Sequence Divergence & The Molecular “Clock”

Sequence Divergence

- simple genetic distance, d = the proportion of sites that differ between two aligned, homologous sequences
- given a constant mutation/substitution rate, d should provide a measure of time since divergence
  - but this is complicated by **multiple hits** (homoplasy)
  - corrected distance metrics account for the fact that there are **not** an infinite number of sites in a sequence

Expected sequence divergence

- for neutral polymorphisms, substitution rate = mutation rate
- thus, for two diverging lineages… $k = 2T\mu$
  - where k = the number of substitutions observed between two species and T is the time since divergence
  - note that T and μ can be measured either in years or generations
- solving for T… $T = \dfrac{k}{2\mu}$
  - note that 2μ is often expressed as the "rate of sequence divergence" (i.e., twice the per lineage rate)

Rates and Dates: Divergence Time Estimates

- requires calibration with fossil or geological events
- typically assumes a "molecular clock"
  - Zuckerland & Pauling (1962)
- but new methods allow a relaxation of the molecular clock assumption

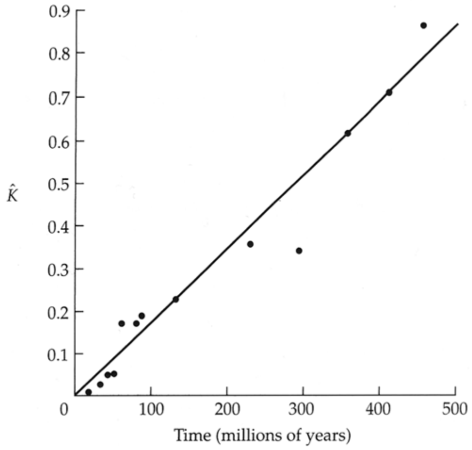


FIGURE 7.6 Relation between estimated number of amino acid substitutions in α-globin ($\hat{K}$) between pairs of the vertebrate species, against time since each pair diverged from a common ancestor. The straight line is expected based on a uniform rate of amino acid substitution during the entire period. (From Kimura 1983.)

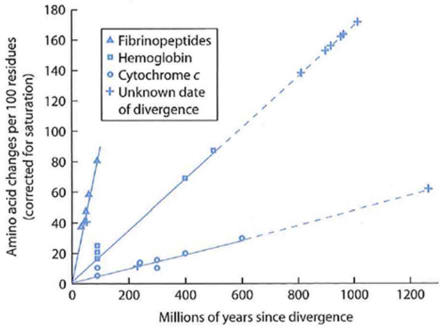


Figure 8.13 (*left*) Rates of protein evolution as amino acid changes per 100 residues in fibrinopeptides, hemoglobin, and cytochrome *c* over very long periods of time. Rates of divergence are linear over time for each protein, as expected for a molecular clock. Different proteins have different clock rates due to different mutation rates and degrees of functional constraint imposed by natural selection. Amino acid changes between pairs of taxa with unknown divergence times are plotted on dashed lines with the same slope as lines through points for taxa with estimated divergence times. The six points with unknown divergence times for hemoglobin represent divergence of ancestral globins into hemoglobins and myoglobins in the earliest animals, events that the molecular clock estimates to have happened between 1.1 billion and 800 million years ago. Data from Dickerson (1971).

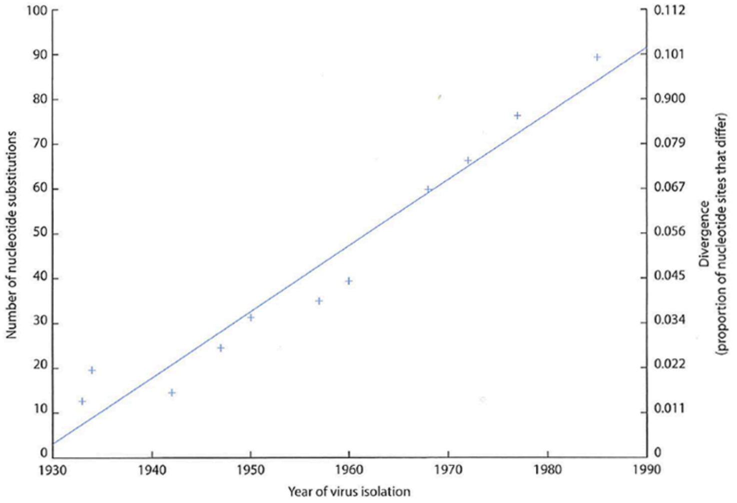


Figure 8.12 Rates of nucleotide change in the NS gene that codes for "nonstructural" proteins based on 11 human influenza A virus samples isolated between 1933 and 1985. The number of years since isolation and DNA sequence divergence from an inferred common ancestor are positively correlated. The pattern of increasing substitutions as time since divergence increases is expected under the molecular clock hypothesis. The observed rate of substitution was approximately 1.9×10^{-3} substitutions per nucleotide site per year, a very high rate compared to most genes in eukaryotes. The line is a least-squares fit. Data from Buonagurio et al. (1986).

Fleischer *et al*. 1998. Evolution on a volcanic conveyor belt: using phylogeographic reconstructions and K-Ar based ages of the Hawaiian Islands to estimate molecular evolutionary rates. *Mol. Ecol*. 7:533-545.

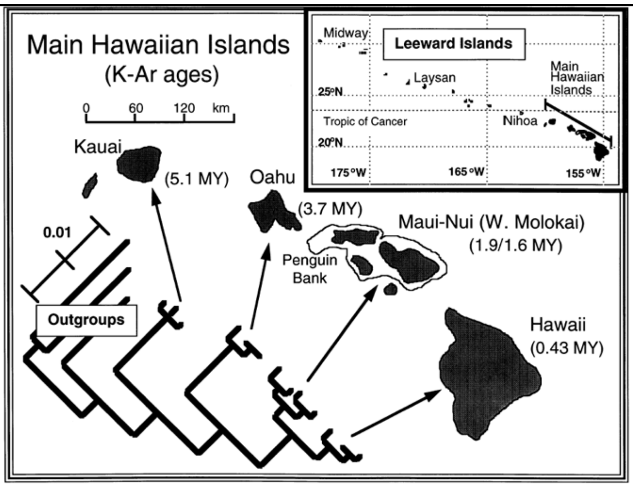


Fig. 1 Map of the main Hawaiian Islands (plus inset map of main and leeward Hawaiian Islands). Ages of the oldest rocks from the main islands based on K-Ar dating are noted (from Carson & Clague 1995). Maui-nui is composed of the islands of Maui, Lanai, Molokai and Kahoolawe, all of which were connected until ≈ 0.3–0.4 Ma (and again at times during Pleistocene periods of low sea level). Oahu and West Molokai were putatively connected via the Penguin Bank for the first ≈ 0.3 Myr after West Molokai formed. Below the map is a neighbour-joining tree (Saitou & Nei 1987) based on Kimura 2-parameter and Γ-corrected distances that shows relationships and relative differentiation among four island populations of amakihi based on 675 bp of cytochrome *b* sequence. Included in the tree are three Kauai amakihi (*Hemignathus kauaiensis*) and 10 common amakihi of three subspecies: *Hemignathus virens chloris*, *H. vs. wilsoni* and *H. vs. virens*, from Oahu, Maui and Hawaii, respectively. Parsimony analyses produce a tree with the same topology as the neighbour-joining tree, as do analyses using sequences of mtDNA ATPase 6/8 (R. C. Fleischer *et al*. unpublished), ND2 and control region (C. Tarr, unpublished). Trees constructed from mtDNA restriction sites show sister taxon status between a clade containing Kauai and Oahu clades, and one containing Maui and Hawaii clades (Tarr & Fleischer 1993). In all of our analyses to date, most for which we have larger samples per island than the case reported here, each island population of amakihi has been found to be monophyletic.

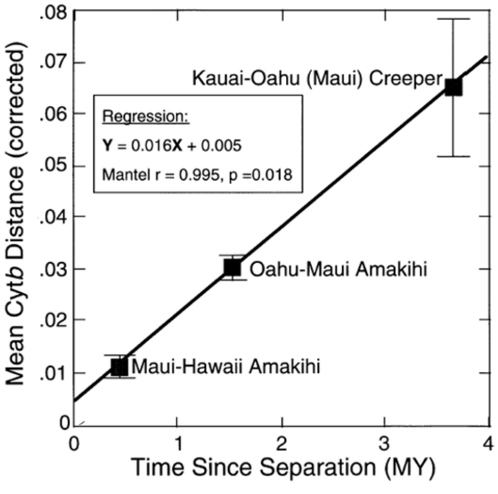


Fig. 3 Rate calibration for drepanidine mitochondrial cytochrome *b*. The two amakihi points are the mean of all pairwise Kimura 2-parameter and Γ-corrected distances, with intrapopulation variation from the parental population subtracted and unbiased standard error bars (Steel *et al*. 1996). The standard error bar for the single Maui and Kauai creeper pairwise comparison was calculated in MEGA (Kumar *et al*. 1993). The regression was significant based on 1000 permutations in a Mantel test (Rohlf 1990).

