

Detecting positive selection with Markov models of codon substitution

Maria Anisimova

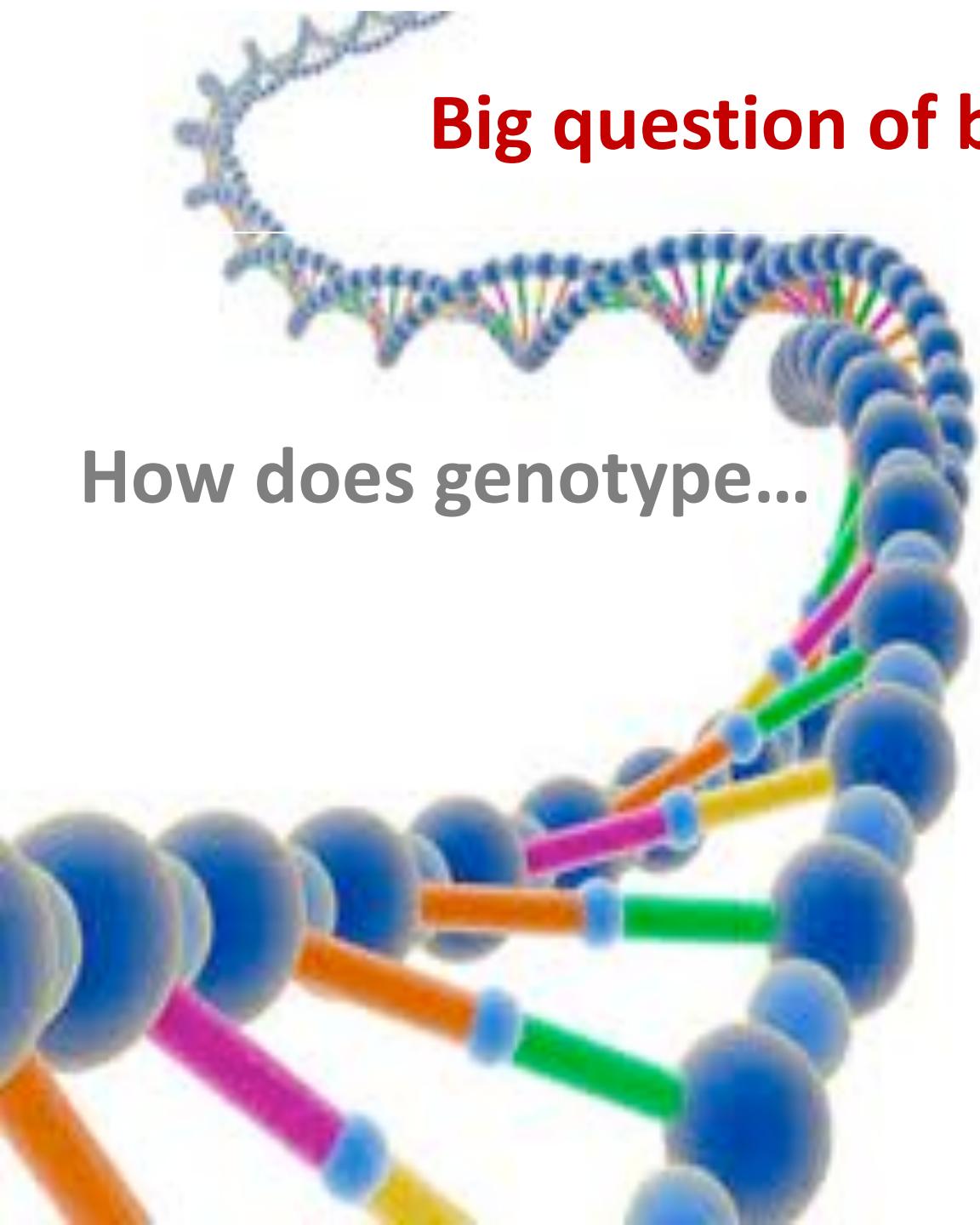
Applied Computational Genomics Team – ACGT
Institute of Applied Simulation
Zürich University of Applied Sciences

maria.anisimova@zhaw.ch

Zürcher Hochschule
für Angewandte Wissenschaften

Why study natural selection

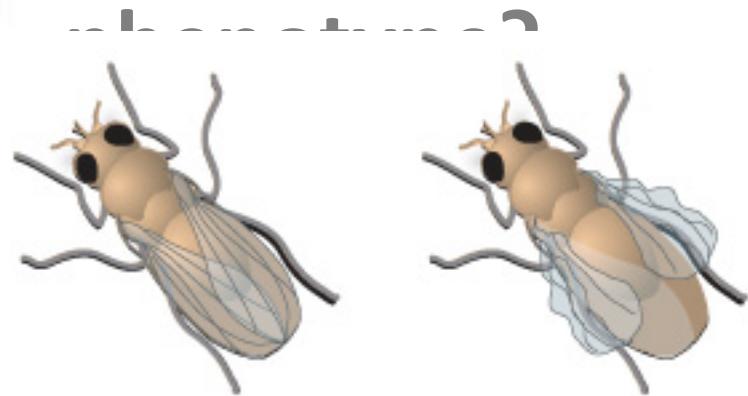




Big question of biology

How does genotype...

... shape



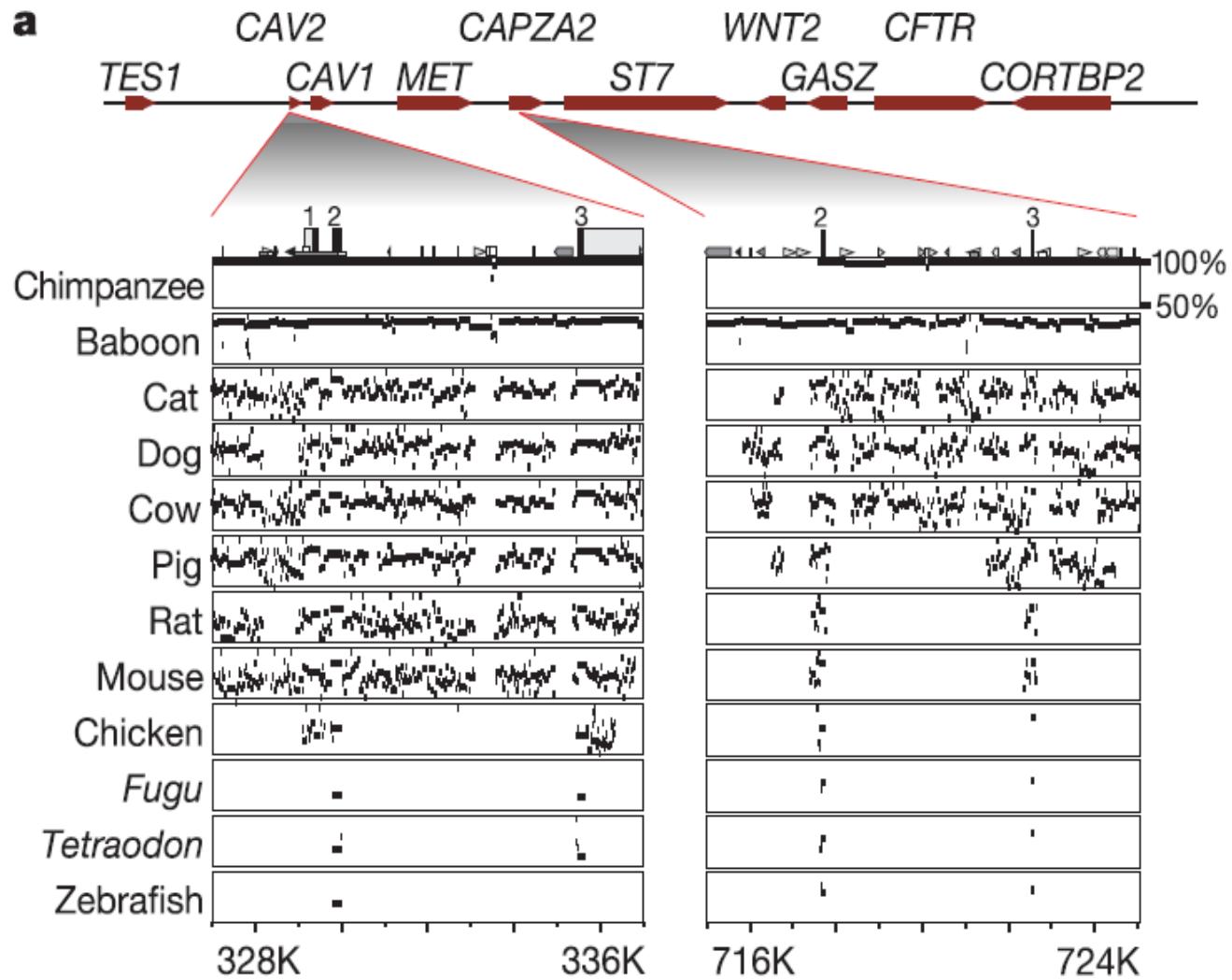
Normal Wings

Wrinkled Wings

Evolutionary conservation means function

Genomic regions conserved across diverse species most likely have some functional significance

Conservation
↓
function



Percentage identity when human is aligned with another species.
Close species are effective in identifying regulatory elements while distant species
are effective in identifying coding regions.

High variability may also mean functional significance, if the variability is driven by selection

Evolutionary biologists are more interested in positive selection because fixations of advantageous mutations in the genes or genomes are responsible for evolutionary innovations and species divergences.

There are two main explanations for genetic variation observed within a population or between species:

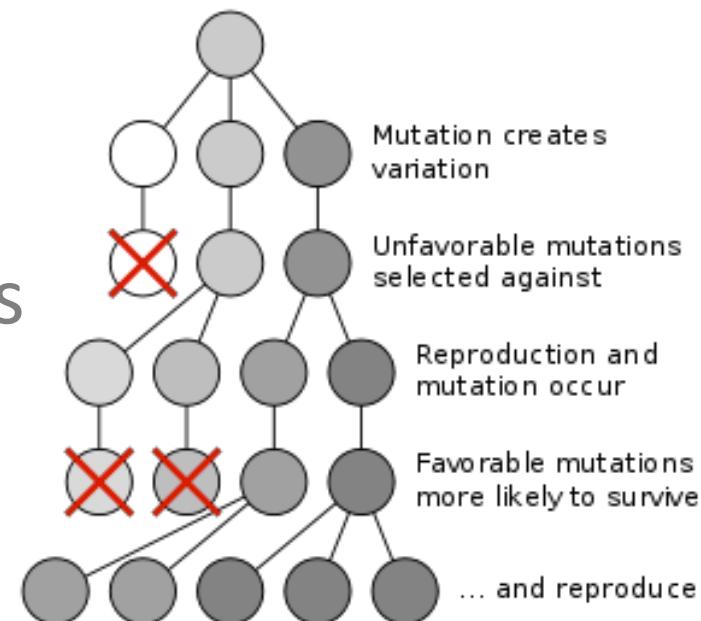
Natural selection (survival of the fittest)
Mutation and drift (survival of the luckiest)

Gillespie, J.H. 1998. *Population genetics: a concise guide*. John Hopkins University Press, Baltimore.

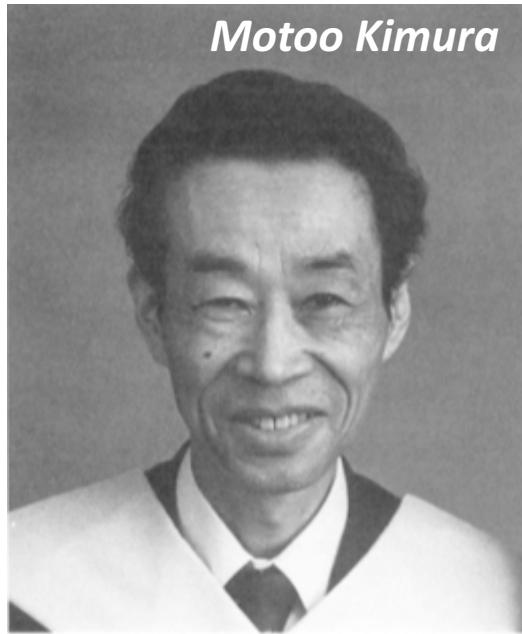
Hartl, D.L., and A.G. Clark. 1997. *Principles of population genetics*. Sinauer Associates, Sunderland, Massachusetts.

The neo-Darwinian theory of evolution

- Natural selection shapes the genetic makeup
- Most mutations are *deleterious*, removed by purifying selection
- Substitutions ≠ polymorphisms
- Substitutions are acquired by *positive selection*
- Polymorphisms are kept by *balancing selection*



The neutral theory of molecular evolution



Motoo Kimura



- Most mutations are deleterious
- Most changes: random fixation of neutral mutations
- The fate of alleles is determined by random genetic drift
- Substitution rate = neutral mutation rate (molecular clock)
- Selection may operate; but is too weak to influence
- Substitution = polymorphism
- Morphological traits evolve by natural selection

The impact of the neutral theory

- The neutral theory makes simple and testable predictions about what we should observe: provided *a falsifiable null hypothesis*
- Strengthened the connection between molecular biology and population genetics
- Availability of such null hypothesis prompted the development of neutrality tests

s = selection coefficient

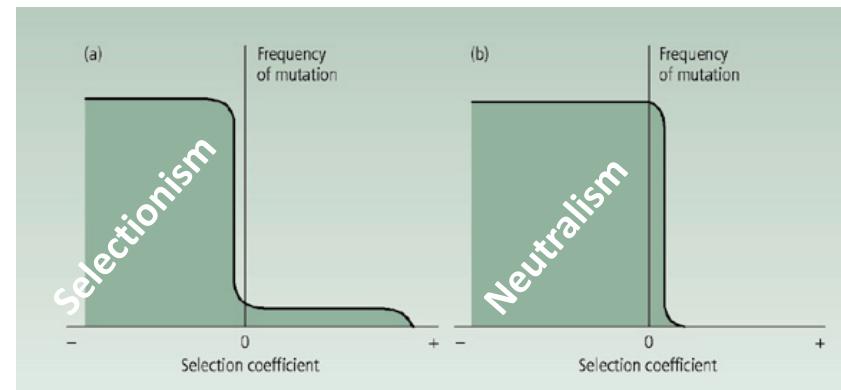
s describes relative fitness of mutant a vs. wild-type A .

