Introduction to population genetics

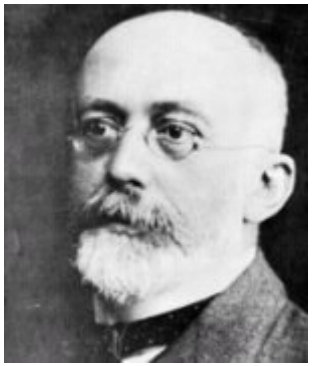
Population genetics

- Studies genetic variation in population a processes that affect it
- Understand evolution through mechanisms that change allel frequencies.

Hardy-Weinberg law

Godfrey Hardy

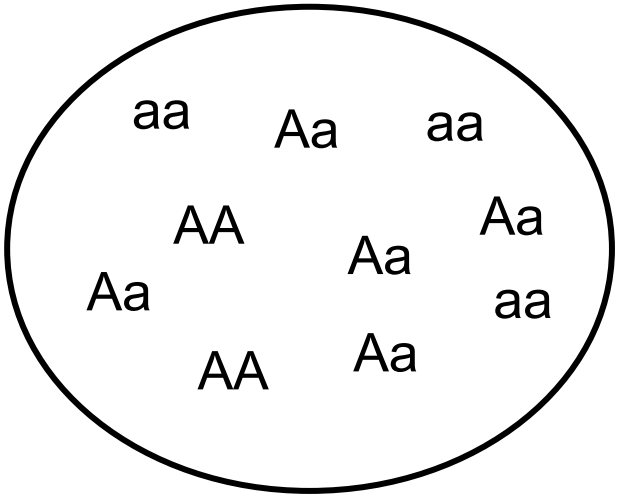
$$p^2 + 2pq + q^2 = 1$$



Wilhelm Weinberg

p …… frequency of alele A
q …… frequency of alele a

p^2 …… frequency of genotype AA
q^2 …… frequency of genotype aa
2pq …… frequency of genotype Aa

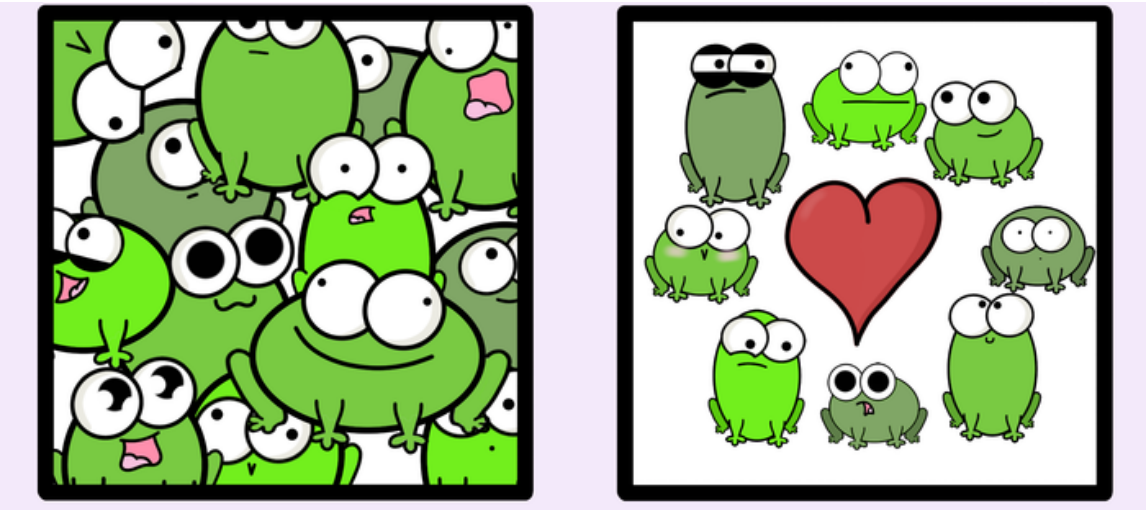


$p^2 = 2/10 = 0.2$
$q^2 = 3/10 = 0.3$
$2pq = 5/10 = 0.5$

$p = \sqrt{0.2} = 0.45$
$q = \sqrt{0.3} = 0.55$

Assumptions of Hardy-Weinberg equilibrium

- Large population.
- Random mating (in respect to particular genotype)
- No selection
- One generation of random mating can restore HW equilibrium.

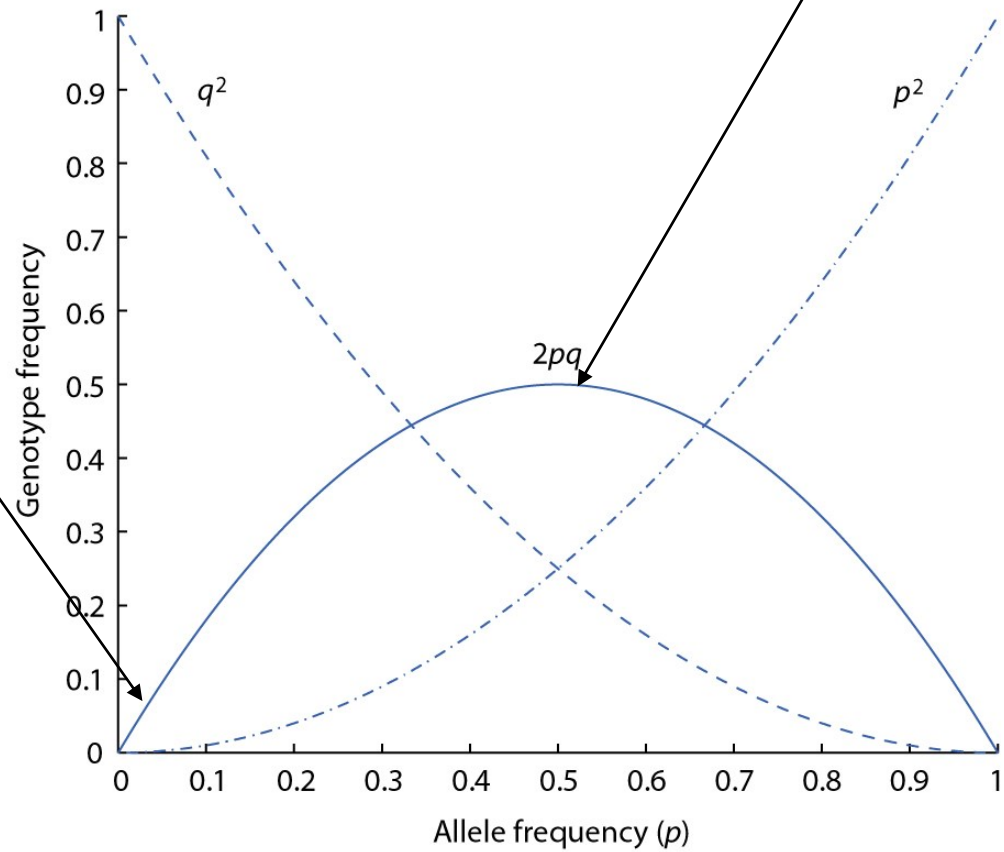


Hardy-Weinberg law

The recessive disadvantageus alleles can persist in population in very low frequencies.

The highest frequency of heterozygotes

The rare alleles hidden in heterozygotes.



Cystic fibrosis

Heterozygous carriers up to 1 from 25 individuals

Only about one from 3000 newbords affected

Hardy-Weinberg law

2 alleles

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

$$\boxed{p^2 + 2pq + q^2 = 1}$$

3 alleles

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

$$\boxed{p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1}$$

4 alleles

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

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

etc.

How to test whether population is in the Hardy-Weinberg equilibrium?

Genotypes	Observed numbers
AA	125
Aa	550
aa	325

Test of Hardy-Weinberg equilibrium

Genotypes	Observed numbers
AA	125
Aa	550
aa	325
all	1000

Calculation of allele frequencies

$$f_A = \frac{2 \times 125 + 550}{2 \times 1000} = 0.4$$

$$f_a = 1 - A = 0.6$$

Estimation of expected numbers of genotypes

AA	$1000 \times (0.4)^2 = 160$
Aa	$1000 \times 2(0.4)(0.6) = 480$
aa	$1000 \times (0.6)^2 = 360$

Difference between observed and expected no. of genotypes

AA	125 – 160 = -35
Aa	550 – 480 = 70
aa	325 – 360 = -35

Chi-square statistics (X^2)

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Degree of freedom (df)
number of different genotypes
– 1 (no. of estimated variables)
– 1

Chi-square test

$$X^2 = \frac{(-35)^2}{160} + \frac{(70)^2}{480} + \frac{(-35)^2}{360} = 21.27$$

df = 3-1-1=1 **$p < 0.01$**

Population is not in HW equilibrium.

