Mutation

BIOL 434/509

Mutation

A mutation is a genetic change causing the genetic material of offspring to differ from the material it inherits from its parent.

Mutations occur because of errors in copying, followed by a failure to correct that error.

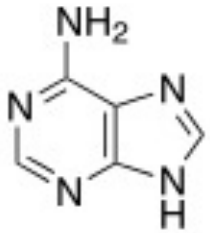
Point mutations

A point mutation or single nucleotide mutation is a mutation which causes one nucleotide to change to another.

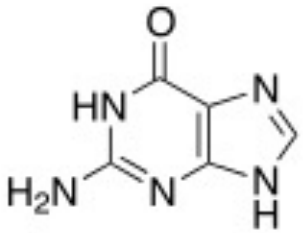
Transitions

Changes from:

one purine to another (A → G or G→A), or
one pyrimidine to another (C→T or T→C).

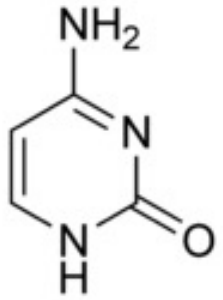


adenine

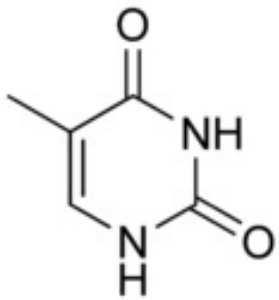


guanine

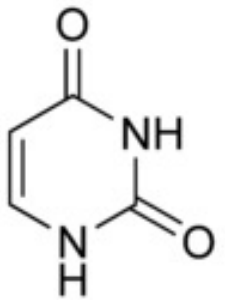
Purines



Cytosine (**C**)



Thymine (**T**)

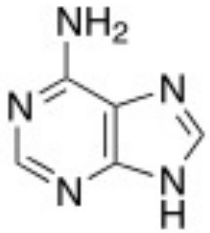


Uracil (**U**)

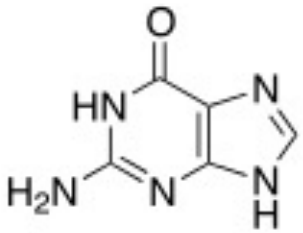
Pyrimidines

Transversions

Changes from:
a purine to a pyrimidine (A→C, A→T, G→C, G→T) or
from a pyrimidine to purine (C→A, C→G, T→A, T→G).

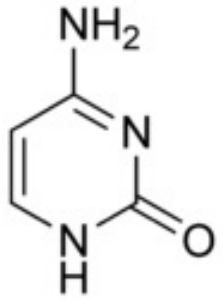


adenine

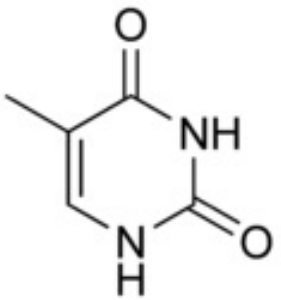


guanine

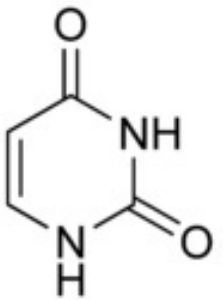
Purines



Cytosine (**C**)



Thymine (**T**)



Uracil (**U**)

Pyrimidines

Transitions are much more common than transversions

Estimated mutation rates in humans

(Nachman and Crowell 2003, Genetics):

Mutation type	Mutation rate
Transition at CpG	1.6×10^{-7}
Transversion at CpG	4.4×10^{-8}
Transition at non-CpG	1.2×10^{-8}
Transversion at non-CpG	5.5×10^{-9}
All nucleotide substitutions	2.3×10^{-8}
Length mutations	2.3×10^{-9}
All mutations	2.5×10^{-8}

As estimated by substitutions in pseudogenes.

Insertions and deletions

Insertions are stretches of DNA inserted into the parental sequence.

