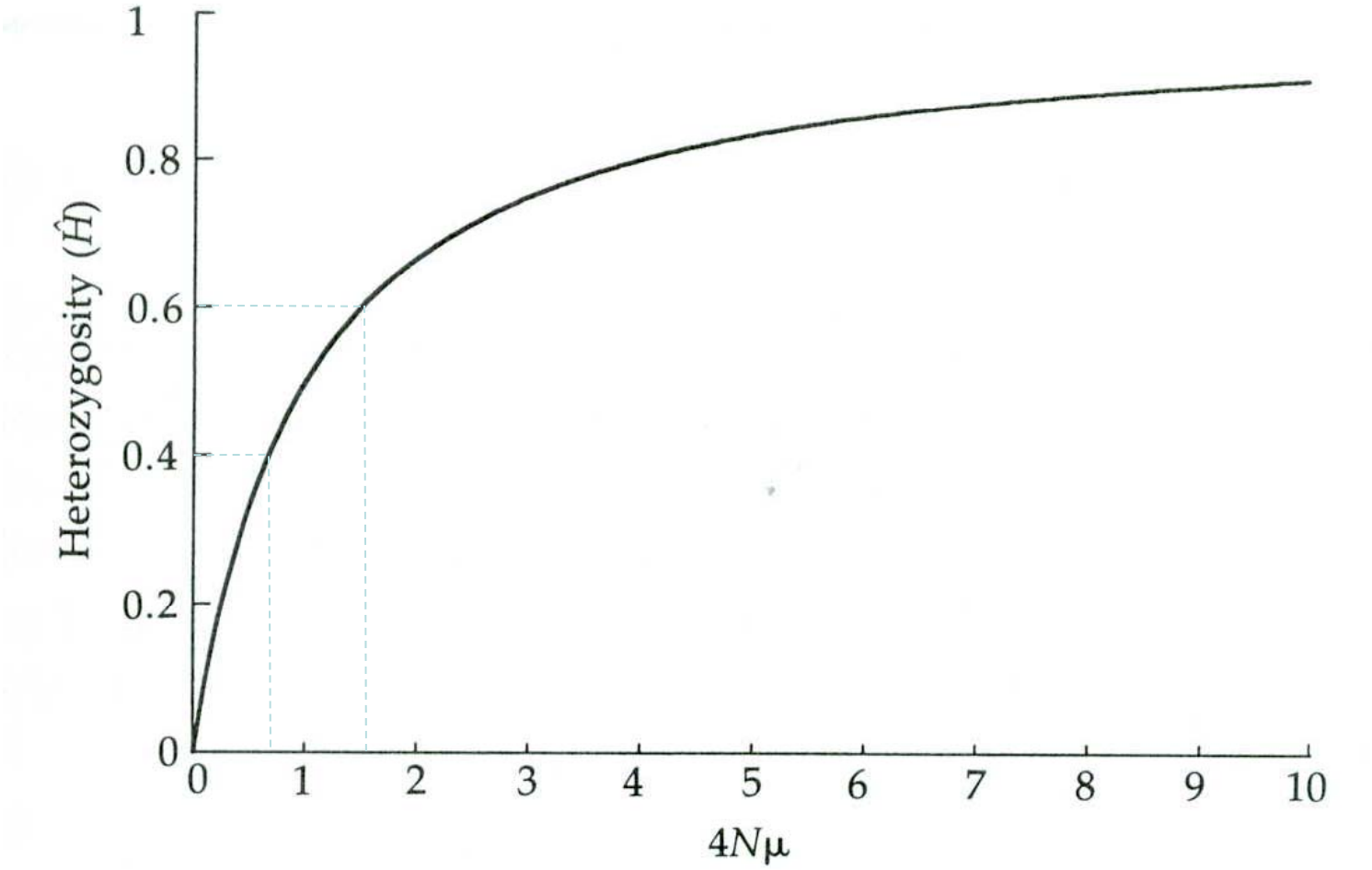
$$F = p_1^2 + p_2^2 + \ldots + p_n^2$$

INFINITE ALLELES MODEL

Since $F = \Sigma p_i^2$, then H = 1-F and its value is given by

$$H = 4N\mu / (4N\mu + 1)$$

INFINITE ALLELES MODEL



INFINITE ALLELES MODEL

Implications of mutation-drift equilibrium: not only H reaches the equilibrium but it is also possible to demonstrate that the configuration of allelic distributions also remains constant.