1.6% divergence per MY

0.8% per lineage per MY

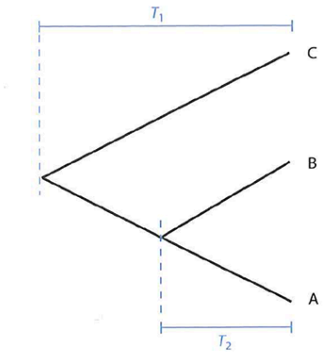


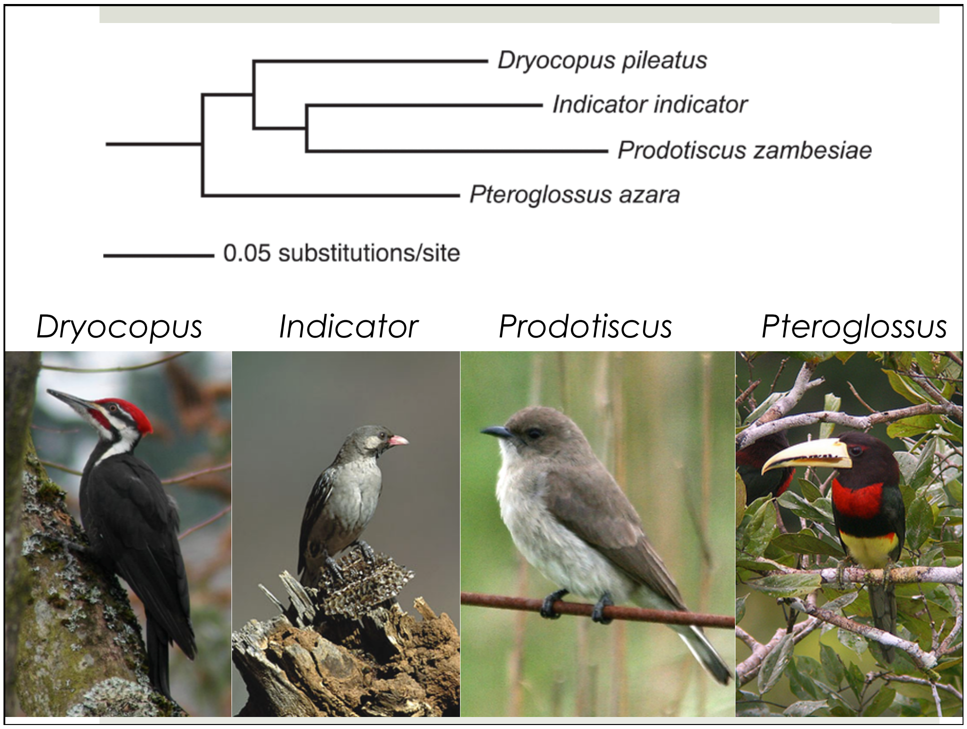
Figure 8.14 A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock. T_1 is the time in the past when species C and the ancestor of species A and B diverged. T_2 is the time in the past when species A and B diverged. If either T_1 or T_2 are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

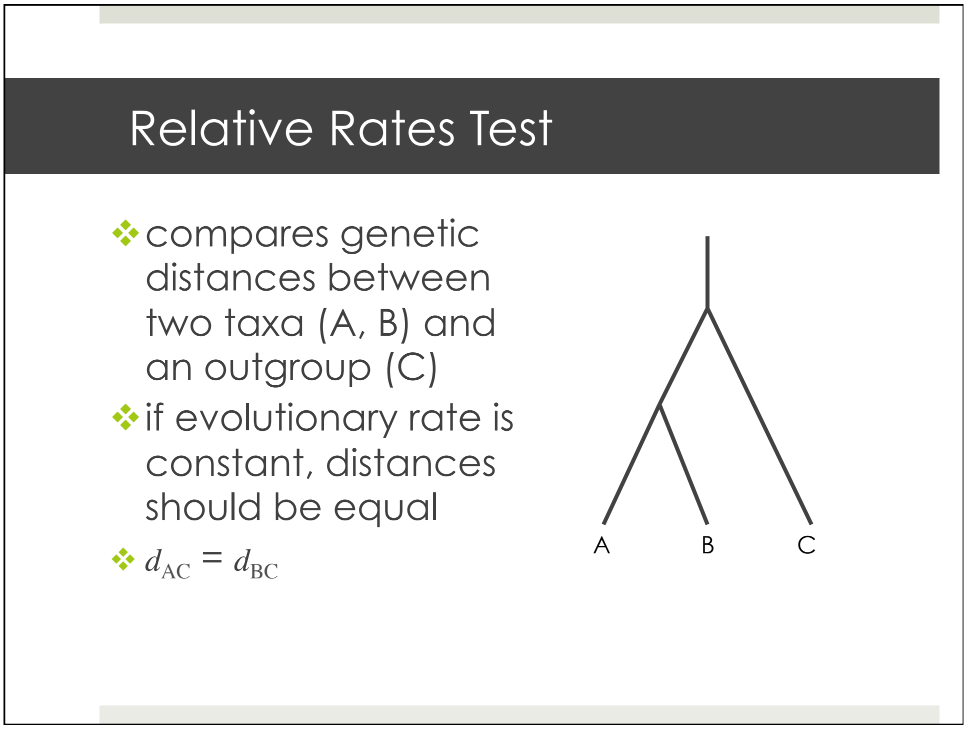
suppose T_1 is known...

$$\mu=\frac{1}{2}\left(\frac{K_{AC}}{2T_1}+\frac{K_{BC}}{2T_1}\right)$$

Problems with dating…

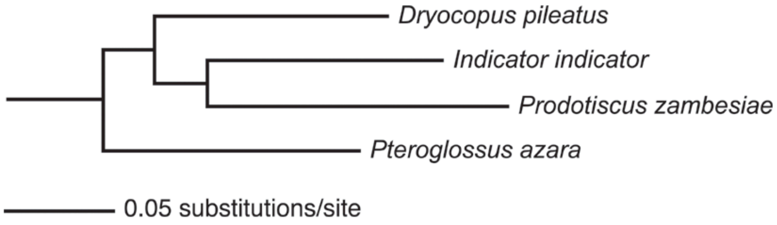
- uncertainty in calibration points
- fossil evidence provides lower bound on age only
- variance of genetic distance estimates
- “saturation” of genetic distances
- extrapolation outside of calibrated range
- ancestral polymorphism
- **variation in substitution rate among lineages**





Relative Rates Test

- compares genetic distances between two taxa (A, B) and an outgroup (C)
- if evolutionary rate is constant, distances should be equal
- $d_{AC} = d_{BC}$

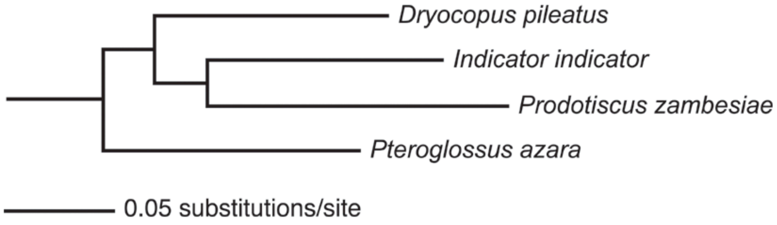


Dryocopus

Indicator

Prodotiscus

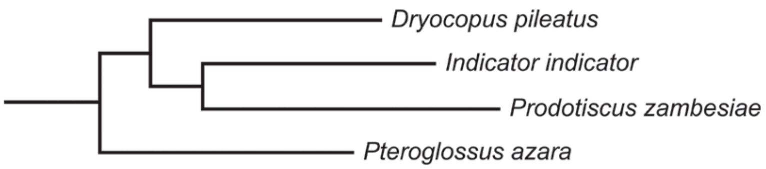
Pteroglossus

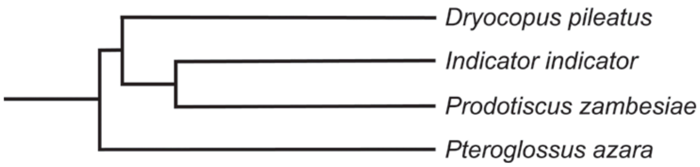


		Differences							
Comparison	Sites	**AG**	**CT**	**AC**	**AT**	**CG**	**GT**	All	TVs
Dryocopus vs *Indicator*	8991	323	754	360	149	61	30	1677	600
Dryocopus vs *Prodotiscus*	8991	322	772	458	157	78	44	1831	737

Likelihood ratio test for rate constancy

- compare the likelihood (probability) of the data when a molecular clock is enforced versus the likelihood when all branches are free to vary in length (product of time and mutation rate)



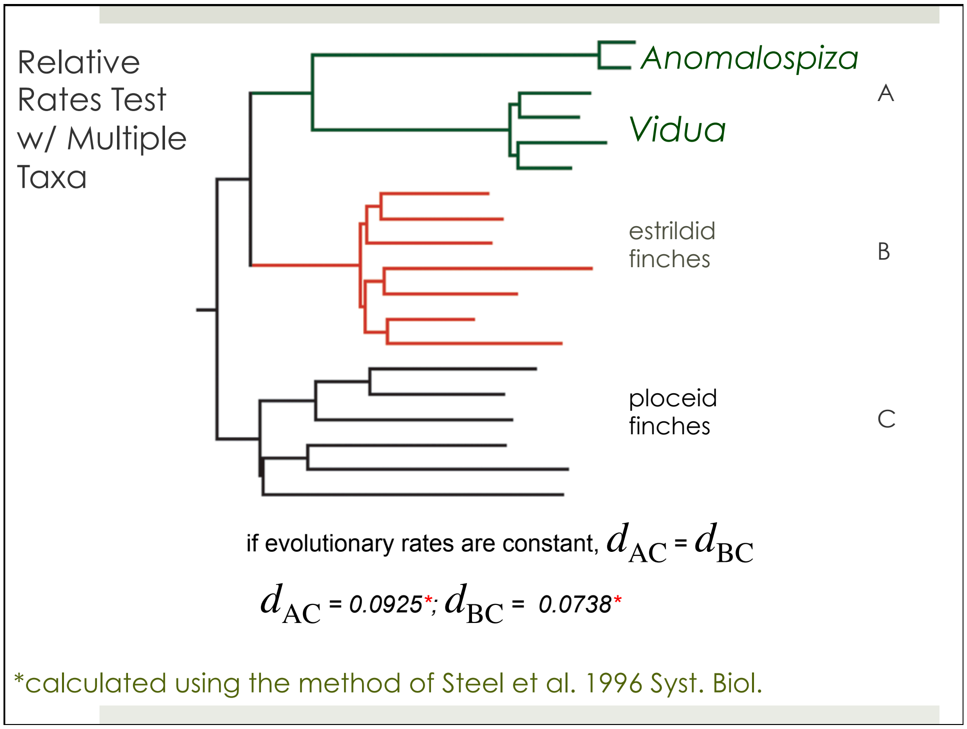


No clock: -ln L = 27859.36

Clock: -ln L = 27904.29

2 x ΔlnL = 89.86, $p < 0.0001$

- test statistic: 2 * Δ ln L is distributed approximately as X^2 (chi-square) with n-2 degrees of freedom, where n = number of terminal taxa
  - unconstrained tree: 2n-3 branch lengths
  - constrained tree: n-1 branch lengths

