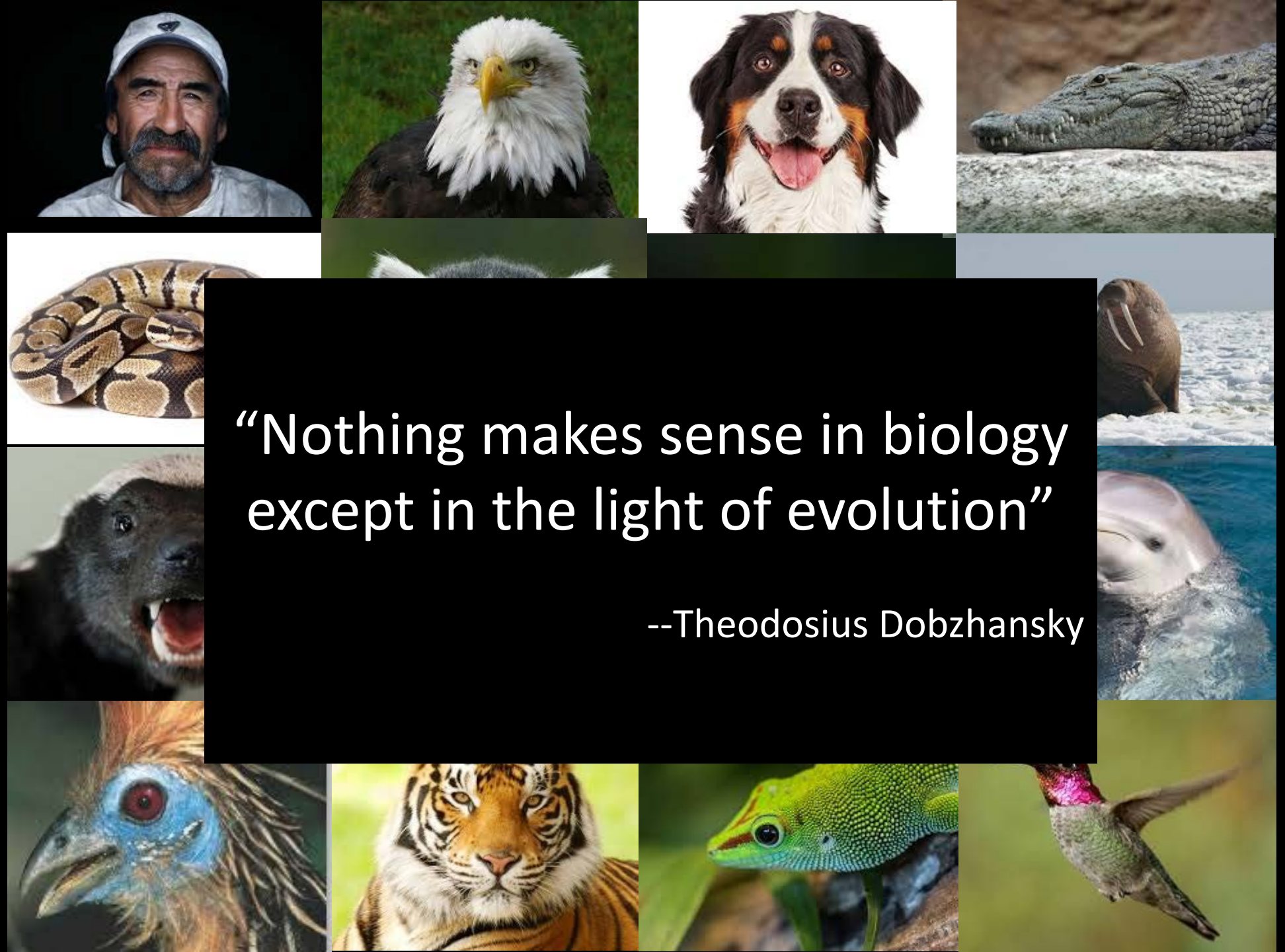
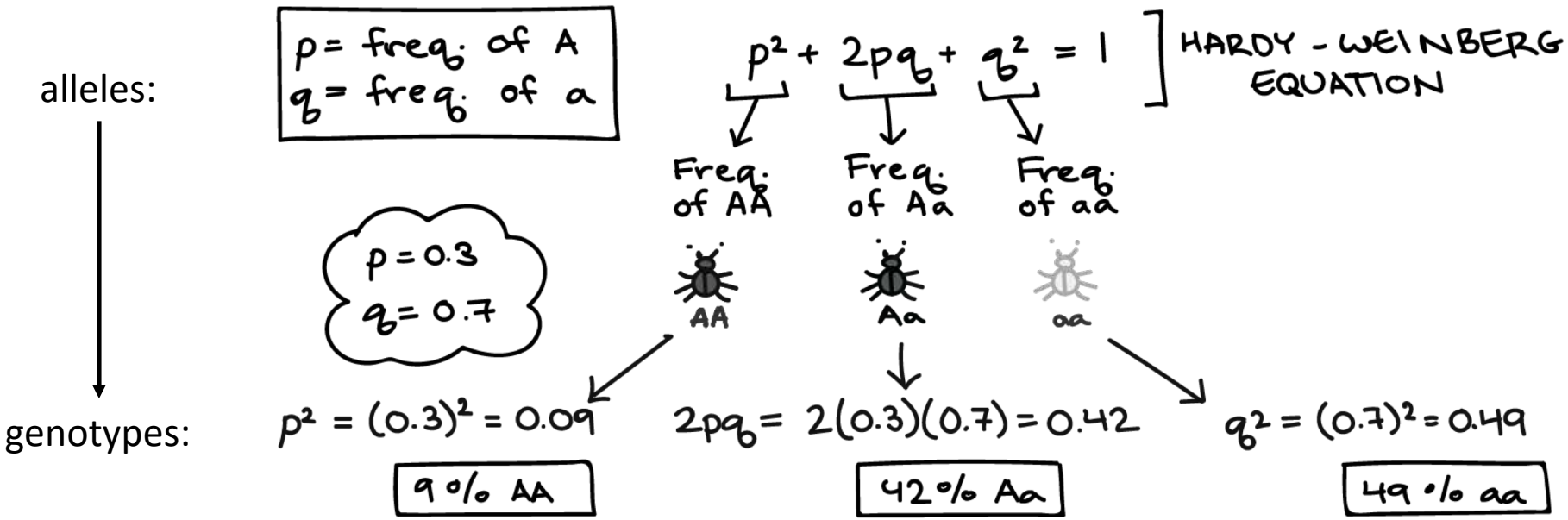
Tests for Selection: Codon Models



"Nothing makes sense in biology except in the light of evolution"

--Theodosius Dobzhansky

Where does Evolution Begin?