Ewens formula

$$\Pr(a_1, a_2, \ldots, a_k) = [n!\theta^k / \theta(\theta+1)\ldots(\theta+n-1)] \prod_{1}^{n} (1/i^{a_i} a_i!)$$

where: a_i is the nr. of alleles present i times, n is the sample size, k is the number of alleles and $\theta = 4N\mu$

INFINITE ALLELES MODEL

From Ewens formula one can verify the existence of a relationship between the sample size (n) and the number of alleles (k)

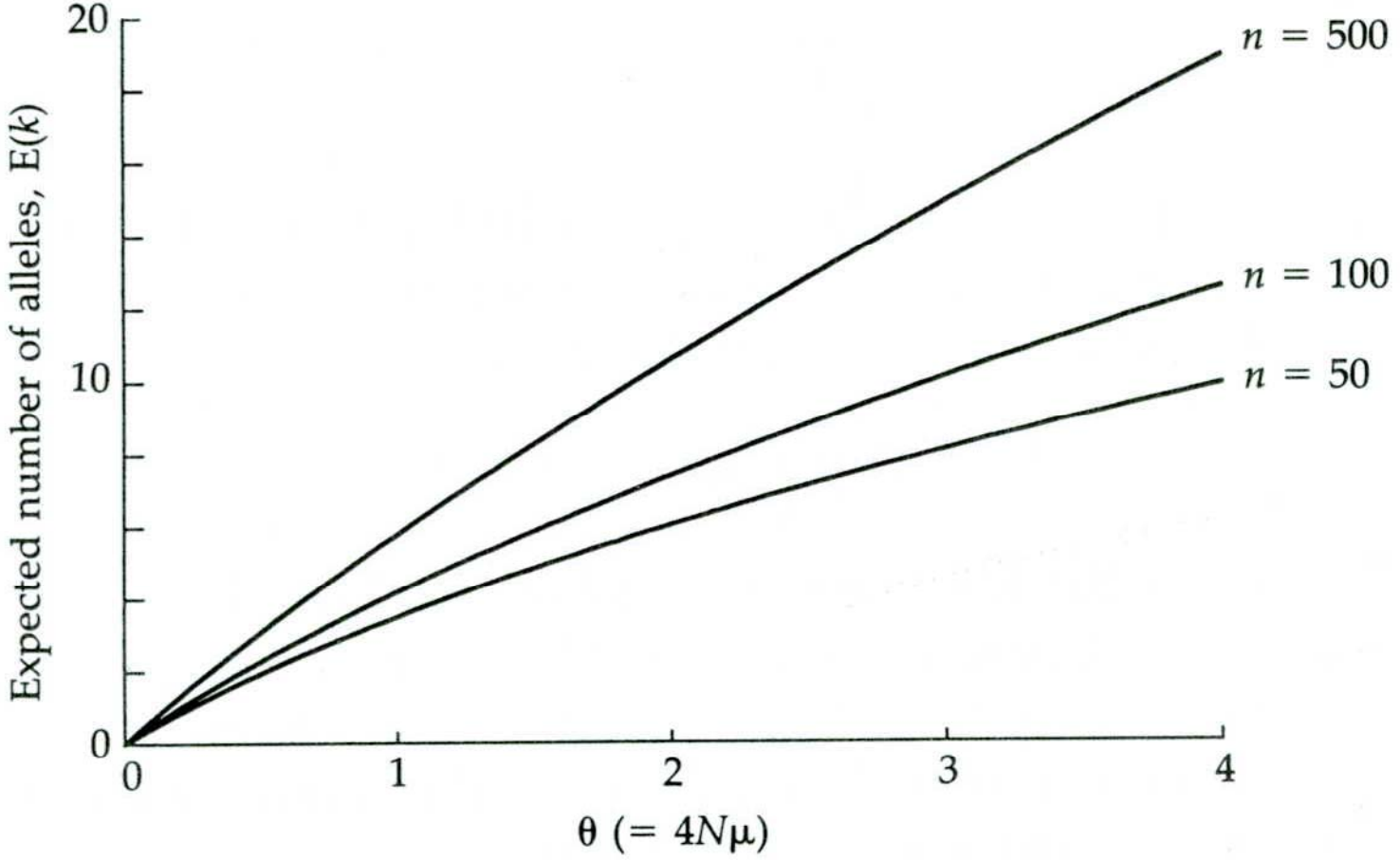
$$E(k) = 1 + \theta/(\theta+1) + \theta/(\theta+2) + \ldots + \theta/(\theta+n-1)$$