Mechanisms causing deviations from HW equilibrium

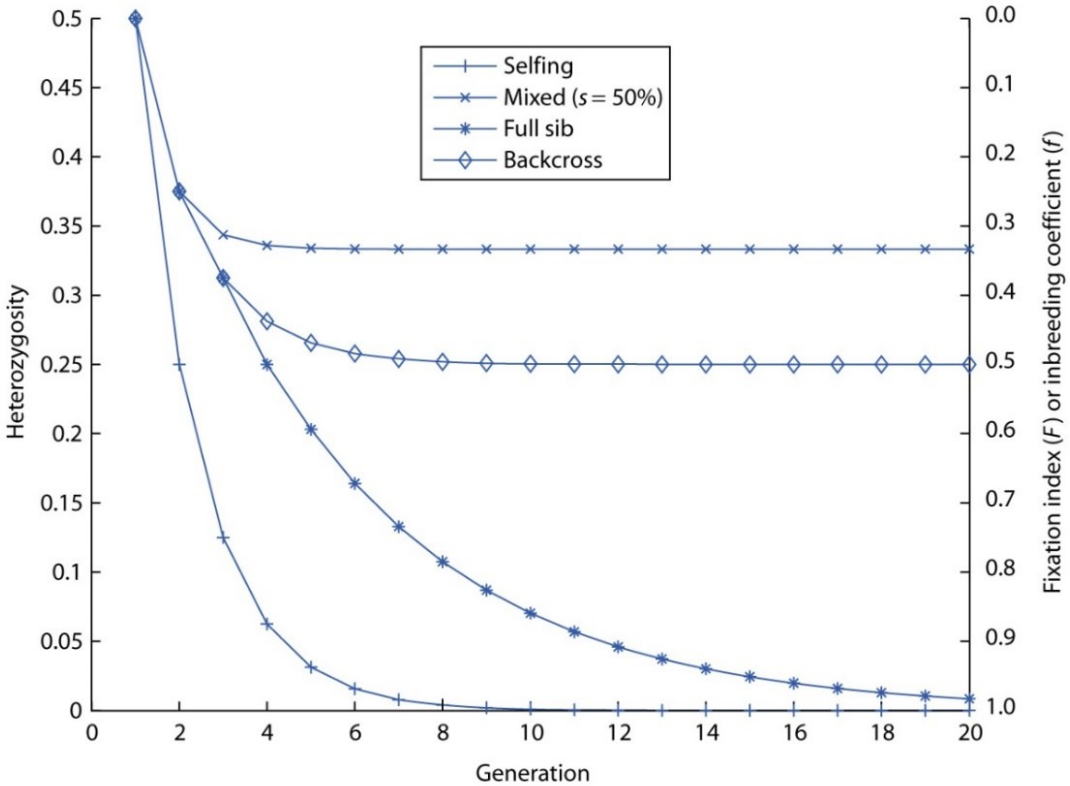
Assortative mating

Positive assortative mating

- Individuals with similar phenotypes (genotypes) mate with one another more frequently than would be expected under a random mating.
- Inbreeding
- Excess of homozygotes in population.
- Recesive disadvantageous alleles can be manifested (inbreeding depression).

Inbreeding

Establishment of laboratory inbred lines.

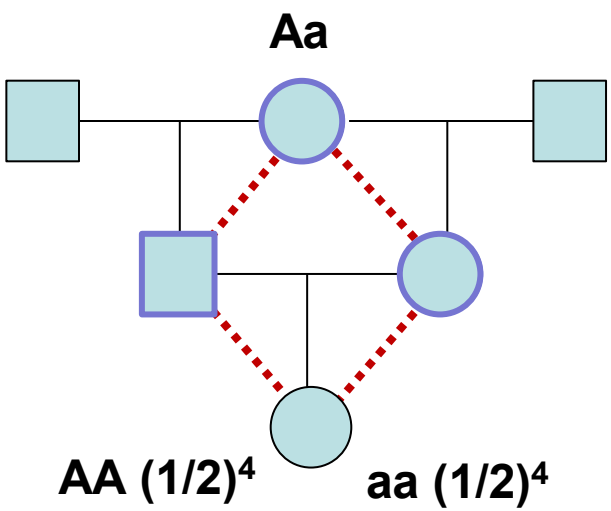


Coeficient of inbreeding (F)

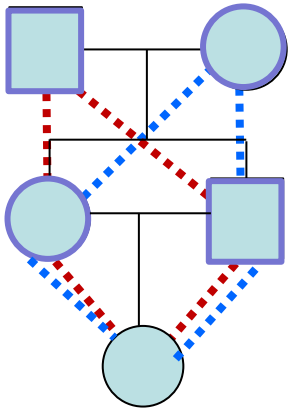
The probability that two alleles at any locus in an individual are identical by descent, i.e. inherited from the common ancestor.

$F = (1/2)^n$

n = number of individuals in the genealogy from the given individual to the common ancestor.



$F = (1/2)^3 = 1/8$



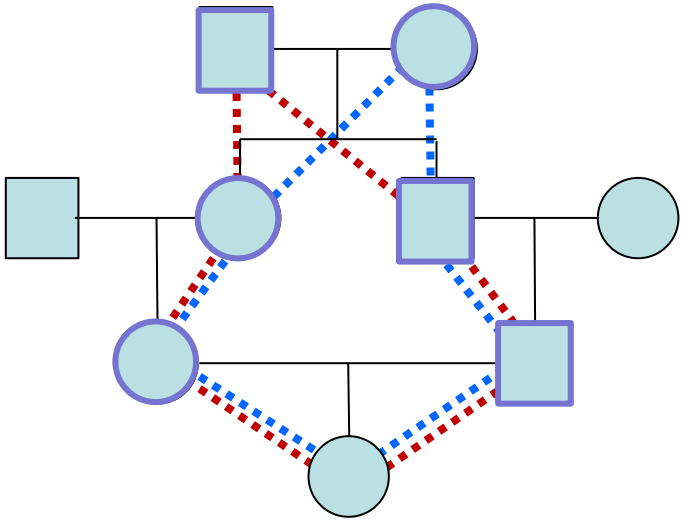
$F = 1/8 + 1/8 = 1/4$

Coeficient of inbreeding (F)

The probability that two alleles at any locus in an individual are identical by descent, i.e. inherited from the common ancestor.

$F = (1/2)^n$

n = number of individuals in the genealogy from the given individual to the common ancestor.

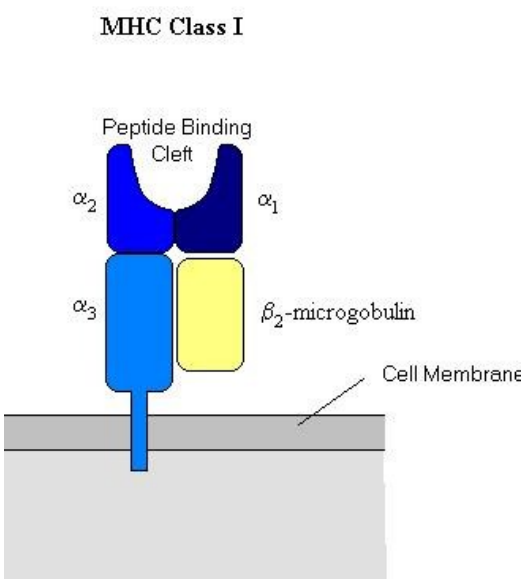


$F = (1/2)^5 + (1/2)^5$

$F = 1/32 + 1/32 = 1/16$

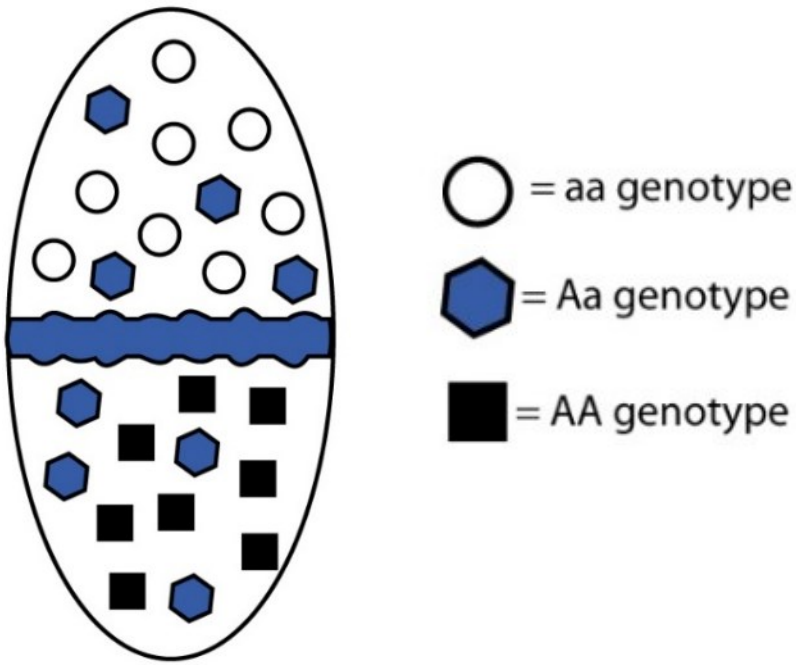
Negative assortative (disassortative) mating

- Individuals with dissimilar phenotypes (genotypes) mate with one another more frequently than would be expected under a random mating.
- Excess of heterozygotes.
- Example: MHC I (*Major Histocompatibility Complex I*) genes



Geographic population structure

- Geographic barriers prevent random mating.
- Reduces frequency of heterozygotes = **Wahlund effect.**



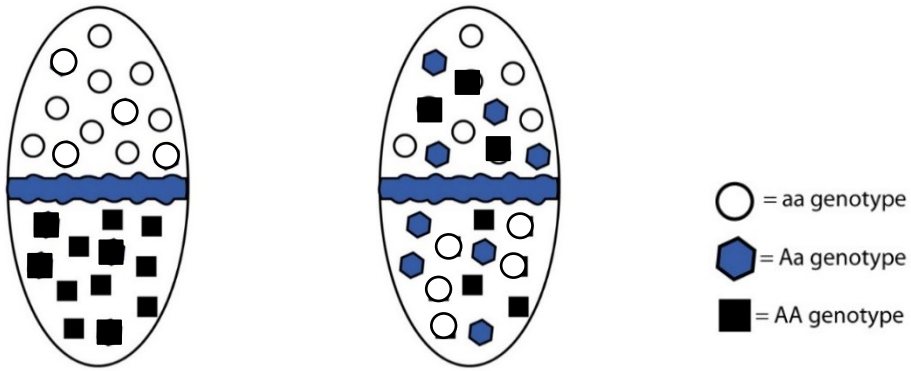
Estimation of levels of population structure

F_{ST} statistics (fixation index)

The expected degree of a reduction in heterozygosity when compared to Hardy–Weinberg expectation.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

H_T expected heterozygosity for whole population
H_S expected heterozygosity for subpopulations



$F_{ST} = 1$ $F_{ST} = 0$

Calculation of F_{ST} between two subpopulations

	Genotype numbers				Allele frequencies	
	AA	**Aa**	**aa**		**p**	**q**
subpopulation 1	25	50	25	100	0.5	0.5
subpopulation 2	49	42	9	100	0.7	0.3
whole population	74	92	34	200	0.6	0.4

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$H_T = 2pq = 2 \,.\, 0.6 \,.\, 0.4 = 0.48$

