Week 5

tree tests, clustering, PCA

Molecular Evolution

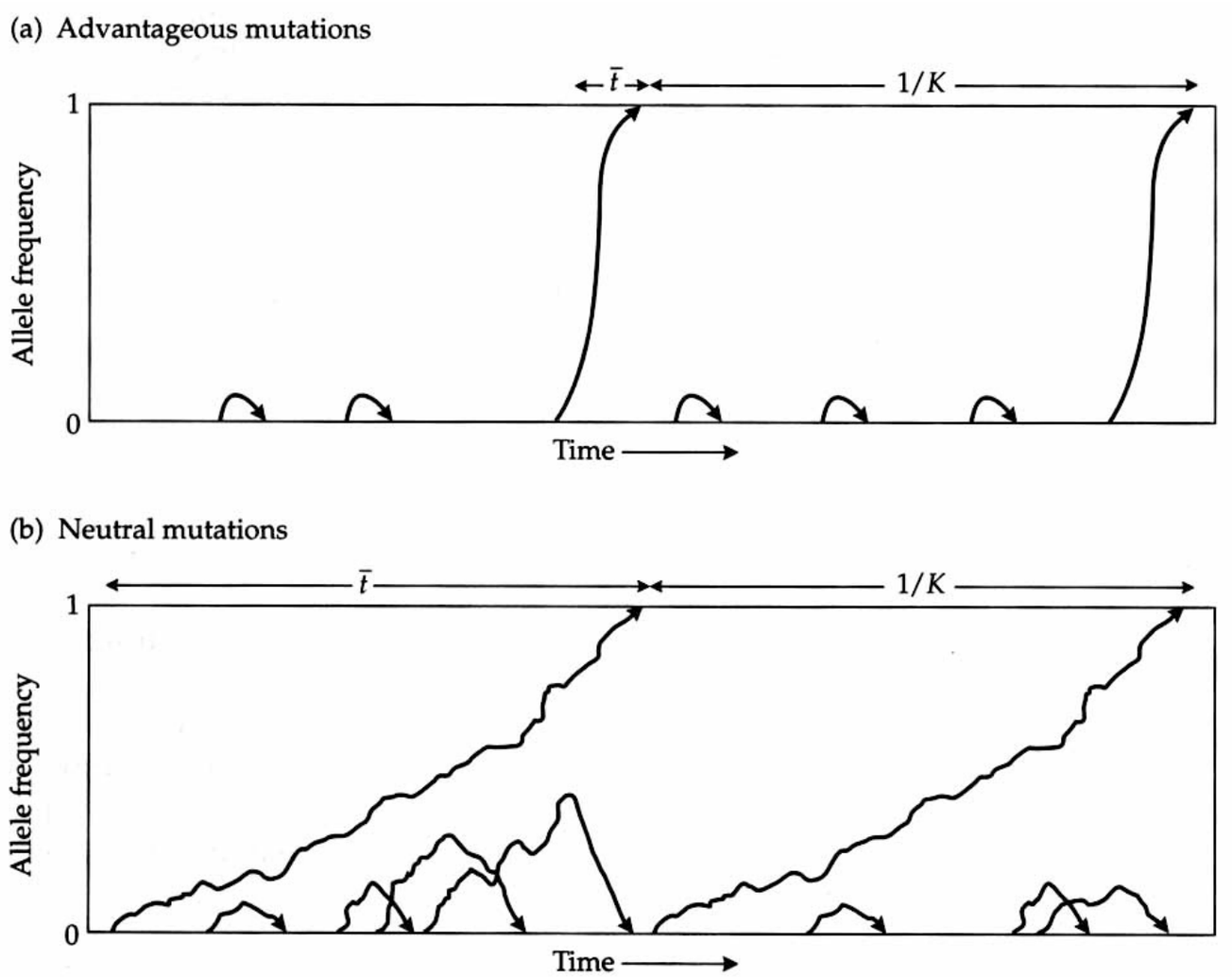
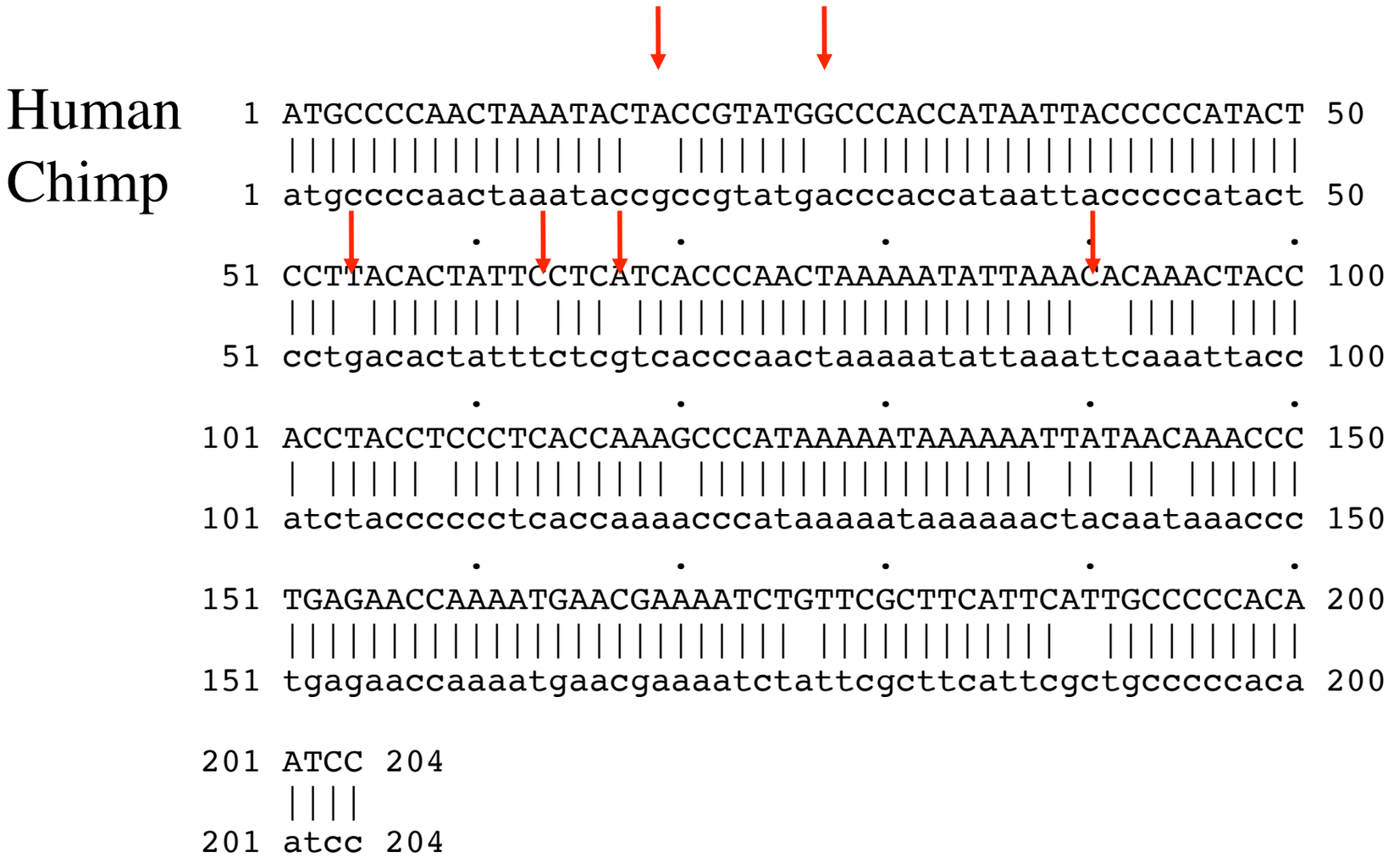


FIGURE 2.7 Schematic representation of the dynamics of gene substitution for (a) advantageous and (b) neutral mutations. Advantageous mutations are either quickly lost from the population or quickly fixed, so that their contribution to genetic polymorphism is small. The frequency of neutral alleles, on the other hand, changes very slowly by comparison, so that a large amount of transient polymorphism is generated. The conditional fixation time is $\bar{t}$, and $1/K$ is the mean time between two consecutive fixation events. Modified from Nei (1987).