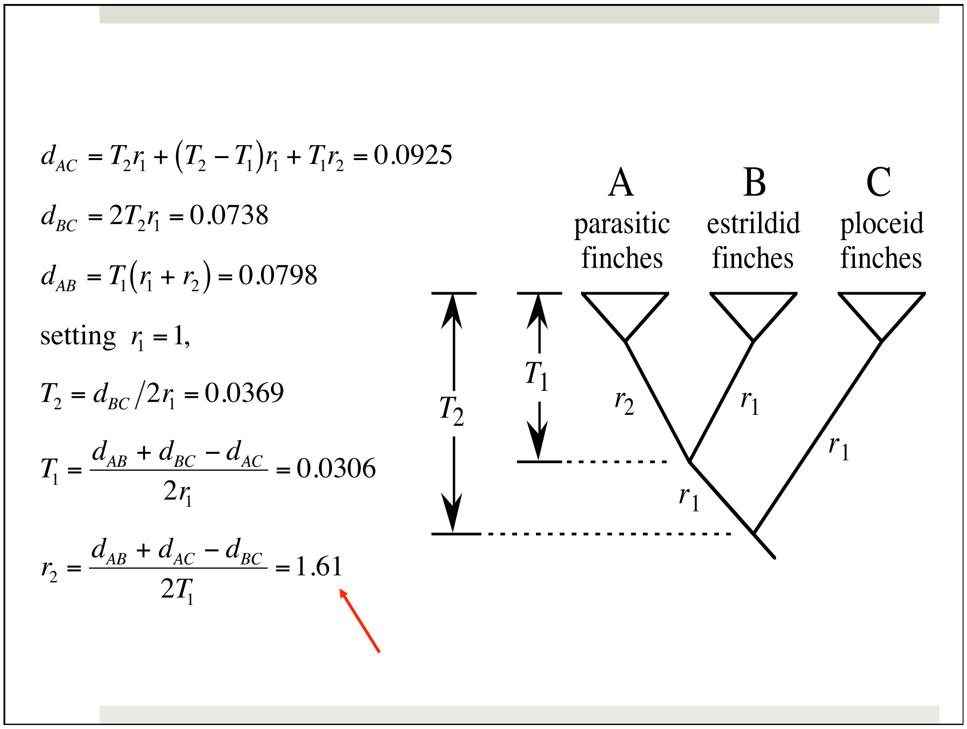


Relative Rates Test w/ Multiple Taxa

if evolutionary rates are constant, $d_{AC} = d_{BC}$

d_{AC} = *0.0925**; d_{BC} = *0.0738**

*calculated using the method of Steel et al. 1996 Syst. Biol.



$$d_{AC} = T_2 r_1 + (T_2 - T_1) r_1 + T_1 r_2 = 0.0925$$

$$d_{BC} = 2T_2 r_1 = 0.0738$$

$$d_{AB} = T_1(r_1 + r_2) = 0.0798$$

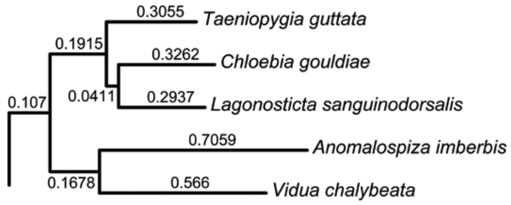
setting $r_1 = 1$,

$$T_2 = d_{BC}/2r_1 = 0.0369$$

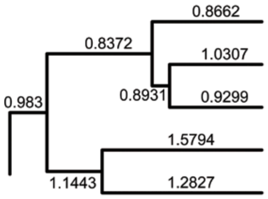
$$T_1 = \frac{d_{AB} + d_{BC} - d_{AC}}{2r_1} = 0.0306$$

$$r_2 = \frac{d_{AB} + d_{AC} - d_{BC}}{2T_1} = 1.61$$

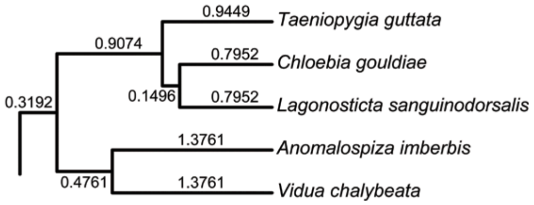
MrBayes: branch lengths (product of time and rate)



BEAST: separate estimates of rate and time



rate



time

Variation in Evolutionary Rate

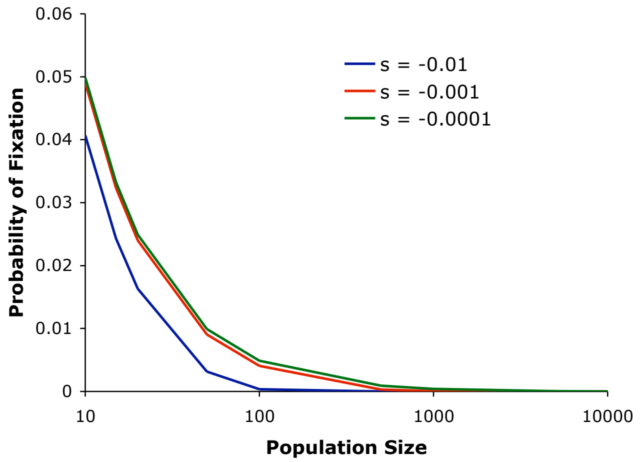
- rates may vary among lineages due to…
  - differences in life history
    - especially generation time, metabolic rate
  - diversifying natural selection
    - but likely limited to few sites in few genes
  - population history
    - the rate of neutral evolution does not depend on population size
    - the rate of nearly neutral evolution does!

Nearly Neutral Theory

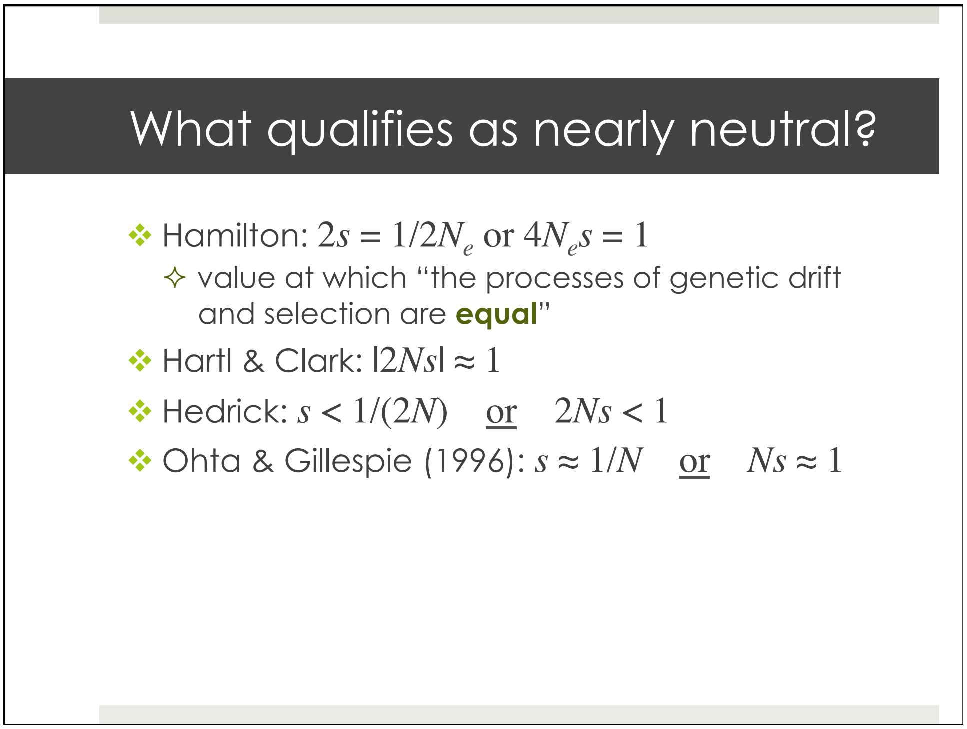
- what happens in small populations when selection is weak?
  - changes in allele frequency due to drift and selection are approximately equal $$|2Ns| \approx 1$$
- probability of fixation for a new, “nearly neutral” allele:

$$\Pr\left(A \text{ fixed}\right) = \frac{2s}{1-e^{-4Ns}}$$

$$w_{AA} = 1 + \underline{s}, \quad w_{Aa} = 1 + \underline{s/2}, \quad w_{aa} = 1$$

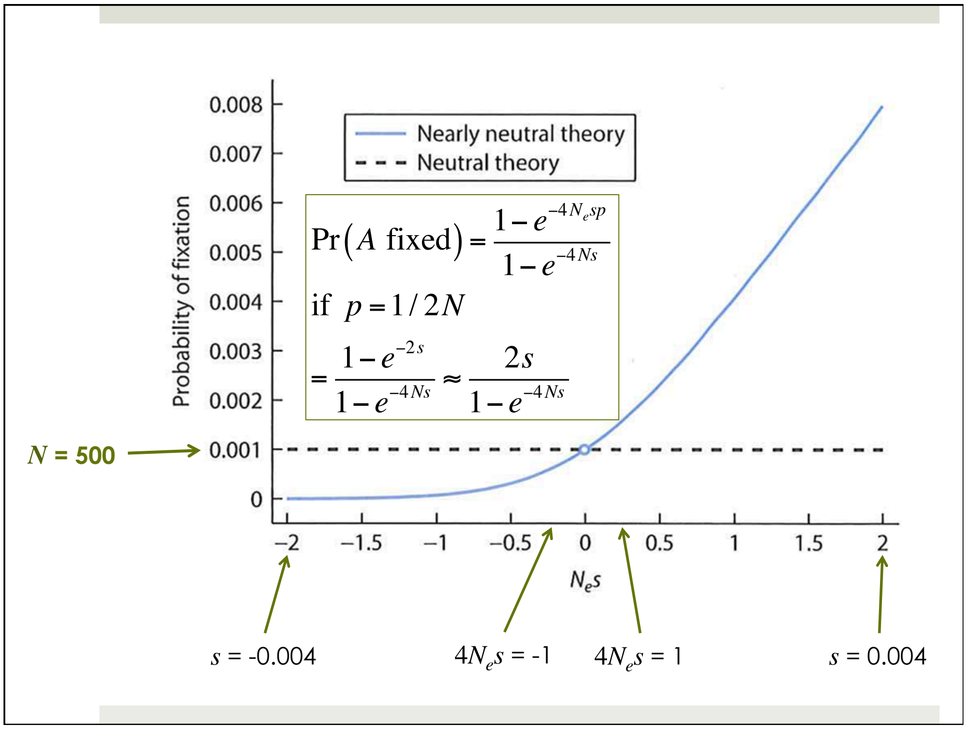


$$\Pr\left(A \text{ fixed}\right) = \frac{2s}{1-e^{-4Ns}}$$



What qualifies as nearly neutral?