Genotype fitness:

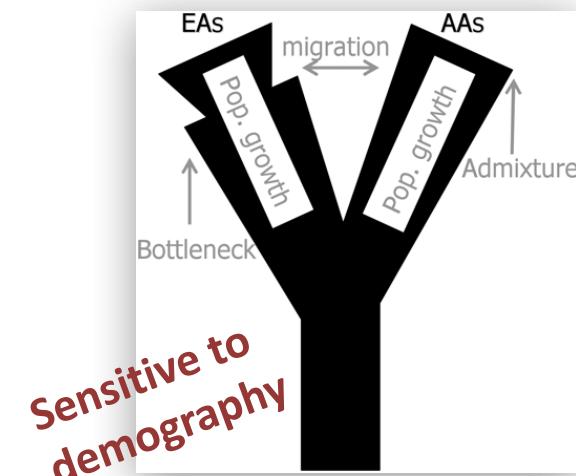
1 for AA , $1+s$ for Aa , $1+2s$ for aa

$s > 0$ positive selection

$s < 0$ negative selection



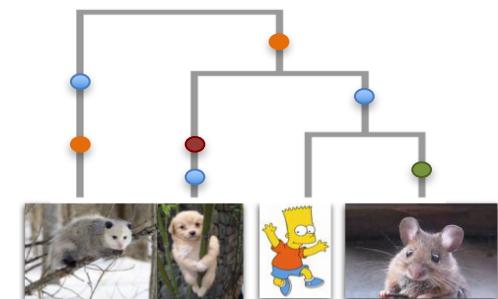
Neutrality and selection tests



- Mutational frequency spectrum (eg, Tajima's D, Tajima 1989)
- Population subdivision
- LD & haplotype structure
- Within/between species variability (HKA test, Hudson, Kreitman, Aguade 1987)

Account for codon structure:

- Within/between species variability (MK test, McDonald-Kreitman 1991)
- Based on codon models



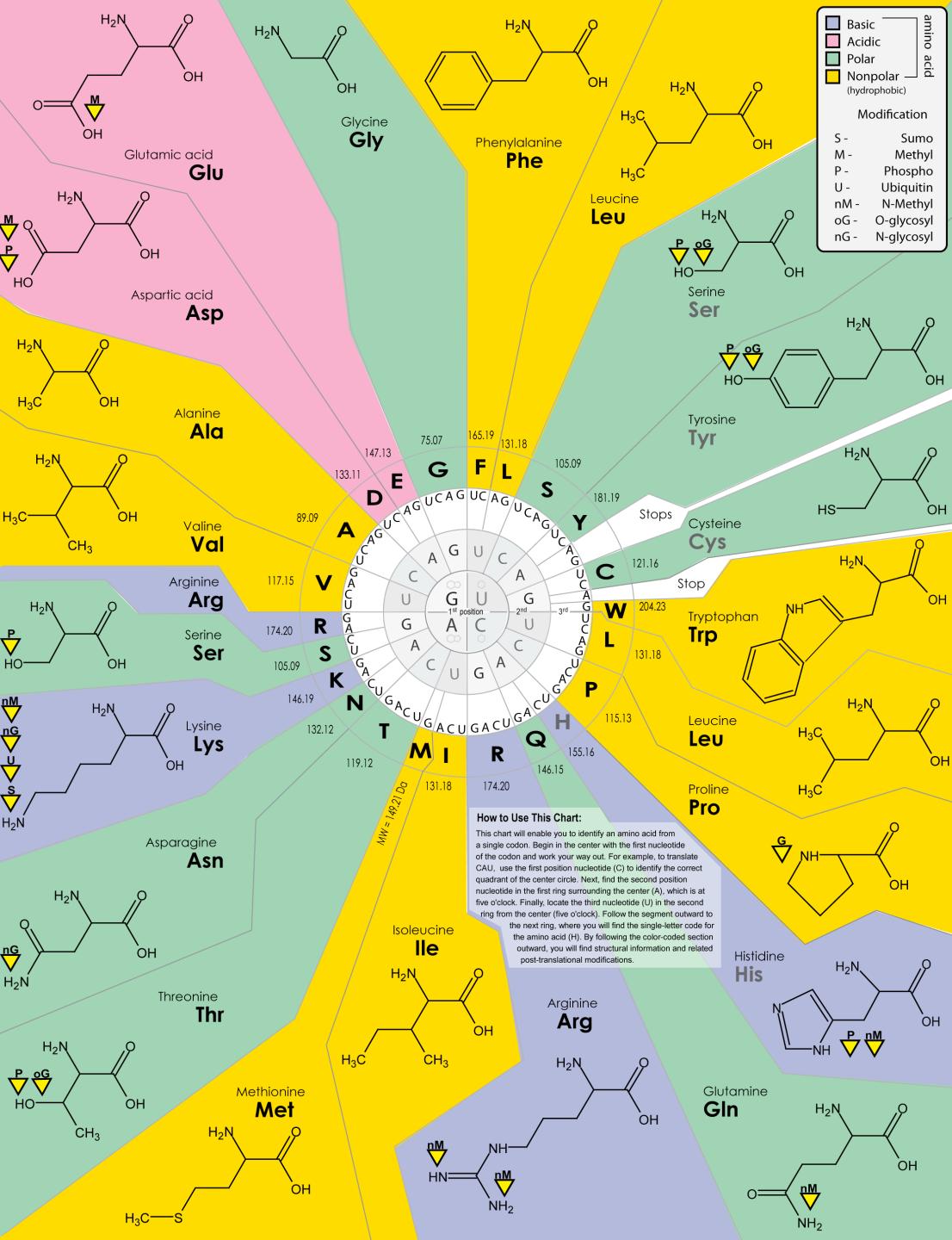
Standard genetic code

The genetic code determines how random changes to the gene brought about by the process of mutation will impact the function of the encoded protein

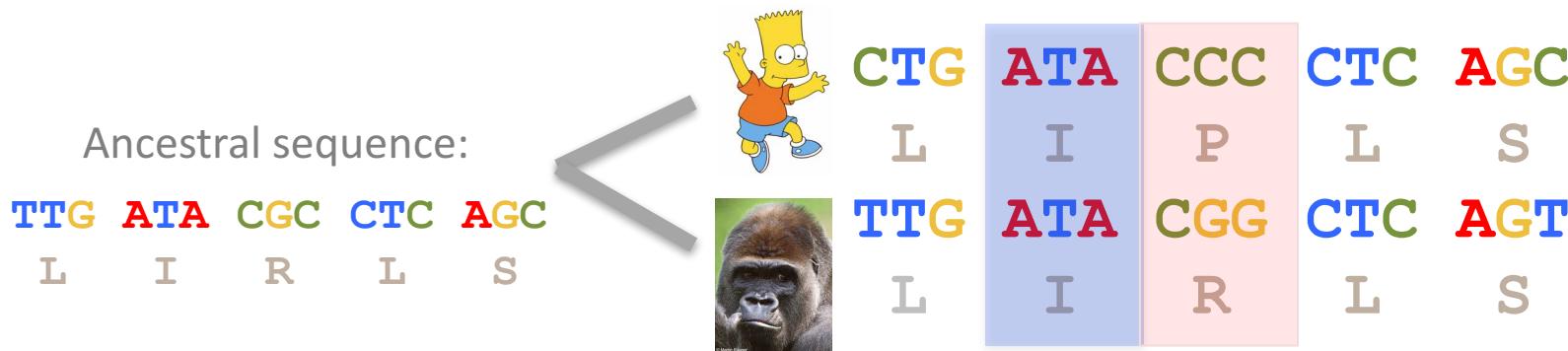
Types of codon changes

Synonymous (silent):
 TTC (Phe) → TTT (Phe)

Nonsynonymous:
 TTC (Phe) → TTA (Leu)



Measuring selection on the protein



Synonymous rate = d_s (also K_s)

Nonsynonymous rate = d_N (also K_A)

$$\omega = d_N/d_s$$

$\omega > 1$ Positive selection

$\omega = 1$ Neutral evolution

$\omega < 1$ Negative selection

Why not counts but rates?

Example:

Pairwise alignment of 500 codons

Observed differences:

5 synonymous differences

5 nonsynonymous differences

Conclusion: Neutral evolution?

Hint: Need to know how many sites are synonymous and how many are nonsynonymous

Evolution at the three codon positions

Relative proportion of different types of mutations in hypothetical protein coding sequence.				
Type	Expected number of changes (proportion)			
	All 3 Positions	1 st positions	2 nd positions	3 rd positions
Total mutations	549 (100)	183 (100)	183 (100)	183 (100)
Synonymous	134 (25)	8 (4)	0 (0)	126 (69)
Nonsynonymous	392 (71)	166 (91)	176 (96)	57 (27)
nonsense	23 (4)	9 (5)	7 (4)	7 (4)

Modified from Li and Graur (1991). Note that we assume a hypothetical model where all codons are used equally and that all types of point mutations are equally likely.

Note: by framing the counting of sites in this way we are using a “mutational opportunity” definition of the sites. Not everyone agrees that this is the best approach. For an alternative view see **Bierne and Eyre-Walker 2003 Genetics 168:1587-1597**.

Why not counts but rates?

Example:

Pairwise alignment of 500 codons (or 3x500 nt)

5 syn. differences, 25.5% syn. sites:

$$S = 500 \times 3 \times 25.5\% = 382.5, \text{ so } d_S = 5/382.5 = 0.013$$

5 nonsyn. differences, 74.5% nonsyn. sites:

$$N = 500 \times 3 \times 74.5\% = 1117.5, \text{ so } d_N = 5/1117.5 = 0.0045$$

$$d_N/d_S = 0.0045/0.013 = 0.35 < 1$$

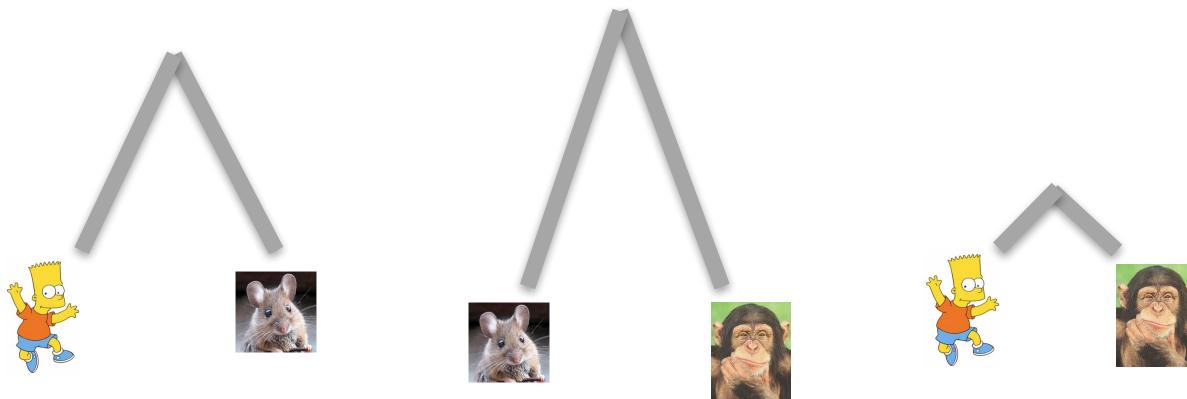
Conclusion: Purifying selection

Pairwise estimation of dN and dS

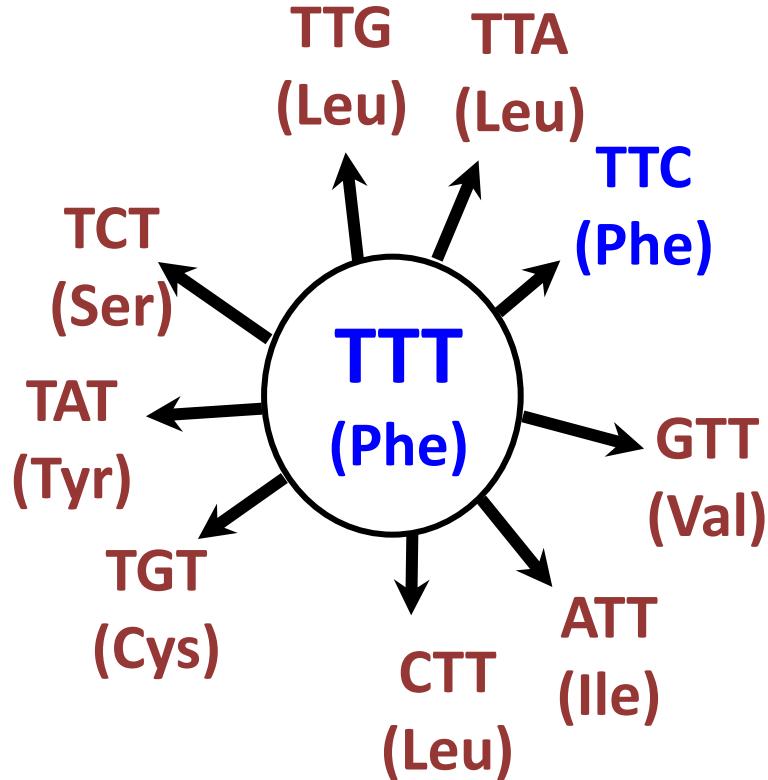
1. Count synonymous and nonsynonymous sites (S and N)
2. Count synonymous and nonsynonymous *differences*
3. Calculate the proportion of differences, then d_N and d_S
4. Correct for *multiple hits*