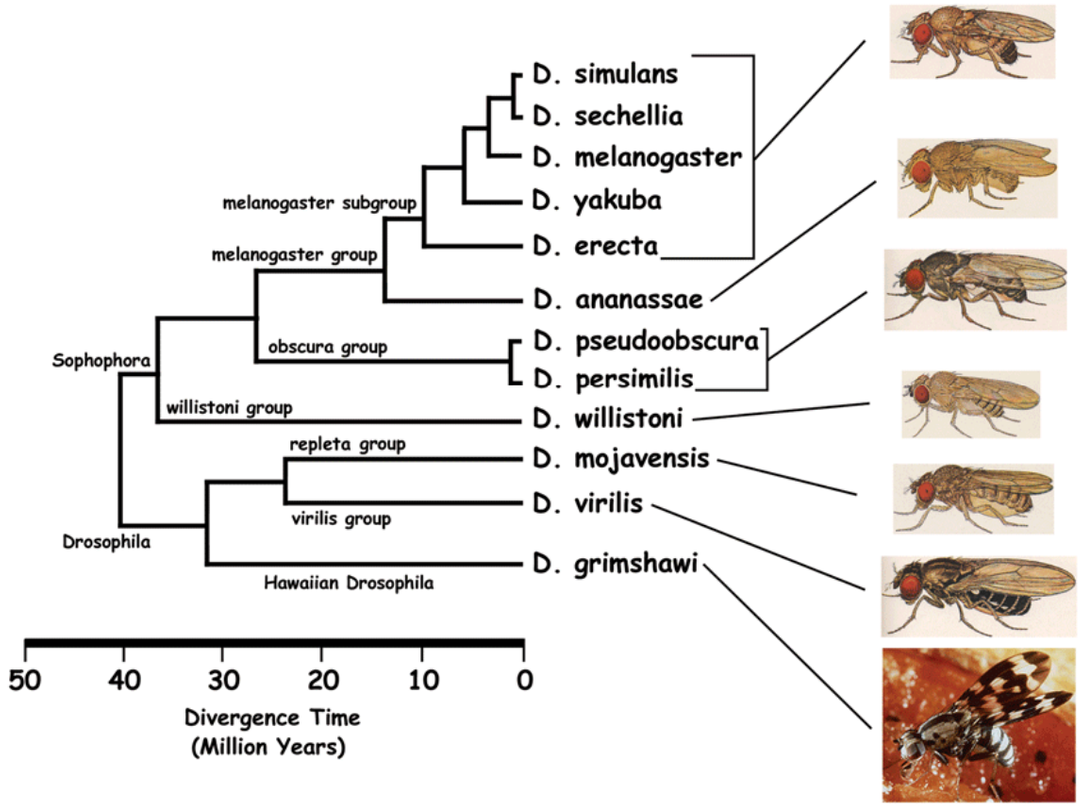
Graur and Li (2000) Fundamentals of Molecular Evolution

How does DNA evolve?



Measuring DNA Evolution

1. Align sequences
2. Determine length of sequences
3. Count number of differences
4. Divergence = (number of differences) / (length of sequence)
5. Rate = (sequence divergence) / 2 x (age of common ancestor)



Complex IV, COI

	402									471
D. mel	TLNNKWLKSH	FIIMFIGVNL	TFFPQHFLGL	AGMPRRYSDY	PDAYTTWNIV	STIGSTISLL	GILFFFFIIW
D. sim	.....Q	.....	.....	.....	.....V.	.....	.....Y...
D. sec	.....Q	..T.....	.....	.....	.....	.....	.....
D. mau	.....Q	.....	.....	.....	.....V.	.....	.....Y...
D. yak	.....Q	.....	.....	.....	.....V.	.....	.....Y...
D. ere	.....Q	.L.....	.....	.....	.....V.	.....	.....Y...
D. ana	...V.....Q	.....	.....	.....	.....V.	.....	.....Y.V.
D. per	.M...L...Q	.V.....	.....	.....	.....V.	.....S....	.....Y...
D. wil	...A.....Q	.....	.....	.....	.....	.....S....	.....
D. moj	...S.....Q	.....	.....	.....	VI	.....S....	.....Y...
D. vir	.M.....Q	.....	.....	.....	VI	.....S....	.....Y...
D. gri	.....Q	.....	.....	.....	VI	.....	.....Y...

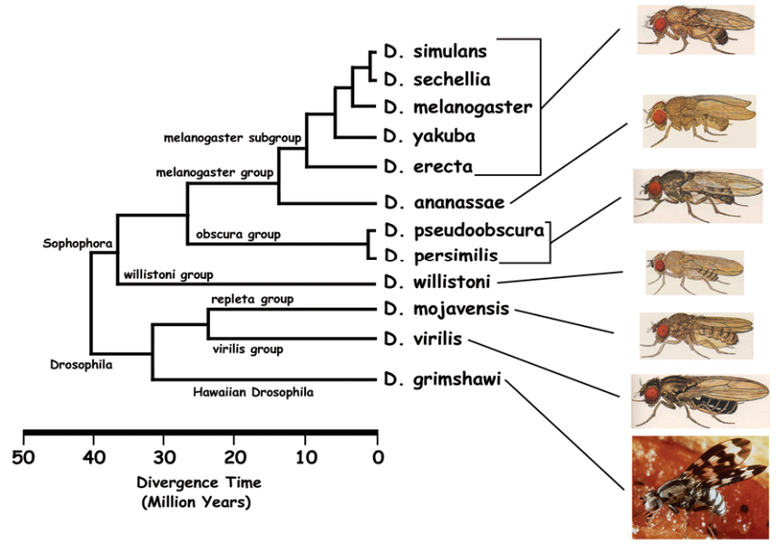
Compare *D. melanogaster to D. simulans*

3 Amino Acid differences at COI

Measuring DNA Evolution

Complex IV, COI

	402								471
D. mel	TLNNKWLKSH	FIIMFIGVNL	TFFPQHFLGL	AGMPRRYSDY	PDAYTTWNIV	STIGSTISLL	GILFFFFIIW
D. sim	.........Q	....	....	....	.......V.	....	......Y...
D. sec	.........Q	..T.......	....	....	....	....	....
D. mau	.........Q	....	....	....	.......V.	....	......Y...
D. yak	.........Q	....	....	....	.......V.	....	......Y...
D. ere	.........Q	.L........	....	....	.......V.	....	......Y...
D. ana	...V.....Q	....	....	....	.......V.	....	......Y.V.
D. per	.M...L...Q	.V........	....	....	.......V.	.....S....	......Y...
D. wil	...A.....Q	....	....	....	....	.....S....	....
D. moj	...S.....Q	....	....	....	VI	.....S....	......Y...
D. vir	.M.......Q	....	....	....	VI	.....S....	......Y...
D. gri	.........Q	....	....	....	VI	....	......Y...



$$D = \frac{3 \text{ diffs}}{70 \text{ AAs}} = 0.043 \text{ diffs / site}$$

$$\rho = \frac{3 \text{ diffs}}{2 \times 5 \text{ million years}} = 3 \times 10^{-7} \text{diffs / year}$$

$$\rho = \frac{3 \times 10^{-7} \text{diffs / year}}{70 \text{ AAs}} = 4.29 \times 10^{-9} \text{ diffs / year / site}$$

Rate of Substitution: Neutrality

Recall the probability of fixation of new mutation:

$$Prob\{\text{fix}\} = \frac{1}{2N}$$

Define u as Prob(mutation/gamete/generation):

$$\text{Average \# mutants / generation} = 2Nu$$

Rate of Substitution:

$$\begin{aligned} \rho &= \frac{1}{2N} \times 2Nu \\ &= u \end{aligned}$$

Rate of Substitution: Neutrality

$$\rho = u$$

Under neutrality, no dependence on population size!

Among species might expect to see generation time effect...

This gives rise to idea of “molecular clock”

First Molecular Clocks

Protein Polymorphism as a Phase of Molecular Evolution

MOTOO KIMURA & TOMOKO OHTA

National Institute of Genetics, Mishima, Shizuoka-ken

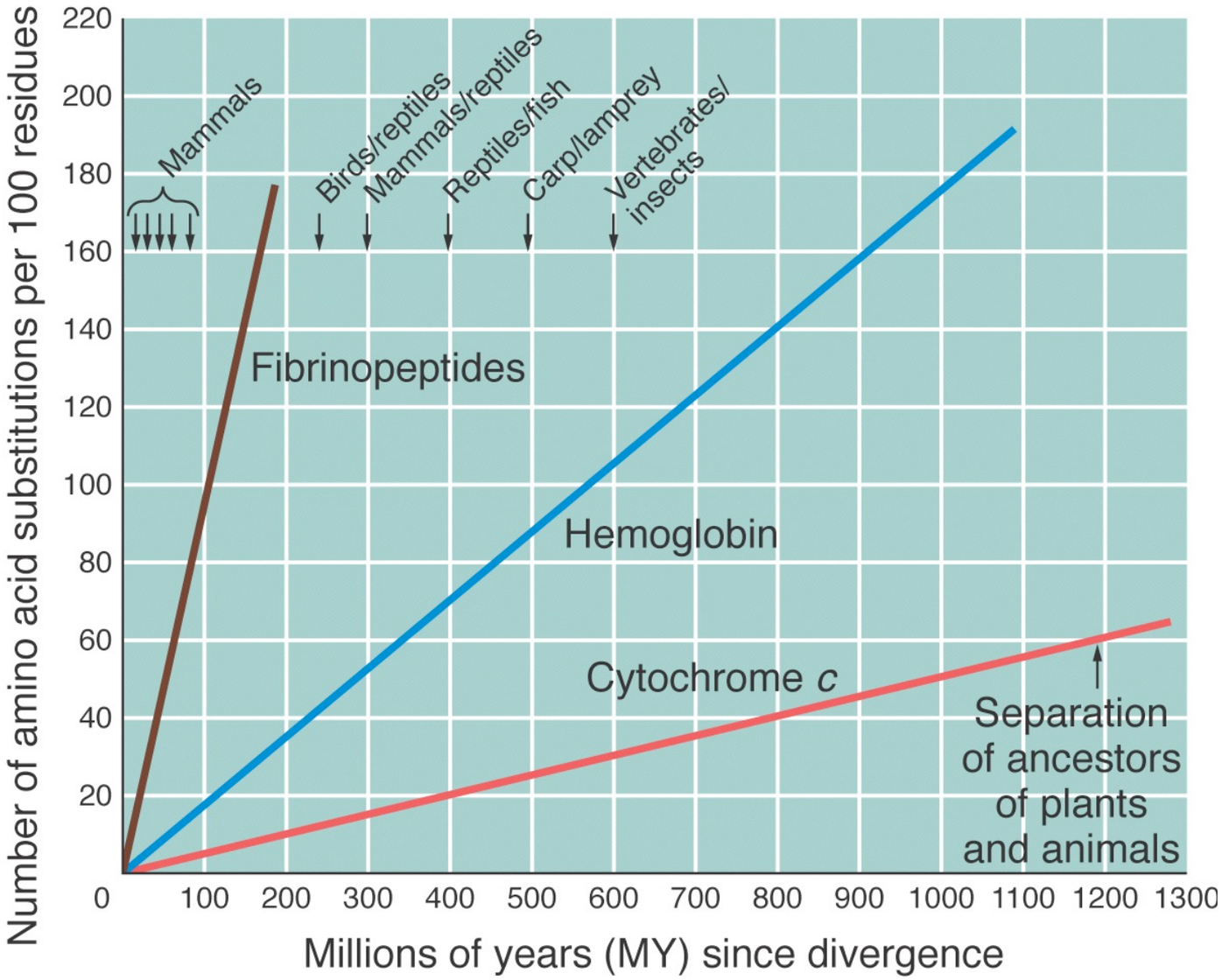
http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/profiles/ohta.html