- Hamilton: $2s = 1/2N_e$ or $4N_e s = 1$
  - value at which "the processes of genetic drift and selection are **equal**"
- Hartl & Clark: $|2Ns| \approx 1$
- Hedrick: $s < 1/(2N)$ or $2Ns < 1$
- Ohta & Gillespie (1996): $s \approx 1/N$ or $Ns \approx 1$



$$\Pr(A \text{ fixed}) = \frac{1-e^{-4N_e sp}}{1-e^{-4Ns}}$$

if $p = 1/2N$

$$= \frac{1-e^{-2s}}{1-e^{-4Ns}} \approx \frac{2s}{1-e^{-4Ns}}$$

$N = 500$

s = -0.004

$4N_e s$ = -1

$4N_e s$ = 1

s = 0.004

Nearly Neutral Theory - Summary

- the rate of neutral evolution is independent of population size $\cancel{2N}\mu \times \frac{1}{\cancel{2N}} = \mu$
  - substitution rate equals mutation rate
- in contrast, the fate of nearly neutral mutations depends on population size
  - $|2Ns| \approx 1$ when N is small, the effect of genetic drift can be comparable to that of selection, making slightly deleterious mutations "effectively neutral"
- thus, lineages experiencing small population size should accumulate both neutral and nearly neutral mutations, leading to a faster rate of sequence evolution

Testing the Nearly Neutral Theory

- how to distinguish neutral and nearly neutral mutations?
  - synonymous (silent) versus non-synonymous (replacement) substitutions?
  - synonymous likely to be neutral
  - non-synonymous more likely to be deleterious

- ❖ Ohta (1994) - generation time effect differs between synonymous and non-synonymous mutations

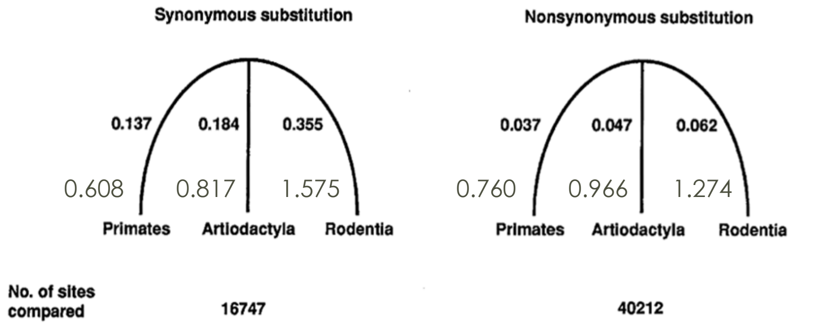
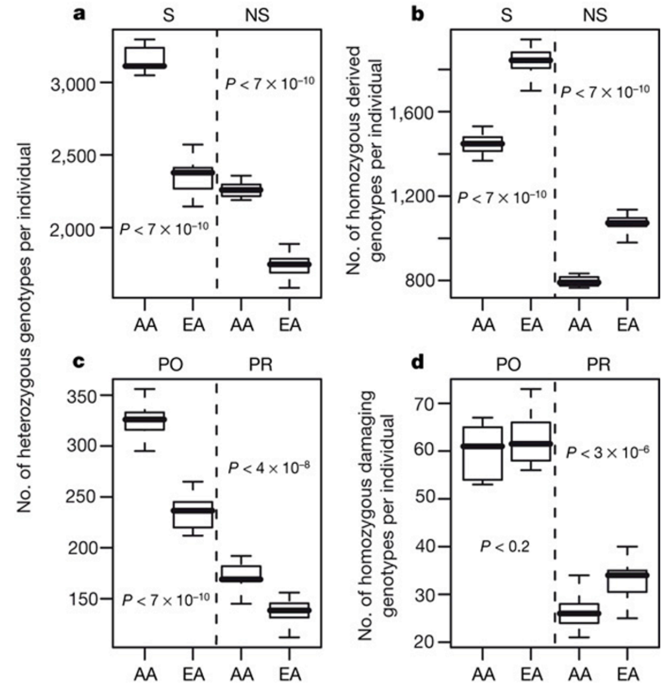


Fig. 1. Star phylogenies of 49 genes. Figures beside each branch are the estimated numbers of substitutions per site.

  - ✧ interpreted as consequence of nearly neutral evolution
  - ✧ inverse correlation between population size and body size/generation time

Lohmueller et al. 2007. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994-997.

- ❖ used protein structure prediction to estimate the number of functionally consequential SNPs carried by each of 15 African Americans (AA) and 20 European Americans (EA)
- ❖ higher heterozygosity in AA, but…
- ❖ the proportion of SNPs that are non-synonymous is significantly higher in the EA sample (55.4%) than in the AA sample (47.0%)
- ❖ same result for SNPs that were inferred to be 'probably damaging' (15.9% in EA; 12.1% in AA)



Lohmueller et al. 2007. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994-997.

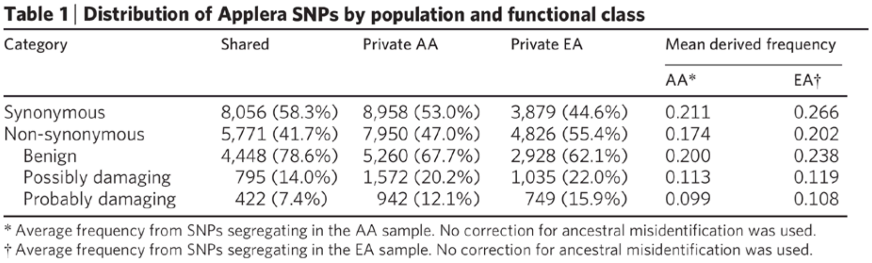


Table 1 | Distribution of Applera SNPs by population and functional class

Category	Shared	Private AA	Private EA	Mean derived frequency	
				AA*	EA†
Synonymous	8,056 (58.3%)	8,958 (53.0%)	3,879 (44.6%)	0.211	0.266
Non-synonymous	5,771 (41.7%)	7,950 (47.0%)	4,826 (55.4%)	0.174	0.202
Benign	4,448 (78.6%)	5,260 (67.7%)	2,928 (62.1%)	0.200	0.238
Possibly damaging	795 (14.0%)	1,572 (20.2%)	1,035 (22.0%)	0.113	0.119
Probably damaging	422 (7.4%)	942 (12.1%)	749 (15.9%)	0.099	0.108

* Average frequency from SNPs segregating in the AA sample. No correction for ancestral misidentification was used.
† Average frequency from SNPs segregating in the EA sample. No correction for ancestral misidentification was used.

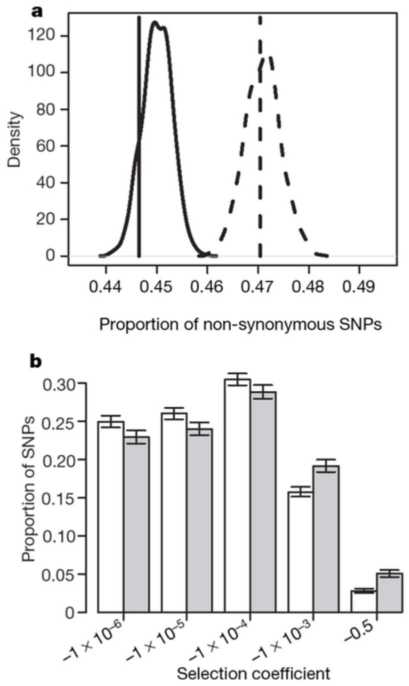


Figure 2 | Demography and selection can cause a proportional excess of non-synonymous SNPs in Europeans. a, b, Results of forward simulations of a population that expanded (AA 2 in Supplementary Table 2), to represent the AA population and a population that experienced a bottleneck to represent the EA population (EA 1 in Supplementary Table 2). **a**, Distribution of the proportion of non-synonymous SNPs segregating in samples simulated under European (dashed curve) and African (solid curve) demographic models. Vertical lines show the observed proportions in the Applera data set. **b**, Distribution of selection coefficients for simulated SNPs in the AA (white bars) and the EA (grey bars) samples. The labels on the *x* axis are the more negative limits of the bins. Error bars denote 95% intervals on the proportion of SNPs in each group.

Lohmueller et al. 2007. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994-997.