CTG	ATA	CCC	CTC	AGC
TTA	ATA	CCC	CTC	AGC
CTG	ATA	TGT	CTA	GGA



Counting sites (S and N)



1/3 synonymous sites

8/3 nonsynonymous
sites

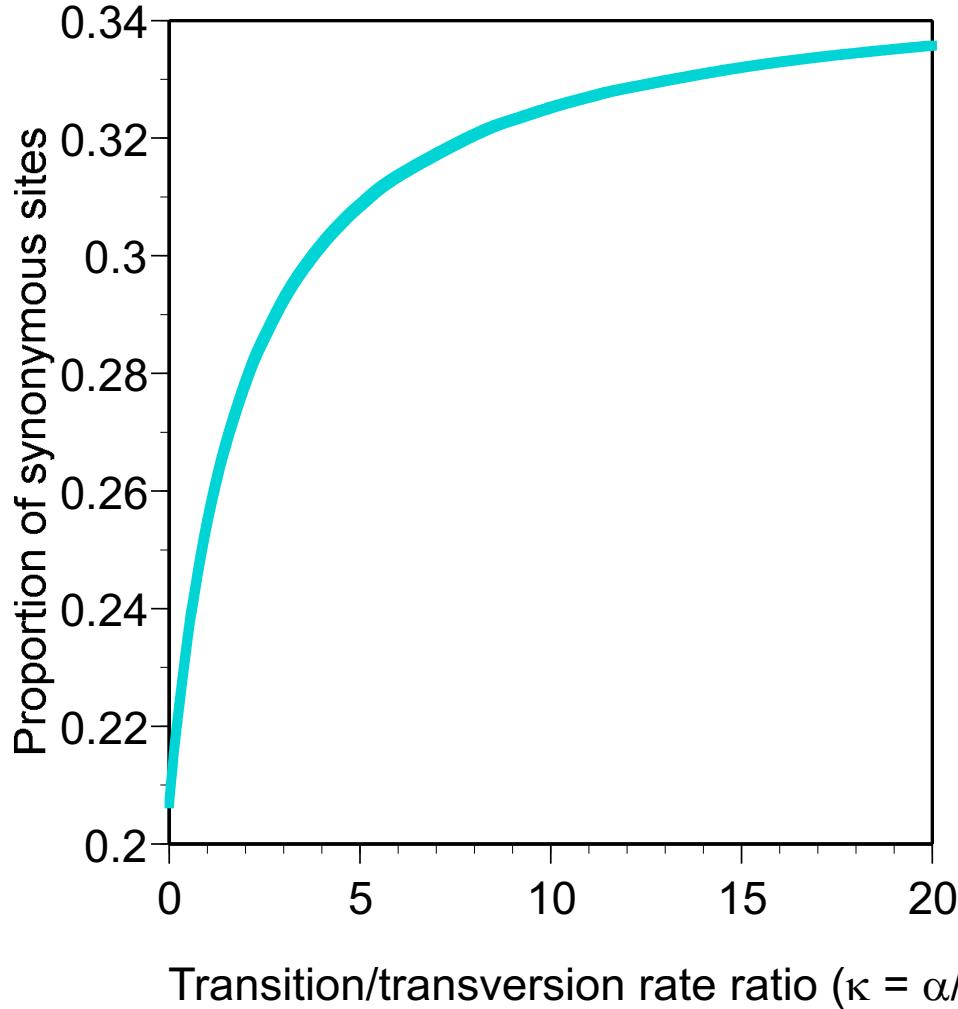
Sites are defined as mutational opportunities

Counting differences

How many differences between CCT and CAG?

Pathways between CCT and CAG	Syn	Nonsyn
CCT (Pro) \leftrightarrow CAT (His) \leftrightarrow CAG (Gln)	0	2
CCT (Pro) \leftrightarrow CCG (Pro) \leftrightarrow CAG (Gln)	1	1
Average	0.5	1.5

The impact of κ



At 3d positions,
transitions are
more likely to be
synonymous than
transversions

Codon usage bias

Analysis of real genes suggests that codon usage bias leads to reduced number of synonymous sites

(the opposite effect to the κ bias)

Correcting for multiple hits

Ad hoc correction using DNA models, which assume that a nonsynonymous site has equal rate of changing into 3 other nonsynonymous nucleotides (Lewontin 1989).

Numerous counting methods of increasing sophistication

1. Perler, F. et al. 1980. *Cell* 20: 555-566
2. Miyata, T. & T. Yasunaga. 1980. *JME* 16:23-36
3. Li, W.-H., C.-I. Wu, & C.-C. Luo. 1985. *MBE* 2:150-174
4. Nei, M. & T. Gojobori. 1986. *MBE* 3: 418-426
5. Li, W.-H. 1993. *JME* 36:96-99
6. Pamilo & Bianchi 1993 *MBE* 10:271-281
7. Ina, Y. 1995. *JME* 40:190-226
8. Comeran, J. M. 1995. *JME* 41:1152-1159
9. Moriyama, E. N. & F. R. Powell, 1997. *JME* 45:378-391
10. Yang, Z., and R. Nielsen. 2000. *MBE* 17:32-43.

- no ts/tv bias + no codon bias
- ts/tv bias + no codon bias
- ts/tv bias + codon bias

Human & orangutan α 2-globin genes: 142 codons

Method/Model	κ	S	N	d_N	d_S	d_N/d_S
NG86	1	109.4	316.6	0.0095	0.0569	0.168
Ina95	2.1	119.3	299.9	0.0101	0.0523	0.193
YN00	6.1	61.7	367.3	0.0083	0.1065	0.078
ML (GY94)						
(1) ML F _{equal} , $\kappa = 1$	1	108.5	317.5	0.0093	0.0557	0.167
(2) ML F _{equal} , κ estimated	3.0	124.6	301.4	0.0099	0.0480	0.206
(7) ML F ₆₁ , $\kappa = 1$ fixed	1	58.3	367.7	0.0082	0.1145	0.072
(8) ML F ₆₁ , κ estimated	5.3	55.3	370.7	0.0082	0.1237	0.066

Base frequencies at 3rd position:
T = 9%, C = 52%, A = 1%, G = 37%
(Yang & Bielawski 2000. *TREE* 15:496–503)

Software

Methods

Counting
methods

NG86

Li93

Comeron 95

YN00

ML methods

GY94

Software

MEGA; codeml & yn00 in
PAML

MEGA, DAMBE, codeml

DIVERGE by Comeron

yn00 in PAML

codeml

Detecting selection based on π

- **From pairwise comparisons**

Best known examples:

Adaptation in primate lysozyme (Messier & Stewart 1997)

Adaptation in human MHC (Hughes & Nei 1988)

- **From MSAs using underlying phylogeny**

° Using ancestral reconstruction and counting at each site

(HA gene from flu, Fitch et al. 1997, Suzuki & Gojobori 1999)

° Markov models of codon evolution detect positive selection

at individual sites in the protein

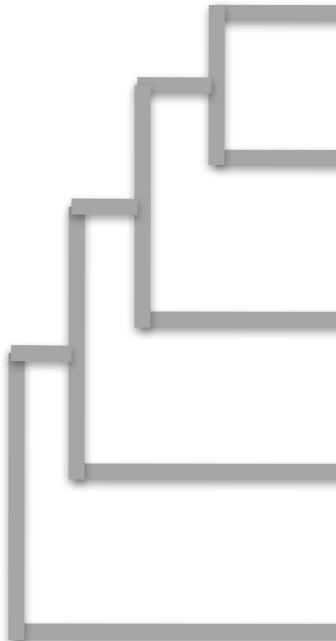
in individual lineages

at individual sites & lineages (episodic selection)

Types of codon substitution models

- **Branch models** to test positive selection on lineages on the tree
(Yang 1998. *Mol. Biol. Evol.* 15:568-573)
- **Site models** to test positive selection affecting individual sites
(Nielsen & Yang. 1998. *Genetics* 148:929-936;
Yang, *et al.* 2000. *Genetics* 155:431-449)
- **Branch-site models** to detect positive selection at a few sites on a particular lineage
(Yang & Nielsen. 2002. *Mol. Biol. Evol.* 19:908-917;
Yang, *et al.* 2005. *Mol. Biol. Evol.* 22:1107-1118)

Measuring selection on the protein



	CTG	ATA	CCC	CTC	AGC
Bart Simpson	L	I	P	L	S
Chimpanzee	TTA	ATA	CCC	CTC	AGC
Gorilla	L	I	P	L	S
Orangutan	TTG	ATA	CGG	CTC	AGT
Mouse	L	I	R	L	S
Human	TTA	ATA	TGG	CTC	AGC
	L	I	W	L	S
	CTG	ATA	TGT	CTA	GGA
	L	I	C	L	G

synonymous rate: d_S nonsynonymous rate: d_N

$\omega = d_N/d_S > 1$ positive selection

$\omega < 1$ negative selection

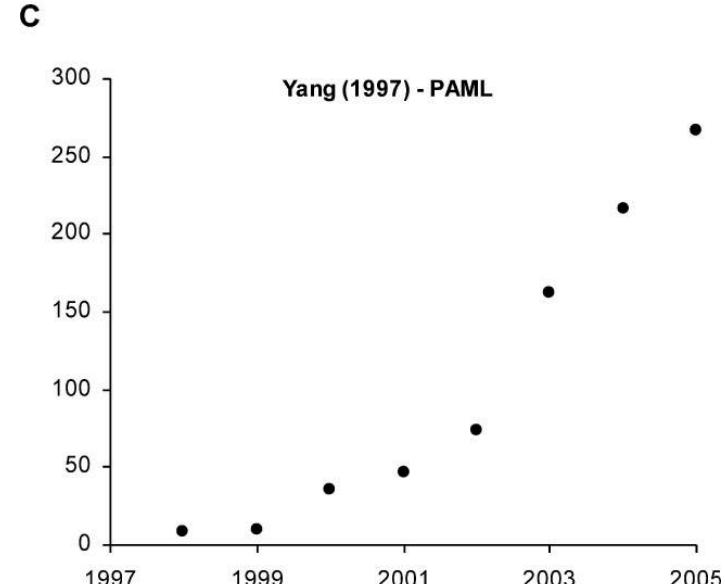
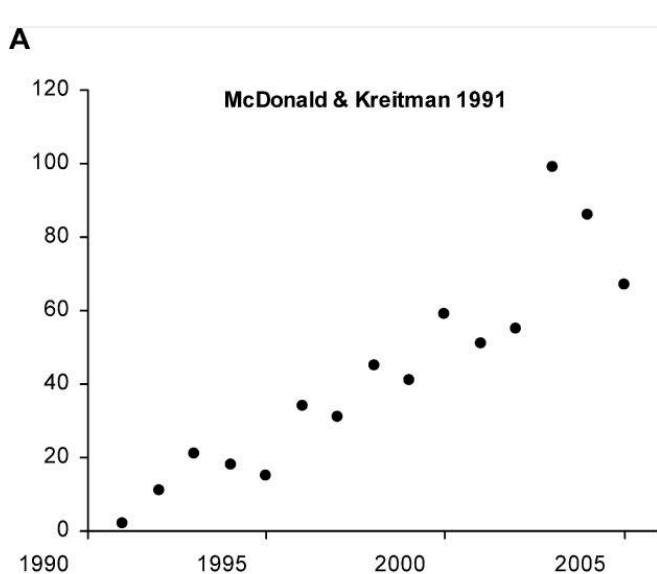
Why Markov codon models

- Take phylogeny into account
- Estimate evolutionary parameters
- Correct for multiple hits
- Account for all possible evolutionary pathways between codons and weight them based on a model

Markov codon models: a success story

- Rigorous statistical framework for hypothesis testing
- Explicitly incorporates evolutionary parameters
- Extensively tested in simulation and on real data:
 - Low false positive rate
 - Much more powerful tests

(eg, Anisimova *et al.* 2001, 2002, 2003; Anisimova & Yang 2007)



Markov model of codon evolution

Instantaneous substitution matrix $Q = \{q_{ij}\}$:

MG-type model	Type of change	GY-type model
0	2 or 3 nt changes	0
f_x^p	Synonymous transversion	π_j
Kf_x^p	Synonymous transition	$K\pi_j$
ωf_x^p	Nonsynonymous transversion	$\omega\pi_j$
ωKf_x^p	Nonsynonymous transition	$\omega K\pi_j$

$\omega = d_N/d_S$ (selection on protein)

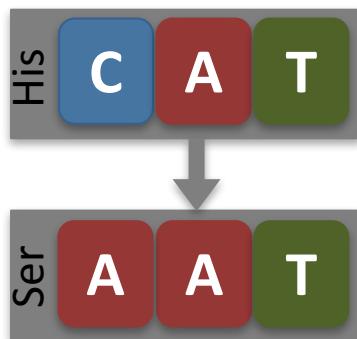
K = transition/transversion ratio

π_j = frequency of codon j

f_x^p = frequency of nucleotide x at codon position p

Defining instantaneous rates

There are many ways to define instantaneous rates:



Exchangeabilities based on	MG-type frequencies	GY-type frequencies
HKY85	$\omega K f_A^1$	$\omega K \pi_{AAT}$
GTR	$\omega r_{C \rightarrow A} f_A^1$	$\omega r_{C \rightarrow A} \pi_{AAT}$
Codon-based	$R_{CAT \rightarrow AAT} f_A^1$	$R_{CAT \rightarrow AAT} \pi_{AAT}$

Modeling codon frequencies

All codon models assume reversibility and stationarity

Codon frequencies $\{\pi_j\}$ are the same at any time

Model	His C A T	Ser A A T
Fequal	1/61	1/61
F1x4	$f_C f_A f_T$	$\left(f_A\right)^2 f_T$
F3x4	$f_C^1 f_A^2 f_T^3$	$f_A^1 f_A^2 f_T^3$
F61	π_{CAT}	π_{AAT}

Likelihood function over phylogeny

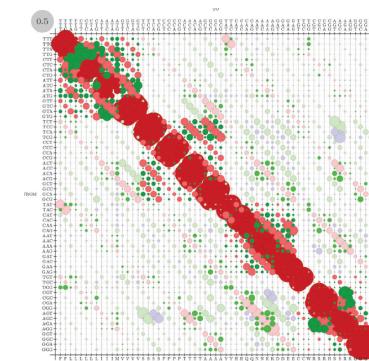
Transition probability matrix over time t : $P(t) = e^{Qt}$
Using $P(t)$ a likelihood $L(\text{Data})$ can be constructed:

Pr(

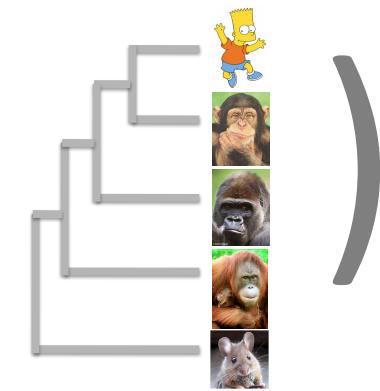


CTC ATA CCC CTC AGC
TTA ATA CCC CTC AGC
TTG ATA CGG CTC AGT
TTA ATA TGG CTC AGC
CTG ATA TGT CTA GGA

↑
Data



↑
Model



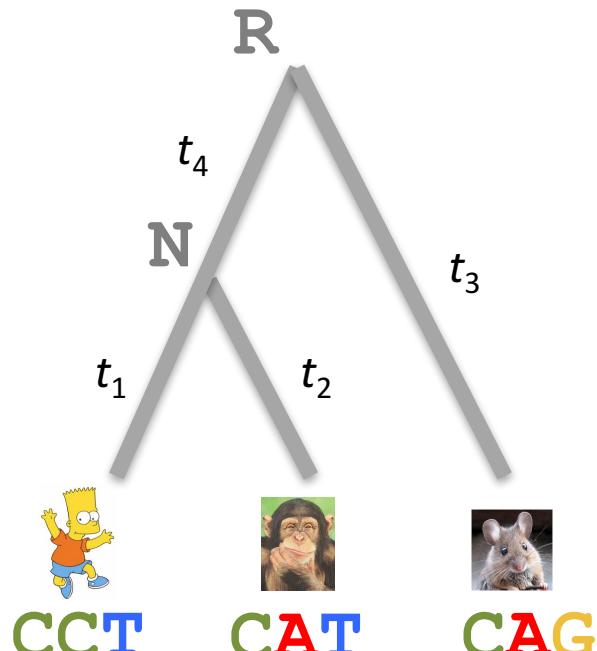
↑
Tree with
branch lengths

Parameters optimized by maximum likelihood

Likelihood function over phylogeny

For each site compute the likelihood:

$$L_h = L \begin{pmatrix} CCT \\ CAT \\ CAG \end{pmatrix} = \sum_R \pi_R p_{R \rightarrow CAG}(t_3) \sum_N p_{R \rightarrow N}(t_4) p_{N \rightarrow CCT}(t_1) p_{N \rightarrow CAT}(t_2)$$



Compute total likelihood assuming independent & identical distribution (i.i.d.) for all sites:

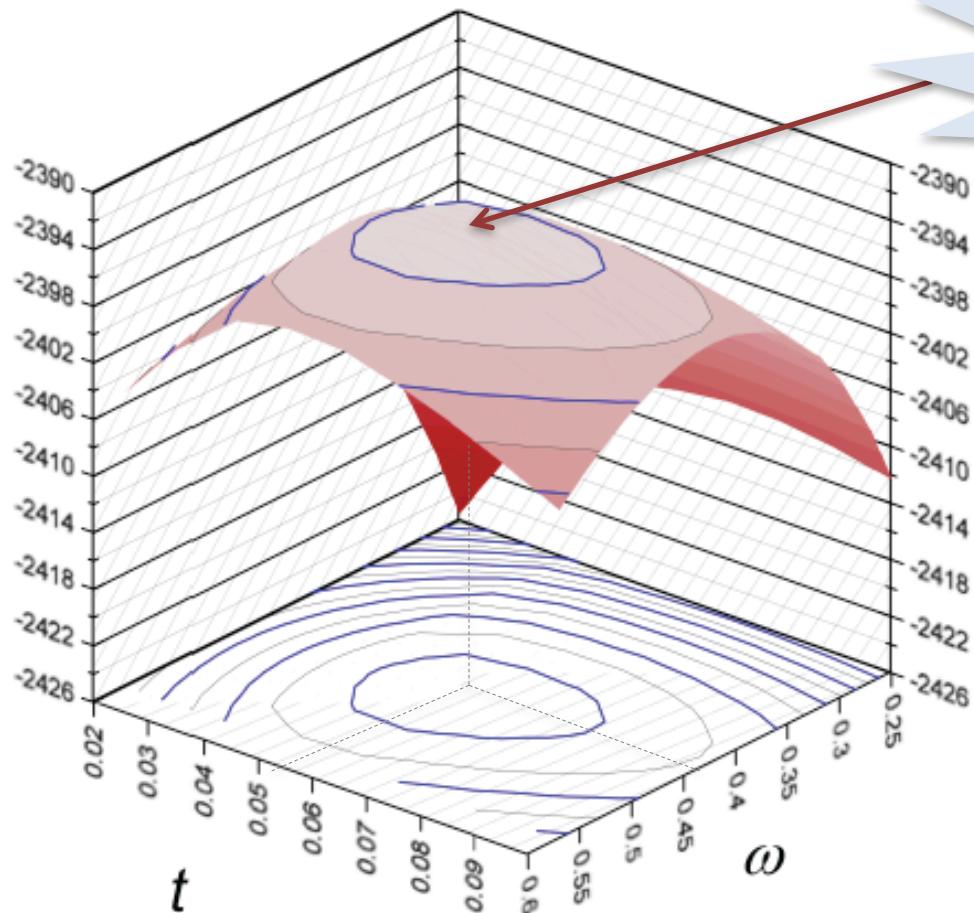
$$L = L_1 \times L_2 \times \dots \times L_n = \prod_{h=1}^n L_h$$

Log-likelihood is optimized (for convenience):

$$\ell = \ln L = \ln L_1 + \ln L_2 + \dots + \ln L_n = \sum_{h=1}^n \ln L_h$$

Unrooted tree – arbitrary root

ML parameter estimation



$\ln L = -2399$

Numerical optimization
by hill-climbing

Example ML estimation
for acetylcholine α receptor
from human and mouse

Exercises with codeml

Focus:

ML estimation with one ω -ratio model M0

Likelihood ratio test for positive selection

Consider two nested models:

Model 0 no positive selection

(H0: λ is always ≤ 1)

Model 1 allows positive selection

(H1: $\lambda > 1$ for some sites or in certain lineages)

LRT statistic: $2\Delta\ell = 2(\ell_1 - \ell_0) \sim \chi^2_{d.f.}$

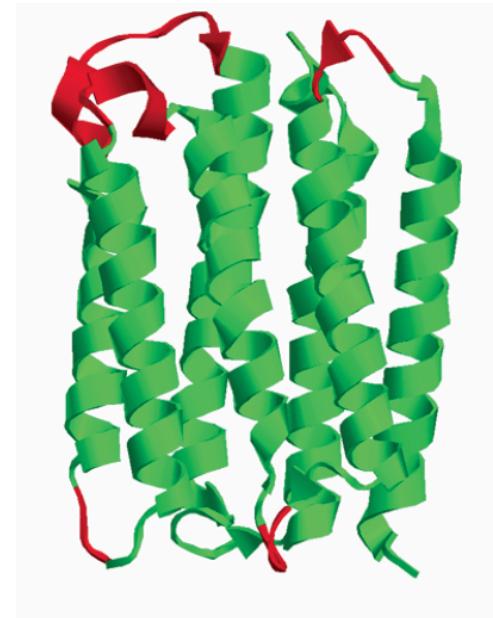
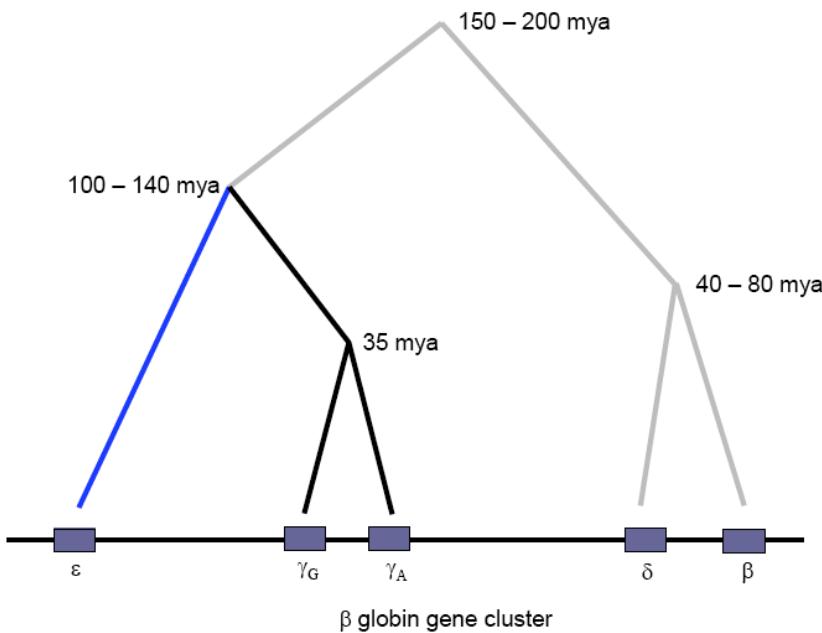
d.f. = difference in numbers of parameters

Modeling selection variability

Assuming *constant selective pressure* across the whole sequence and over the whole phylogeny renders the *power of the test low*

e.g., Endo et al (1996) detected only 17 out of 3595 analyzed genes to be under selection

Positive selection usually affects:
only in a few lineages/branches only few codon sites



Modeling selection variability

By modeling variable ω over time and across sites
we can study:

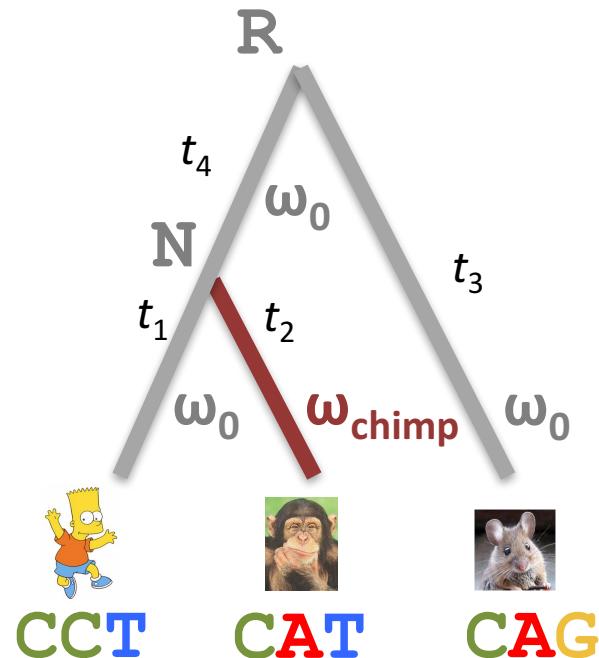
WHEN (in which lineages) did positive selection occur?

WHERE in the sequence did positive selection occur?