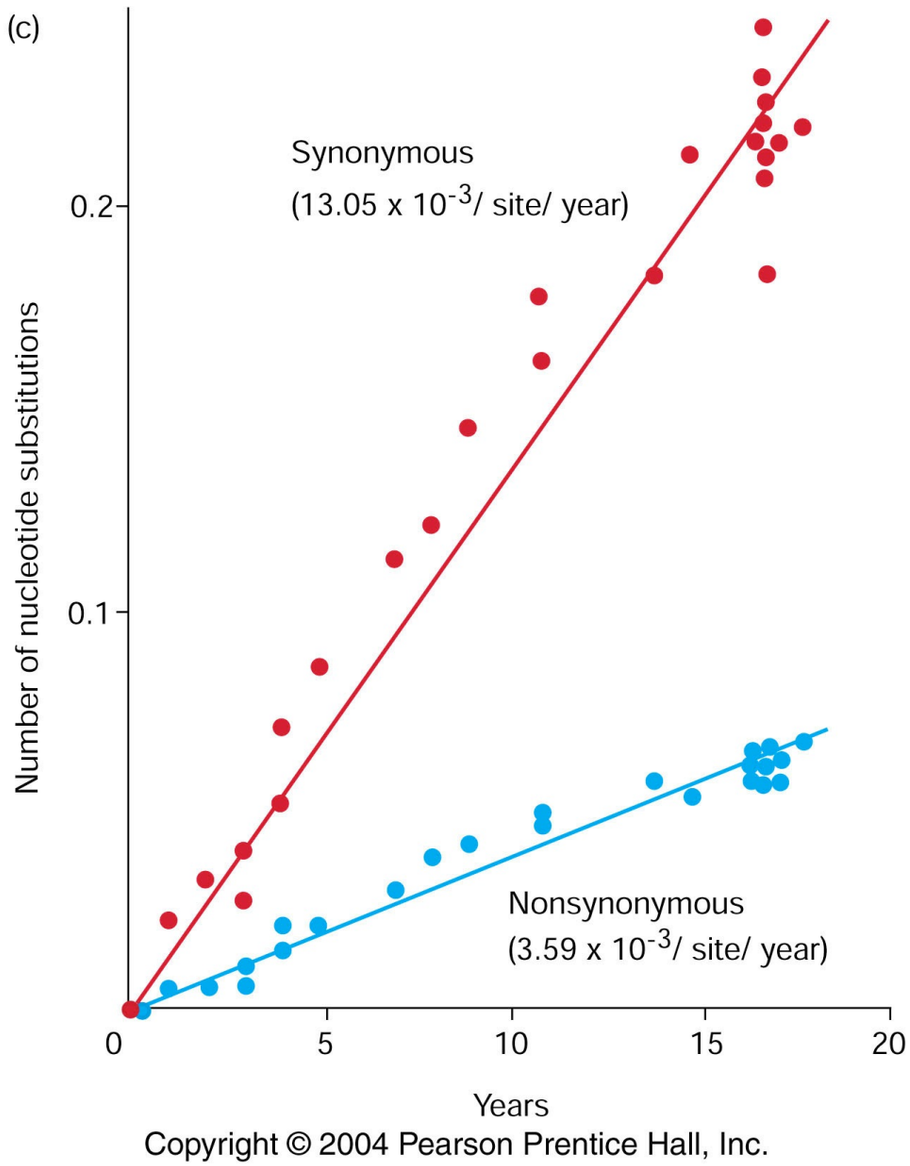
In 1971 noticed that protein evolved at a clock-like pace per year (~10^{-9}/aa site/year)

Argued that this was to due neutrality...

but....

- different genes have different **functional constraints**
- this causes different rates of evolution
- different nucleotide positions evolve at different rates





http://www.bio.miami.edu/dana/250/betaglobinclock.jpg

“Classical” School of Variation

- Low levels of genetic variation in natural populations
- “Wild type” is favored, mutants are eliminated
- “Purifying” selection
- H. J. Muller one proponent