Estimating dN and dS

- dN - non-synonymous divergence
- dS - synonymous divergence

Nei-Gojobori (1986) Method

1. calculate number of potentially synonymous and non-synonymous sites ($s + n = 3$ per codon), disregarding stop codons

$$S = \sum_{j=1}^{n} \sum_{i=1}^{3} f_i$$

$$N = 3n - S$$

e.g., Lysine:
AAA: s = 1/3
AAG: s = 1/3

e.g., Leucine:
TTA: s = 2/3
TTG: s = 2/3
CTA: s = 1-1/3
CTG: s = 1-1/3
CTC: s = 1
CTT: s = 1

Nei-Gojobori (1986) Method

1. calculate number of potentially synonymous and non-synonymous sites ($s + n = 3$ per codon), disregarding stop codons
2. calculate number of synonymous (s_d) and non-synonymous (n_d) differences for each codon
   1. if one difference, then obvious
   2. if two differences…

two differences

- E.g., 2 possible routes…

  (1) TTT (Phe) - GTT (Val) - GTA (Val)

  1 nonsynonymous, 1 synonymous

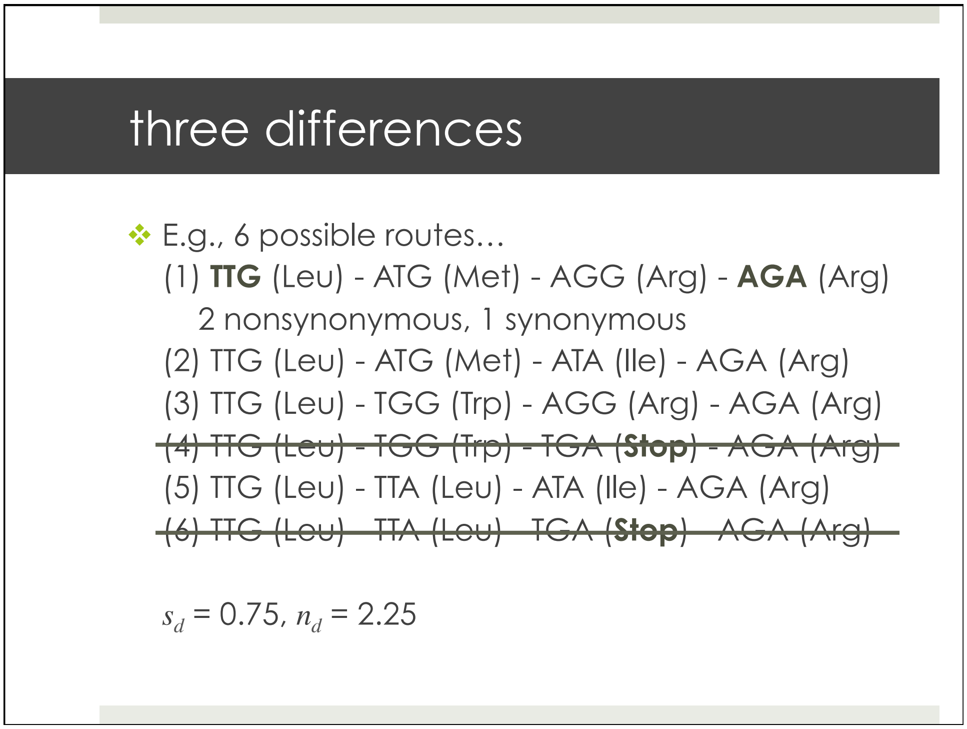
  (2) TTT (Phe) - TTA (Leu) - GTA (Val)

  2 nonsynonymous

  s_d = 0.5, n_d = 1.5

Nei-Gojobori (1986) Method

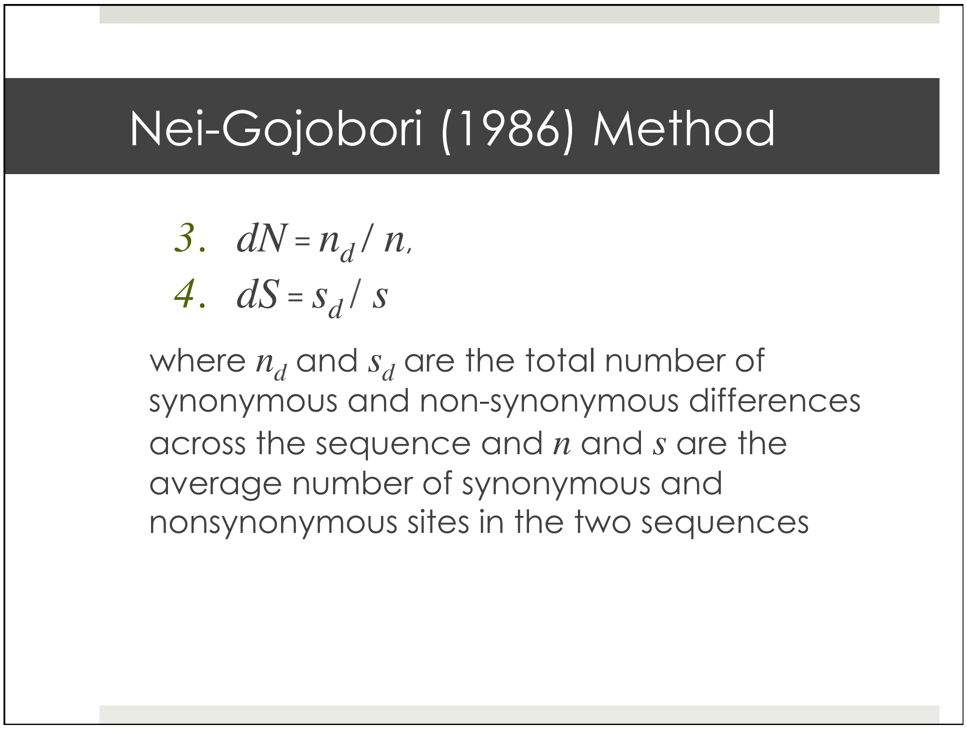
1. calculate number of potentially synonymous and non-synonymous sites ($s + n = 3$ per codon), disregarding stop codons
2. calculate number of synonymous (s_d) and non-synonymous (n_d) differences for each codon
   1. if one difference, then obvious
   2. if two differences…
   3. if three differences…



three differences

- E.g., 6 possible routes…
  (1) **TTG** (Leu) - ATG (Met) - AGG (Arg) - **AGA** (Arg)
  2 nonsynonymous, 1 synonymous
  (2) TTG (Leu) - ATG (Met) - ATA (Ile) - AGA (Arg)
  (3) TTG (Leu) - TGG (Trp) - AGG (Arg) - AGA (Arg)
  ~~(4) TTG (Leu) - TGG (Trp) - TGA (**Stop**) - AGA (Arg)~~
  (5) TTG (Leu) - TTA (Leu) - ATA (Ile) - AGA (Arg)
  ~~(6) TTG (Leu) - TTA (Leu) - TGA (**Stop**) - AGA (Arg)~~

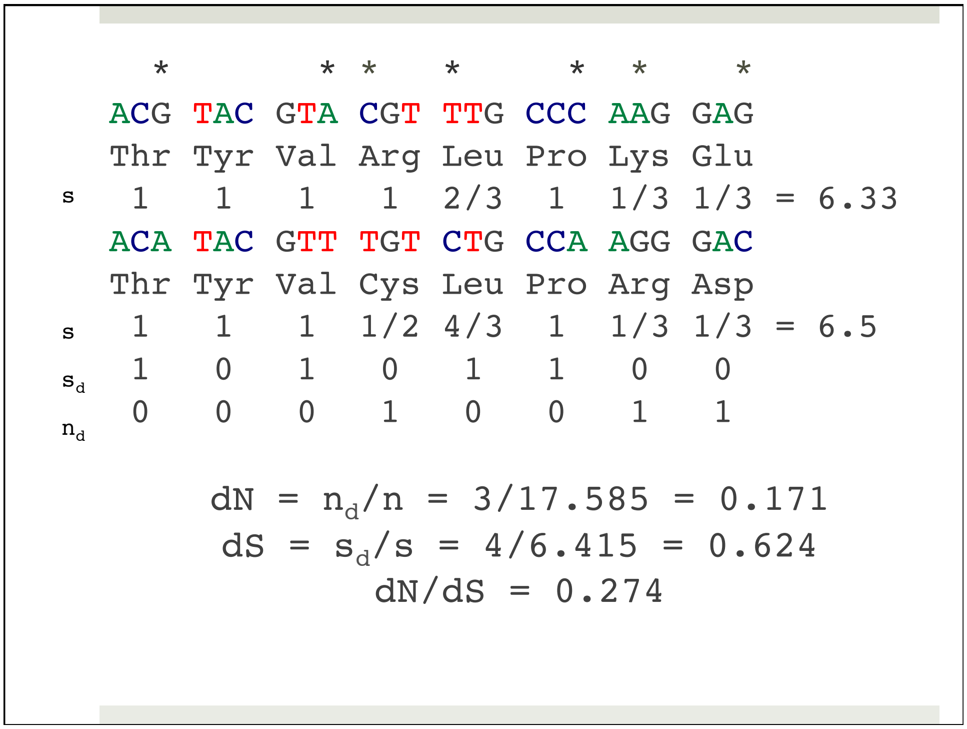
  $s_d = 0.75$, $n_d = 2.25$



Nei-Gojobori (1986) Method

3. $dN = n_d / n,$
4. $dS = s_d / s$

where n_d and s_d are the total number of synonymous and non-synonymous differences across the sequence and n and s are the average number of synonymous and nonsynonymous sites in the two sequences

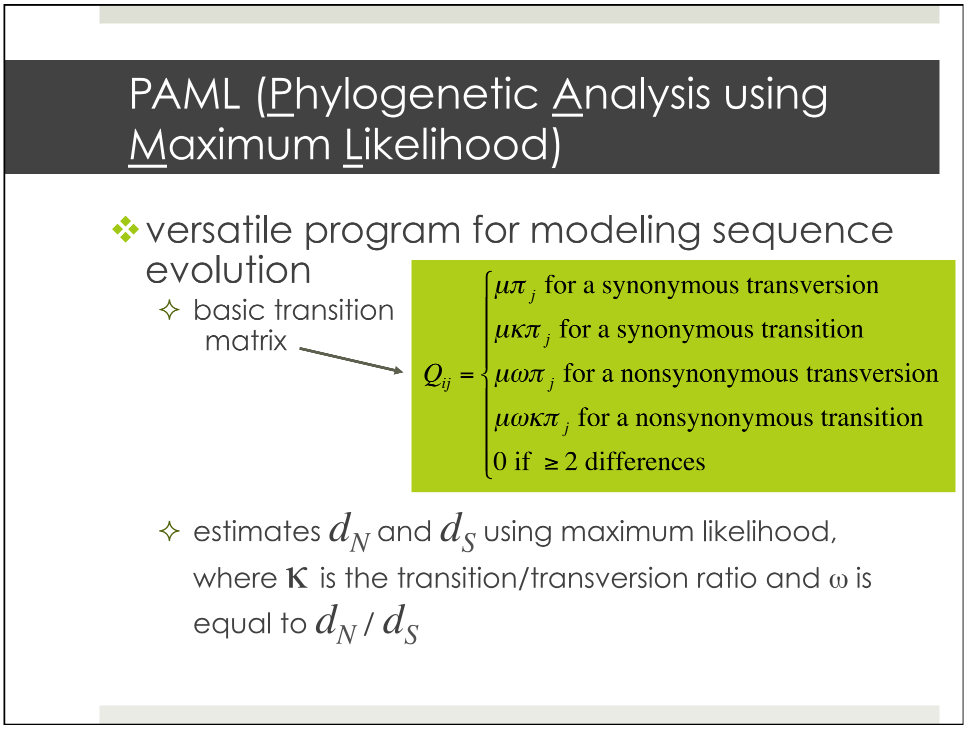


```
      *       *  *   *     *  *    *
    ACG TAC GTA CGT TTG CCC AAG GAG
    Thr Tyr Val Arg Leu Pro Lys Glu
s    1   1   1   1  2/3  1  1/3 1/3 = 6.33
    ACA TAC GTT TGT CTG CCA AGG GAC
    Thr Tyr Val Cys Leu Pro Arg Asp
s    1   1   1  1/2 4/3  1  1/3 1/3 = 6.5
s_d  1   0   1   0   1   1   0   0
n_d  0   0   0   1   0   0   1   1
```

$$dN = n_d/n = 3/17.585 = 0.171$$

$$dS = s_d/s = 4/6.415 = 0.624$$

$$dN/dS = 0.274$$



PAML (Phylogenetic Analysis using Maximum Likelihood)