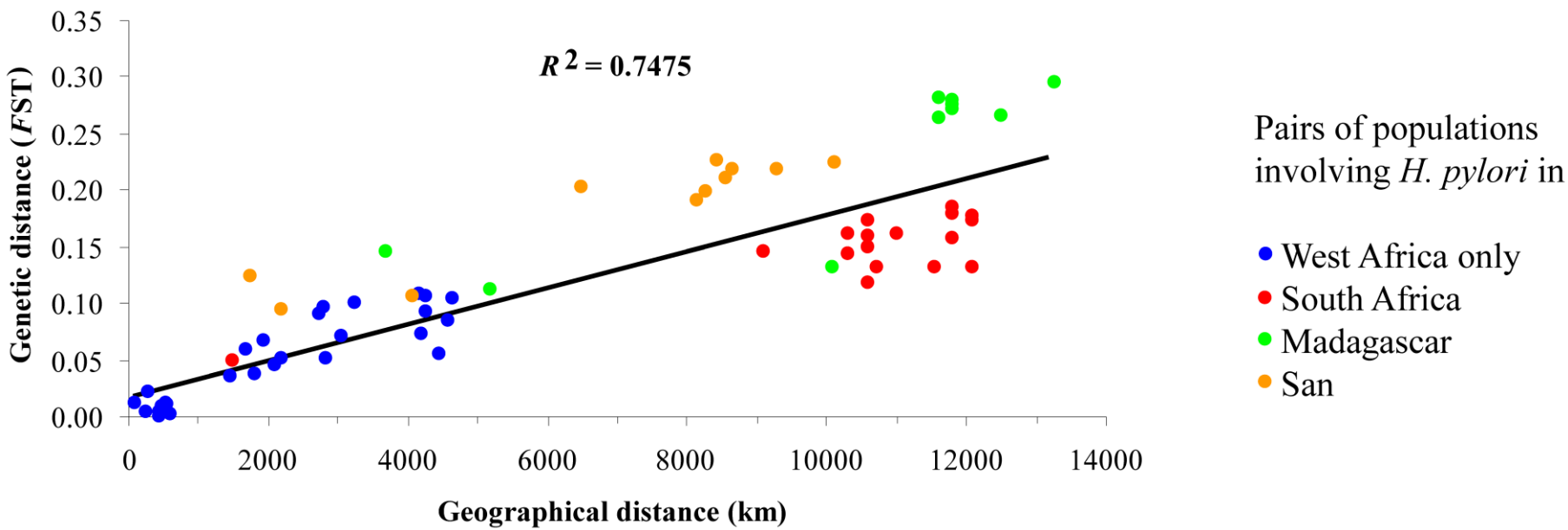
$H_S = p_1q_1 + p_2q_2 = 0.25 + 0.21 = 0.46$

$F_{ST} = 0.0416$

Isolation by distance

Genetic differentiation (Fst) increases with geographic distance.

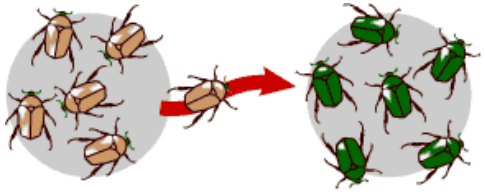
.



Mechanisms changing allele frequencing

mutation

drift

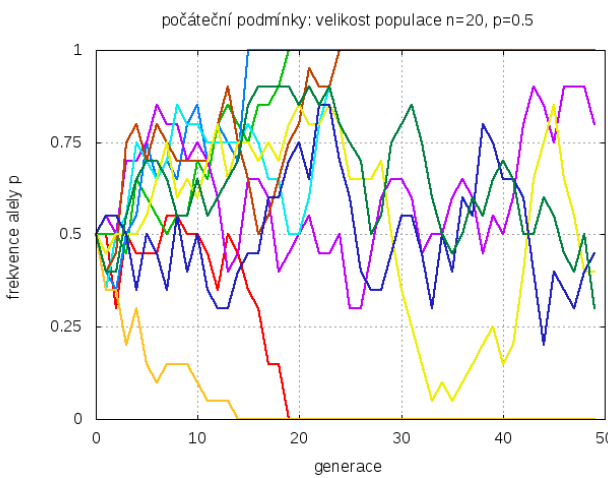


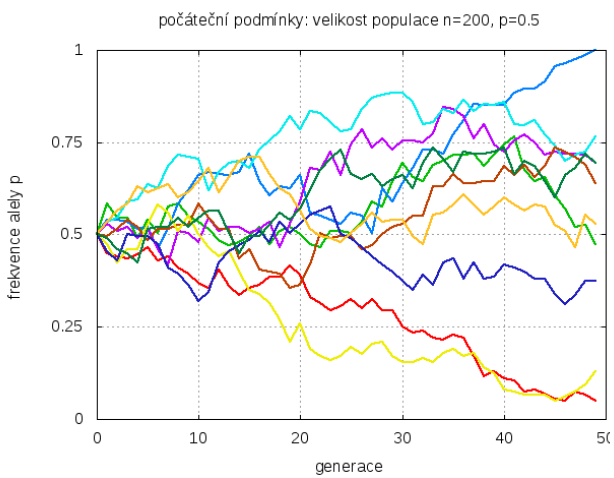
migration

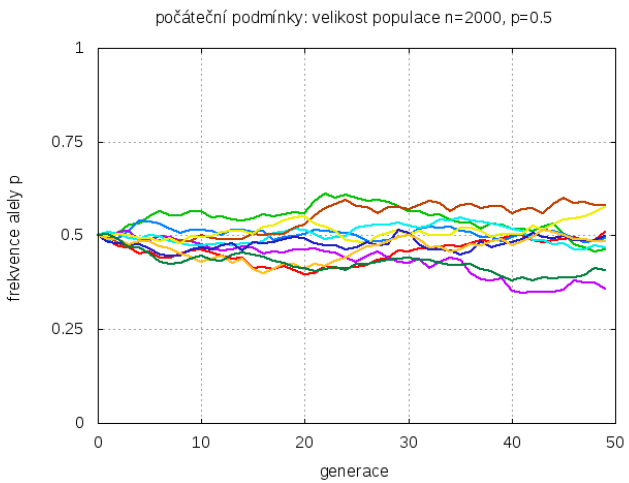
selection

Genetic drift

- Random changes in allele frequencies.
- Stronger in smaller populations.
- Reduces genetic variation.
- Probability of fixation is given by the allele frequency.
- Time to fixation is 4N (generations).