www.orcbs.msu.edu/radiation/radhistory/hermannmuller.html

Kimura's quick fix

Neutrality plus purifying selection due
to functional constraint

$$\text{Average \# mutants / gen} = 2Nuf_0$$

$$Prob\{fix\} = \frac{1}{2N}$$

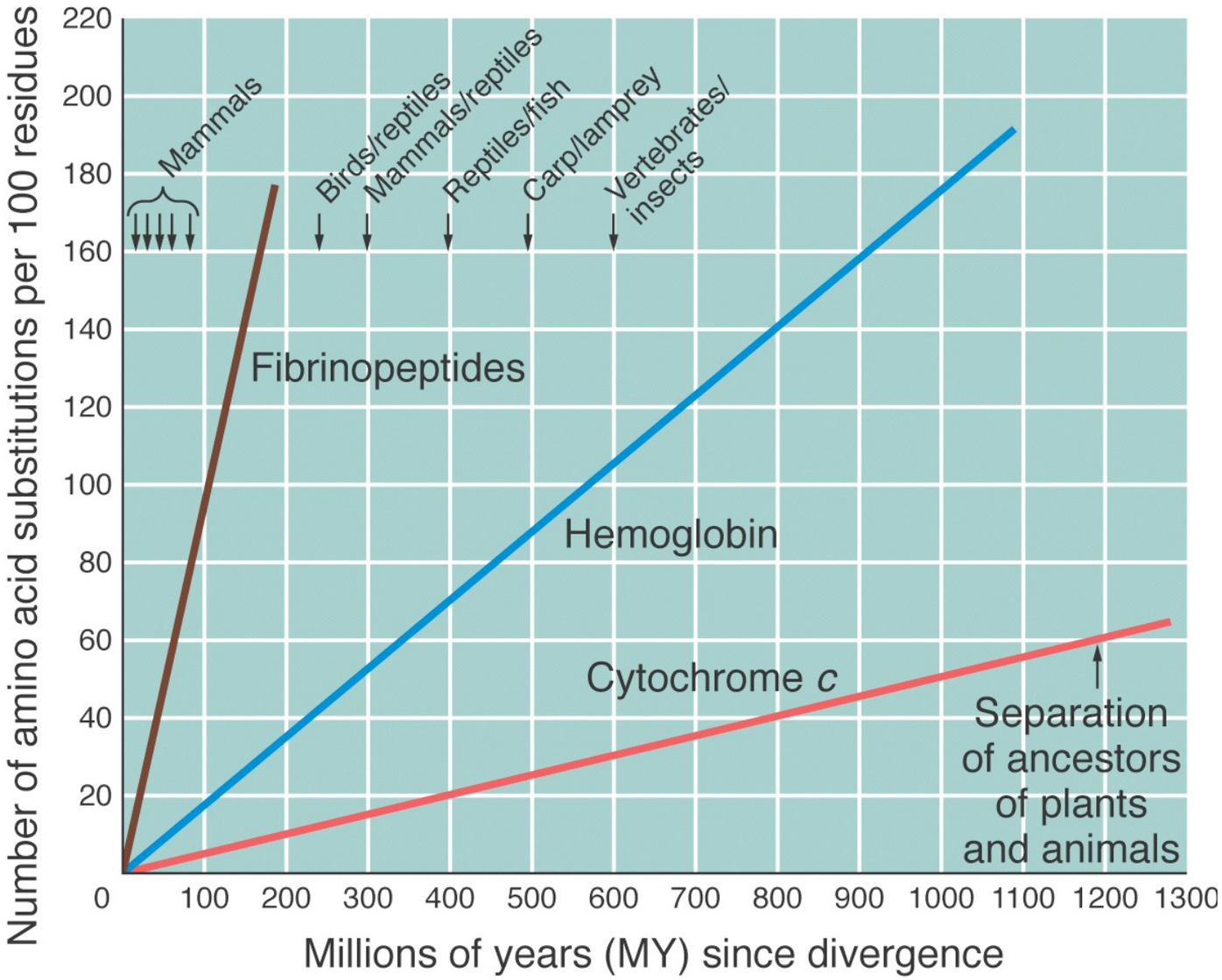
$$f_o = Prob\{\text{neutral}\}$$

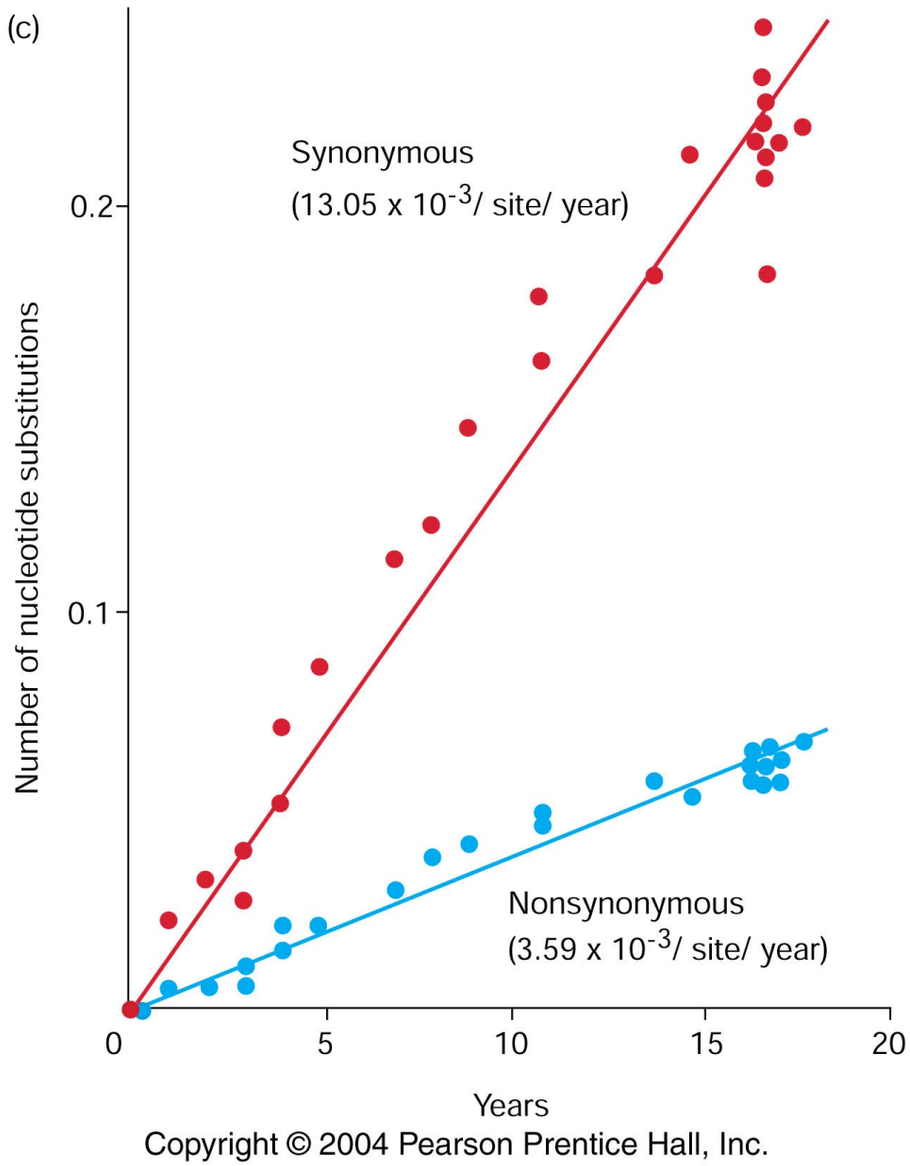
$$\rho = uf_0$$

Now all we need is gene specific f_0

but....

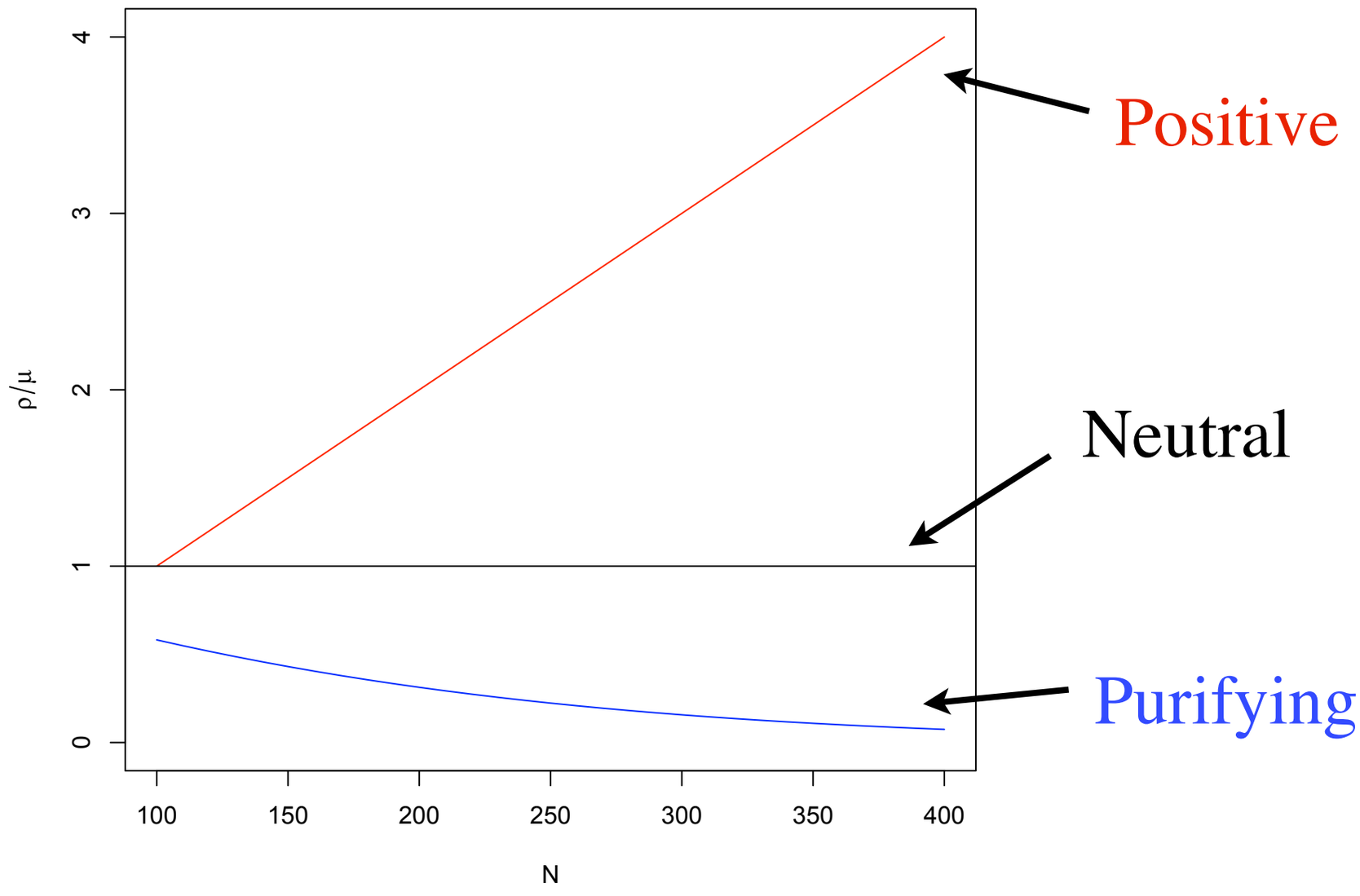
- different genes have different **functional constraints**
- this causes different rates of evolution
- different nucleotide positions evolve at different rates