Khan Academy

When a population is in Hardy Weinberg ***equilibrium***, allele frequencies will not change between generations.

Evolution is when the Hardy Weinberg equilibrium is violated and there is a ***change in allele frequencies***.

Where does Evolution Begin?

There are five mechanisms of evolution that can change allele frequencies.

1. Mutation (the generation of new alleles)
2. Non-random mating
3. Gene flow
4. Genetic drift
5. Natural selection

"Evolution begins as one mutation, on one chromosome, in one individual."

Matthew Hahn, 2019.
First line of Chapter 1, *Molecular Population Genetics*

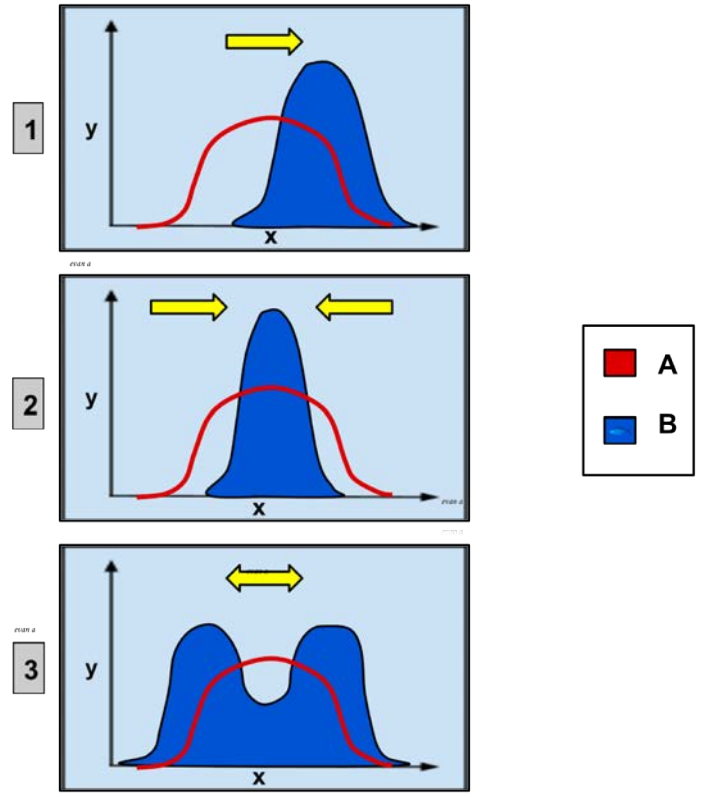
Natural Selection

Differential reproductive success ("fitness") of variants that leads to a shift in the average value of a trait as well as the underlying frequencies of alleles controlling that trait.

New mutations can be ***advantageous***, ***deleterious***, or ***neutral*** relative to the fitness of the ancestral allele.

We define *s* as the ***selection coefficient*** for a new allele.

Selection coefficient	**Interpretation**	**Effect on population frequency of allele**
$s > 0$	advantageous	Increases
$s < 0$	deleterious	Decreases
$s = 0$	neutral	No predictable effect (except when drift is strong)



wikipedia; Ealbert17, CC BY-SA 4.0

What types of selection can a trait (or gene) be under?

Negative selection

- AKA ‘purifying selection
- The vast majority of mutations occurring have a negative effect
- AKA ‘evolutionary constraint’

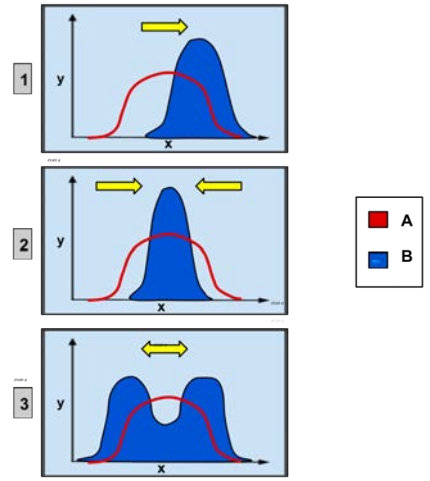
Positive selection

- Advantageous mutations have occurred and are either ***fixed*** in the population or are undergoing ***fixation***.
- AKA ‘Darwinian’ selection
- Remember: *Darwin knew nothing of genetic theories of inheritance.*

“Directional selection”

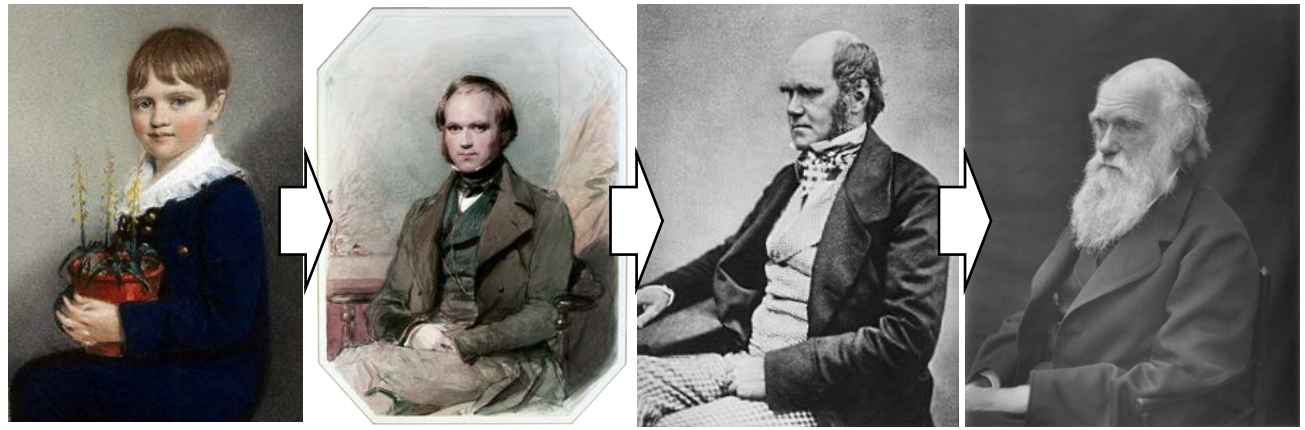
“Stabilizing selection”

“Balancing selection”



wikipedia; Ealbert17, CC BY-SA 4.0

Charles Darwin



wikipedia

Neo-Darwinians: Modern Synthesis

Mendel: Laws of inheritance

19th Century

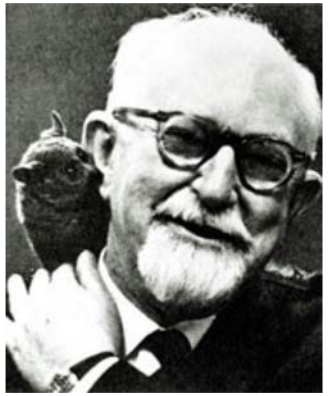
20th Century

Sir Ronald Fisher: He did the math

Sewall Wright: He did the math

J.B.S. Haldane: More math but experiments too

Theodosius Dobzhansky: experiments



George Simpson: Historical biogeography and paleontology

The Molecular Clock

Zuckerlandi and Pauling (1965) suggested the rate of evolution at the molecular level is ***constant through time and among species***.