Modeling variability over time

Assign independent ω parameters to different branches on the tree:

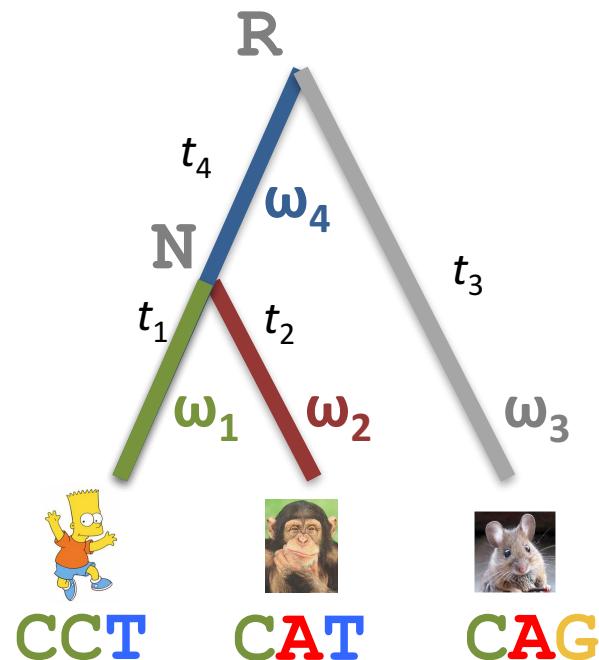
$$L_h = \sum_R \pi_R p_{R \rightarrow CAG}(t_3 | \omega_0) \sum_N p_{R \rightarrow N}(t_4 | \omega_0) p_{N \rightarrow CCT}(t_1 | \omega_0) p_{N \rightarrow CAT}(t_2 | \omega_{\text{chimp}})$$



Modeling variability over time

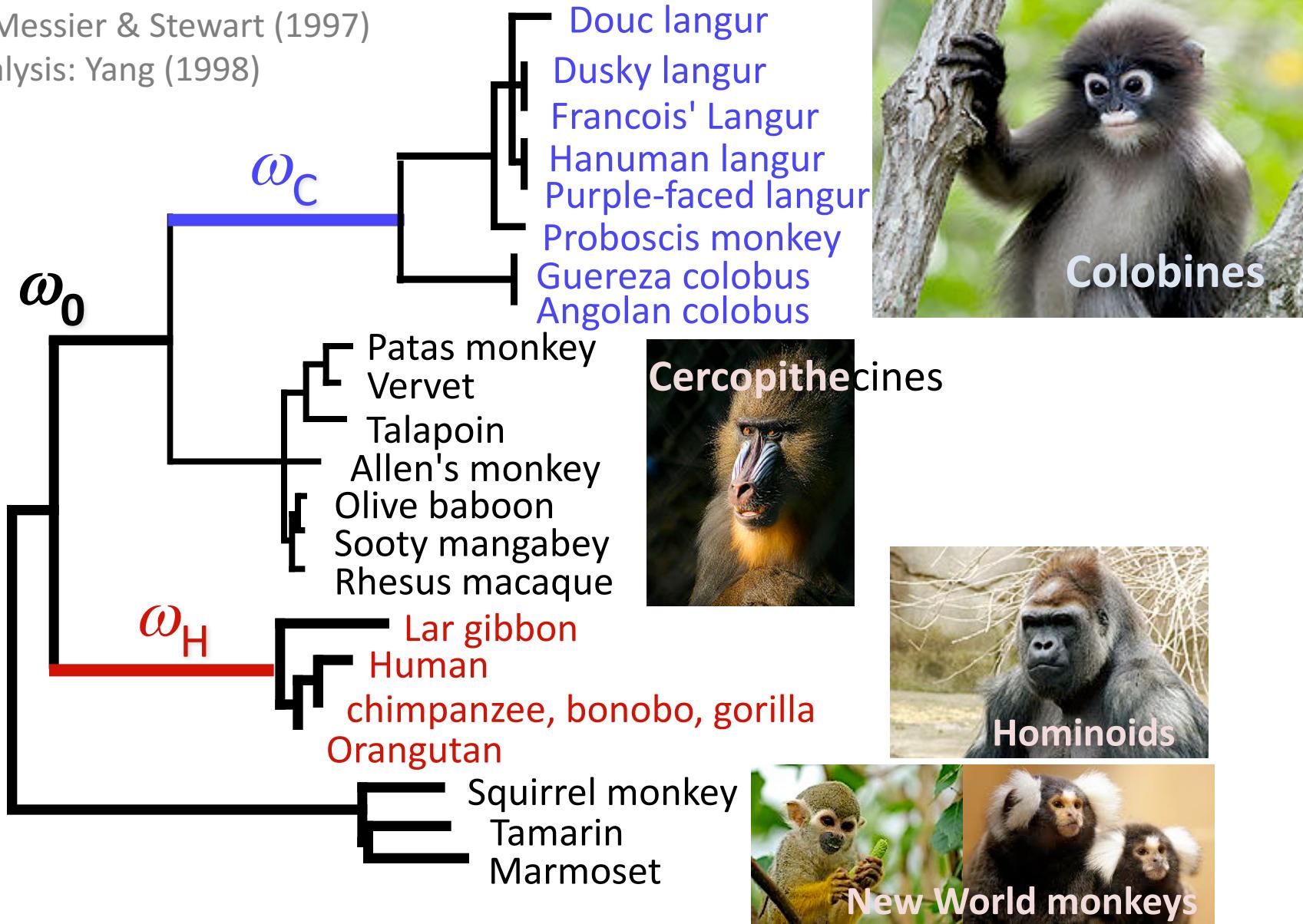
Assign independent ω parameters to different branches on the tree:

$$L_h = \sum_R \pi_R p_{R \rightarrow CAG}(t_3 | \omega_3) \sum_N p_{R \rightarrow N}(t_4 | \omega_4) p_{N \rightarrow CCT}(t_1 | \omega_1) p_{N \rightarrow CAT}(t_2 | \omega_2)$$



Adaptive evolution in primate lyzozyme: ω variability over time

Data: Messier & Stewart (1997)
Re-analysis: Yang (1998)



Primate lysozyme: ML estimates

Model	p	ℓ	ω_0	ω_c
A. 1-ratio: $\omega_0 = \omega_c$	35	-1043.84	0.574	$= \omega_0$
B. 2-ratios: ω_0, ω_c	36	-1041.70	0.489	3.383
C. 2-ratios: $\omega_0, \omega_c = 1$	35	-1042.50	0.488	1 (fixed)

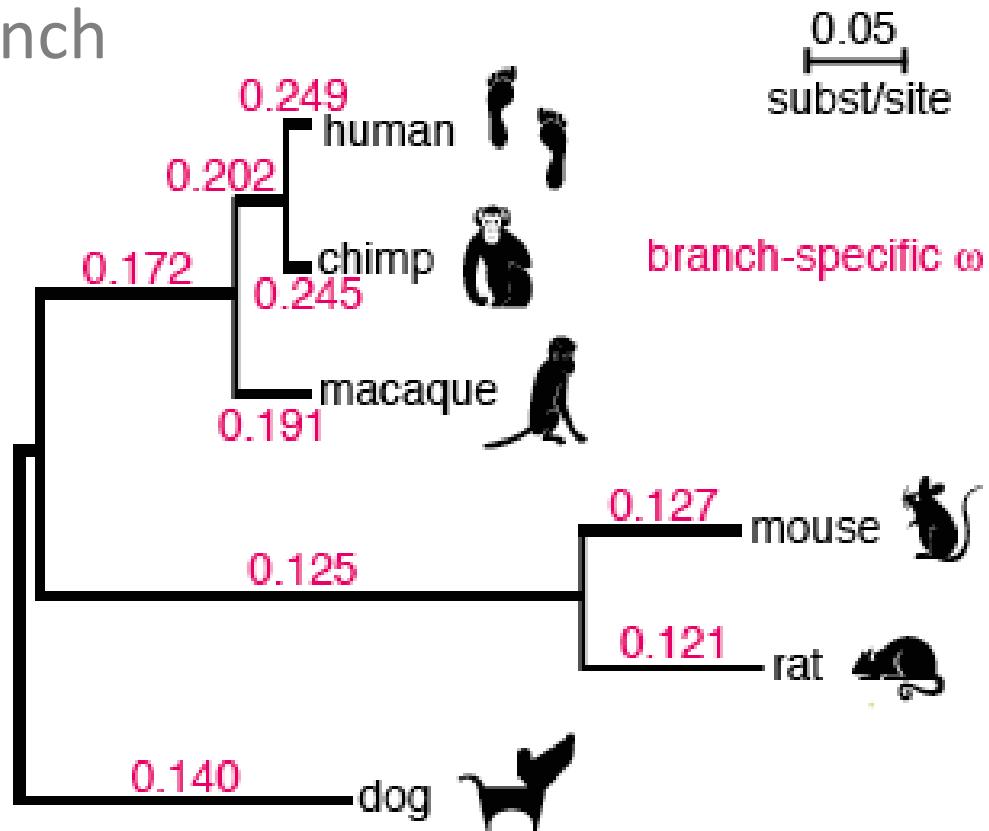
LRT

Null hypothesis	$2\Delta\ell$	d.f.
$\omega_c = \omega_0$	4.24*	1
$\omega_c = 1$	1.60	1

Free ω -ratio LRT with branch model

H_0 : one ω for all branches

H_1 : different ω for each branch



Free ω -ratio LRT with branch model

H_0 : one ω for all branches

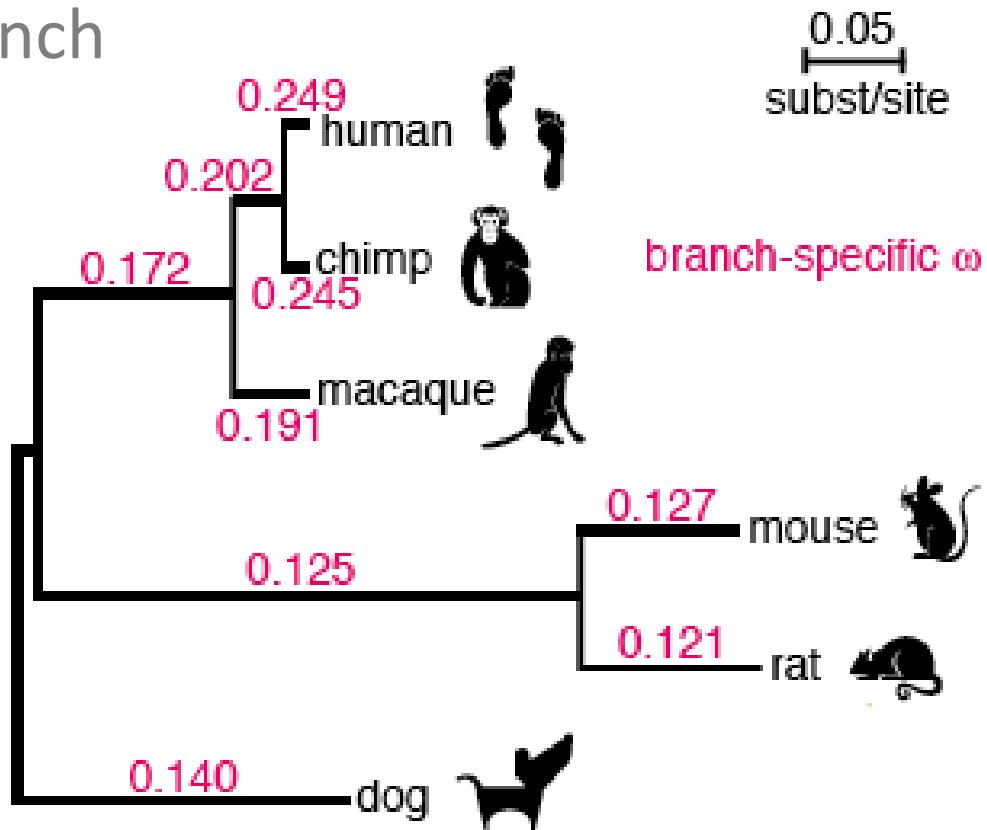
H_1 : different ω for each branch

#branches (for unrooted tree with T leaves):

$$2T-3$$

$$d.f. = (2T-3) - 1 = 2T- 4$$

$$\text{Here: } d.f. = 8$$



Exercises with codeml

Focus:

ML estimation with branch models

Modeling ω variability across sites

M-series models vary only by distributions used to model ω

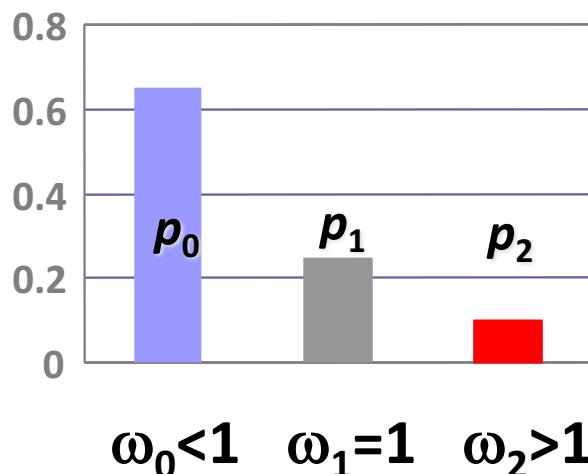
Yang et al. (2000), MBE

It is hard to say what distribution shapes better reflects the data

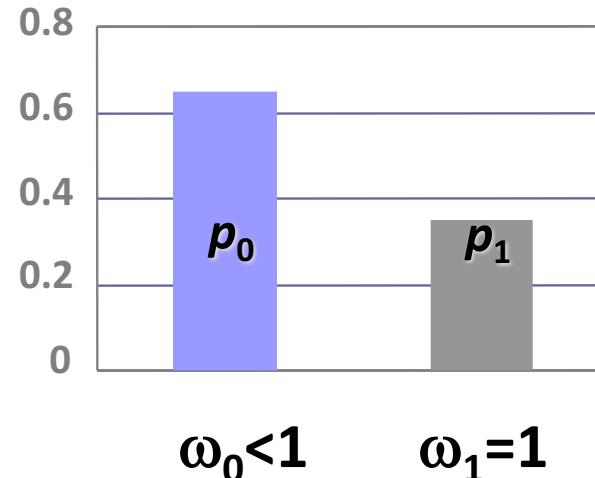
Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1a	2	$p_0, \omega_0,$
Selection	M2a	4	$p_0, p_1, \omega_0, \omega_2$
Discrete	M3	2K-1	p_0, p_1, \dots, p_{K-2} $\omega_0, \omega_1, \dots, \omega_{K-1}$
Frequency	M4	5	p_0, p_1, \dots, p_4
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
Beta	M7	2	p, q
Beta& ω	M8	4	p_0, p, q, ω
Beta&gamma	M9	5	p_0, p, q, α, β
Beta&normal+1	M10	5	p_0, p, q, α, β
Beta&normal>1	M11	5	p_0, p, q, μ, σ
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$