http://www.bio.miami.edu/dana/250/betaglobinclock.jpg

Rate of Substitution

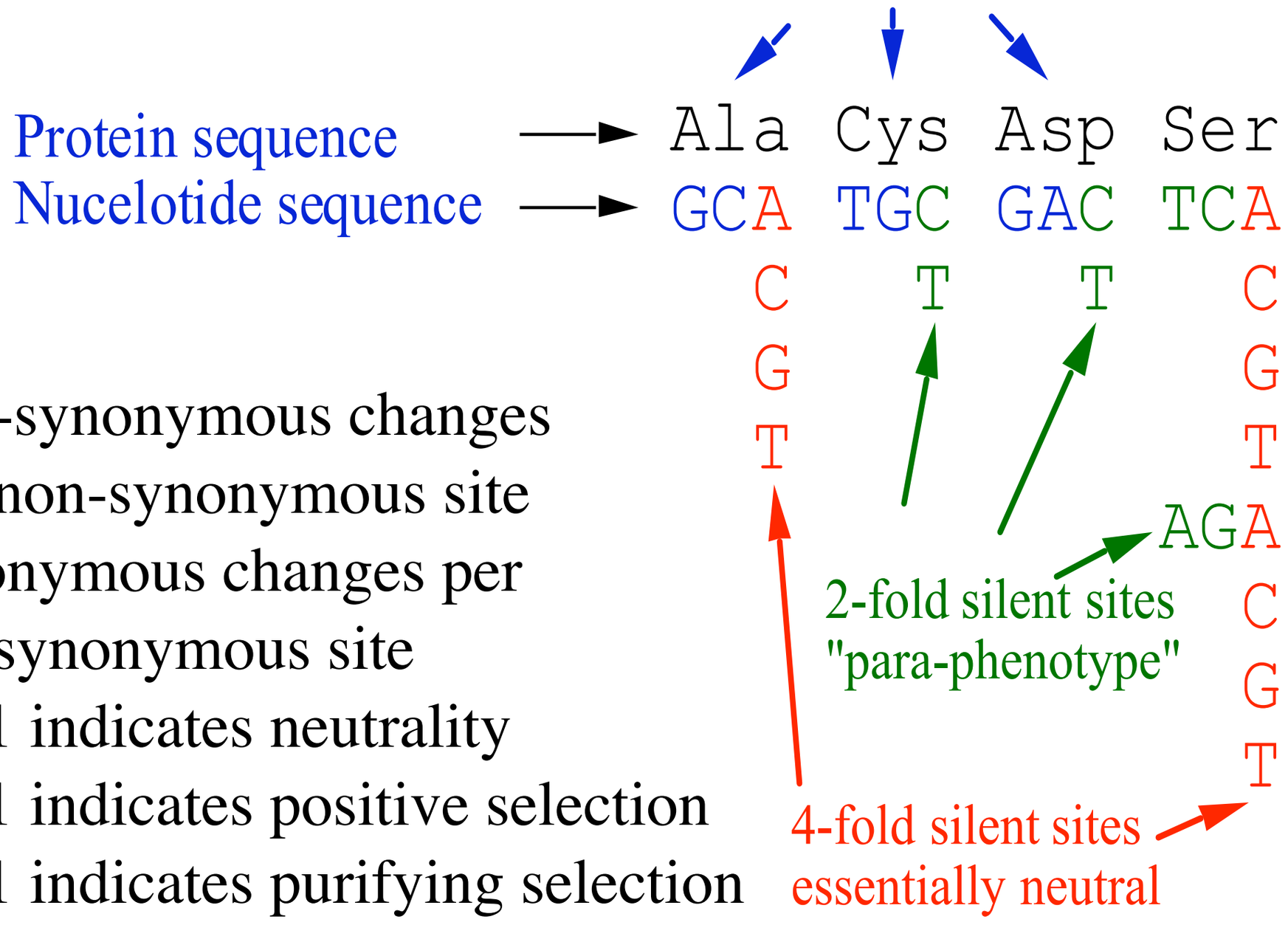


This intuition leads to possible tests of neutrality...

The genetic code and DNA "phenotypes"

Synonymous sites = nucleotide differences between alternative codons

Nonsynonymous = amino acid replacement sites = "Phenotype"



dN = non-synonymous changes per non-synonymous site

dS = synonymous changes per per synonymous site

dN/dS = 1 indicates neutrality

dN/dS > 1 indicates positive selection

dN/dS < 1 indicates purifying selection

Functional constraint and the genetic code

Table 4. Relative frequencies of different types of mutational substitutions in a random protein-coding sequence.

Substitution	Number	Percent
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4
Total in first position	183	100
Synonymous	8	4
Nonsynonymous	175	96
Missense	166	91
Nonsense	9	5
Total in second position	183	100
Synonymous	0	0
Nonsynonymous	183	100
Missense	176	96
Nonsense	7	4
Total in third position	183	100
Synonymous	126	69
Nonsynonymous	57	31
Missense	50	27
Nonsense	7	4

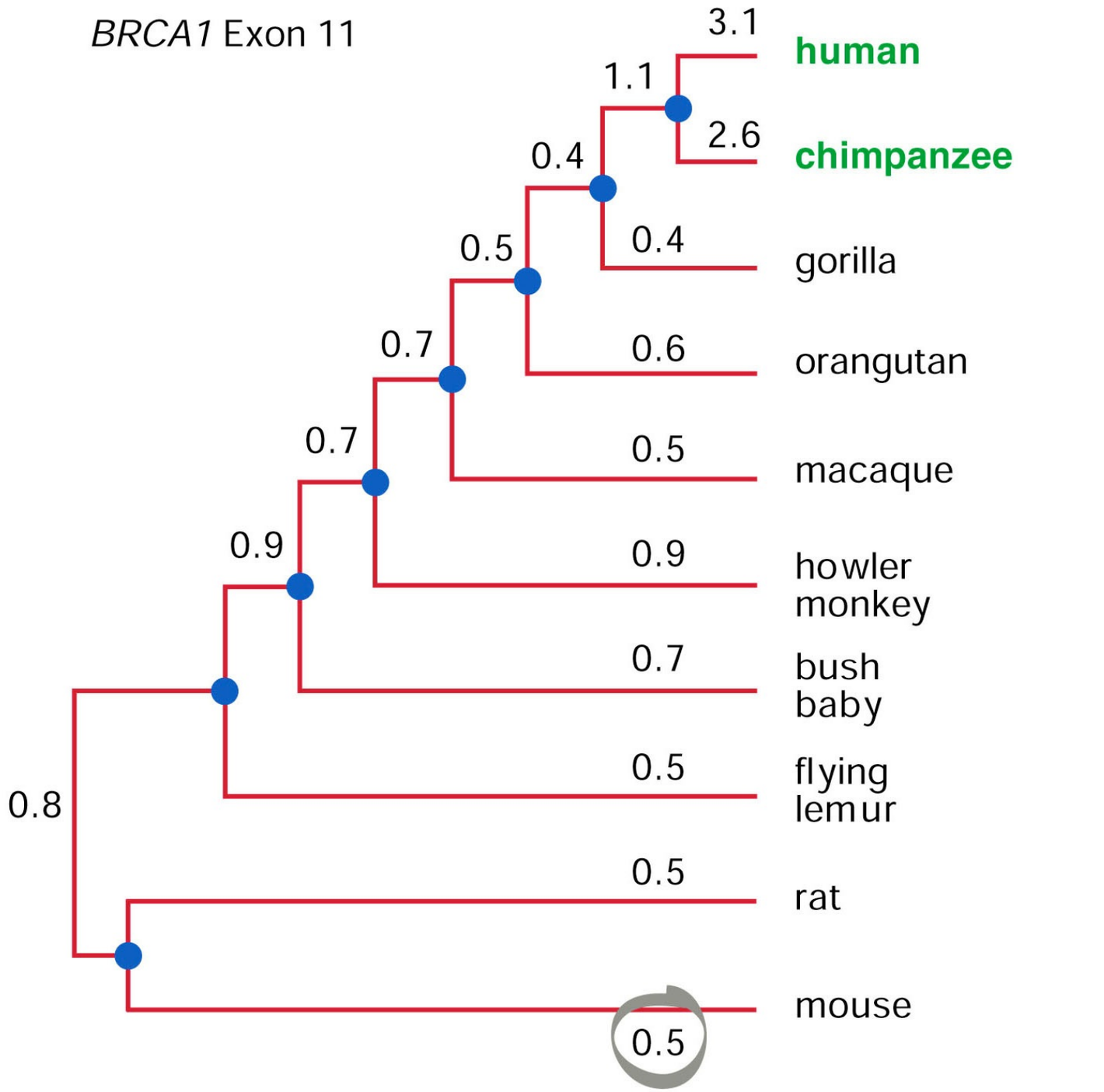
1st position: a few Synonymous

2nd position: ALL Nonsynonymous

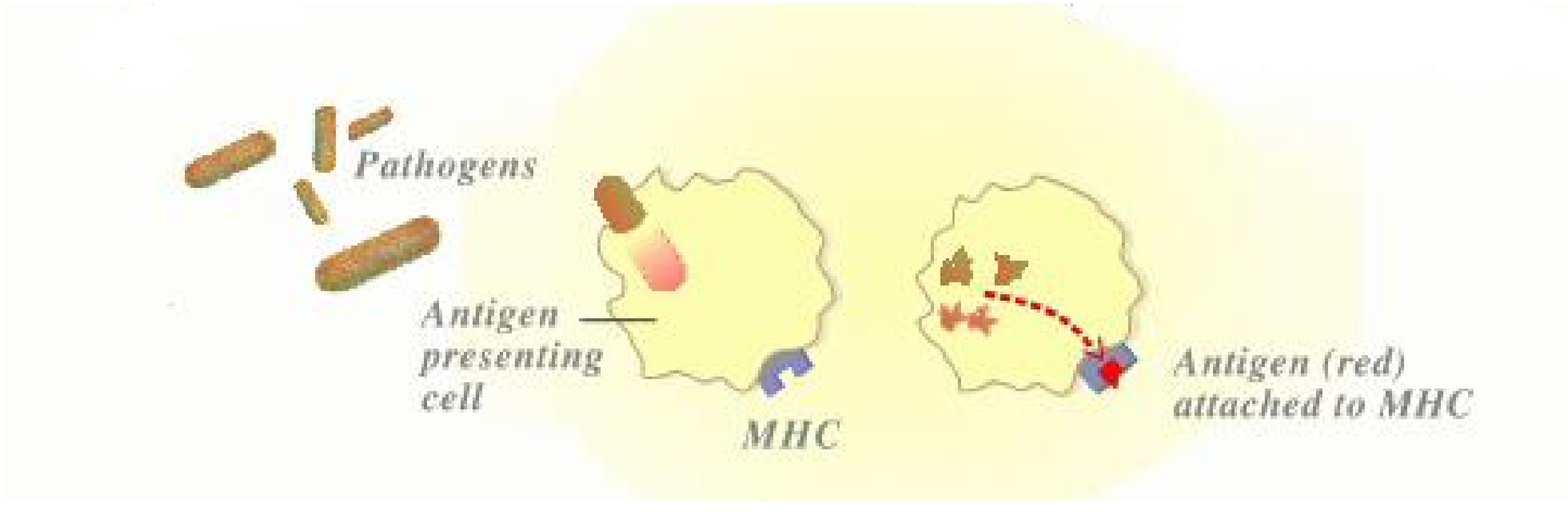
3rd position: mostly Synonymous (not all)