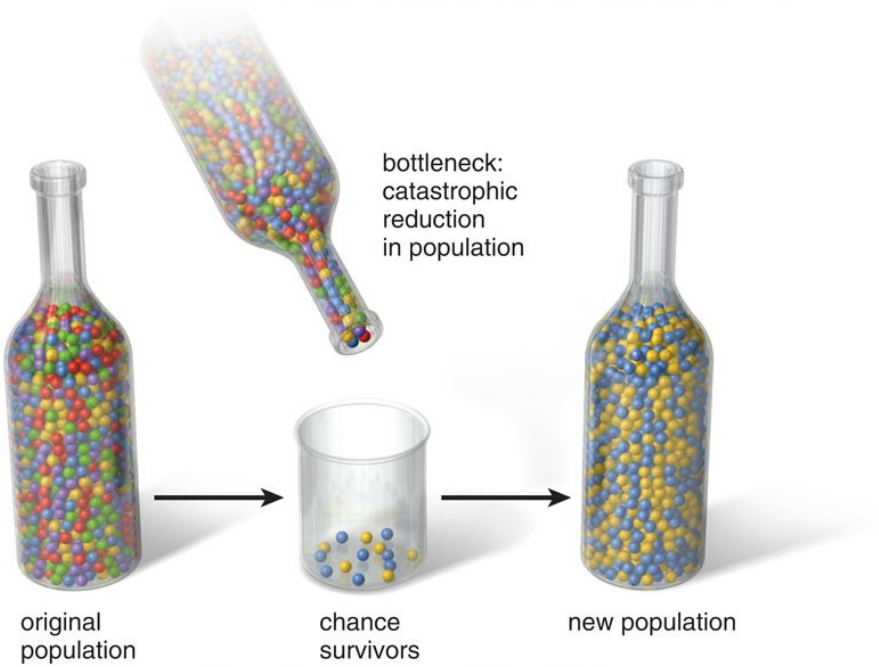




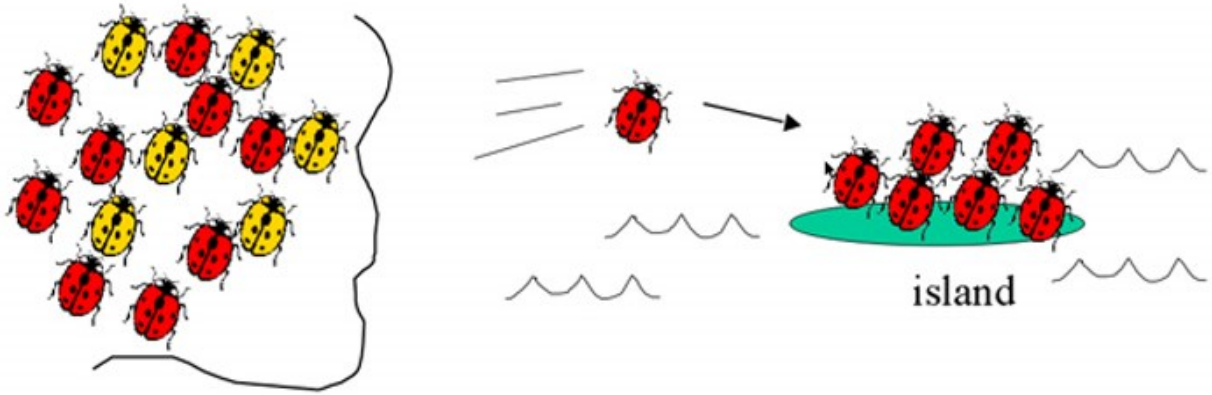


Genetic drift

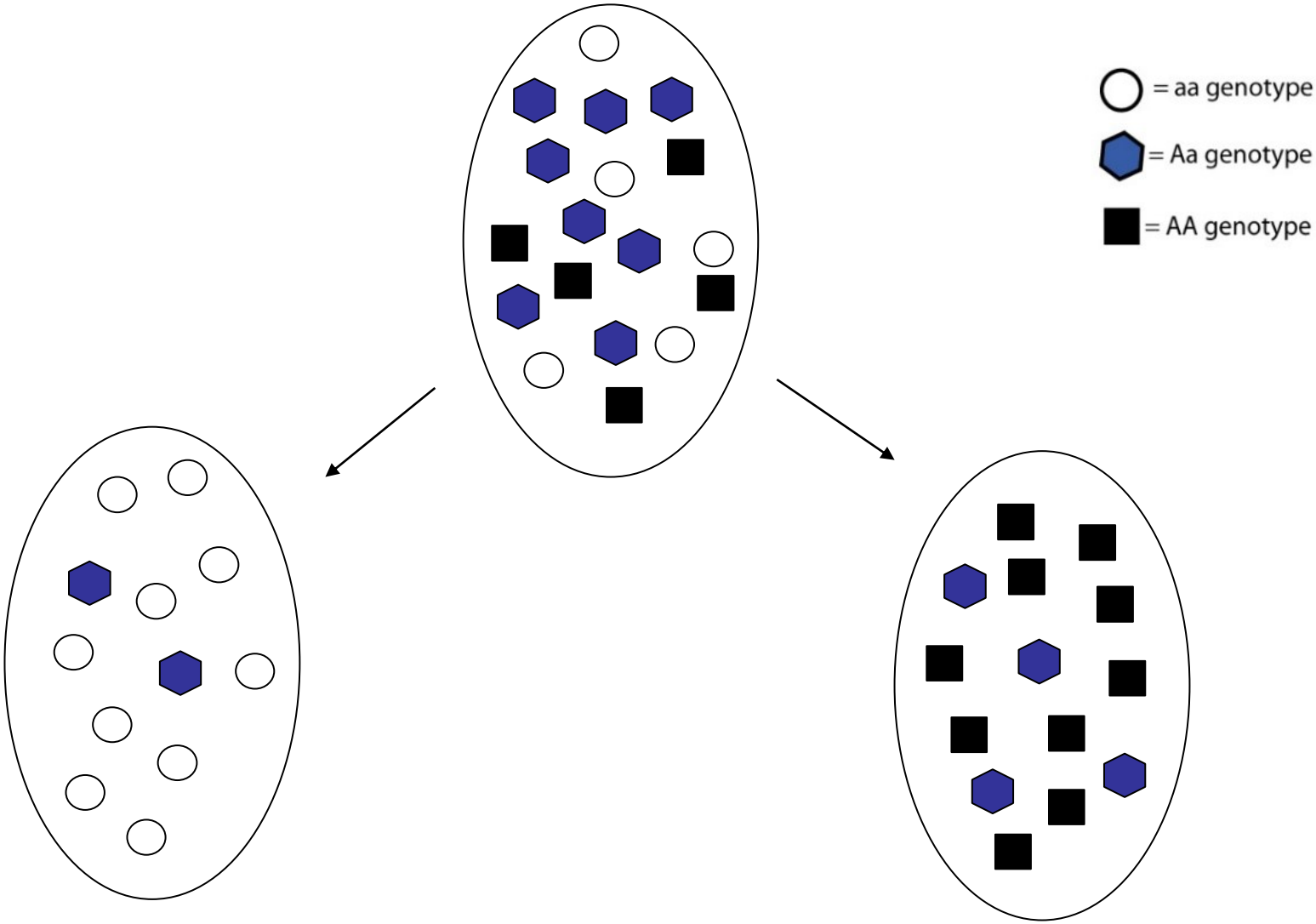
Bottleneck effect



Founder effect



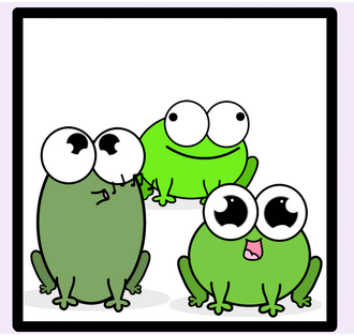
Genetic drift leads to differentiation of isolated populations



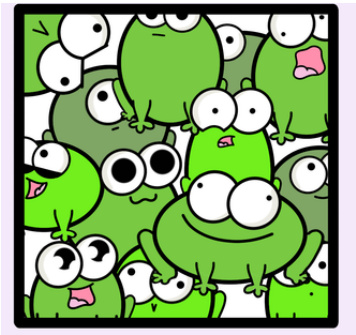
Effective population size (N_e)

- Number of breeding individual in the population.
- Is the size of ideal panmictic population (that meet all the HW assumptions), in which genetic proceses (e.g. genetic drift) have the same effects as in the real population.

Ideal population
N = 100



Real population
?

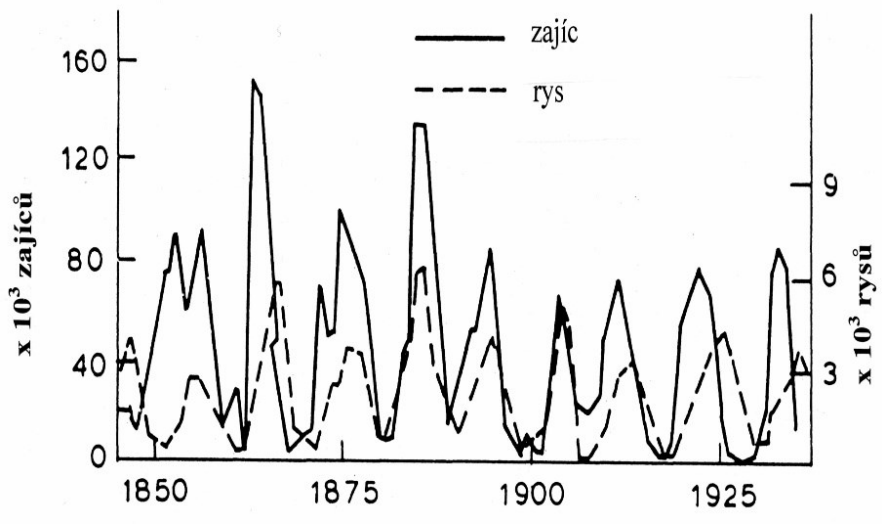


The strength of genetic drift is the same.

Factors affecting N_e

Changes in population size

- N_e reflects past changes in the population size.
- Correspond more to lower population sizes than higher population sizes.



Factors affecting N_e

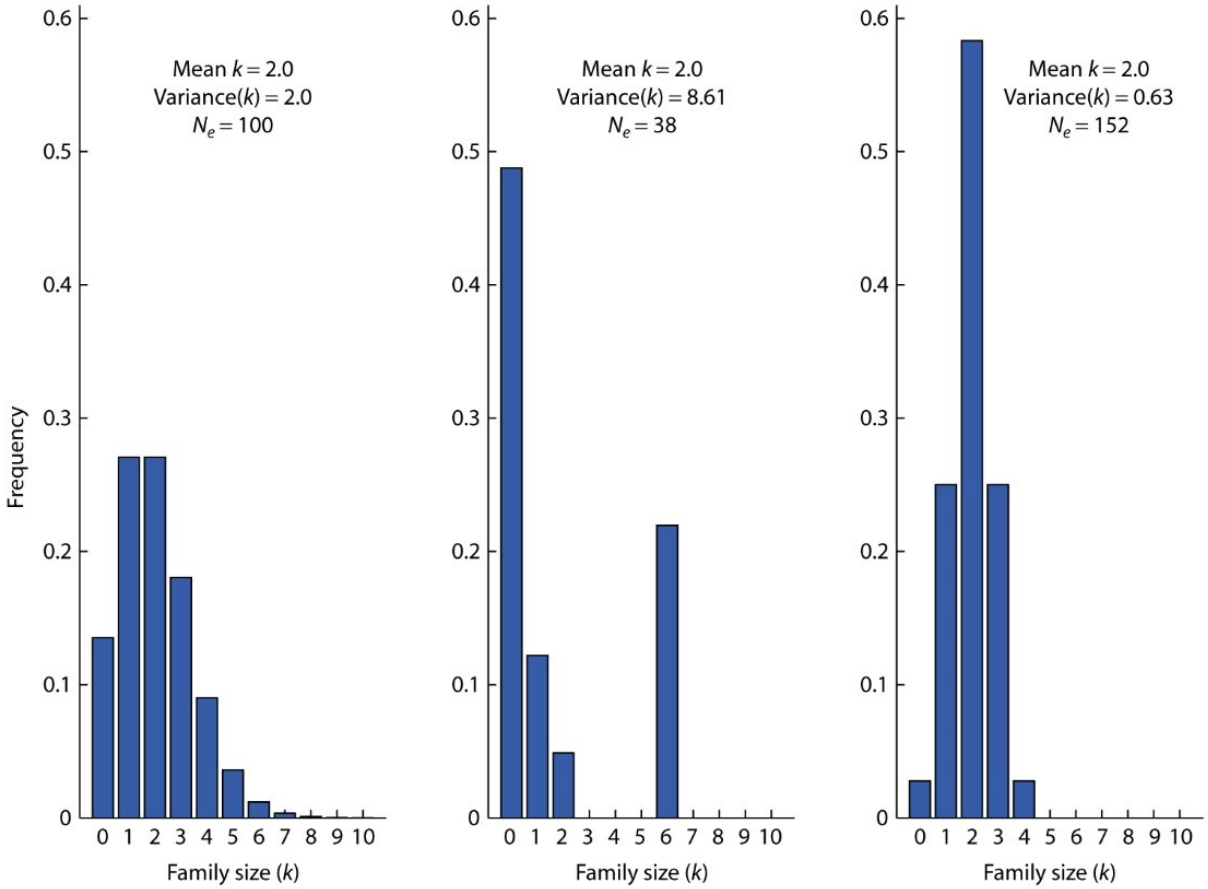
Different number of reproducing males and females