Kimura (1968) suggested most mutations do not affect fitness (neutral theory), will become fixed through genetic drift.

- The rate at which neutral mutations become fixed is the ***substitution rate***.
- If mutation rates are similar among species, then substitution rates should be constant.

Motoo Kimura. 1983.
The Neutral Theory of Molecular Evolution

The Molecular Clock

Zuckerlandi and Pauling (1965) suggested the rate of evolution at the molecular level is ***constant through time and among species***.

Kimura (1968) suggested most mutations do not affect fitness (neutral theory), will become fixed through genetic drift.

- The rate at which neutral mutations become fixed is the ***substitution rate***.
- If mutation rates are similar among species, then substitution rates should be constant.

Key to the neutral theory: the expected amount of genetic variation in a population can be estimated as $\theta = 4N_e\mu$

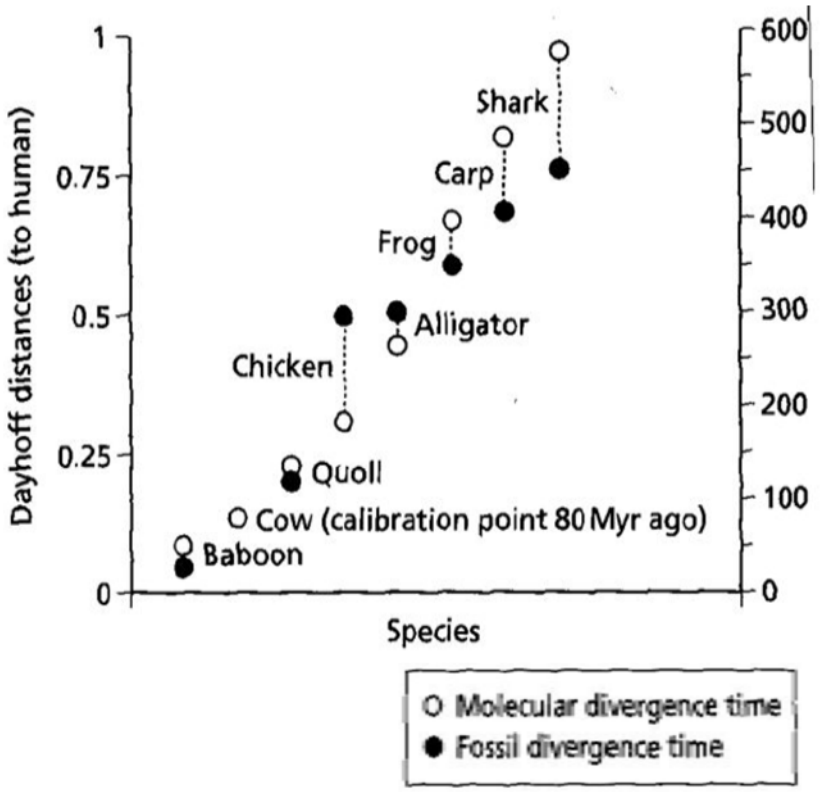
Motoo Kimura. 1983.
The Neutral Theory of Molecular Evolution

Evidence for Neutral Theory

Degeneracy of Genetic Code

	U	C	A	G
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA UCG Ser	UAU Tyr UAC UAA *Stop* UAG *Stop*	UGU Cys UGC UGA *Stop* UGG Trp
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU His CAC CAA Gln CAG	CGU CGC CGA CGG Arg
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA AGG Arg
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU Asp GAC GAA Glu GAG	GGU GGC GGA GGG Gly

Molecular Clock



Zuckerlandi and Pauling 1965

We will debate the merits of the Neutral Theory and how the genomic era has informed interpretations of it later in the course.

Regardless, genomic methods to detect and measure selection often rely on the Neutral Theory as a ***null hypothesis***.

If the null hypothesis of neutrality is ***rejected***, we may be able to infer ***selection*** has acted.

Tests of Neutral Theory

Tajima's D

Difference between avg. pairwise difference and S

Negative selection, balancing selection

Based on data from a single population

HKA Test

Compares divergence in neutral versus functional loci across 2 species

McDonald-Kreitman Test

Compares both within and between species

syn. Vs non-syn. sites, one locus

Based on data from multiple populations or species

Tajima's D

Assuming the neutral theory and an infinite sites model, there are estimators of $\theta = 4N_e\mu$.

One is $\hat{\theta}_{\boldsymbol{S}}$, where ***S*** is the number of segregating sites in a sample of *n* sequences.

Another is $\hat{\theta}_{\boldsymbol{\pi}}$, where π is the average pairwise difference between sequences in the sample.

Under neutrality, $\theta = \pi =$ ***S***, when ***S*** is corrected for the number of sequences and their lengths.

Fumio Tajima

"Fumio Tajima and the Origin of Modern Population Genetics". R. Nielsen, 2016. *Genetics*

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 123(3):585-95

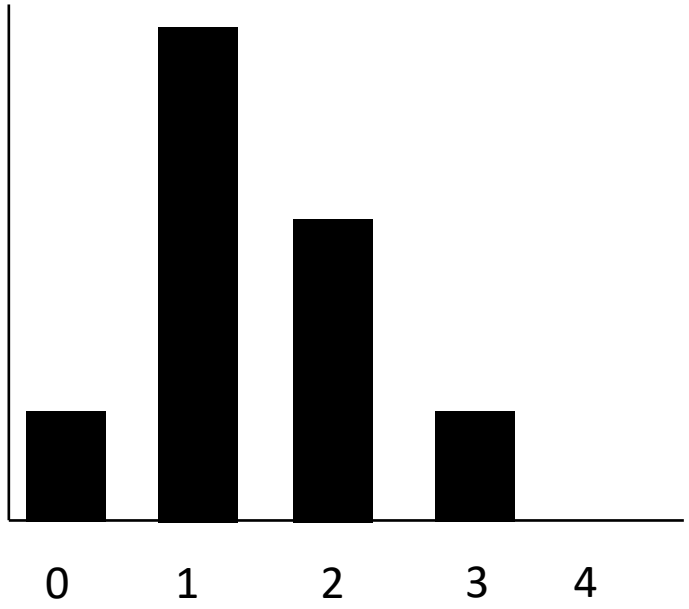
Tajima's D in a *neutrally evolving population*