Examples of nested site models

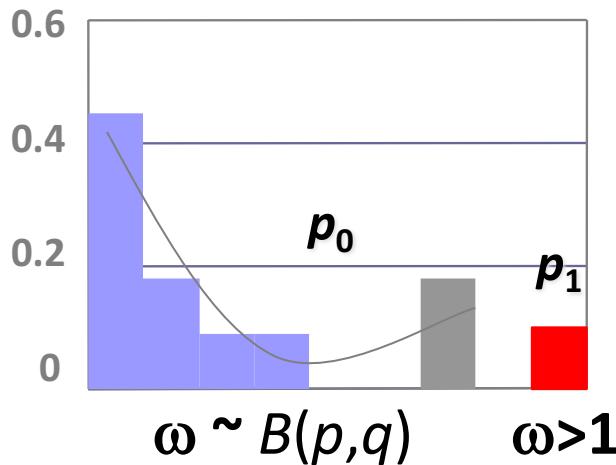
M2



M1

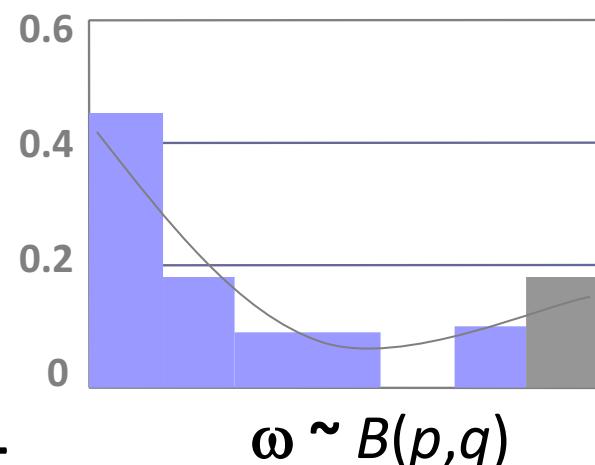


M8



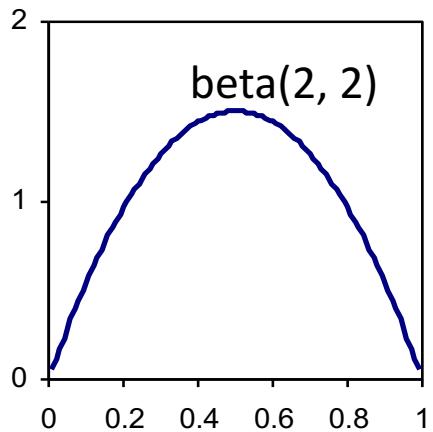
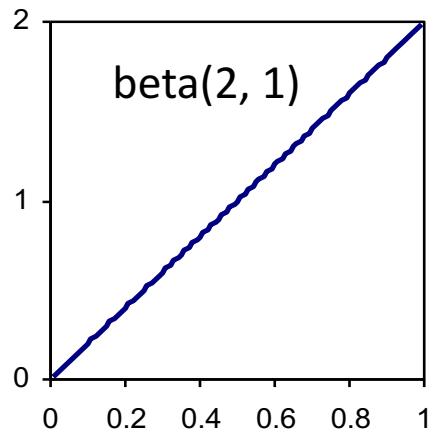
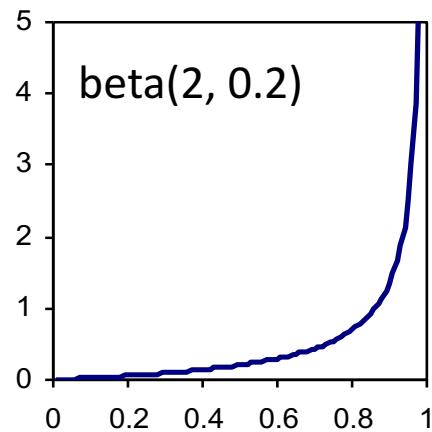
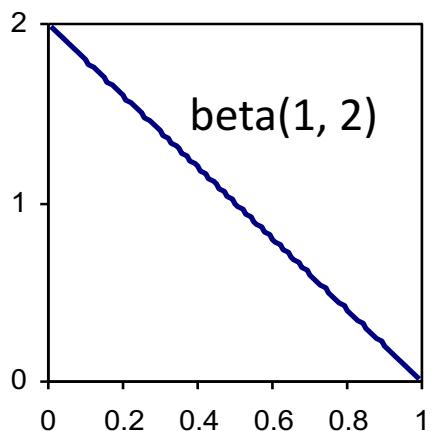
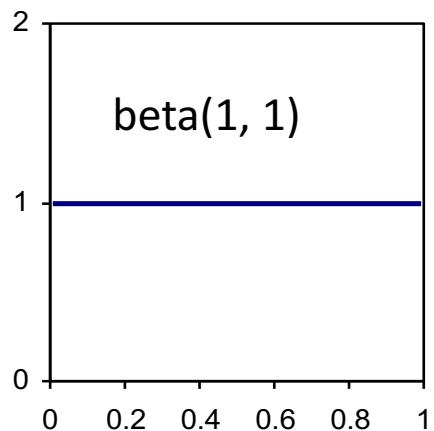
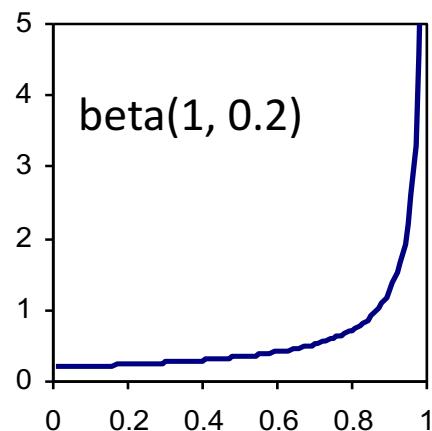
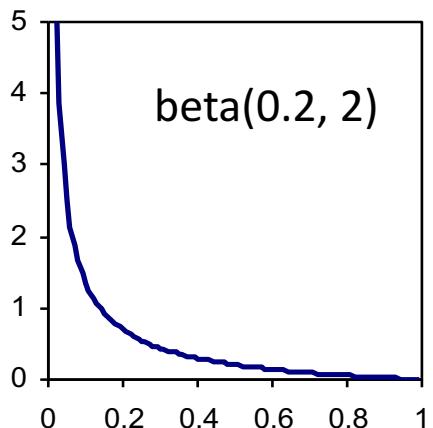
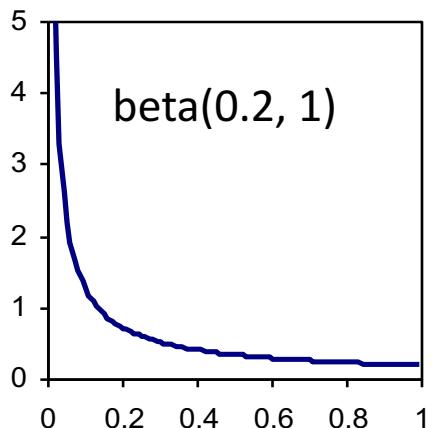
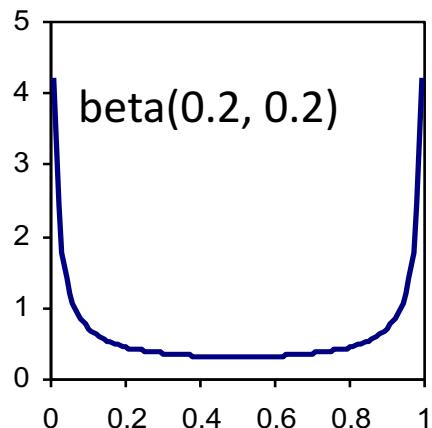
Alternative

M7



Null

$0 \leq B(p,q) \leq 1$



Theoretical distribution of LRT

A. M0 vs. M3 (with 3 classes)

Transition from M3 to M0 requires

$p_0 = p_1 = 0$ (boundary)

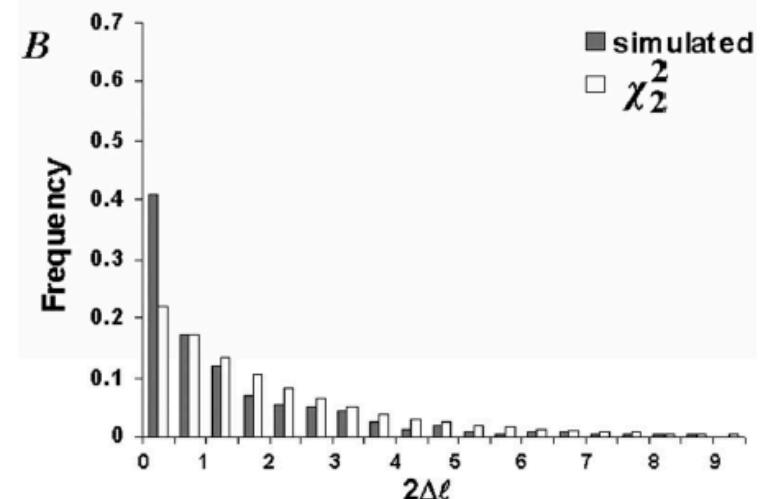
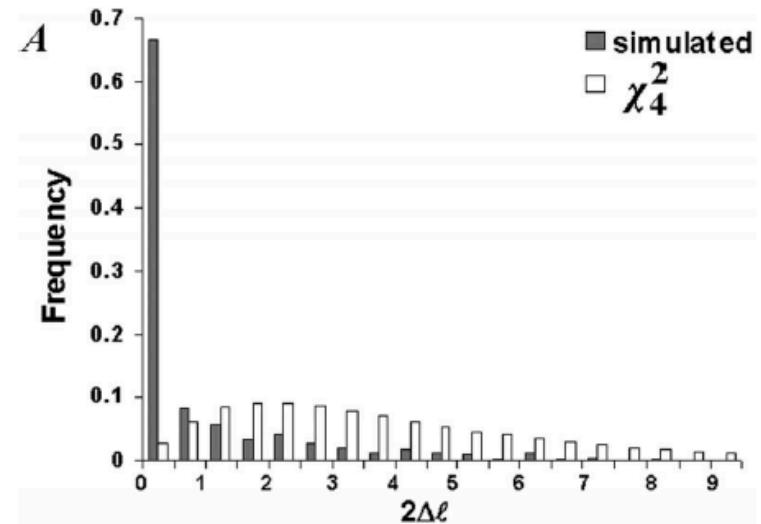
Theoretical distribution makes the test
conservative

B. M7 vs. M8

Transition from M8 to M7 requires

sets $p_0 = 1$ (or $p_1 = 0$, both at the boundary)

Theoretical distribution fits better than in A
(slightly conservative)

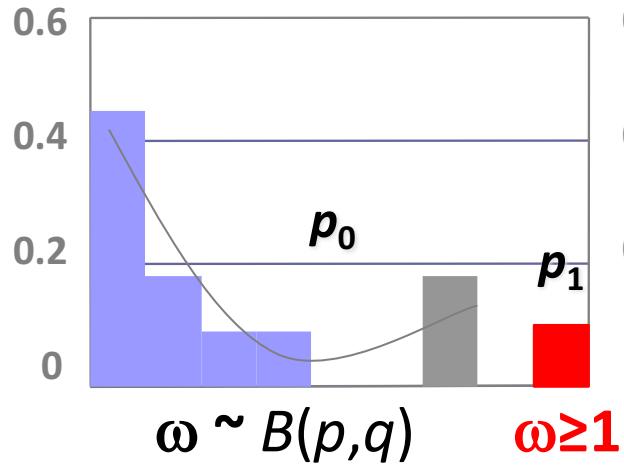


Examples of nested site models

A better defined LRT:

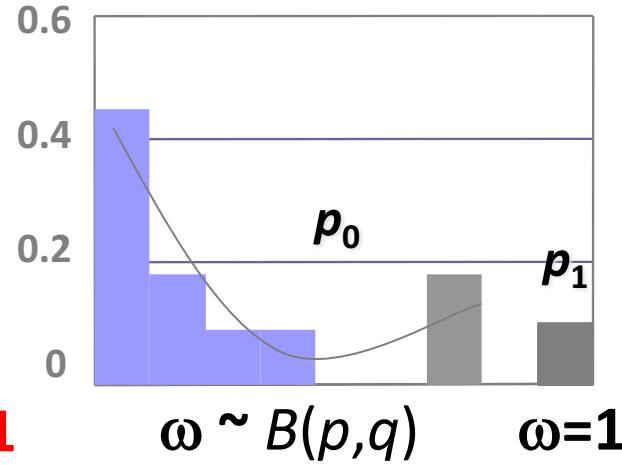
The null is 50:50 χ^2 mixture (with d.f. = 1 and 0)

M8



Alternative

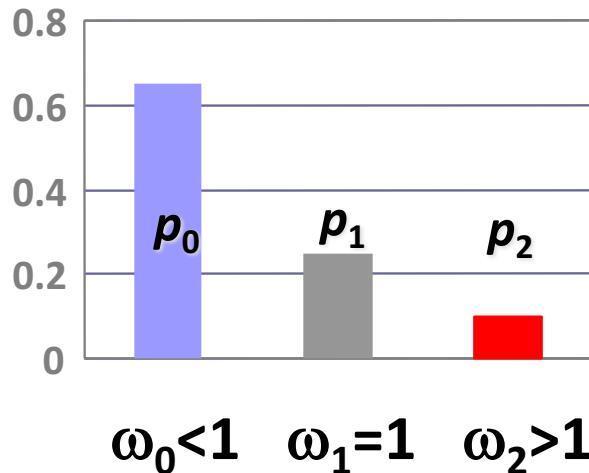
M8a



Null

Examples of nested site-specific models

M2



Likelihood calculation should take into account
that a site may come from a number of different classes:

$$L_h = \Pr(\text{data}_{\text{site}}) = \sum_{\text{class}=1}^K \Pr(\text{data}_{\text{site}} \mid \omega_{\text{site}} = \omega_{\text{class}}) p_{\text{class}}$$