E(k) = expected nr. of alleles

If $\theta \approx 0$, then $E(k) \approx 1$

If θ is large, then $E(k) \approx n$

INFINITE ALLELES MODEL

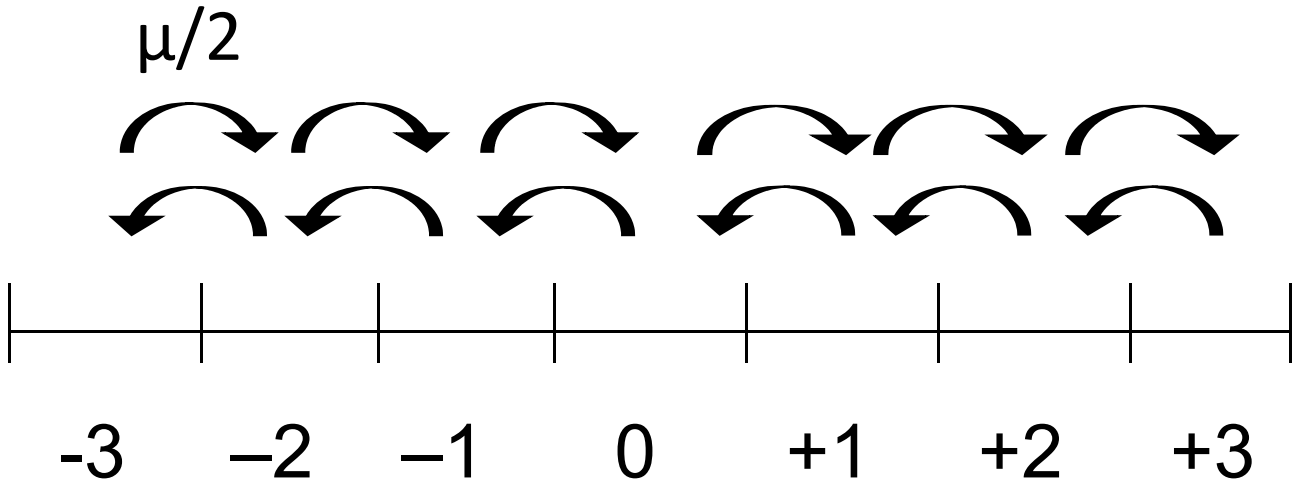


STEPWISE MUTATION MODEL

- Most mutations in microsatellites involve an increase or decrease of a single repeat unit;

- The opportunity for back mutation is thus much greater than for SNPs;

- Increase or decrease allele lenght by one unit with equal probability: STEPWISE MUTATION MODEL.



MICROSATELLITE MUTATION

However…

- Mutation rate tend to increase with array lenght;

- Dinucleotide repeat loci mutate more rapidly than tri- and tetranucleotide repeat loci;

- Pure repeat arrays mutate faster than interrupted arrays;

- Etc…

These factors are not incorporated in the SMM.

More complex models have thus been created (e.g. Two-Phase, Proportional Slippage, K-allele…).

INFINITE SITES MODEL

This model considers that mutation is rare enough that it always occurs in a site that was previously monomorphic. This way, almost all sites in a nucleotide sequence are considered monomorphic and those that vary present only two alelles.

BASE SUBSTITUTION MUTATION RATE

- Base substitutions are 10 times more frequent than insertions/deletions (indels);

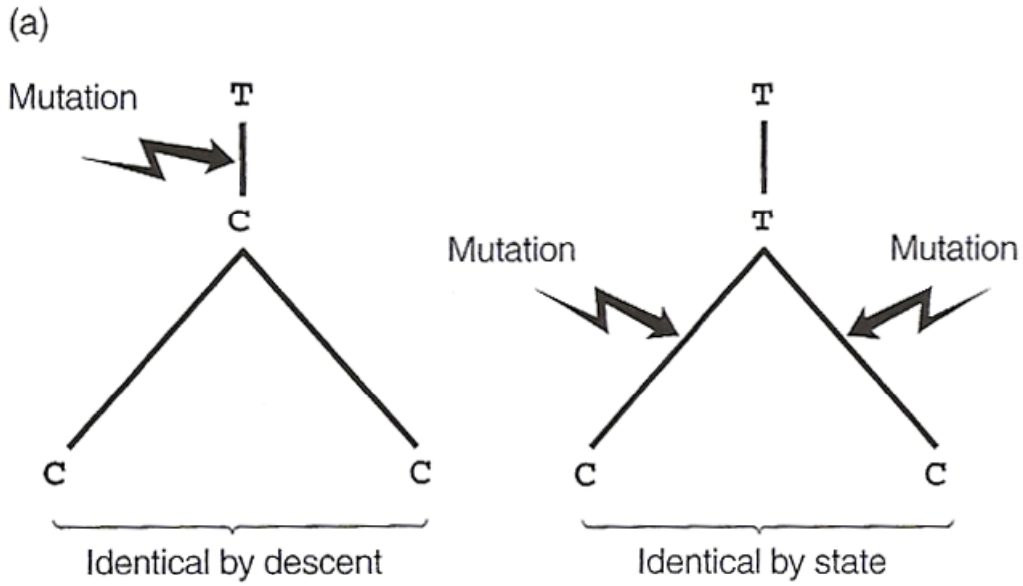
- Transitions are more than twice as frequent than transversions (contrary to the 1:2 expectation; error detection and repair, sequence context, differences in misincorporation rates…);