1	A	G	A	T	C	G	C	T	**G**	C	**A**	A	**T**
2	A	G	A	T	C	G	C	T	**T**	C	**A**	A	**T**
3	A	G	A	T	C	G	C	T	**T**	C	**A**	A	**T**
4	A	G	A	T	C	G	C	T	**T**	C	**G**	A	**T**
5	A	G	A	T	C	G	C	T	**T**	C	**G**	A	**G**

# of segregating sites = 3

Distribution of pairwise differences

π = roughly between one and two



D is the normalized difference between $\hat{\theta}_{\boldsymbol{\pi}}$ and $\hat{\theta}_{\mathbf{s}}$

In this case the observed variation (π) is similar to what is expected (θ)

$$\hat{\theta}_{\boldsymbol{\pi}} - \hat{\theta}_{\mathbf{s}} = between\ 0\ and\ 1$$

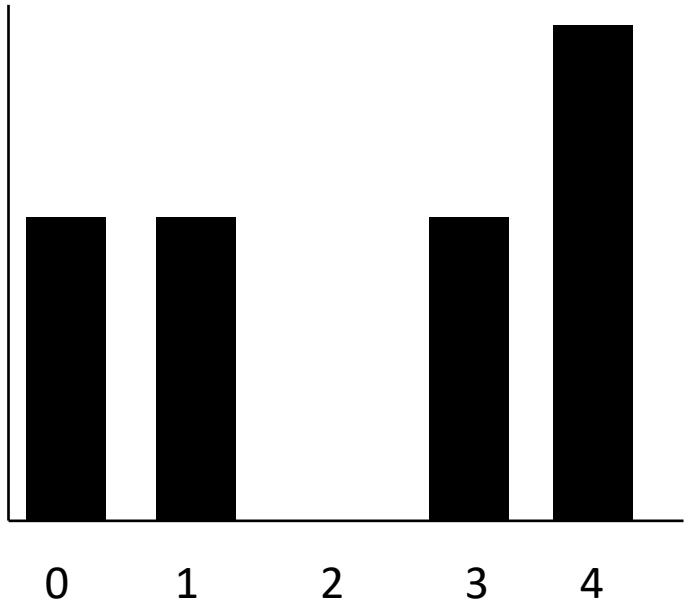
Tajima's D: balancing selection or recent bottleneck

1	A	**G**	**A**	T	C	G	C	T	**T**	C	G	A	**T**
2	A	**G**	**A**	T	C	G	C	T	**T**	C	G	A	**G**
3	A	**G**	**A**	T	C	G	C	T	**T**	C	G	A	**G**
4	A	**C**	**C**	T	C	G	C	T	**A**	C	G	A	**T**
5	A	**C**	**C**	T	C	G	C	T	**A**	C	G	A	**T**

# of segregating sites = 4

Distribution of pairwise differences

Bimodal - sequences are either completely different or belong in one of two clusters



D is the normalized difference between $\hat{\theta}_{\boldsymbol{\pi}}$ and$\hat{\theta}_{\mathbf{s}}$

The observed variation (π) is too high, relative to what is expected (θ)

$$\hat{\theta}_{\boldsymbol{\pi}} - \hat{\theta}_{\mathbf{s}} >> 1$$

Tajima's D is ***positive.***

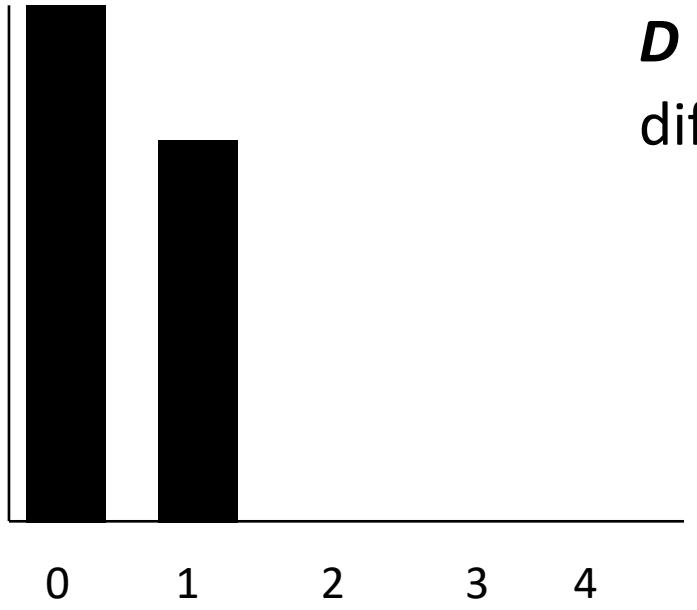
Tajima's D: purifying selection or population expansion

1	A	G	A	T	C	G	C	T	T	C	G	A	**G**
2	A	G	A	T	C	G	C	T	T	C	G	A	**T**
3	A	G	A	T	C	G	C	T	T	C	G	A	**T**
4	A	G	A	T	C	G	C	T	T	C	G	A	**T**
5	A	G	A	T	C	G	C	T	T	C	G	A	**T**

# of segregating sites = 1

Distribution of pairwise differences

Most individuals are identical or differ by only 1



D is the normalized difference between $\hat{\theta}_{\boldsymbol{\pi}}$ and$\hat{\theta}_{\mathbf{s}}$