Example: Human MHC Class I data

192 alleles, 270 codons

Model	ℓ	Parameter estimates
M1a (neutral)	-7,490.99	$p_0 = 0.830, \omega_0 = 0.041$ $p_1 = 0.170, \omega_1 = 1$
M2a (selection)	-7,231.15	$p_0 = 0.776, \omega_0 = 0.058$ $p_1 = 0.140, \omega_1 = 1$ $p_2 = 0.084, \omega_2 = 5.389$

LRT of positive selection:

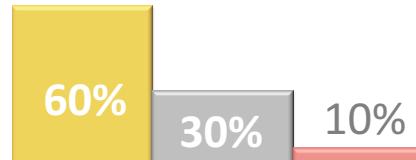
$$2\Delta\ell = 2 \times 259.84 = 519.68, \quad P < 0.000 \text{ (d.f. = 2)}$$

**So far we used
models with variable selection
to test if selection affected the data**

If LRT for positive selection is *significant*
we can proceed inferring WHEN and WHERE...
(but this is more difficult)

Prediction of sites with Bayesian approach

ω site classes (GDD or M3):



$\omega=0.1$ $\omega=1$ **$\omega = 4.3$**

For each site compute posterior probability:

$$P(\text{Red Box} | \text{CTC, TTA, TTG, TTA, CTG}) = \frac{P(\text{Red Box} | \text{CTC, TTA, TTG, TTA, CTG}) P(\text{Red Box})}{P(\text{Red Box} | \text{CTC, TTA, TTG, TTA, CTG}) P(\text{Red Box}) + P(\text{Yellow Box} | \text{CTC, TTA, TTG, TTA, CTG}) P(\text{Yellow Box}) + P(\text{Grey Box} | \text{CTC, TTA, TTG, TTA, CTG}) P(\text{Grey Box})}$$

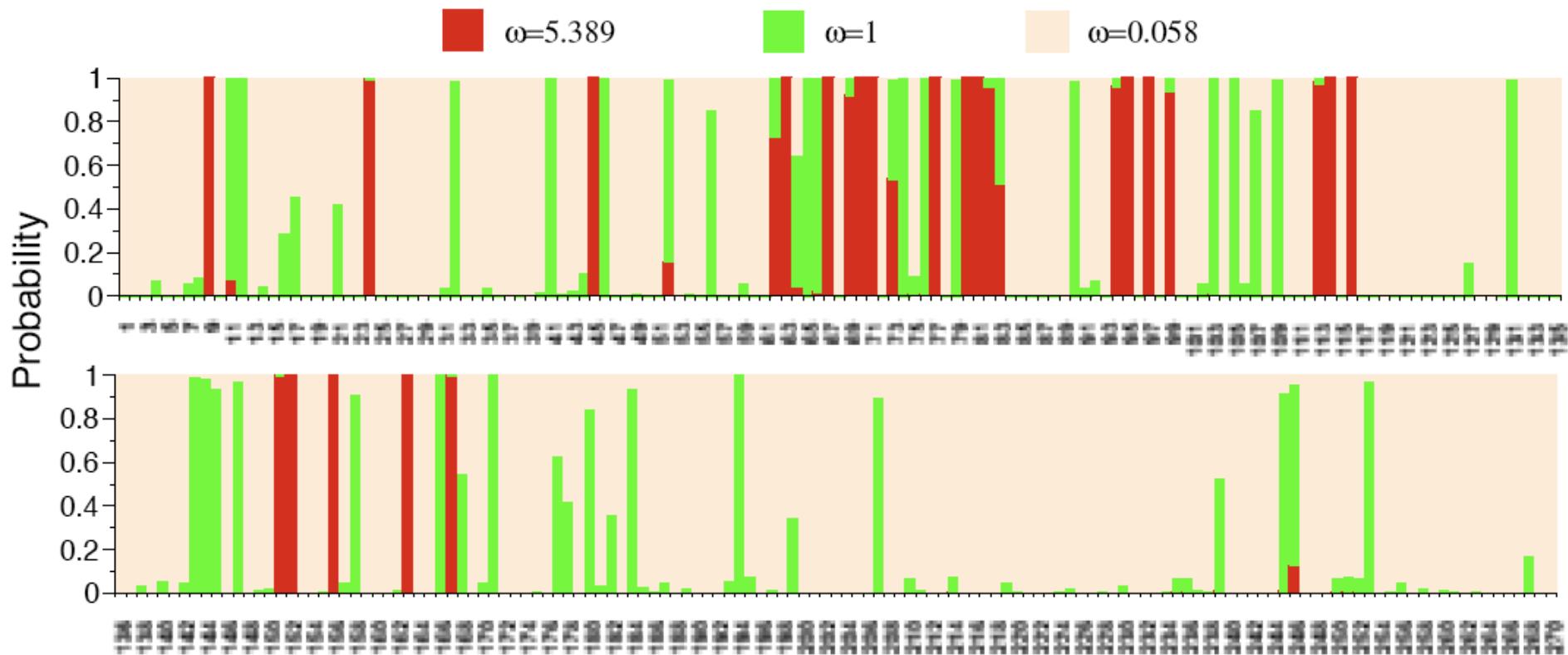
Sites with high posteriors (≥ 0.95)
may be inferred to be under positive selection

Empirical Bayesian calculation of posterior probabilities that a site is under positive selection with $\omega > 1$.

- Naïve Empirical Bayes (NEB) ignores sampling errors in parameter estimates.
- Bayes Empirical Bayes (BEB) accounts for sampling errors by integrating over a prior.

Nielsen & Yang. 1998 *Genetics* **148**
Yang, Wong & Nielsen 2005 *Mol Biol Evol* **22**

Posterior probabilities of ω for MHC (M2a)



$$p(\omega_{\text{site}} = \omega_{\text{class}} | \text{data}_{\text{site}}) = \frac{p(\text{data}_{\text{site}} | \omega_{\text{class}}) p_{\text{class}}}{\sum_j p(\text{data}_{\text{site}} | \omega_j) p_j}$$

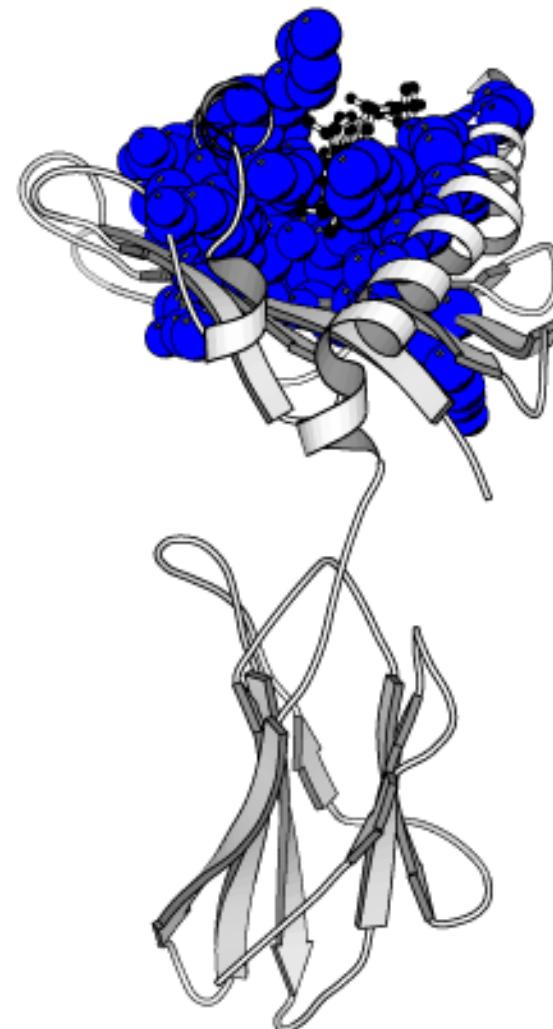
$j = \text{site class}$

Human MHC Class I: 3D structure

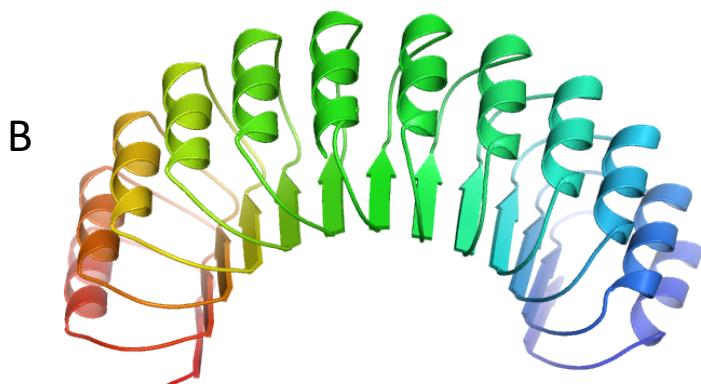
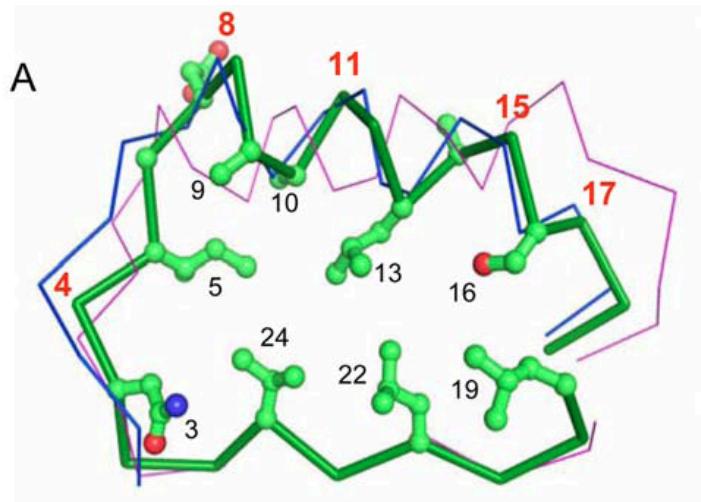
25 sites identified
under M2a

All sites cluster together in
the antigen recognition
domain (blue)

Yang and Swanson (2002)



Positive selection in bacterial GALA



Bacterial GALA (type III effectors) acquired from host plants by LGT: residues under positive selection are found on the convex side of horseshoe & involved in binding

Data from Kajava, Anisimova, Peeters (2008)

Figure 2. Structural model of GALA-LRR. (A) C α -trace superposition of a modeled GALA-LRR and the known CC-LRR from human Skp2 protein [10] and RI-LRR from porcine ribonuclease inhibitor [46]. GALA-LRR model is shown in a ball-and-stick representation, CC-LRR is shown by a blue trace and RI-LRR by a magenta trace. Numbering of the conserved GALA-LRR residues is taken from Figure 1. Numbers in red point to positions inferred to be under positive selection. The carbon atoms are in green, oxygen in red, nitrogen in blue. (B) A ribbon diagram of a structural model of the C-terminal LRR domain of GALA4 type III effector protein from *R. solanacearum* (strain MolK2, region 170 to 460, accession code ZP_00946474). The figure was generated with Pymol [47]. The atomic coordinates of the model are available on request.

With more genomes sequenced, the approach of evolutionary comparison becomes more powerful.

It provides a way of generating interesting biological hypotheses, which can be validated by experimentation.

Ivarsson, Mackey, Edalat, Pearson, and Mannervik (2002)
Identification of residues in glutathione transferase capable of driving functional diversification in evolution: **a novel approach to protein design**. *J. Biol. Chem.* 278:8733-8738.

Bielawski, Dunn, Sabehi, and Beja (2004) **Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment**. *Proc. Natl. Acad. Sci. U.S.A.* 101:14824-14829.

Positive selection of primate *TRIM5 α* identifies a critical species-specific retroviral restriction domain

Sara L. Sawyer*, Lily I. Wu†, Michael Emerman*†, and Harmit S. Malik**‡

Divisions of *Basic Sciences and †Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Communicated by Mark T. Groudine, Fred Hutchinson Cancer Research Center, Seattle, WA, December 29, 2004 (received for review December 8, 2004)

Primate genomes encode a variety of innate immune strategies to defend themselves against retroviruses. One of these, *TRIM5 α* , can restrict diverse retroviruses in a species-specific manner. Thus, whereas rhesus *TRIM5 α* can strongly restrict HIV-1, human *TRIM5 α* only has weak HIV-1 restriction. The biology of *TRIM5 α* restriction

genome defense predates the origin of primate lentiviruses (11, 12) and that many other *APOBEC* cytidine deaminase genes likely participate in defending the primate genome against retroviruses.

Here, we show that the *TRIM5 α* restriction factor has

Rhesus *TRIM5 \checkmark* restricts HIV-1 while human *TRIM5 \checkmark* has only weak restriction.