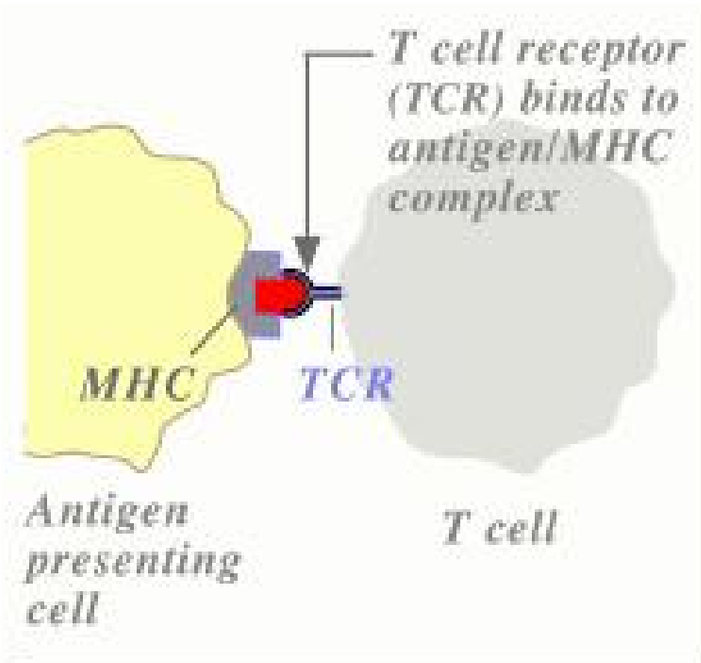
Variation in
dN/dS ratios
Across a phylogeny

Not consistent with
Neutral expectations



Foreign particles (antigens) are presented for elimination by MHC





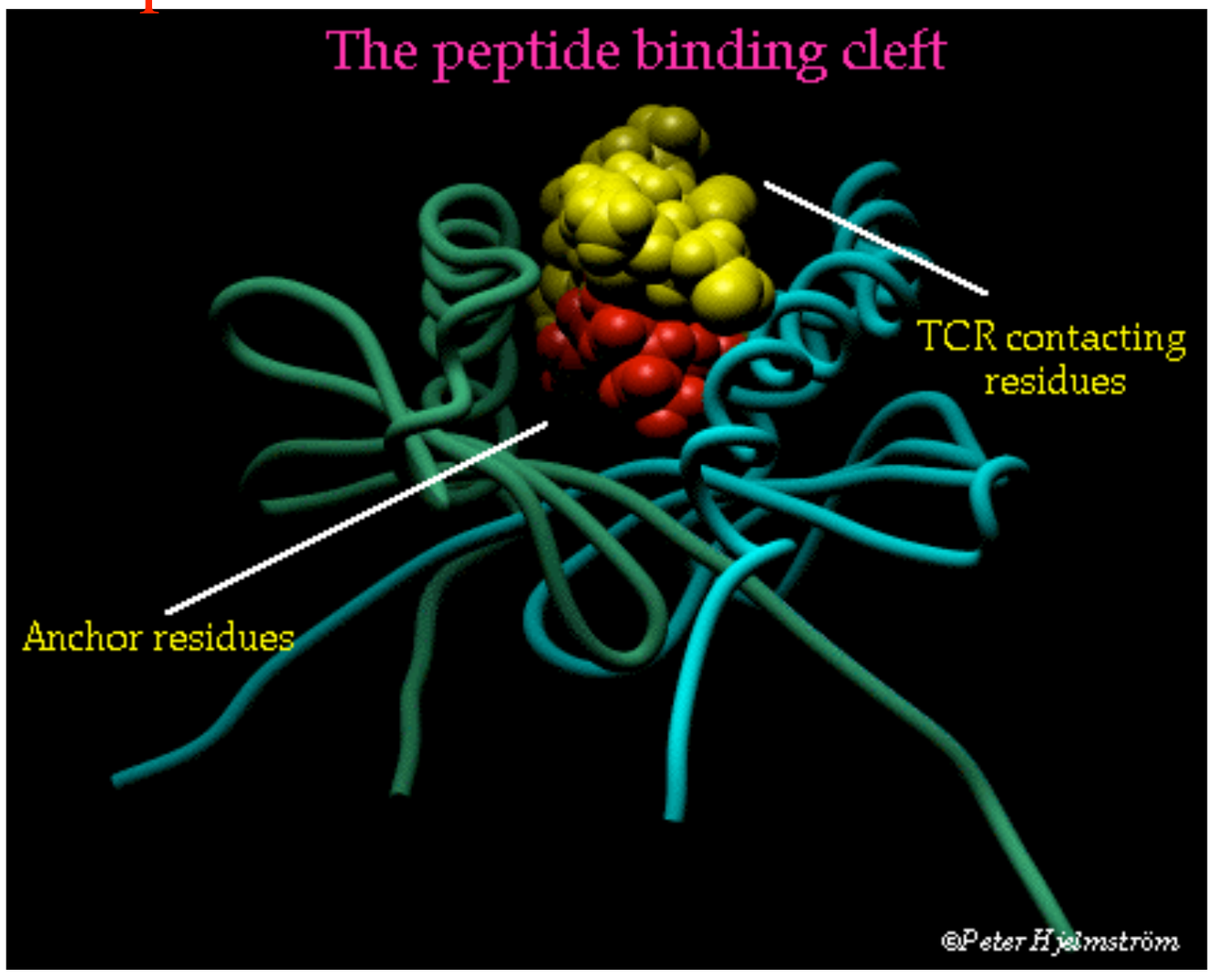
http://www.mifoundation.org/icons/Imm1_tcell_mhc.png

DNA test of neutrality

- Neutral prediction:
- amino acid (nonsynonymous) substitution rate (dN) should be lower than silent (synonymous) substitution rate (dS)
- True for most genes
  - –Follows from functional constraint argument
- Different for Major Histocompatibility Complex (MHC) loci
  - –Antigen recognition sequence shows dN > dS
  - –Rest of molecule shows dN < dS, as expected
- Amino acid mutations are favored in antigen recognition region
- Promotes diversity, better recognition of foreign peptides
- Amino acid mutations are eliminated from rest of molecule

Antigen binding sites: dN/dS > 1

"positive" selection

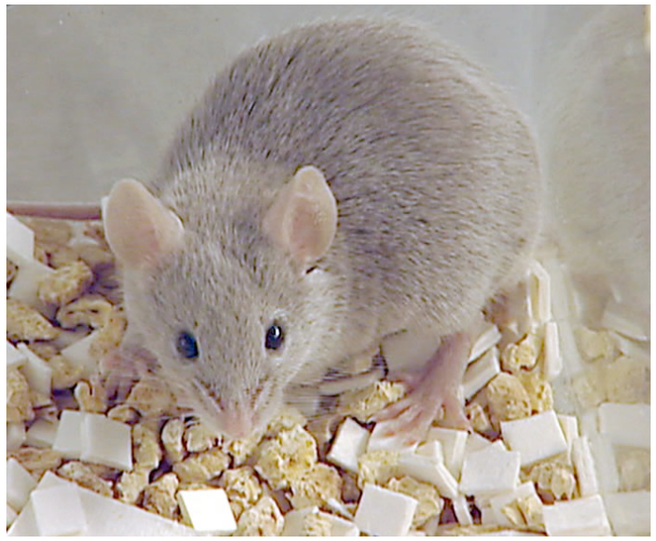


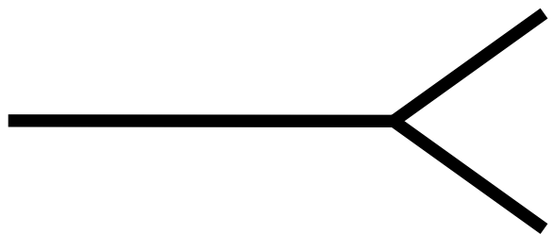
http://depts.washington.edu/rhwlab/dq/3structure.html

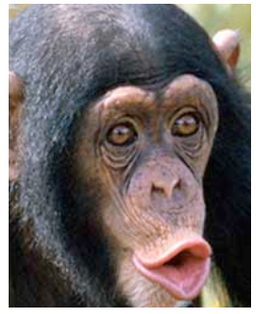
Rest of molecule: dN/dS < 1

Negative (purifying) selection

Can we find human-specific adaptive evolution?