$$\hat{\theta}_{\boldsymbol{\pi}} - \hat{\theta}_{\mathbf{s}} < 0$$

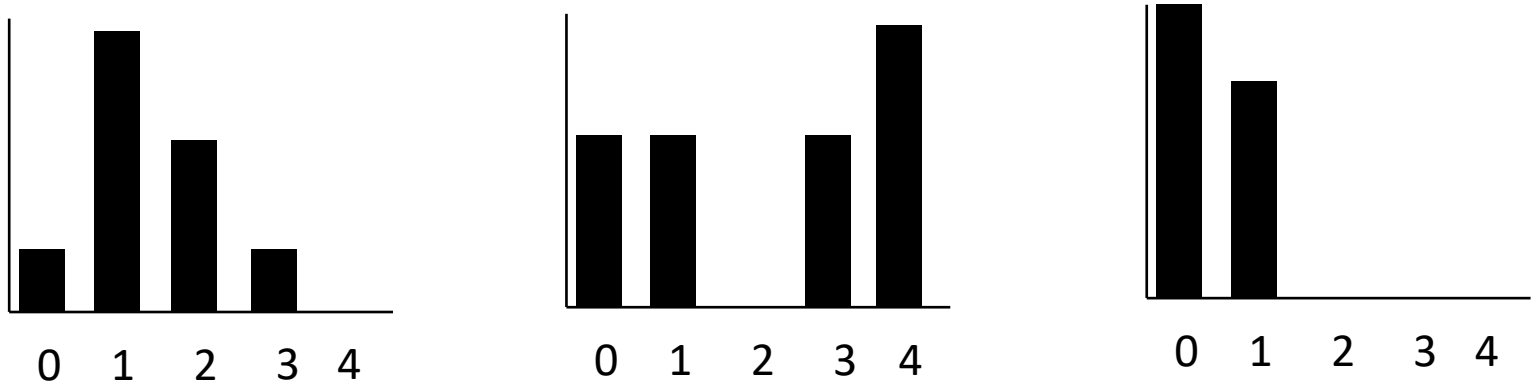
Tajima's D is ***negative.***

Genealogies Under Different Forms of Selection

Aurélien Tellier

Neutral	Balancing selection	Recent sweep
$D = 0$	$D > 0$	$D < 0$
• variety of branch lengths • "balanced" tree	• two clusters separated by long branches • low diversity within clusters	• "star" phylogeny • very short internal branches

Site frequency spectrum:



McDonald-Kreitman Test

Focuses on the number of ***nonsynonymous*** differences per nonsynonymous site (D_N) and the number of ***synonymous*** differences per synonymous site (D_S).

Adaptive differences between species should be ***fixed.***

Non-adaptive differences between species are more likely to be ***polymorphic.***

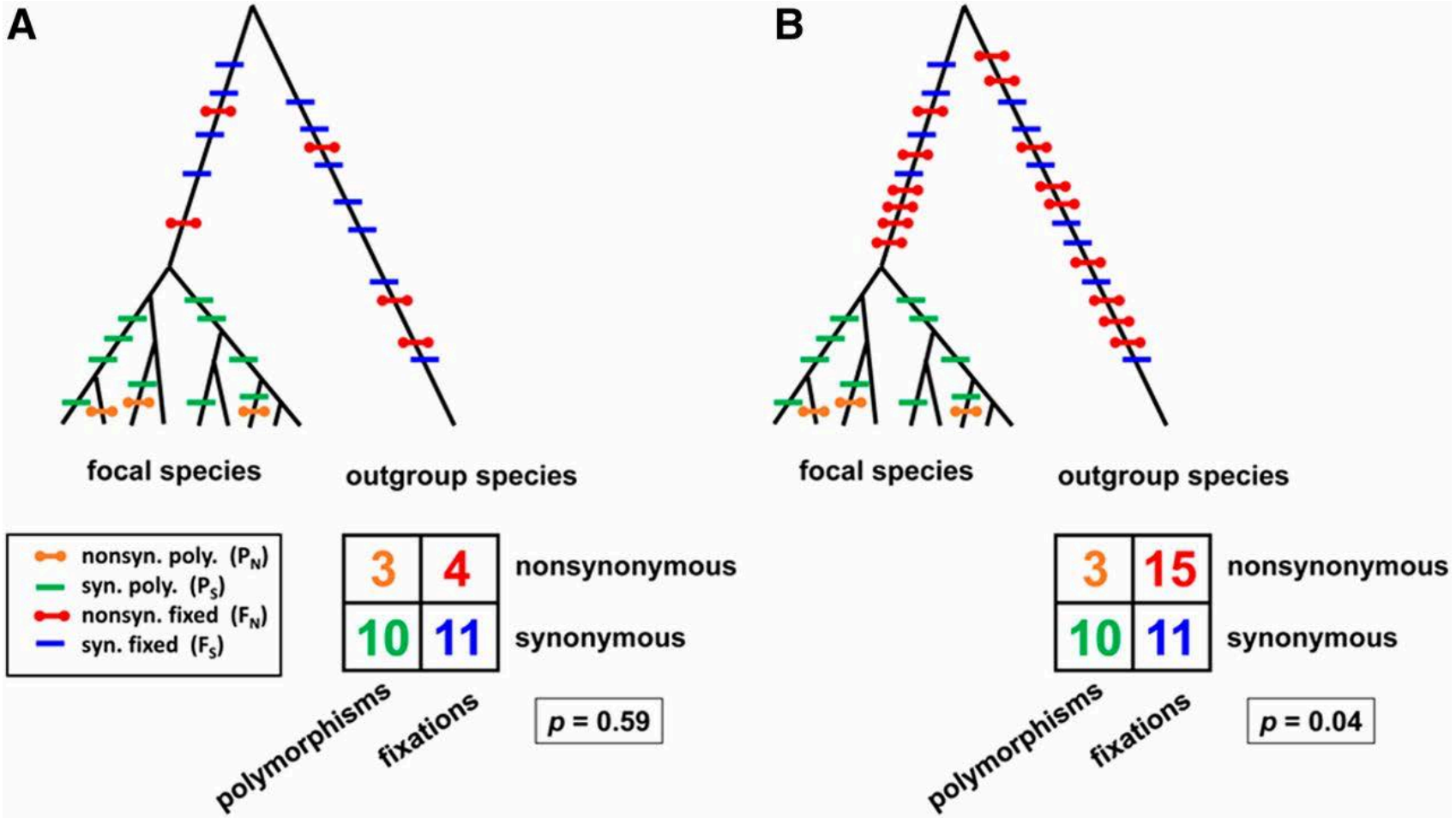
Four counts are needed:

1. The number of nonsynonymous fixed differences
2. The number of synonymous fixed differences
3. The number of nonsynonymous polymorphisms
4. The number of synonymous polymorphisms

	Fixed	Polymorphic
Nonsynonymous	D_N	P_N
Synonymous	D_S	P_S

McDonald and Kreitman. 1991. Adaptive protein evolution at the Adh locus in *Drosophila. Nature*.

McDonald-Kreitman Test



Lazzaro, 2018. Detecting Adaptation with Genome-Scale Molecular Evolutionary Analysis: An Educational Primer for Use with "RNA Interference Pathways Display High Rates of Adaptive Protein Evolution in Multiple Invertebrates. *Genetics.*

Inferring Positive Selection on Protein Coding Genes

d_N	d_S
Nonsynonymous substitutions	Synonymous substitutions
Result in amino acid change	No amino acid change

The ratio of nonsynonymous substitutions per synonymous site and synonymous substitutions per synonymous site is d_N/d_S.