- versatile program for modeling sequence evolution
  - basic transition matrix

$$Q_{ij} = \begin{cases} \mu\pi_j \text{ for a synonymous transversion} \\ \mu\kappa\pi_j \text{ for a synonymous transition} \\ \mu\omega\pi_j \text{ for a nonsynonymous transversion} \\ \mu\omega\kappa\pi_j \text{ for a nonsynonymous transition} \\ 0 \text{ if } \geq 2 \text{ differences} \end{cases}$$

  - estimates d_N and d_S using maximum likelihood, where $\mathbf{K}$ is the transition/transversion ratio and ω is equal to d_N / d_S

dN: *dS* ratio

- when measured in relation to the number of synonymous and non-synonymous sites
  - ✧ dN/dS = 1 if all substitutions are neutral
  - ✧ dN/dS > 1 suggests positive, diversifying selection
  - ✧ dN/dS < 1 suggests purifying selection (i.e., constraints on protein evolution)

Hughes & Nei 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167-170.

- examined dN / dS in human MHC class 1 genes HLA-A, B & C (n = 12 sequences)

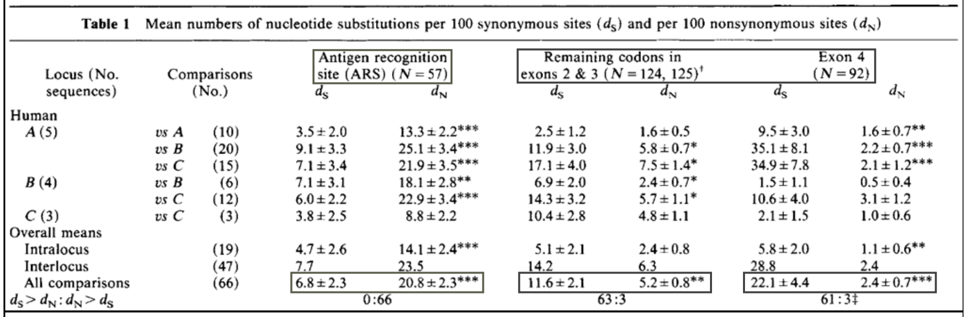


Table 1 Mean numbers of nucleotide substitutions per 100 synonymous sites (d_S) and per 100 nonsynonymous sites (d_N)

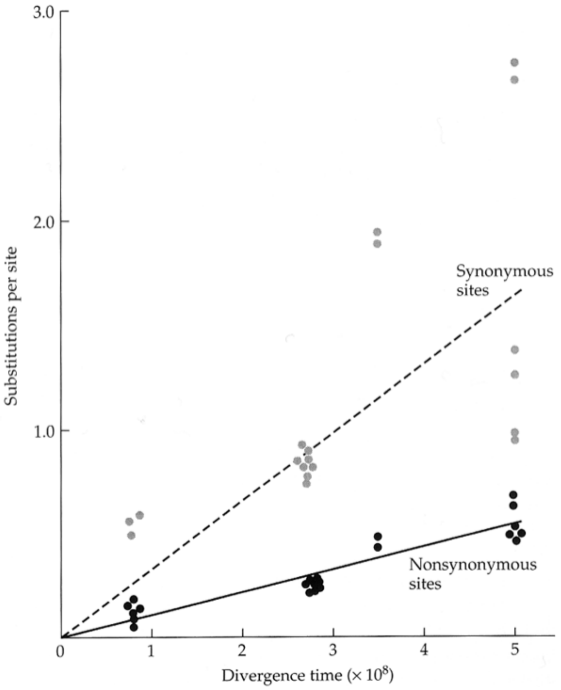
Locus (No. sequences)	Comparisons (No.)		Antigen recognition site (ARS) ($N = 57$)		Remaining codons in exons 2 & 3 ($N = 124, 125$)†		Exon 4 ($N = 92$)	
			d_S	d_N	d_S	d_N	d_S	d_N
Human								
A (5)	*vs A*	(10)	3.5 ± 2.0	13.3 ± 2.2***	2.5 ± 1.2	1.6 ± 0.5	9.5 ± 3.0	1.6 ± 0.7**
	vs B	(20)	9.1 ± 3.3	25.1 ± 3.4***	11.9 ± 3.0	5.8 ± 0.7*	35.1 ± 8.1	2.2 ± 0.7***
	vs C	(15)	7.1 ± 3.4	21.9 ± 3.5***	17.1 ± 4.0	7.5 ± 1.4*	34.9 ± 7.8	2.1 ± 1.2***
B (4)	*vs B*	(6)	7.1 ± 3.1	18.1 ± 2.8**	6.9 ± 2.0	2.4 ± 0.7*	1.5 ± 1.1	0.5 ± 0.4
	vs C	(12)	6.0 ± 2.2	22.9 ± 3.4***	14.3 ± 3.2	5.7 ± 1.1*	10.6 ± 4.0	3.1 ± 1.2
C (3)	*vs C*	(3)	3.8 ± 2.5	8.8 ± 2.2	10.4 ± 2.8	4.8 ± 1.1	2.1 ± 1.5	1.0 ± 0.6
Overall means								
Intralocus		(19)	4.7 ± 2.6	14.1 ± 2.4***	5.1 ± 2.1	2.4 ± 0.8	5.8 ± 2.0	1.1 ± 0.6**
Interlocus		(47)	7.7	23.5	14.2	6.3	28.8	2.4
All comparisons		(66)	6.8 ± 2.3	20.8 ± 2.3***	11.6 ± 2.1	5.2 ± 0.8**	22.1 ± 4.4	2.4 ± 0.7***
$d_S > d_N : d_N > d_S$			0:66		63:3		61:3‡	

dN / dS > 1: evidence of positive, diversifying selection

Positive Selection

- testing at the gene level is a "dull tool"
  - positive selection will usually affect one or a few codons, while the rest of the gene remains constrained (dN/dS << 1)
  - nonetheless, genes associated with immune function and reproduction (self-recognition, sperm competition, sexual conflict) often have dN/dS > 1
- more sophisticated methods are available to identify individual sites under selection

FIGURE 7.12 Synonymous sites and nonsynonymous sites in the *β*-globin gene undergo substitutions at different rates, but to a first approximation both may appear to exhibit a clocklike substitution process. (From Li et al. 1985a.)



purifying selection is the norm: synonymous substitutions are more frequent than non-synonymous substitutions in most genes much greater variation among genes in non-synonymous rate than in synonymous rate

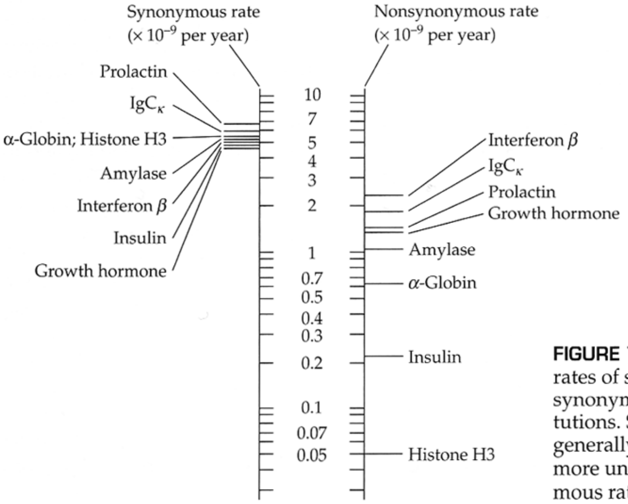
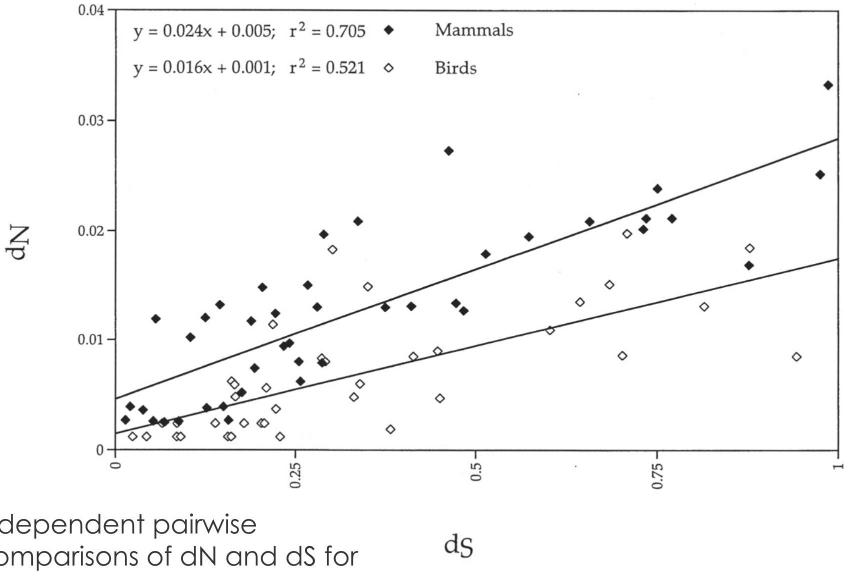


FIGURE 7.14 Comparison of rates of synonymous and nonsynonymous nucleotide substitutions. Synonymous rates are generally much faster and much more uniform than nonsynonymous rates. (From Kimura 1986.)

The avian constraint hypothesis...