Clark, A.G. et al. (2003) Science 302: 1960-1963

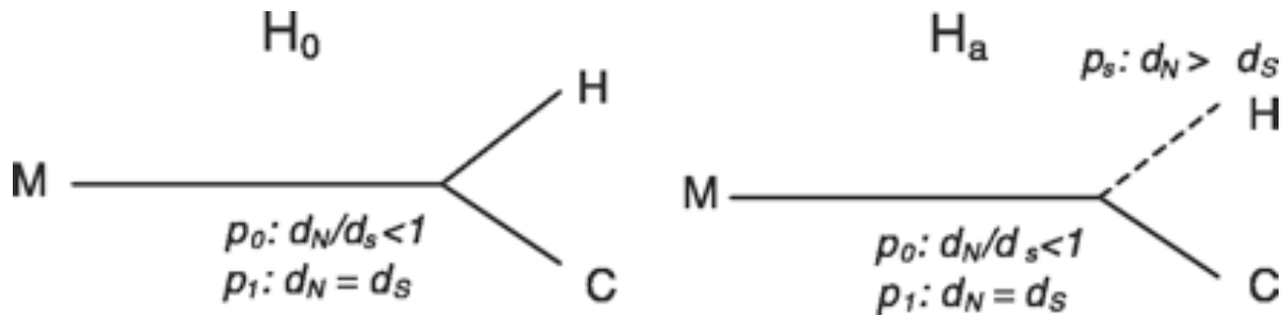
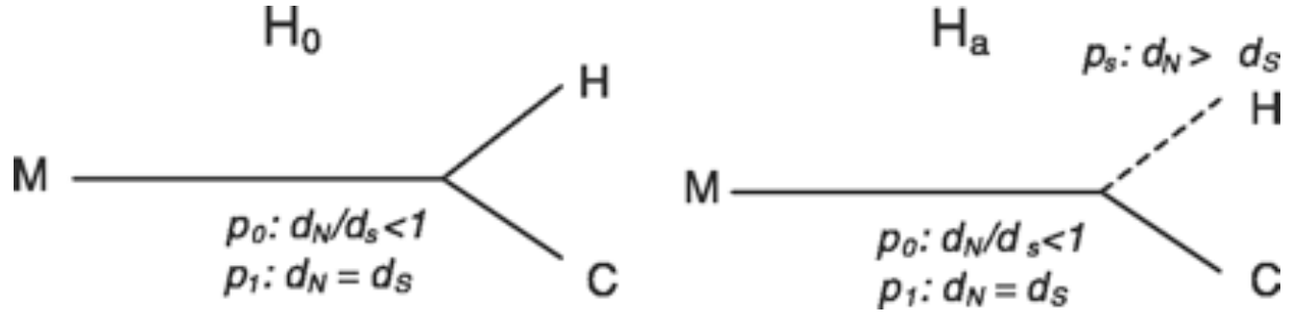


Fig. 1. Graphical representation of the test of positive selection (Model 2). The null hypothesis (H0) assumes all three branches have two classes of amino acid residues: those that are neutrally evolving ($p1$: dN = dS) and those that are under constraint ($p0$: dN/dS<1). The alternative hypothesis (Ha) allows the human lineage to have a subset of sites (pS) with accelerated amino acid substitution (dN > dS).

http://2.bp.blogspot.com/_Gf65NLuwWsU/SZ2r2UQF3AI/AAAAAAAAA74/UDeQ9Rj69m4/s400/9george-bush.jpg

http://www.ornl.gov/info/ornlreview/v37_3_04/images/a08_mouse_full.jpg

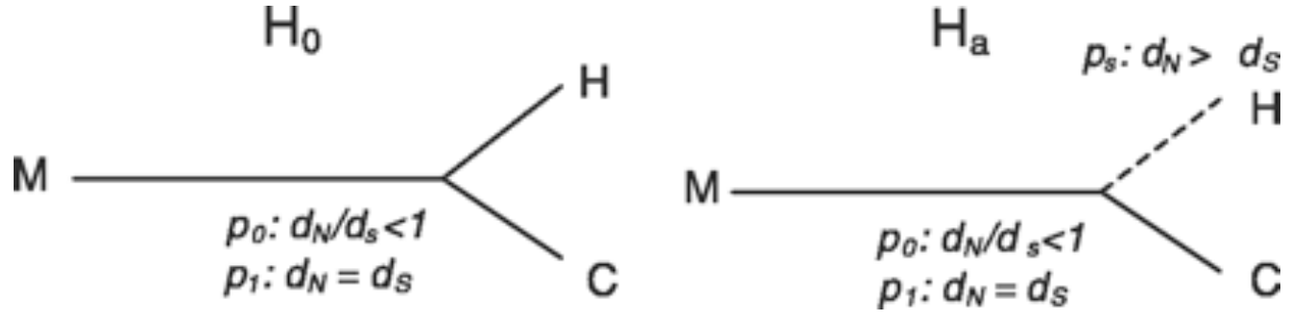
Likelihood Ratio Test



PAML model of rates

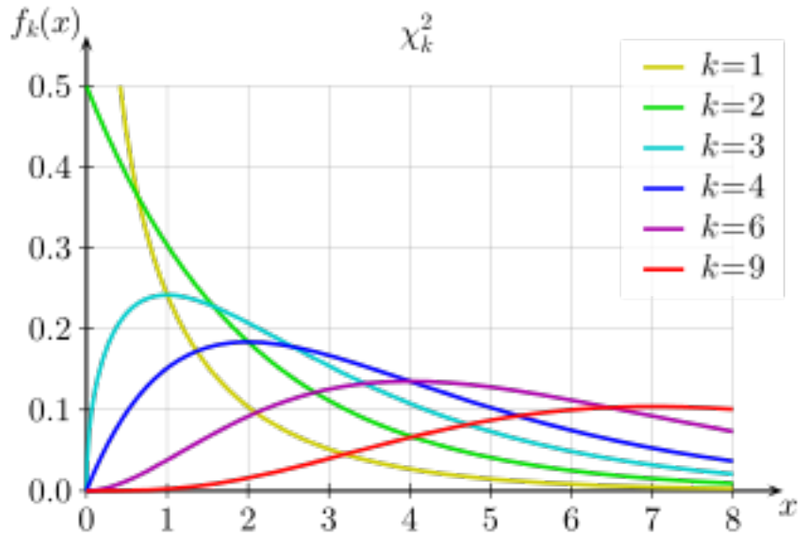
$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