$d_N/d_S = \omega$

if $\omega < 1$; purifying selection

If $\omega = 1$; neutral evolution

If $\omega > 1$; positive selection

If $d_N/d_S > 1$: overabundance of amino acid changes → changes in protein structure → functional relevance → adaptive change

Positive selection.

Likelihood Methods for Inferring Positive Selection

The basic model is a codon substitution model.

This model accounts for:

1. The genetic code
2. Transition/transversion rate differences
3. Different base frequencies at codon positions.

q_{ij} is the instantaneous substitution rate from codon *i* to codon *j*

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

Yang, 1998. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Mol Biol Evol.*

- π_j is the equilibrium frequency of codon *j*
- κ is the transition/transversion bias
- ω is the relative probability of observing nonsynonymous substitutions relative to synonymous substitutions (d_N/d_S)

Likelihood Ratio Tests for Positive Selection

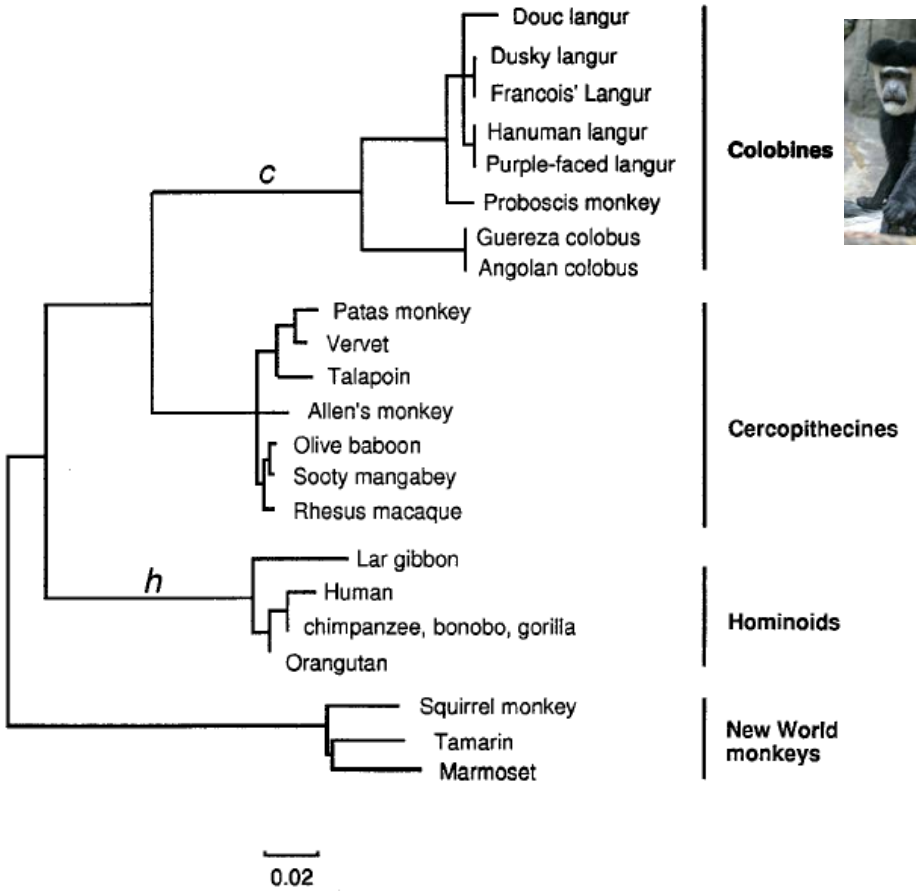
Compare models allowing different levels of heterogeneity in d_N/d_S among lineages in a phylogeny.

1. "one ratio" model (simplest)
2. "free ratio" model (most general)
3. "two ratio" model: ***foreground*** branches(*h* and *c*) versus ***background*** branches (0)
4. "three ratio" model: *h* branch, *c* branch, and background (0)

Log Likelihood Values and Parameter Estimates Under Different Models

	Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_H$	$\hat{\omega}_C$
	Large data set (n = 19)						
Testing for different d_N/d_S among lineages	A. One ratio: $\omega_0 = \omega_H = \omega_C$	35	−1043.84	4.157	0.574	$=\hat{\omega}_0$	$=\hat{\omega}_0$
	B. Two ratios: $\omega_0 = \omega_H, \omega_C$	36	−1041.70	4.163	0.489	$=\hat{\omega}_0$	3.383
	C. Two ratios: $\omega_0 = \omega_C, \omega_H$	36	−1039.92	4.186	0.484	∞	$=\hat{\omega}_0$
	D. Two ratios: $\omega_0, \omega_H = \omega_C$	36	−1037.59	4.199	0.392	7.166	$=\hat{\omega}_H$
	E. Three ratios: $\omega_0, \omega_H, \omega_C$	37	−1037.04	4.196	0.392	∞	3.516
Testing that d_N/d_S > 1 in foreground branches	F. Two ratios: $\omega_0 = \omega_H, \omega_C = 1$	35	−1042.50	4.074	0.488	$=\hat{\omega}_0$	1
	G. Two ratios: $\omega_0 = \omega_C, \omega_H = 1$	35	−1042.29	4.058	0.484	1	$=\hat{\omega}_0$
	H. Two ratios: $\omega_0, \omega_H = \omega_C = 1$	35	−1040.32	3.974	0.392	1	1
	I. Three ratios: $\omega_0, \omega_H, \omega_C = 1$	36	−1037.92	4.101	0.392	∞	1
	J. Three ratios: $\omega_0, \omega_H = 1, \omega_C$	36	−1039.49	4.063	0.392	1	3.448

NOTE.—p, number of parameters in the model not including the nine parameters for codon frequencies (π_j's in eq. 1). Parameters ω_H, ω_C, and ω_0 are the d_N/d_S ratios for branches *h*, *c*, and all other branches, respectively (see figs. 1 and 2). Estimates of branch lengths are not shown.



Yang, 1998. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Mol Biol Evol.*

Likelihood Ratio Tests for Positive Selection

Using the ***likelihood ratio test***, these models can be compared to examine hypotheses.