AGGTGAC → AGGCTCTTGAC

Deletions are stretches of DNA that are missing compared to the parental sequence.

AGGTGAC → AGGAC

Both insertions and deletions, if the occur inside a protein coding region, can cause frameshift mutations (if the length of the insertion or deletion is not a multiple of 3.) With a frameshift mutation, every codon downstream of the mutation is now disrupted.

Frameshift mutations can be caused by insertions or deletions

Ser|His|Phe|Arg|Cys|Glu|Leu|Leu|Arg|Arg|Gly|Val

TCA CAC TTC CGC TGT GAG TTG CTG AGG CGA GGT GTC

Deletion of a C

TCA CAC TTC GCT GTG AGT TGC TGA GGCGAGGTGTC…

Ser|His|Phe|Ala|Val|Ser|Cys|STOP|

Example is from a patient with Tay-Sachs;

α-Subunit of β-Hexosaminidas

Lau and Neufeld, 1989, J. Biol Chem.

How are mutation rates measured?

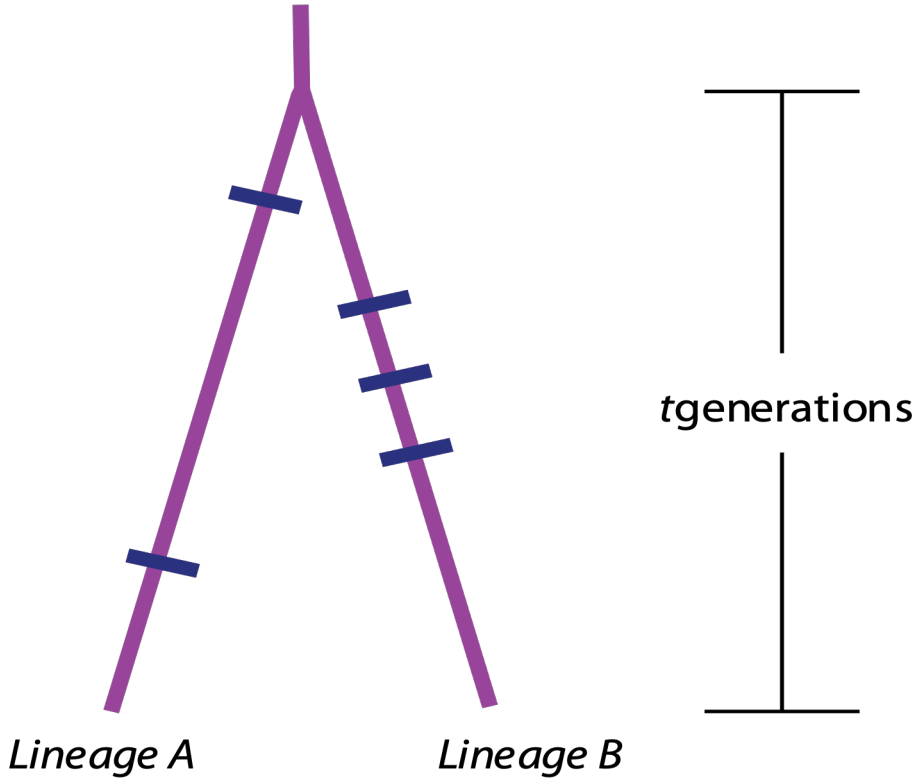
Direct estimates

e.g., comparing sequence of parents and offspring

Estimates per base pair per generation:

Humans:	~1×10^{-8}
Flycatchers:	5×10^{-9}
Mice:	5×10^{-9}
Drosophila:	2×10^{-8}
HIV:	4×10^{-3}
E. coli:	2.2×10^{-10}

Mutation accumulation



Mutation accumulation requires minimizing selection

Otherwise new mutations will be lost due to selection

- Pseudogenes *presumably* experience little selection
- Extremely small population size allows drift to overwhelm effects of moderate or weak selection

Mutation accumulation:

Mukai's experiment with *Drosophila*

Used balancer chromosomes to isolate single chromosomes.