- Rates of mutations at CpG dinucleotides are one order of magnitude higher (methylation, deamination, repair).

Important bearings to the construction of models of sequence evolution.

MODELS OF DNA SEQUENCE EVOLUTION

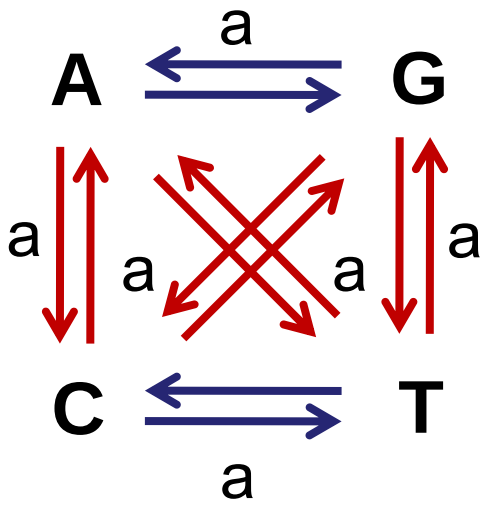
- When we are considering evolution over long time scales we may need to consider e.g. the possibility of multiple hits.

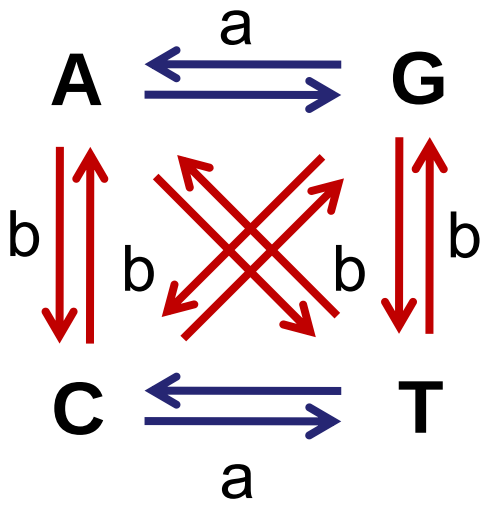


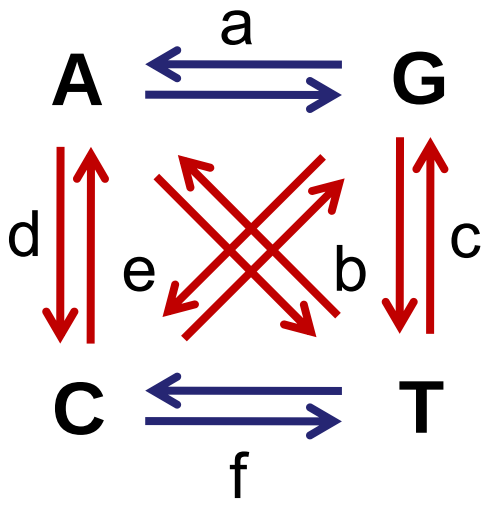
MODELS OF DNA SEQUENCE EVOLUTION

- When we are considering evolution over long time scales we may need to consider e.g. the possibility of multiple hits.

- There are numerous mutation models that consider particular rates for each nucleotide change.







MODELS OF DNA SEQUENCE EVOLUTION

- The frequency of each nucleotide can also influence the probability of nucleotide change: base composition.

- Other parameters such rate variation among sites within a sequence or the proportion of invariant sites can also be incorporated in the models.

- These models are particularly important for long evolutionary scales, where sequence divergence underestimates real divergence.

BASE SUBSTITUTION MUTATION RATE

- mtDNA has generally a much higher mutation rate than nuclear DNA.

- Reasons for this may include:

  - high concentration of mutagenic oxygen free radicals;

  - more replications per unit of time;

  - mechanism of replication implies long periods as single-stranded form;

  - absence of histones;

  - less effective repair systems.