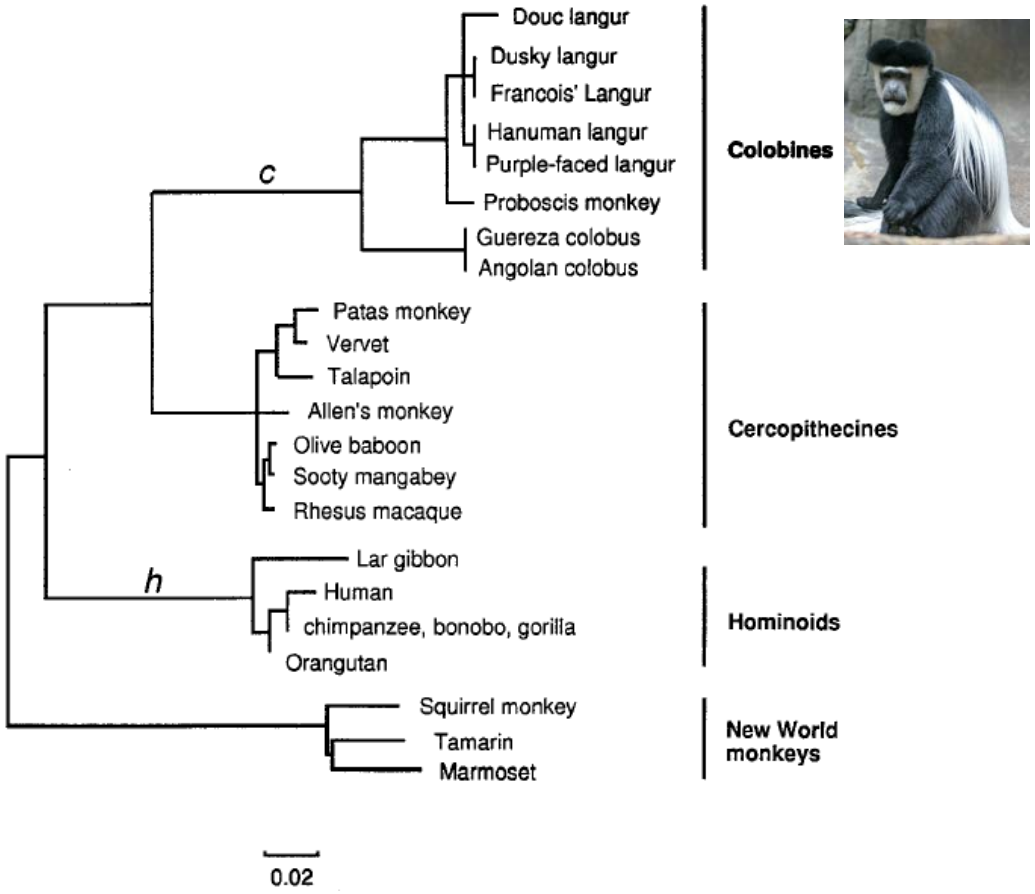
1. Comparing the one-ratio model to the free-ratio model can test if there are different d_N/d_S ratios among lineages.
2. Comparing the one-ratio model and the two-ratio model can test whether some lineages have different d_N/d_S from other lineages.
3. Comparing iterations of the two-ratio model with and without the constraint $\omega_1 \leq 1$ can test whether the branches of interest are evolving under positive selection.
4. We can also test whether the foreground branches have a d_N/d_S greater than one.

Likelihood Ratio = 2(lnL_1 – lnL_2)

	Null Hypothesis Tested	Assumption Made	Models Compared	Large Data Set ($n = 19$)
Testing for different d_N/d_S among lineages	A. $(\omega_H = \omega_C) = \omega_0$	$\omega_H = \omega_C$	A and D	12.50**
	B. $\omega_C = \omega_0$	$\omega_H = \omega_0$	A and B	4.28*
	C. $\omega_C = \omega_0$	ω_H free	C and E	5.76*
	D. $\omega_H = \omega_0$	$\omega_C = \omega_0$	A and C	7.84**
	E. $\omega_H = \omega_0$	ω_C free	B and E	9.32**
Testing that d_N/d_S > 1 in foreground branches	A′. $(\omega_H = \omega_C) \leq 1$ ..	$\omega_H = \omega_C$	D and H	5.46*
	B′. $\omega_C \leq 1$	$\omega_H = \omega_0$	B and F	1.60
	C′. $\omega_C \leq 1$	ω_H free	E and I	1.76
	D′. $\omega_H \leq 1$	$\omega_C = \omega_0$	C and G	4.74*
	E′. $\omega_H \leq 1$	ω_C free	E and J	4.90*

* Significant ($P < 5\%$; $\chi_1^2 = 3.84$).
** Extremely significant ($P < 1\%$; $\chi_1^2 = 6.63$).

Critical value follows a chi-squared distribution.



Yang, 1998. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Mol Biol Evol.*

More Power: The Branch-Site Test

Parameters in the Branch-Site Models

Site Class	Proportion	Background ω	Foreground ω
0	p_0	ω_0	ω_0
1	p_1	ω_1	ω_1
2	$p_2 = (1 - p_0 - p_1)p_0/(p_0 + p_1)$	ω_0	ω_2
3	$p_3 = (1 - p_0 - p_1)p_1/(p_0 + p_1)$	ω_1	ω_2

NOTE.—In model A, $\omega_0 = 0$ and $\omega_1 = 1$ are fixed, whereas in model B they are free to vary.

highly constrained sites (negative selection)

weakly constrained sites (neutral)