Balancer chromosomes:

- Contain inversions, which suppress recombination
- Have dominant alleles for some easily seen phenotype
- Have recessive lethal allele (to kill homozygotes of the balancer)

Mutation accumulation:

Mukai's experiment with *Drosophila*

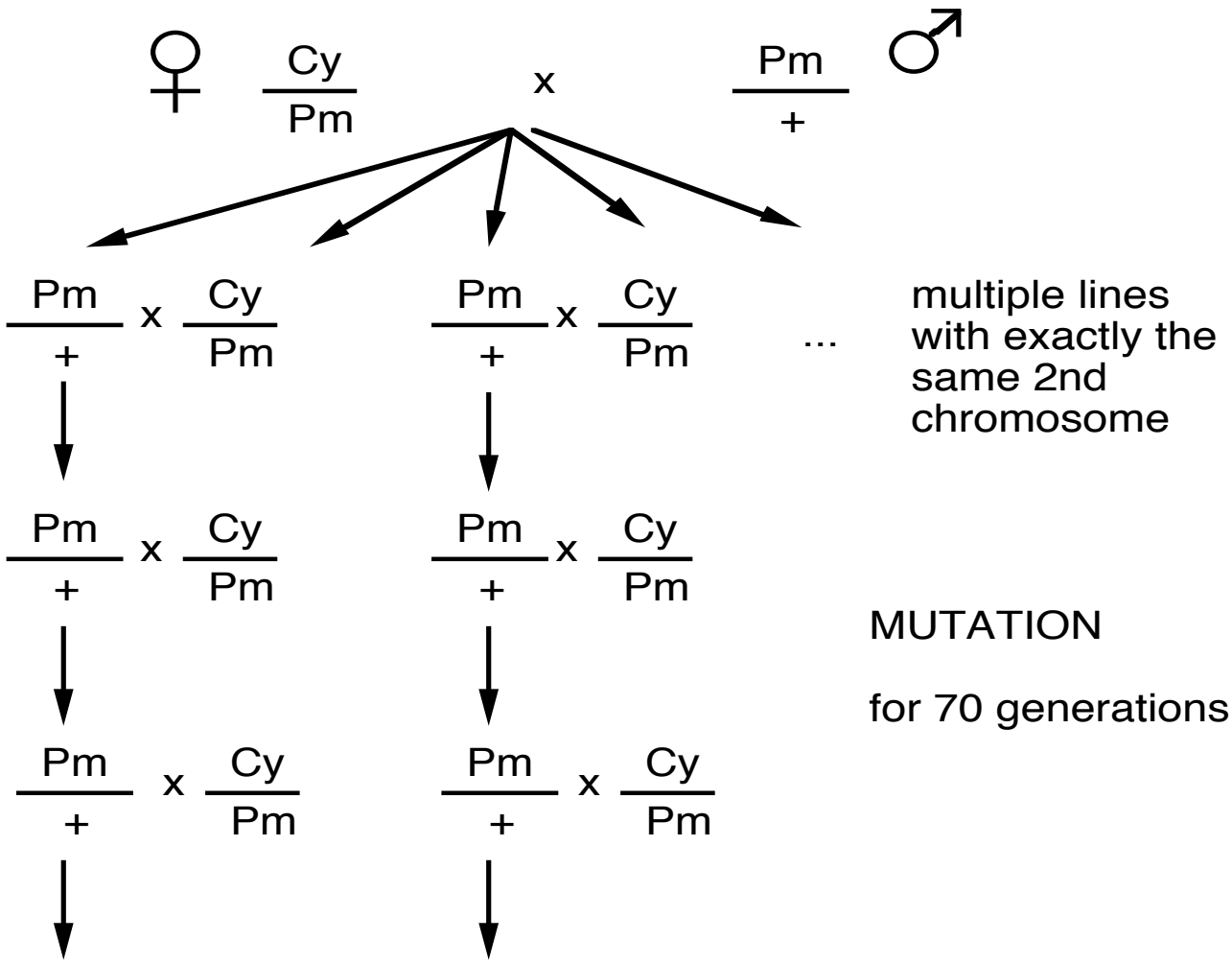
Cy (*Curly*) is a balancer for the 2nd chromosome in *Drosophila melanogaster*.