Likelihood Ratio Test



Compute likelihood of H_0 vs H_a

$$\lambda = -2 \times \{\ln(H_a) - \ln(H_0)\}$$



$$\lambda \sim \chi^2_{df}$$

Olfaction and amino acid catabolism genes evolve rapidly on the human lineage

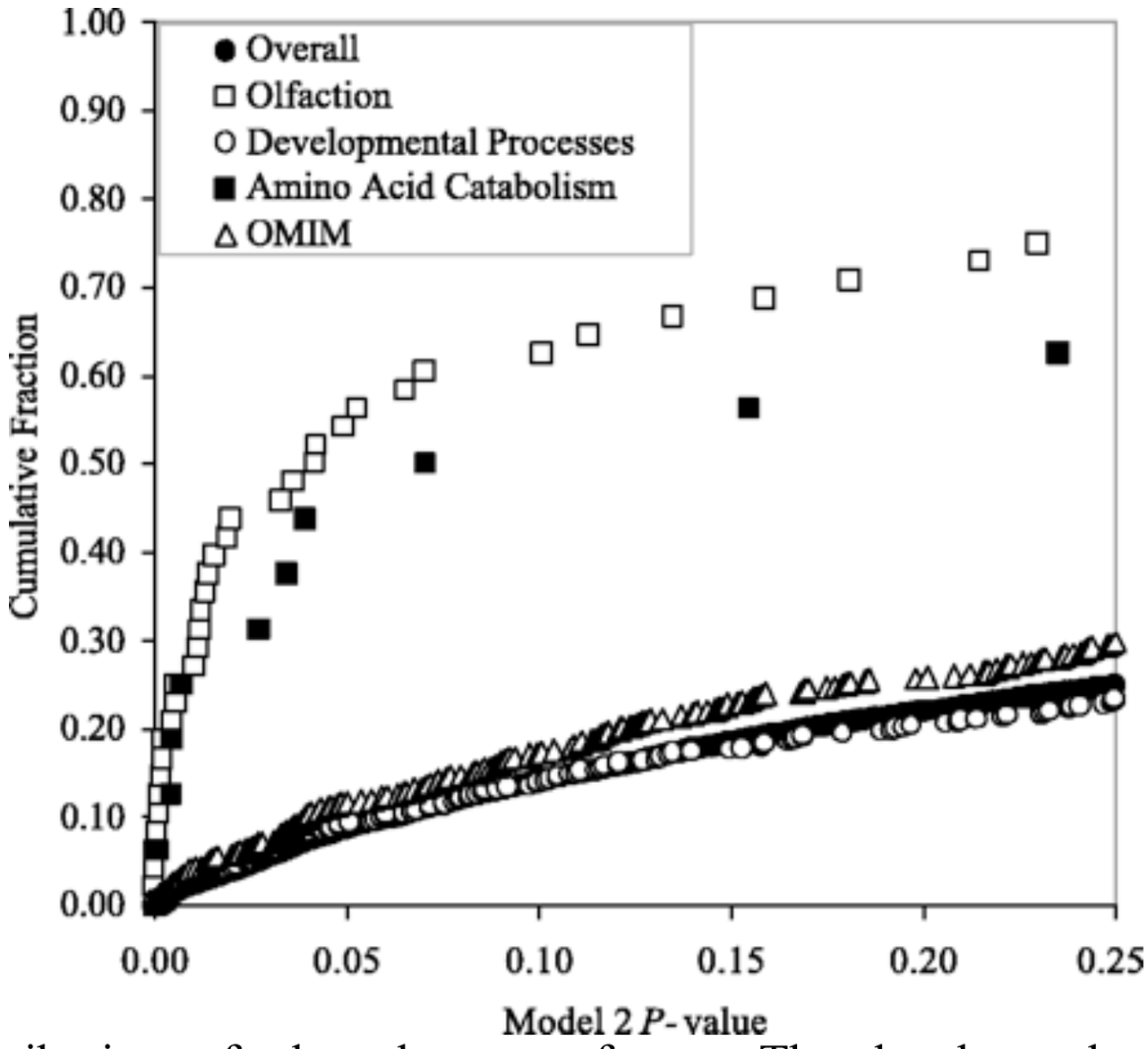
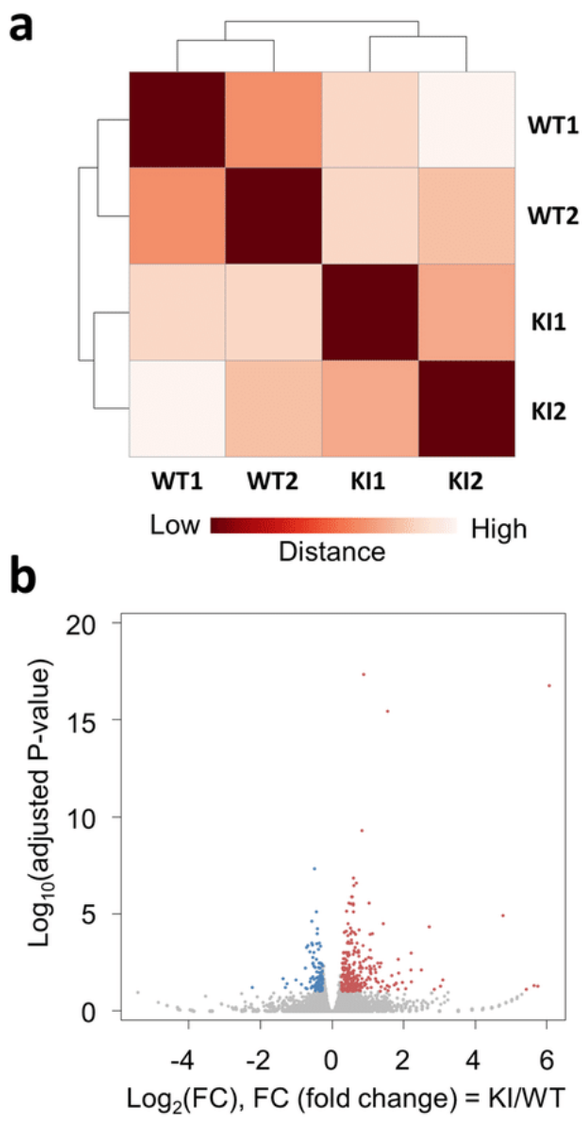
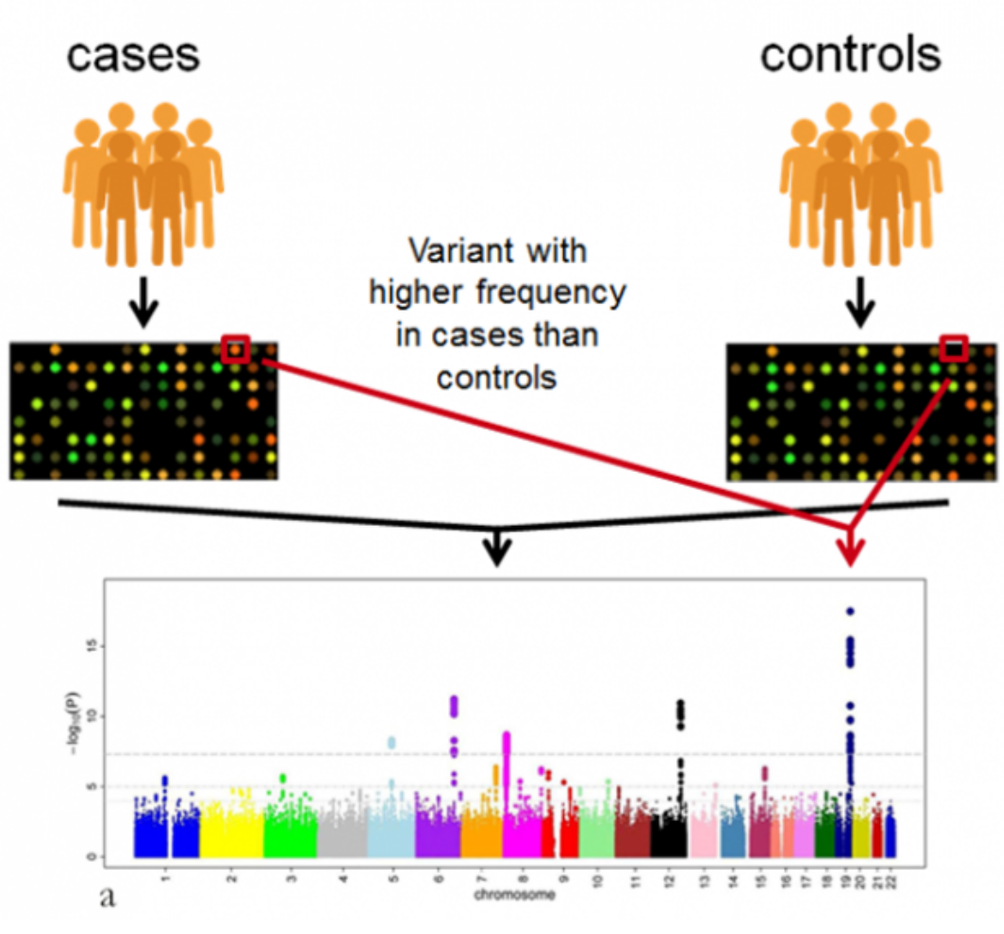


Fig. 2. *P*2-value distributions of selected groups of genes. The plot shows the cumulative fraction of selected biological processes showing the excess of cases of significant positive selection in genes for olfaction, amino acid catabolism, and Mendelian disease genes (OMIM) relative to the overall distribution of genes. The distribution of developmental genes that do not show a significant excess is shown for comparison.

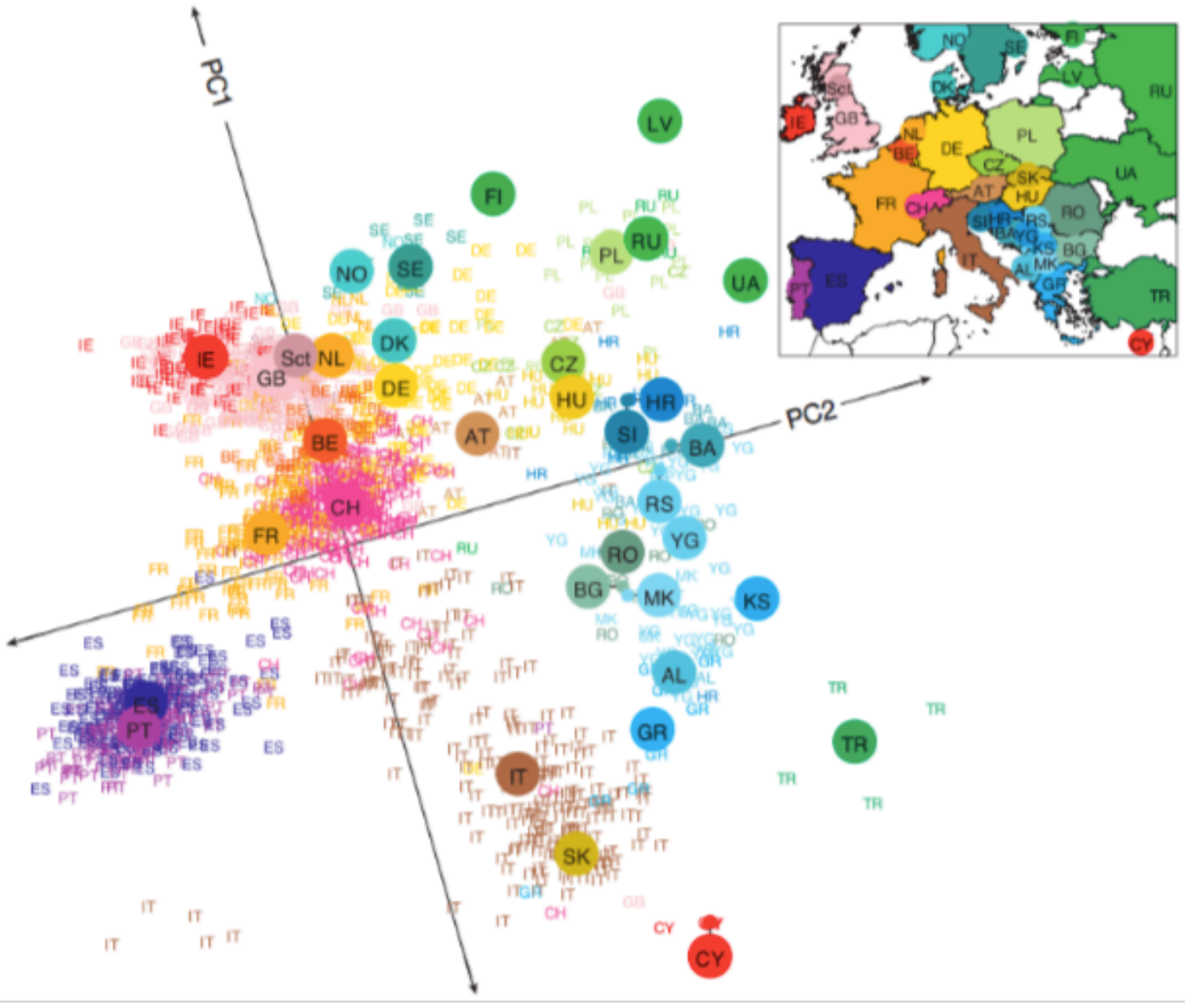
Making Sense of Data





Modern data has a lot of dimensions!

Making Sense of Data



Explore with PCA

Making Sense of Data

Iris setosa

Iris versicolor

Iris virginica

Anderson’s Iris dataset