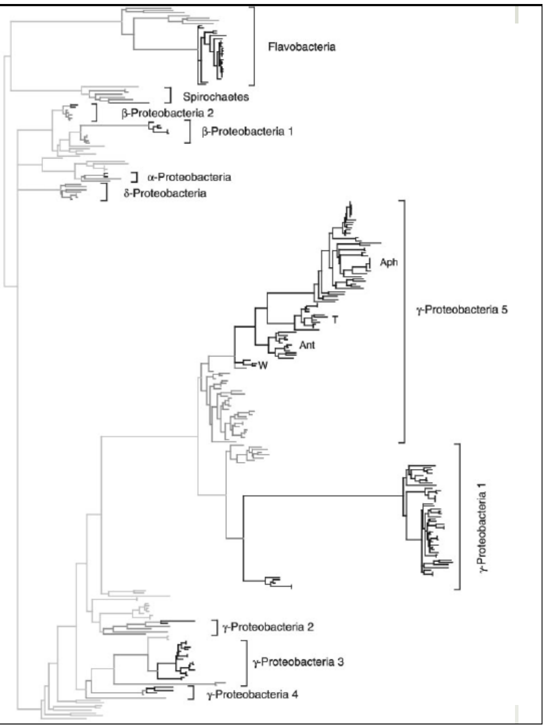


independent pairwise comparisons of dN and dS for mtDNA protein-coding genes

Stanley & Harrison MBE 1999

Woolfit & Bromham 2003 Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *MBE* 20:1545-1555.

- higher rate in endosymbiotic bacteria interpreted as a consequence of nearly neutral evolution
- if so, these organisms should also have a higher dN/dS ratio…



Moran (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *PNAS* 93: 2873-2878

- *Buchnera*: endosymbiont of aphids
- higher dN/dS than free living bacteria

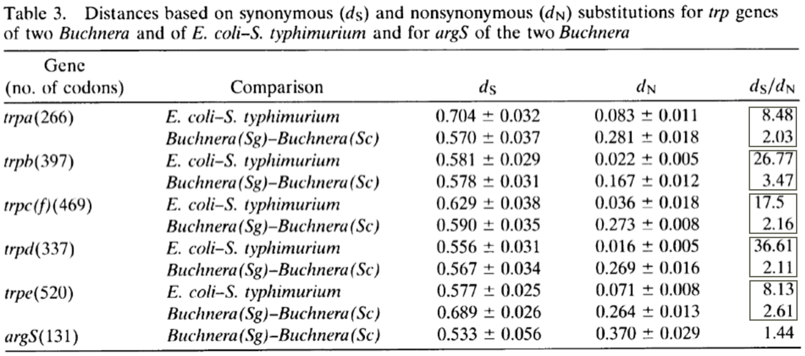
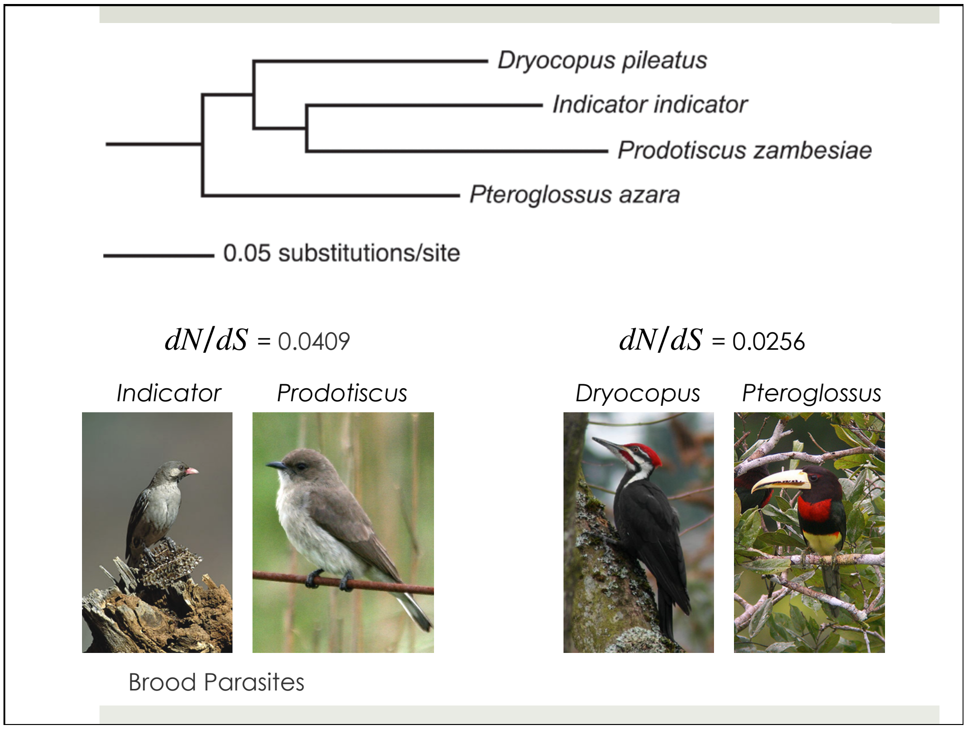


Table 3. Distances based on synonymous (d_S) and nonsynonymous (d_N) substitutions for *trp* genes of two *Buchnera* and of *E. coli–S. typhimurium* and for *argS* of the two *Buchnera*

Gene (no. of codons)	Comparison	d_S	d_N	d_S/d_N
trpa(266)	*E. coli–S. typhimurium*	0.704 ± 0.032	0.083 ± 0.011	8.48
	Buchnera(Sg)–Buchnera(Sc)	0.570 ± 0.037	0.281 ± 0.018	2.03
trpb(397)	*E. coli–S. typhimurium*	0.581 ± 0.029	0.022 ± 0.005	26.77
	Buchnera(Sg)–Buchnera(Sc)	0.578 ± 0.031	0.167 ± 0.012	3.47
trpc(f)(469)	*E. coli–S. typhimurium*	0.629 ± 0.038	0.036 ± 0.018	17.5
	Buchnera(Sg)–Buchnera(Sc)	0.590 ± 0.035	0.273 ± 0.008	2.16
trpd(337)	*E. coli–S. typhimurium*	0.556 ± 0.031	0.016 ± 0.005	36.61
	Buchnera(Sg)–Buchnera(Sc)	0.567 ± 0.034	0.269 ± 0.016	2.11
trpe(520)	*E. coli–S. typhimurium*	0.577 ± 0.025	0.071 ± 0.008	8.13
	Buchnera(Sg)–Buchnera(Sc)	0.689 ± 0.026	0.264 ± 0.013	2.61
argS(131)	*Buchnera(Sg)–Buchnera(Sc)*	0.533 ± 0.056	0.370 ± 0.029	1.44



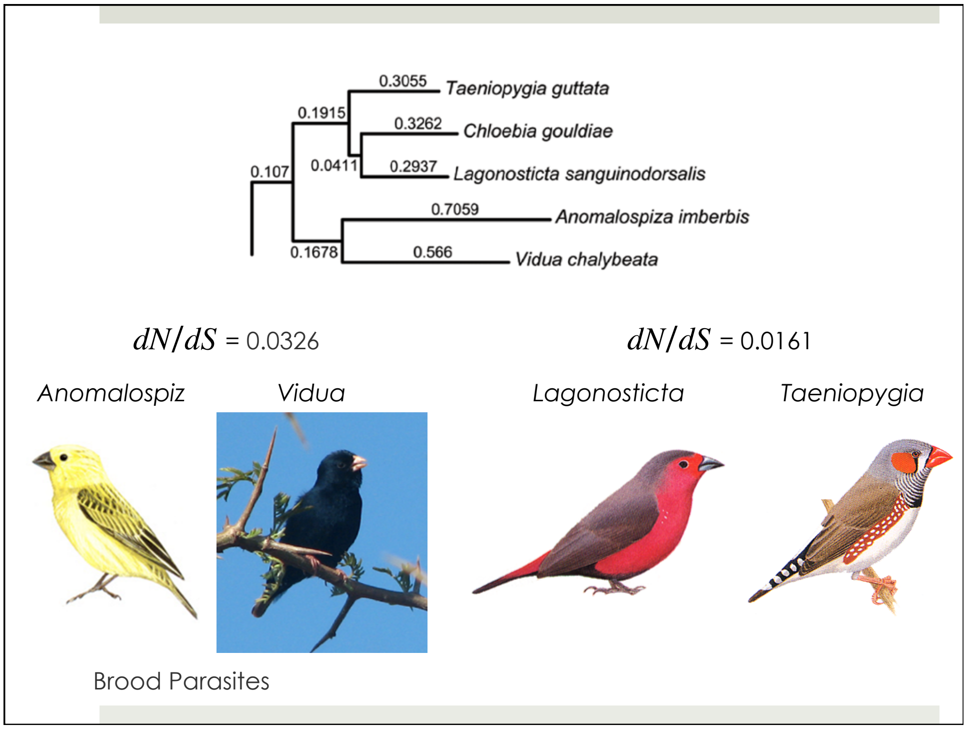
dN/dS = 0.0409

dN/dS = 0.0256

Indicator *Prodotiscus*

Dryocopus *Pteroglossus*

Brood Parasites



dN/dS = 0.0326

dN/dS = 0.0161

Anomalospiz *Vidua*

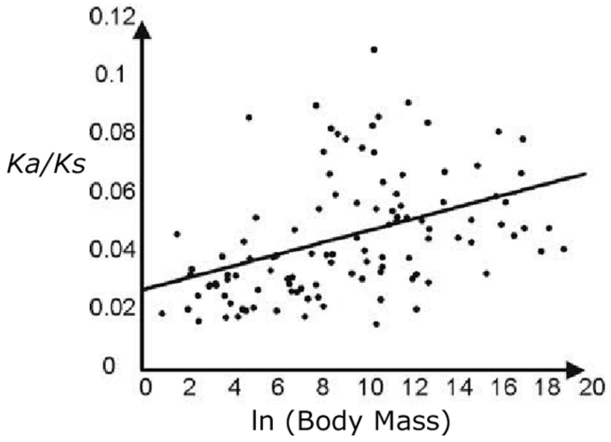
Lagonosticta *Taeniopygia*

Brood Parasites

“Constancy” of the Molecular Clock

- data on sequence divergence suggest that evolutionary rate in many organisms is roughly constant when time is measured in years even if generation times vary
- if nearly neutral evolution is important, then the inverse correlation between generation time and population size may help to explain the relative constancy of rate among organisms with different life histories
  - small critters have shorter generation time resulting in a higher rate of neutral evolution
  - large critters have longer generation time but also smaller populations, resulting in a higher rate of nearly neutral evolution

Popadin et al. (2007) Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *PNAS* **104**, 13390-13395.