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

Factors affecting N_e

Selection (variation in number of progeny among individuals)

- Large variation in number of progeny among individuals reduces N_e



Estimates of N_e

~ 10 000

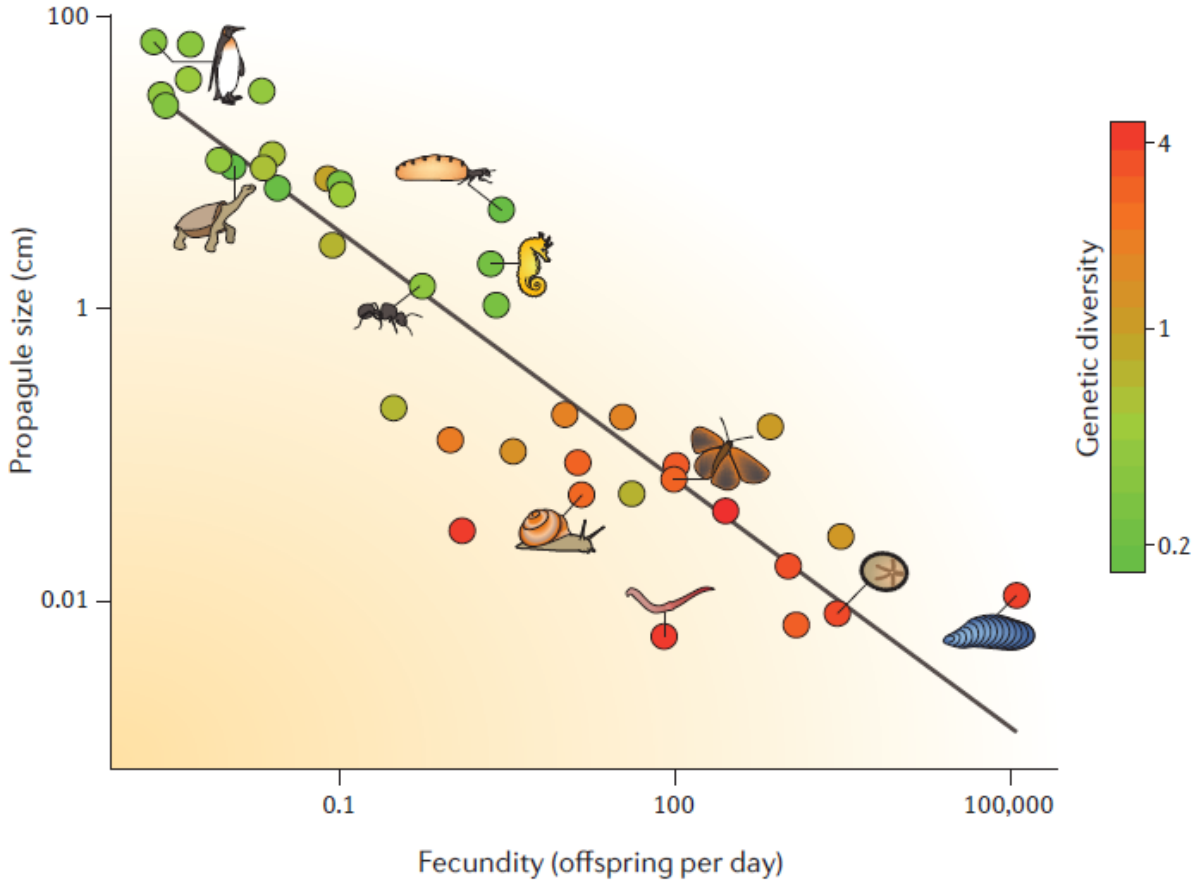
~ 30 000

~ 2 000 000

~ 1 000 000

~ 100 000

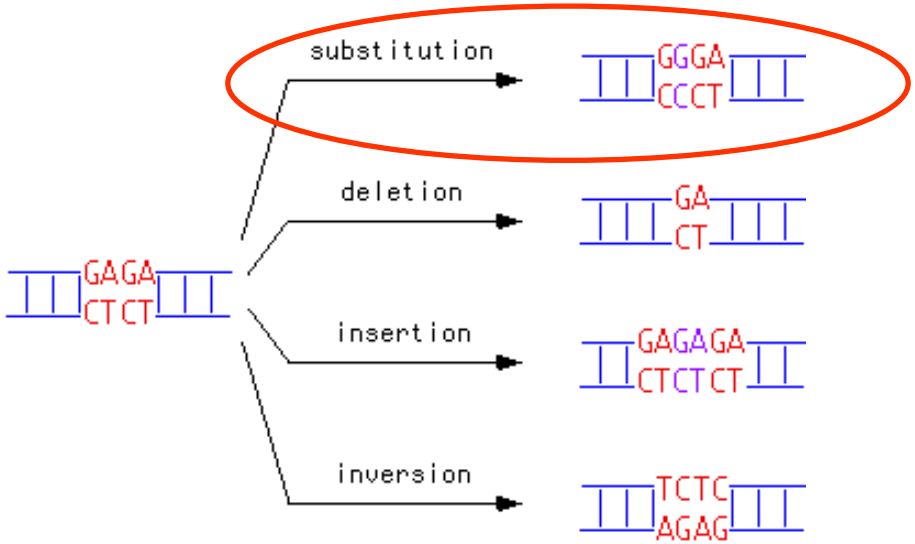
Effective population size and life strategies



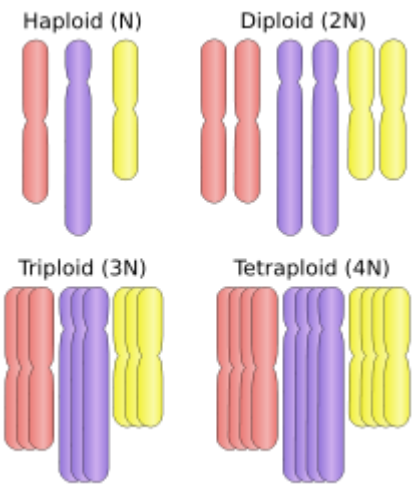
Ellegren and Galtier, 2016

Mutations: the source of genetic variation

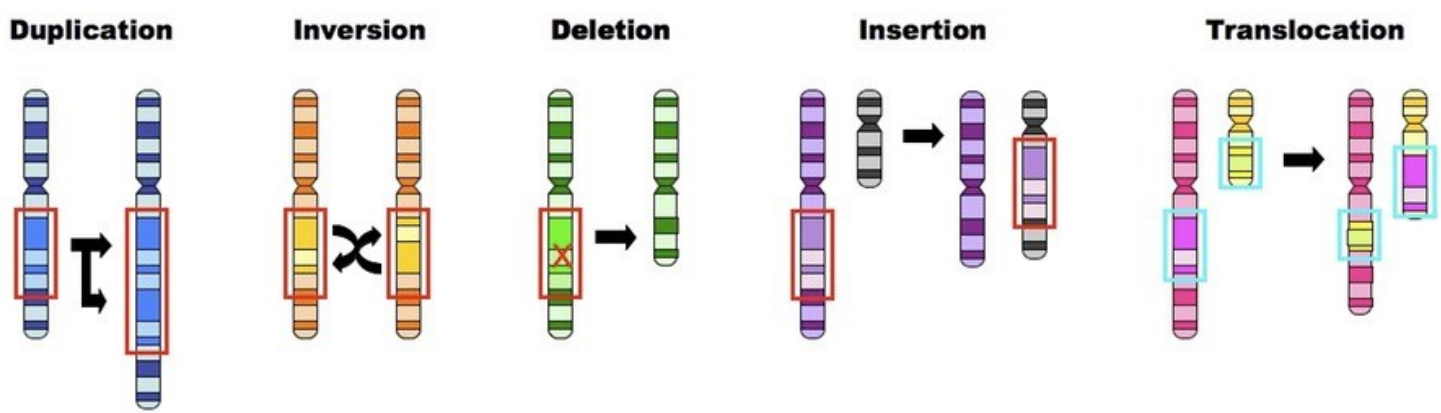
Point/gene mutations



Genome mutations



Chromosomal mutations



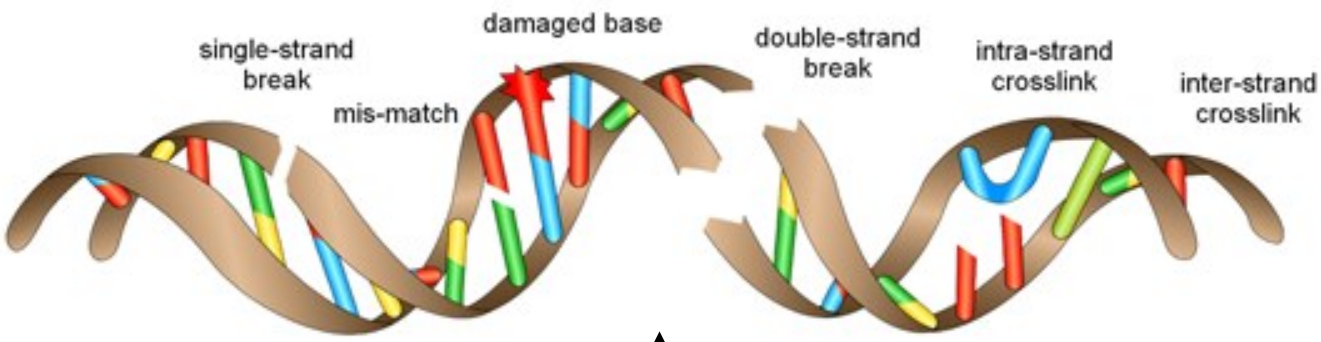
Mutations

Cell metabolism

Physical mutagens (UV light, gamma rays)