$\omega_2 > 1$

$\omega_2 > 1$

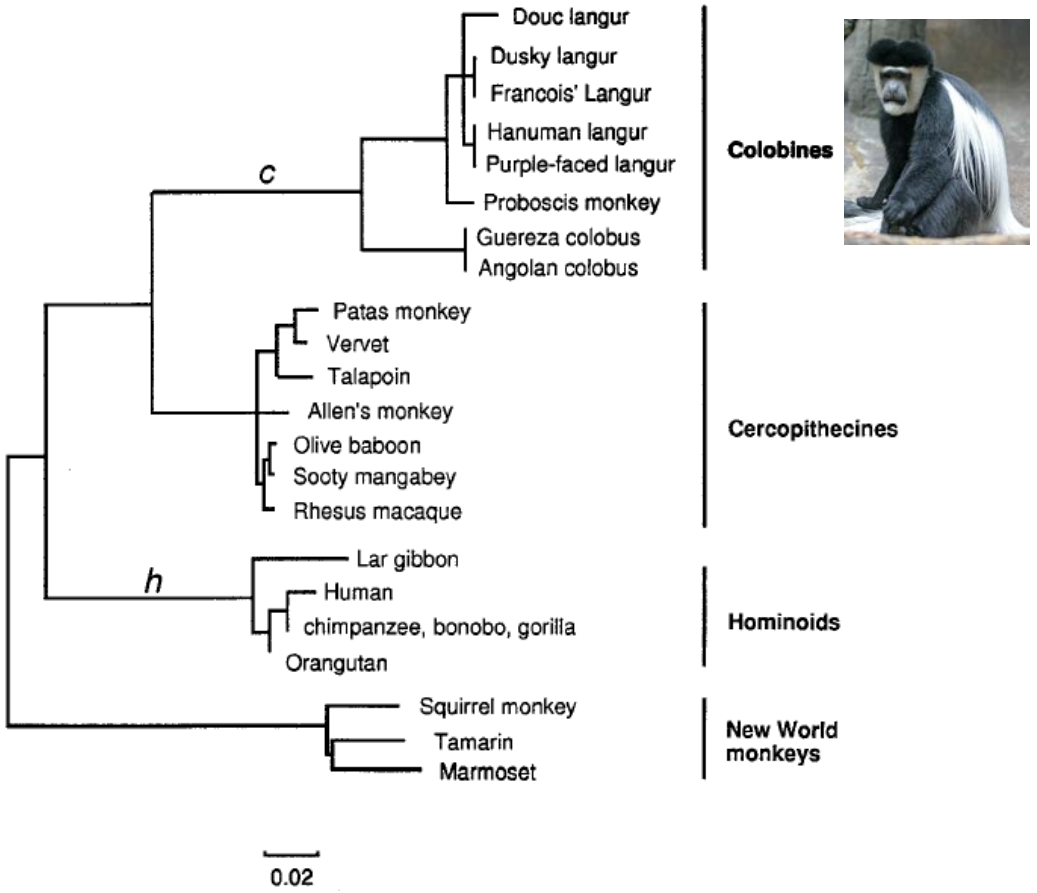
for both of these classes, some sites have come under positive selection.

Parameter Estimates for the Lysozyme Data

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0: one-ratio	1	−1,043.83	$\hat{\omega} = 0.574$	None
Branch-specific models (Model B in table 1 of Yang 1998)				
Two-ratios.....	2	−1,041.70	$\hat{\omega}_0 = 0.489$, $\hat{\omega}_0 = \mathbf{3.383}$	N/A
Site-specific models				
M1: neutral.....	1	−1,037.21	$\hat{p}_0 = 0.502$ ($\hat{p}_1 = 0.498$)	Not allowed
M2: selection	3	−1,035.83	$\hat{p}_0 = 0.498$, $\hat{p}_1 = 0.430$ ($\hat{p}_2 = \mathbf{0.072}$) $\hat{\omega}_2 = \mathbf{3.710}$	15L, 17M, 37G, 41R, 50R, 101R (at $0.5 < P < 0.8$)
M3: discrete ($K = 2$) ..	3	−1,035.23	$\hat{p}_0 = 0.823$ ($\hat{p}_1 = \mathbf{0.177}$) $\hat{\omega}_0 = 0.237$, $\hat{\omega}_1 = \mathbf{2.629}$	37G, 41R (at $P > 0.99$) 15L, 50R, 101R, 114N (at $P > 0.95$)
M3: discrete ($K = 3$) ..	5	Same at $K = 2$		
M7: beta	2	1,037.21	$\hat{p} = 0.011$, $\hat{q} = 0.011$	Not allowed
M8: beta&ω	4	1,035.56	$\hat{p}_0 = 0.788$, $\hat{p} = 99.65$, $\hat{q} = 298$ $\hat{p}_1 = \mathbf{0.212}$, $\hat{\omega} = \mathbf{2.538}$	37G, 41R (at $P > 0.99$) 15L 17M 50R 101R 114N (at $P > 0.95$)
Branch-site models				
Model A	3	−1,035.53	$\hat{p}_0 = 0.327$, $\hat{p}_1 = 0.269$ ($\hat{p}_2 + \hat{p}_3 = \mathbf{0.404}$) $\hat{\omega}_2 = \mathbf{4.809}$	Site for foreground lineage: 14R 21R 23I 87D (at $P > 0.9$) 41R 50R 126Q (at $P > 0.7$)
Model B	5	−1,034.27	$\hat{p}_0 = 0.611$, $\hat{p}_1 = \mathbf{0.157}$ ($\hat{p}_2 + \hat{p}_3 = \mathbf{0.232}$) $\hat{\omega}_0 = 0.166$, $\hat{\omega}_1 = \mathbf{2.319}$, $\hat{\omega}_2 = \mathbf{4.322}$	Sites for background ω_1: 15L 17M 37G 82S 101R 114N 125V ($0.7 < P < 0.8$) Sites for foreground ω_2: 14R 21R 23I 87D ($0.7 < P < 0.85$)

NOTE.—p is the number of free parameters for the ω ratios. Parameters indicating positive selection are presented in boldtype. Those in parentheses are presented for clarity only but are not free parameters; for example, under M8 (beta&ω), $p_1 = 1 - p_0$. Sites potentially under positive selection are identified using the human lysozyme sequence as the reference. Estimates of κ range from 4.1 to 4.6 among models.

Can also estimate the probability that a site has evolved under positive selection.



Yang and Nielsen, 2002. Codon Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. ***Mol Biol Evol.***

Assumptions and Pitfalls of the Branch-Site Test

Problem	Solution	Citation
Assumes that d_N is fixed between but not within species	best to stick with inter-species comparisons	Yang, 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*
Saturation at long evolutionary timescales	test for saturation; stick to comparisons within ~100 MY divergence	Gharib and Robinson-Rechavi 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol.*
Statistical power increases with taxonomic sampling	plan to sample n ≥ 7 species	Anisimova et al. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol.*
Species tree ambiguity can force erroneous changes along branches	use gene trees not species trees as input	Yang, 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*
Alignment errors are not taken into account	ignore results from “gappy” regions of alignments	Anisimova et al. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.*
Trinucleotide substitutions contribute to false positives	manual curation of results to ensure putative positively selected sites do not occur in tandem	Venkat et al. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol.*
False positives are more common in large numbers of statistical tests (such as all the annotation orthologs in a genome)	P-value correction for multiple testing such as Bonferroni, false discovery rate	Anisimova and Yang, 2007. Multiple hypothesis testing to detect lineages under positive selection. *Mol Biol Evol.*

Tools for Gene, Branch, and Site Models of Positive Selection

PAML: Phylogenetic Analysis using Maximum Likelihood

Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591 (http://abacus.gene.ucl.ac.uk/software/paml.html).

hyphy.org