http://flybase.org/reports/FBal0002196.html

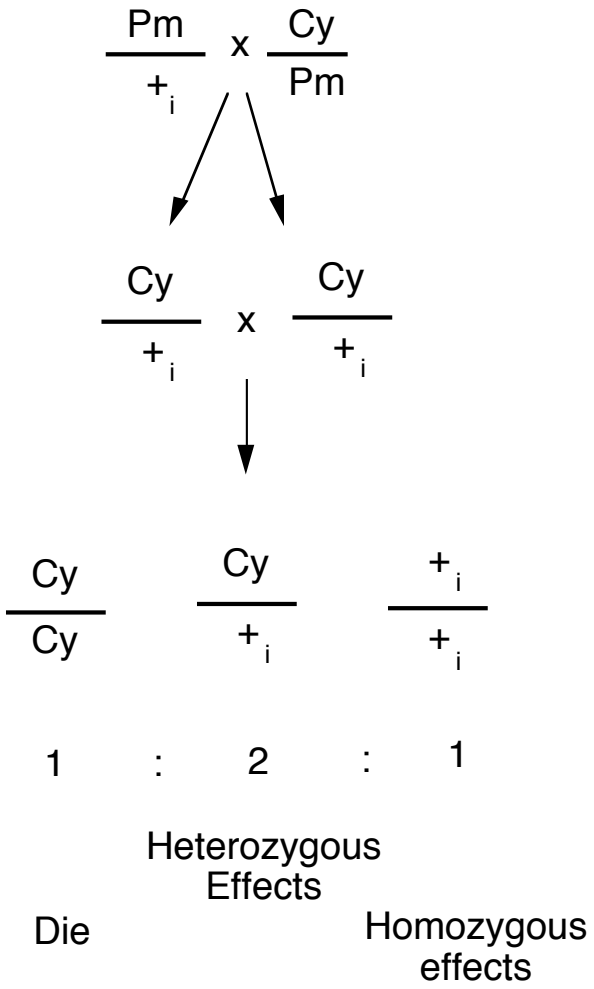
Pm (Plum) is a dominant mutation in *D. melanogaster*, also on the 2nd chromosome.

(Also important: Male *Drosophila* don't have recombination.)