But wait, are synonymous substitutions really neutral?

- codon-bias
  - “favored” codons (corresponding to more abundant tRNAs) are present at higher frequency in highly expressed genes than in genes with lower expression levels
- base composition bias
  - significant and sometimes substantial differences between lineages
    - e.g., birds have > GC content than mammals

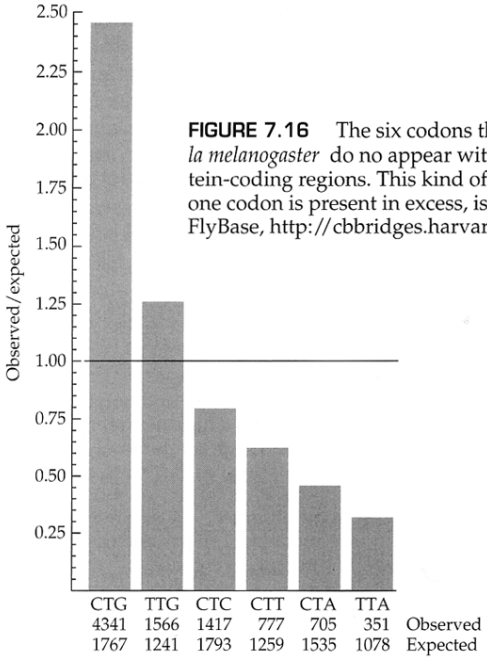
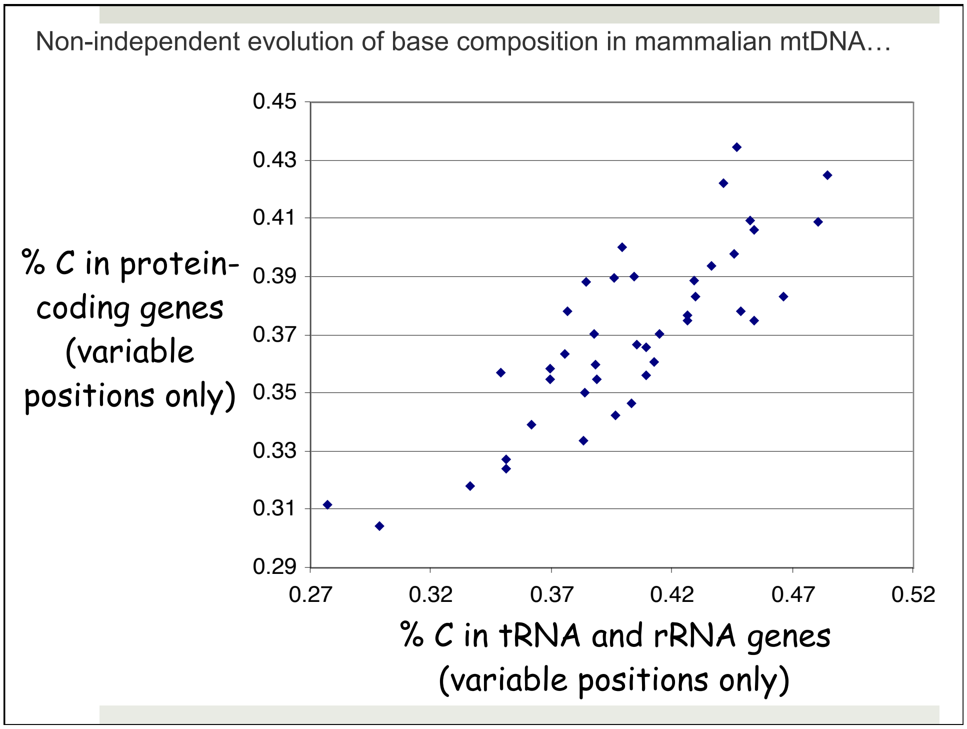
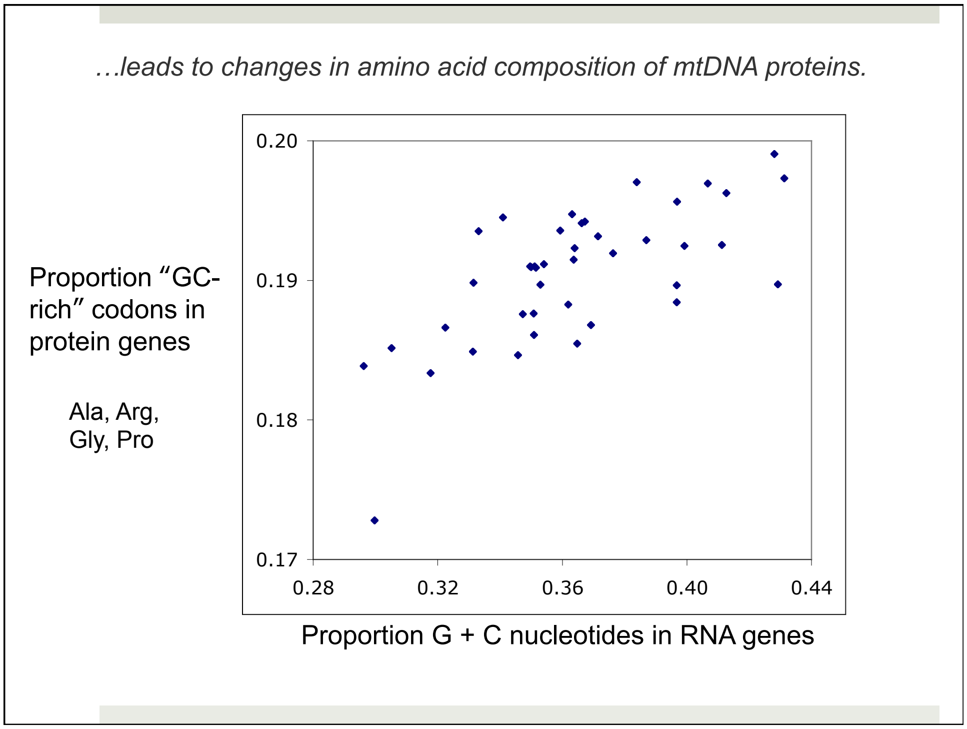


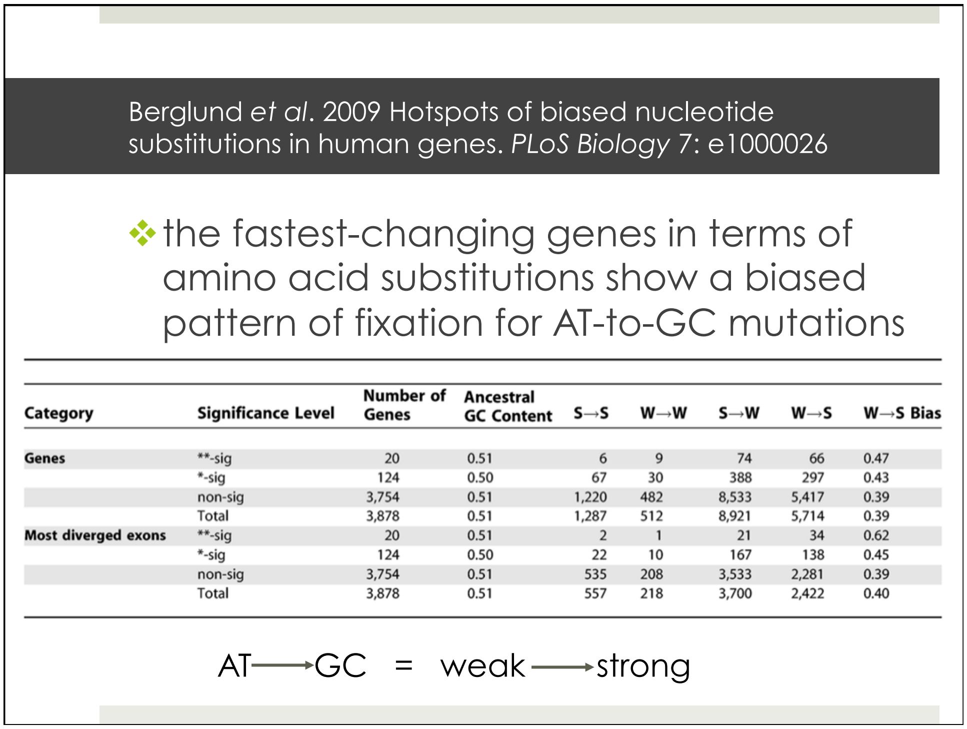
FIGURE 7.16 The six codons that encode leucine in *Drosophila melanogaster* do no appear with uniform frequency in protein-coding regions. This kind of codon-usage bias, in which one codon is present in excess, is often observed. (Data from FlyBase, http://cbbridges.harvard.edu:7081.)



Non-independent evolution of base composition in mammalian mtDNA…



…leads to changes in amino acid composition of mtDNA proteins.



Berglund *et al*. 2009 Hotspots of biased nucleotide substitutions in human genes. *PLoS Biology* 7: e1000026

- the fastest-changing genes in terms of amino acid substitutions show a biased pattern of fixation for AT-to-GC mutations

Category	Significance Level	Number of Genes	Ancestral GC Content	S→S	W→W	S→W	W→S	W→S Bias
Genes	**-sig	20	0.51	6	9	74	66	0.47
	*-sig	124	0.50	67	30	388	297	0.43
	non-sig	3,754	0.51	1,220	482	8,533	5,417	0.39
	Total	3,878	0.51	1,287	512	8,921	5,714	0.39
Most diverged exons	**-sig	20	0.51	2	1	21	34	0.62
	*-sig	124	0.50	22	10	167	138	0.45
	non-sig	3,754	0.51	535	208	3,533	2,281	0.39
	Total	3,878	0.51	557	218	3,700	2,422	0.40

AT → GC = weak → strong