Chemical mutagens

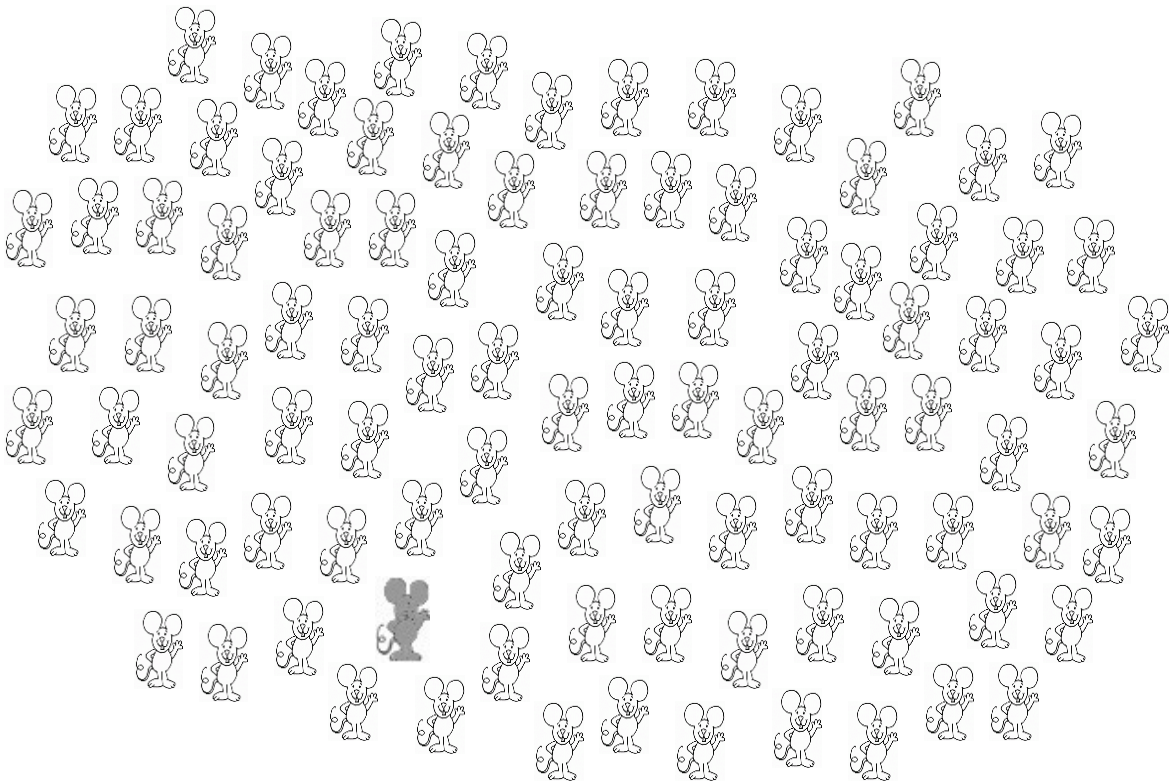
Replication errors



DNA reparation mechanisms

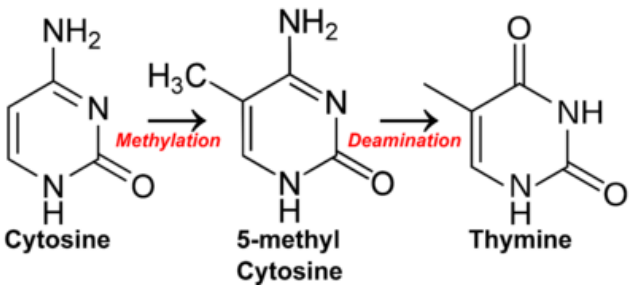
Mutation rate (μ)

- Frequency of new mutations per generation
- μ = 0.01 (1 mutation per 100 individuals in 1 generation)



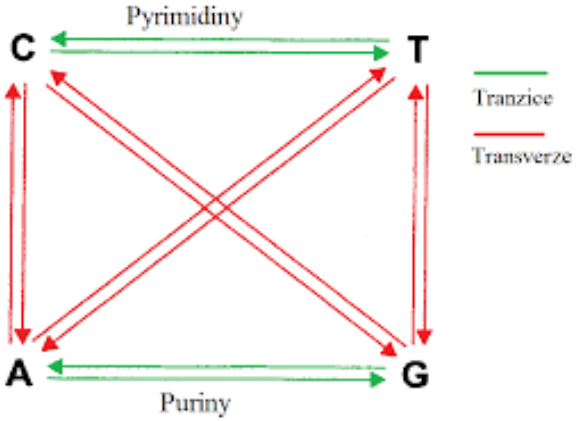
- Mutation rate (for nucleotide substitutions) in humans is cca 1.10^{-8}.
  (cca 100 new substitutions in the genome of the individual each generation)

Mutation rate

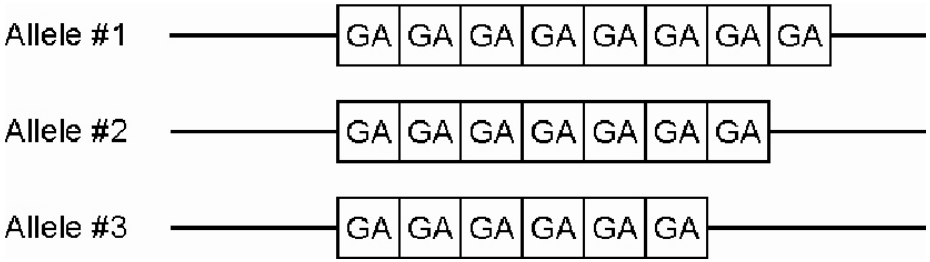


Different for various mutation types

- Transitions are more frequent than transversions
- mutation „hotspots“: CpG islands, mikrosatelites



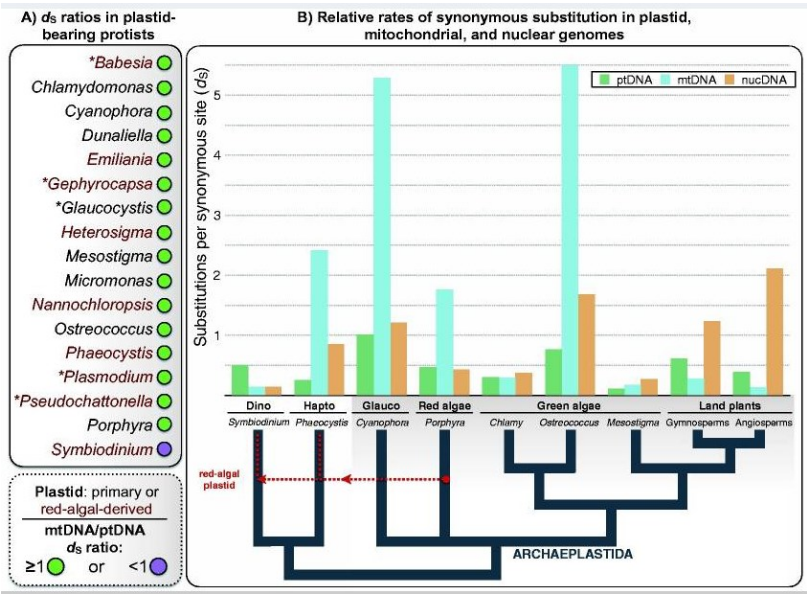
mikrosatelit



Mutation rate

Different for various genes in the genome

- Affected for example by transcription rate, distance to replication origin, nucleosome position etc.
- In animals cca 10x higher in mtDNA than in nuclear DNA.



GBE

Mutation Rates in Plastid Genomes: They Are Lower than You Might Think

David Roy Smith*

Department of Biology, University of Western Ontario, London, ON, Canada

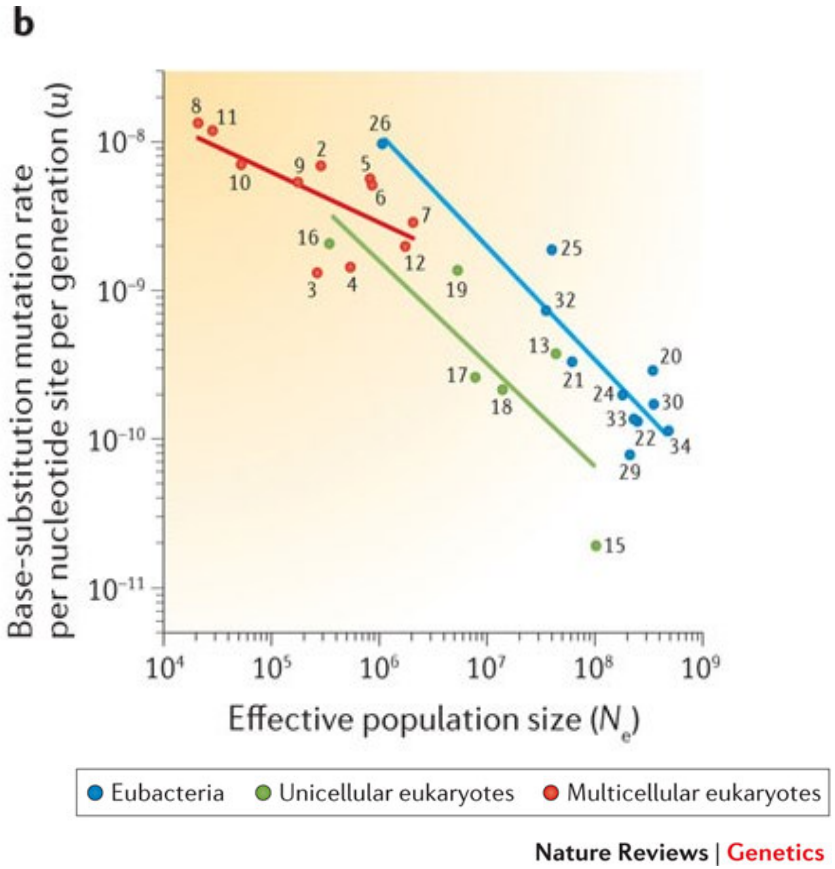
*Corresponding author: E-mail: dsmit242@uwo.ca.