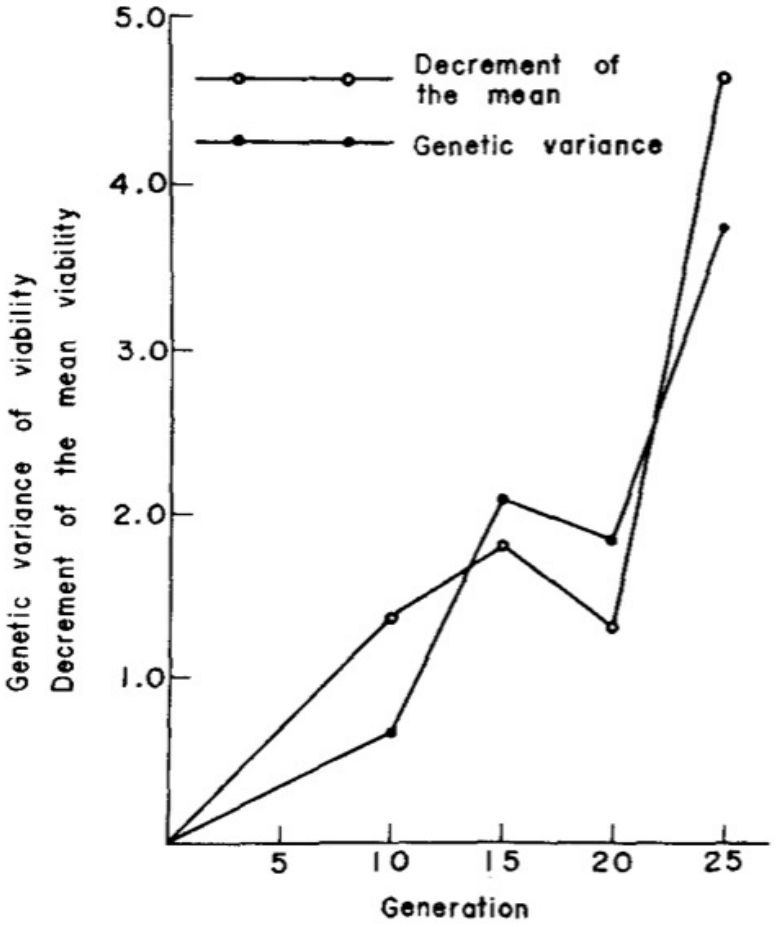
Mukai's experimental design



Measuring relative fitness



Mukai's results



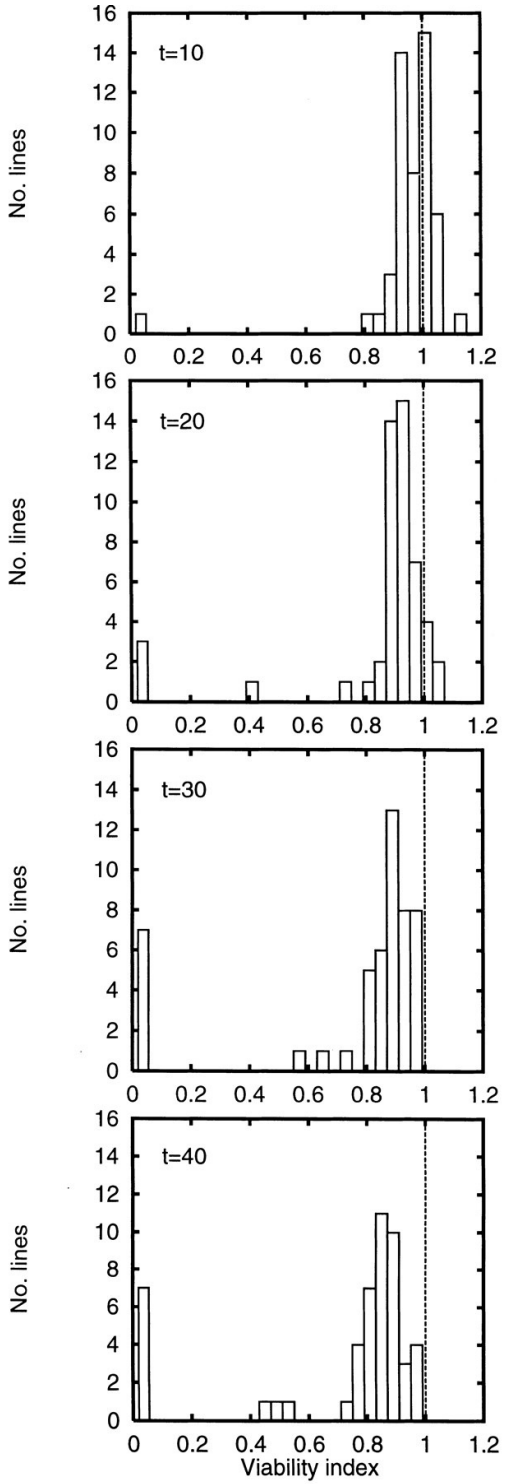


Figure from Keightley and Eyre-Walker 1999. GENETICS 153:515-523
Mukai 1964 Genetics **50:** 1-19

Drift-mutation balance

μ : the mutation rate per gene per individual per generation.

Infinite alleles model

Assumes that each new mutation is unique and not present in the population already.

(This is not so far off if we consider alleles to be defined by a stretch of DNA or if the product $N\mu \ll 1$.)

Infinite alleles model

Modelling identity in state (of two randomly chosen alleles from the population) for neutral mutations:

$$F' = \left[\frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right) F\right] (1 - \mu)^2$$

Infinite alleles model

At equilibrium between mutation and drift:

$$\hat{F} = \left[\frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right)\hat{F}\right](1-\mu)^2$$

When the mutation rate is small, this can be approximated by:

$$\hat{F} = \frac{1}{4N_e\mu + 1}$$

Heterozygosity at mutation-drift balance

The heterozygosity at the equilibrium between drift and mutation is

$$\hat{H} = 1 - \hat{F} = \frac{4N_e\mu}{4N_e\mu + 1} \approx 4N_e\mu$$

The approximation assumes that $4N_e\mu$ is small.

The **scaled mutation rate** is sometimes called $\theta = 4N_e\mu$

Watterson's θ

$$\hat{\theta}_W = \frac{S}{L \sum_{i=1}^{n-1} \frac{1}{i}}$$

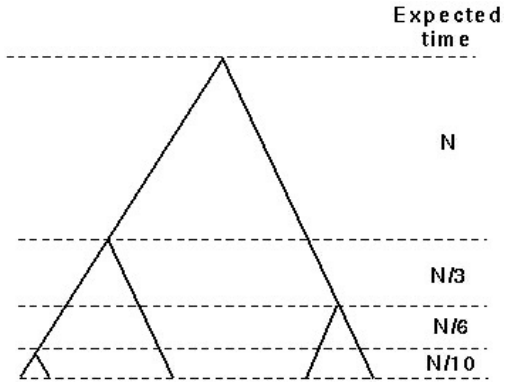
S is the number of segregating sites.

L is the number of sites examined.

Watterson's θ

$\hat{\theta}_W$ should be equal to $4N_e\mu$ if the mutations are all neutral and if the population has a constant size.

$E[S] = \mu E[T_{tot}]$,
where T_{tot} is the sum of all branch lengths
(Assume Neutrality and constant population size)



Substitution rates

A substitution is the fixation in one species of an allele not present in the ancestor.

A substitution is a *population-level process* (while mutation is an *individual level process*).

Neutral substitutions should occur in proportion to the rate of new neutral mutations appearing in the population times the probability that any one of them fixes.

Substitution rates

New mutations appear at rate 2μ per locus per diploid individual per generation.

If there are *N* individuals in the population, this means new mutations appear at rate: $2N\mu$ per population per generation.

Each of these mutations starts at frequency 1/(2*N*). For neutral alleles, the probability of fixation is equal to the starting allele frequency. Therefore

probability of fixation of new allele: 1/(2*N*)

Therefore the rate of substitution is (substitutions per population per generation):