Accepted: April 9, 2015

Mutation rate

Different in various organisms

- RNA viruses, DNA viruses, multicelular eukaryota, bacteria, unicellular eukaryota



Lynch et al. 2016

Germline mutation rate

Table 1. Mutation rates per nucleotide site ($\times 10^{-9}$) in different tissues[a]

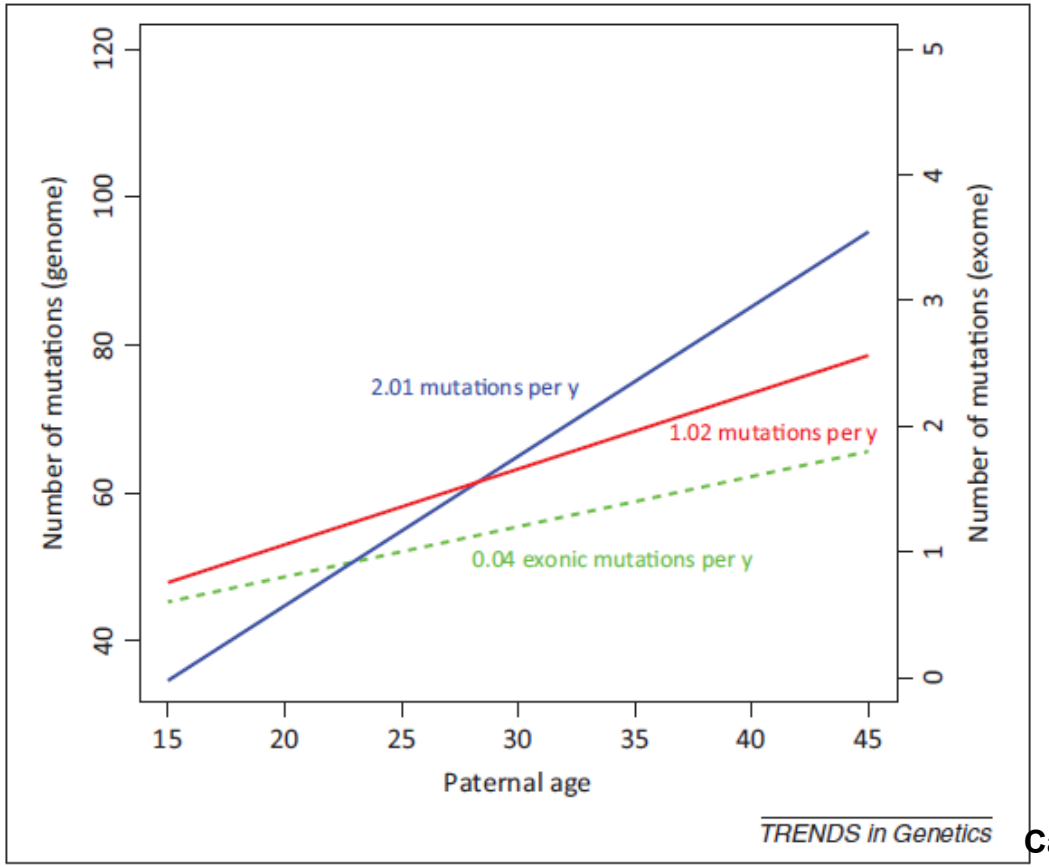
Species	Tissue	Cell divisions per generation[a]	Mutation rates[b]	
			Per generation	Per cell division
Homo sapiens	Germline	216	12.85	0.06
	Retina	55	54.45	0.99
	Intestinal epithelium	600	162.00	0.27
	Fibroblast (culture)			1.34
	Lymphocytes (culture)			1.47
Mus musculus	Male germline	39	38.00	0.97
	Brain		76.94	
	Colon		83.35	
	Epidermis		90.38	
	Intestine		117.69	
	Liver		237.88	
	Lung		166.83	
	Spleen		130.00	
Rattus norvegicus	Colon		178.38	
	Kidney		167.45	
	Liver		179.92	
	Lung		223.22	
	Mammary gland		57.70	
	Prostate		448.90	
	Spleen		101.62	
Drosophila melanogaster	Germline	36	4.65	0.13
	Whole body		380.92	
Caenorhabditis elegans	Germline	9	5.60	0.62
Arabidopsis thaliana	Germline	40	6.50	0.16
Saccharomyces cerevisiae		1	0.33	0.33
Escherichia coli		1	0.26	0.26

[a]References to data on numbers of germline cell divisions: human [Crow 2000]; *D. melanogaster* and mouse [57]; *C. elegans* [58]; and *A. thaliana* [59]. Numbers of cell divisions are unknown for the mouse and rat rates.

[b]Mammalian tissue-specific rates are given only for tissues in which at least two independent estimates have been acquired. All data on human mutation rates are taken from Lynch [36]. Data for somatic mutation rates in mouse and rat are derived from references contained within the supplementary material online. References to data on germline mutation rates are: *D. melanogaster* [5], *C. elegans* [4], *A. thaliana* [Ossowski *et al.*, 2009], *S. cerevisiae* [3], and *E. coli* [24].

Most mutations are inherited from fathers

- Male driven evolution
- Paternal age effect.
- Different mutation rate on sex chromosomes and autosomes.

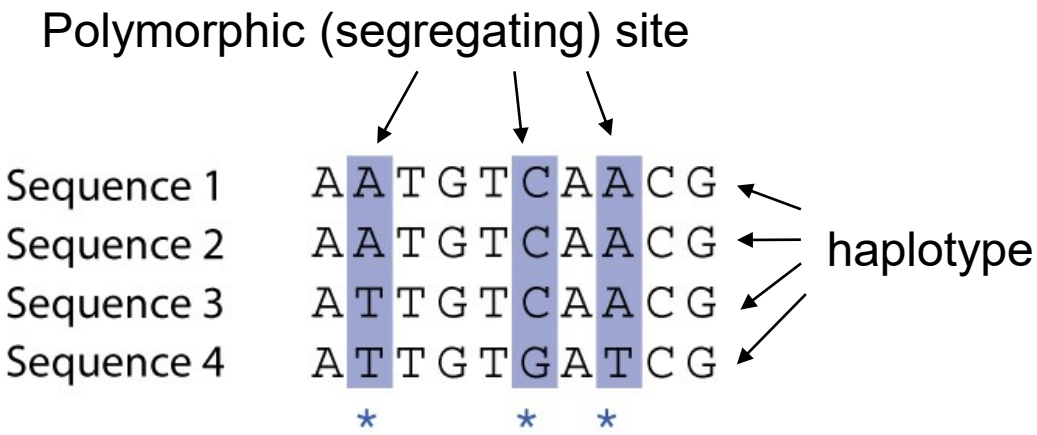


Campbell and Eicher, 2013

Genetic diversity - genetic polymorfism (θ)

For neutral sequence:

$$\theta = 4N_e\mu$$



- Levels of genetic diversity in population increases with increasing population size and mutation rate

Empirical estimates of genetic diversity

Nucleotide diversity (average heterozygosity) (π)

Sequence 1	AATGTCAACG
Sequence 2	AATGTCAACG
Sequence 3	ATTGTCAACG
Sequence 4	ATTGTGATCG
Site number	1 2 3 4 5 6 7 8 9 10

Nucleotide diversity (π):

1 AATGTCAACG
2 AATGTCAACG $d_{12} = 0$

1 AATGTCAACG
3 ATTGTCAACG $d_{13} = 1$

2 AATGTCAACG
3 ATTGTCAACG $d_{23} = 1$

1 AATGTCAACG
4 ATTGTGATCG $d_{14} = 3$

2 AATGTCAACG
4 ATTGTGATCG $d_{24} = 3$

3 ATTGTCAACG
4 ATTGTGATCG $d_{34} = 2$

$$\Sigma d_{ij} = 0 + 1 + 3 + 1 + 3 + 2 = 10$$

Number of pairs of sequences compared = $[n(n-1)]/2 = [4(3)]/2 = 6$
$\hat{\pi}$ = 10 differences/6 pairs = 1.67 average pairwise differences
$\hat{\pi}$ = 1.67 avg. differences/10 sites = 0.167 pairwise differences per site

π = 0.01 on average one polymorphic site per 100 bp

Proportion of polymorfic sites (θ_W)

Sequence 1	A	A	T	G	T	C	A	A	C	G
Sequence 2	A	A	T	G	T	C	A	A	C	G
Sequence 3	A	T	T	G	T	C	A	A	C	G
Sequence 4	A	T	T	G	T	G	A	T	C	G
		*				*		*		
Site number	1	2	3	4	5	6	7	8	9	10

Segregating sites (S and p_S):

Sites 2, 6, and 8 have variable base pairs among the four sequences (columns marked with *). These are segretating sites. Therefore, for these sequences $S = 3$ segregating sites and $p_S = 3/10 = 0.3$ segregating sites per nucleotide site examined.