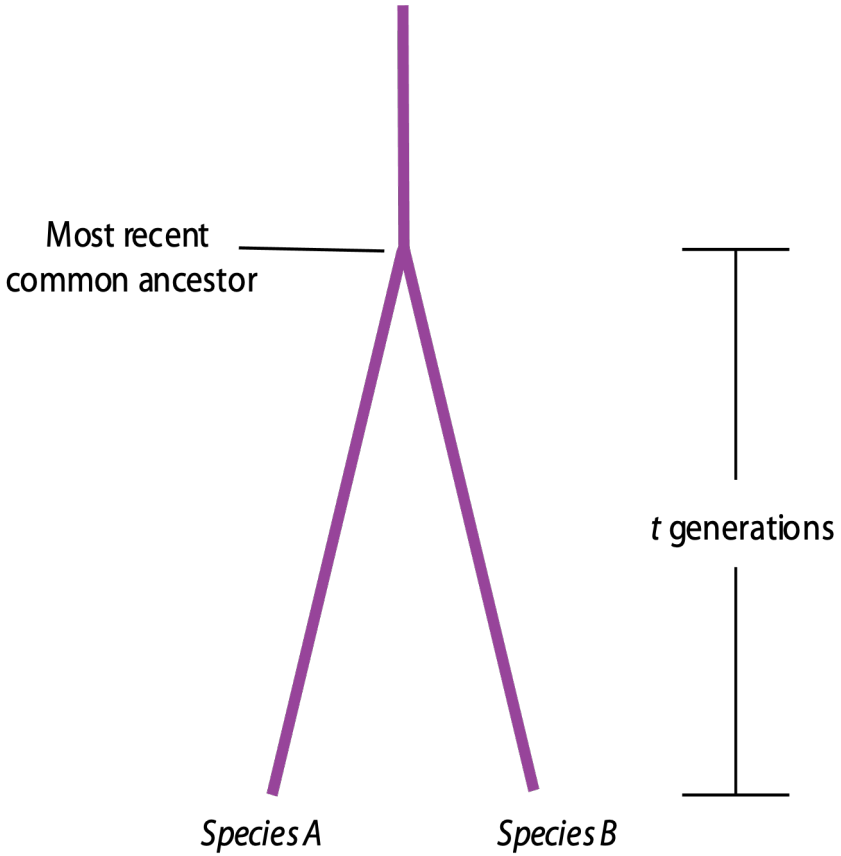
(rate of new mutations) x (probability of fixation) = $2N\mu\left(\frac{1}{2N}\right) = \mu$

Neutral divergence between species

Two lineages

Each has substitutions occurring at rate μ

Total neutral divergence: $2t\mu$



dN/dS

dN is the fraction of possible nonsynonymous substitutions which differ between the species.

dS is the fraction of possible synonymous substitutions which differ between the species.

dN/dS

If there is no selection on that protein, we would expect *dN/dS* to be 1.

If selection is typically purifying *dN/dS* should be <<1.

If selection is strongly diversifying between the species *dN/dS* should be > 1.

Neutral theory

The neutral theory was proposed by Motoo Kimura.

To explain the large amount of genetic variation observed in the genome, Kimura proposed that the vast majority of mutations (including in protein regions) are either deleterious or selectively neutral.

McDonald-Kreitman test

The McDonald-Kreitman test uses protein coding sequence variation within and between species to test for selective divergence between the species.

Compares patterns for synonymous and non-synonymous sites.

Probability of differences per site:

	Within species (Polymorphism or heterozygosity)	Between species (Divergence)
Synonymous	$4N_e \mu_S$	$2t \mu_S$
Non-synonymous	$4N_e \mu_N (1 - C)$	$2t \mu_N (1 - C)$

where

N_e is the effective population size
μ_S is the mutation rate to synonymous codons
μ_N is the mutation rate to non-synonymous codons
C is the **constraint**: the fraction of mutations which are deleterious
t is the number of generations since the common ancestor of the species

	Within species (Polymorphism or heterozygosity)	Between species (Divergence)
Synonymous	$4N_e \mu_S$	$2t \mu_S$
Non-synonymous	$4N_e \mu_N (1 - C)$	$2t \mu_N (1 - C)$

Under this null model, both within- and between-species ratios of synonymous over non-synonymous differences should be $\mu_S / \mu_N (1–C)$).

Analyze numbers of difference with χ^2 test

Example: *Adh* in *Drosophila melanogaster* and *D. simulans*

	Within species (Polymorphism)	Among species (Divergence)
Synonymous	42	17
Non-synonymous	2	7

Significant excess of non-synonymous changes between species ($P = 0.006$) and so evidence of adaptive evolution at this locus.

Proportion of divergence due to selection

Using the logic of the MK test, it is possible to estimate the fraction of non-synonymous substitutions that are caused by positive selection (α).

Estimates range from 10- 50%, depending on the species pair.