$$\theta = S / n / H_{k-1}$$

$$H_{k-1} = 1 + \frac{1}{2} + \frac{1}{3}$$

$$\theta = 3 / 10 / 1{,}83 = 0{,}164$$

S … number of segregating sites

n … number of nucleotides in the sequence

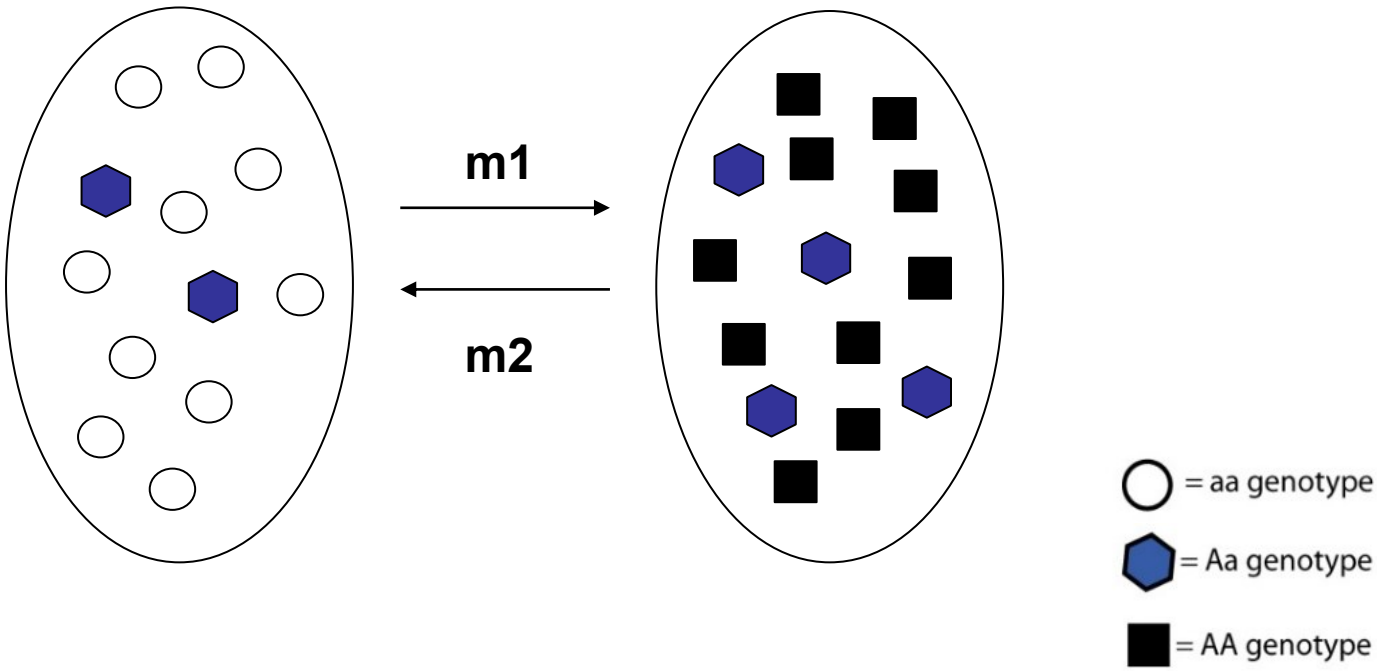
k … number of sequences

H_{k-1}… harmonic number

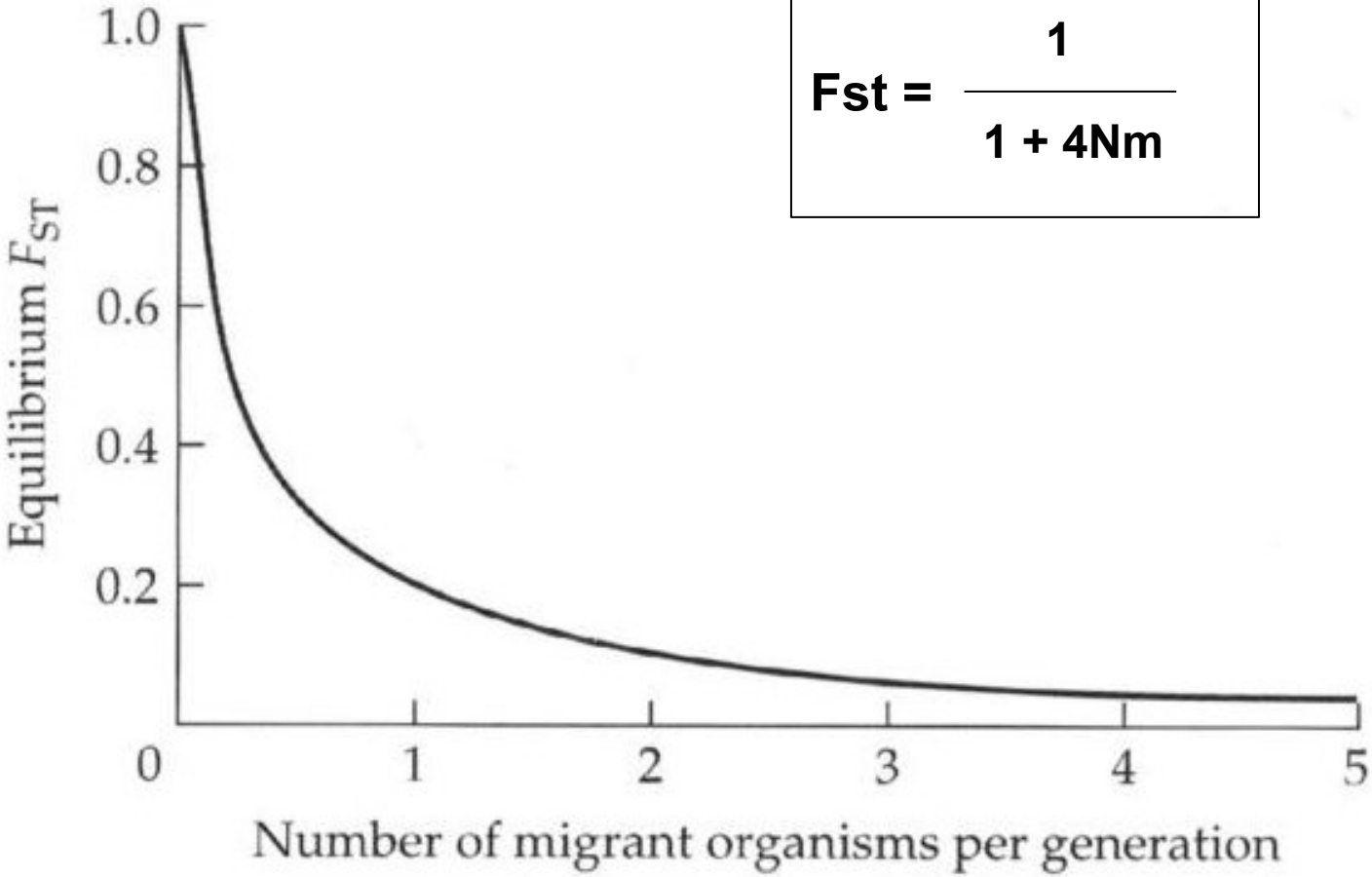
Program: DnaSP

Migration (m)

- **Migration rate (m):** probability that individual from one population will substitute individual from another population (per generation).
- **Population migration rate (Nm):** number of migrants per generation

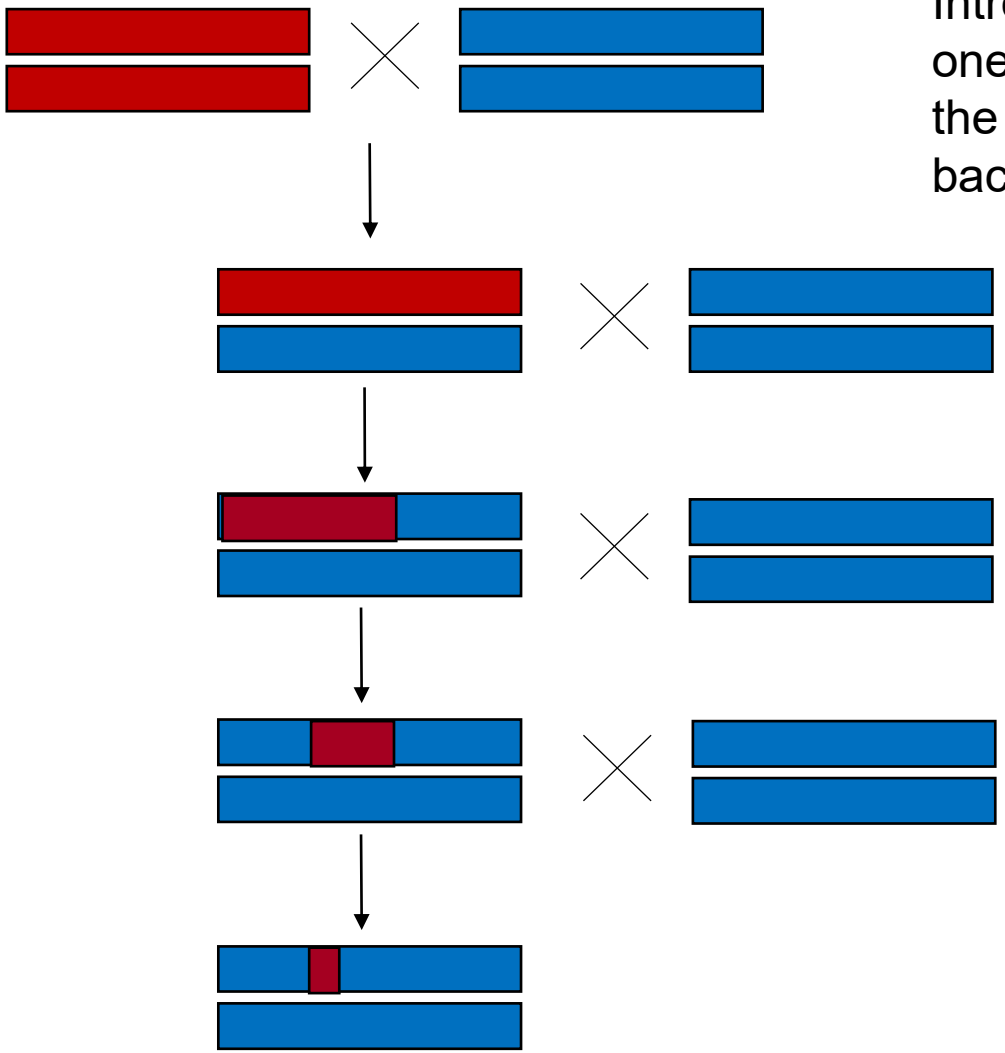


$$F_{st} = \frac{1}{1 + 4Nm}$$



- Nm > 1 prevents genetic differentiation caused by genetic drift

Genetic introgression (gene flow)



Introgression of genes from one population/species to the other. By repeating backcrossing.

Adaptive introgresion

House mouse (*Mus musculus*) acquired a gene for resistance against warfarin from *Mus spretus*.

Adaptive introgression: What have we inherited from Neandertals?

- Non-Africans 2-4% of their genome from Neanderthals.
- Neandertal variant of EPAS1 gene facilitate life in high altitudes in Tibet.
- Inuits have variations of WARS2 a TBX15 genes from Denisovans. Adaption to cold temperatures (body fat distribution, differentiation of adipocytes, generation of heat).
- Adaptive introgression of genes affecting quality and color of the skin and hair, immunity genes.
- Some genes from Neandertals are responsible for the diseases (e.g. obesity, depression).