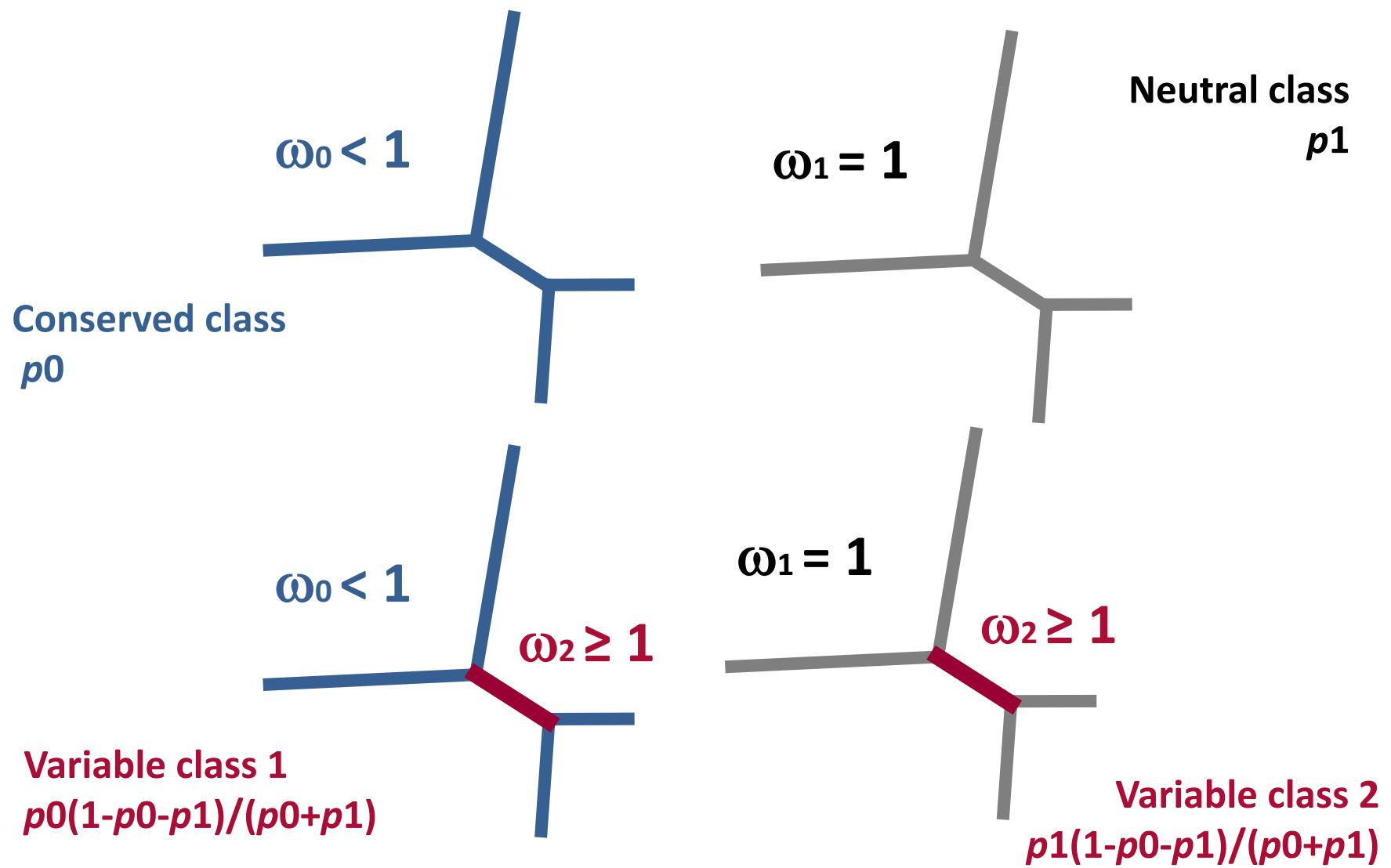
Phylogenetic analysis identified a 13-aa patch with many positive-selected sites. Functional studies of chimeric *TRIM5 \checkmark* genes demonstrated that the patch was largely responsible for the difference in function. (Sawyer et al 2005)

Exercises with codeml

Focus:

ML estimation with site models

Branch-site codon model A (Yang et al 2005)



LRT for positive selection based on branch-site codon model

Null:
Model A
 $\omega_2 = 1$ fixed

Alternative:
Model A
 $\omega_2 \geq 1$ estimated



$$\text{LRT statistic } 2(\ell_0 - \ell_1) \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$$

Foreground branches (with ω_2) are defined *a priori*

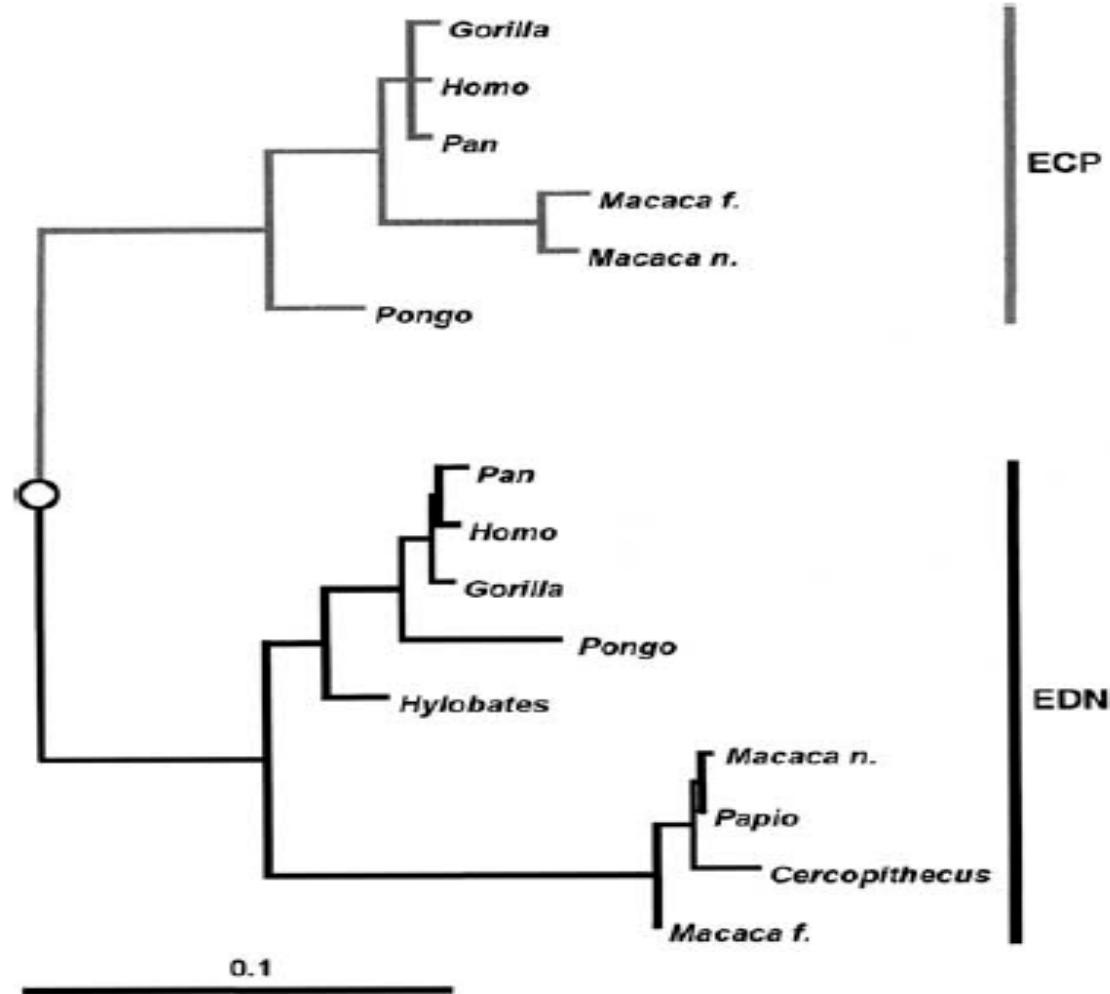
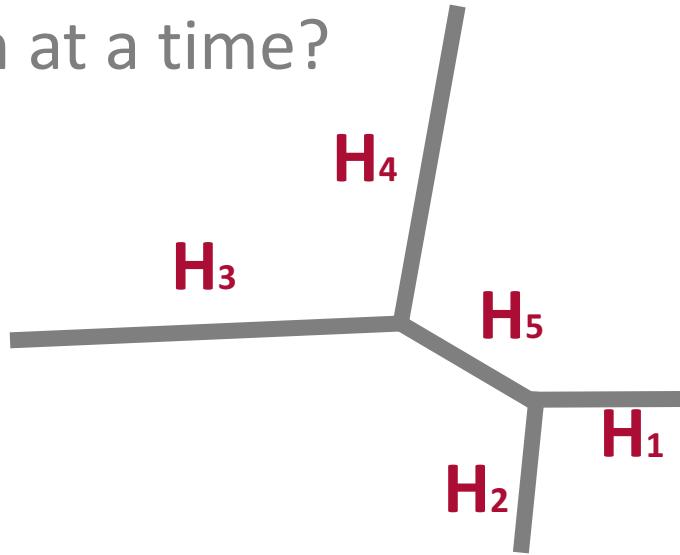


Fig. 3. Gene tree for 15 sequences from the ECP–EDN gene family. The topology was obtained by using maximum likelihood analysis under the HKY85 substitution matrix combined with a correction for among-site rate variation (discrete gamma model). The scale bar indicates the mean number of substitutions per nucleotide site. The open circle indicates the duplication event that gave rise to the ECP and EDN genes. Under Model D, a fraction of sites was allowed to evolve under divergent selection pressure, with ω_{1A} and ω_{1B} for the two paralogous clades, respectively.

To test for selection after gene duplication: branches of one clade following the duplication event are set as foreground

Testing multiple hypotheses

Test one branch at a time?



Are p_1, p_2, p_3, p_4, p_5 significant at an overall threshold α ?

Adjust individual thresholds $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$

so overall type I error rate $\leq \alpha$

Multiple testing correction: FWER or FDR?

Family-Wise Error Rate (FWER): overall type I error (FP rate)

FWER = Pr (reject at least one null when it's true)

For n independent true null hypotheses tested at α :

$$\text{FWER} = 1 - (1 - \alpha)^n$$

e.g. testing 10 hypotheses at 5% each we may get FWER=40%!

If in some cases the null hypotheses is expected to be wrong,
small percentage of false rejections is tolerable

FDR = False Discovery Rate

FDR = $E(\# \text{ false rejections}/\# \text{ all rejections})$

Example: how do FWER and FDR compare

100 simulated datasets with first 6 null hypotheses true

For each sample, test 10 hypotheses, making 1 error per sample

Test results: 1=sign / 0=not sign

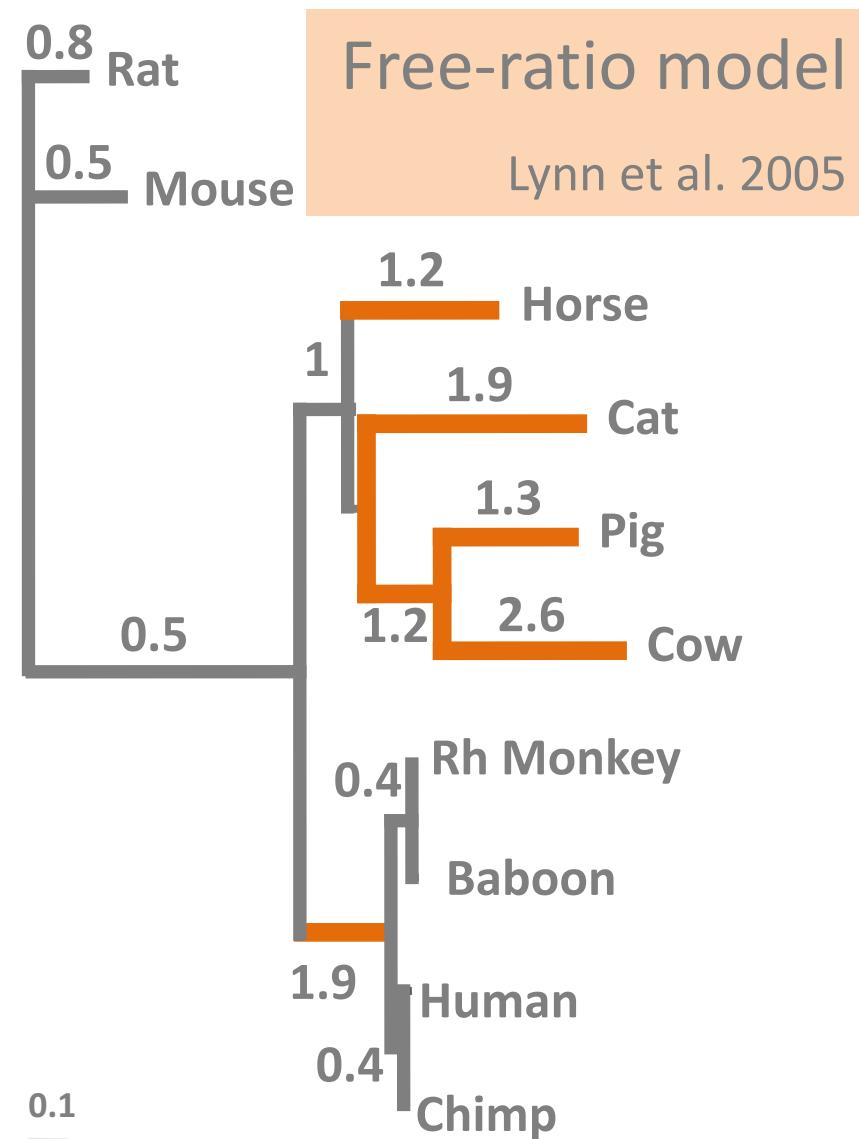
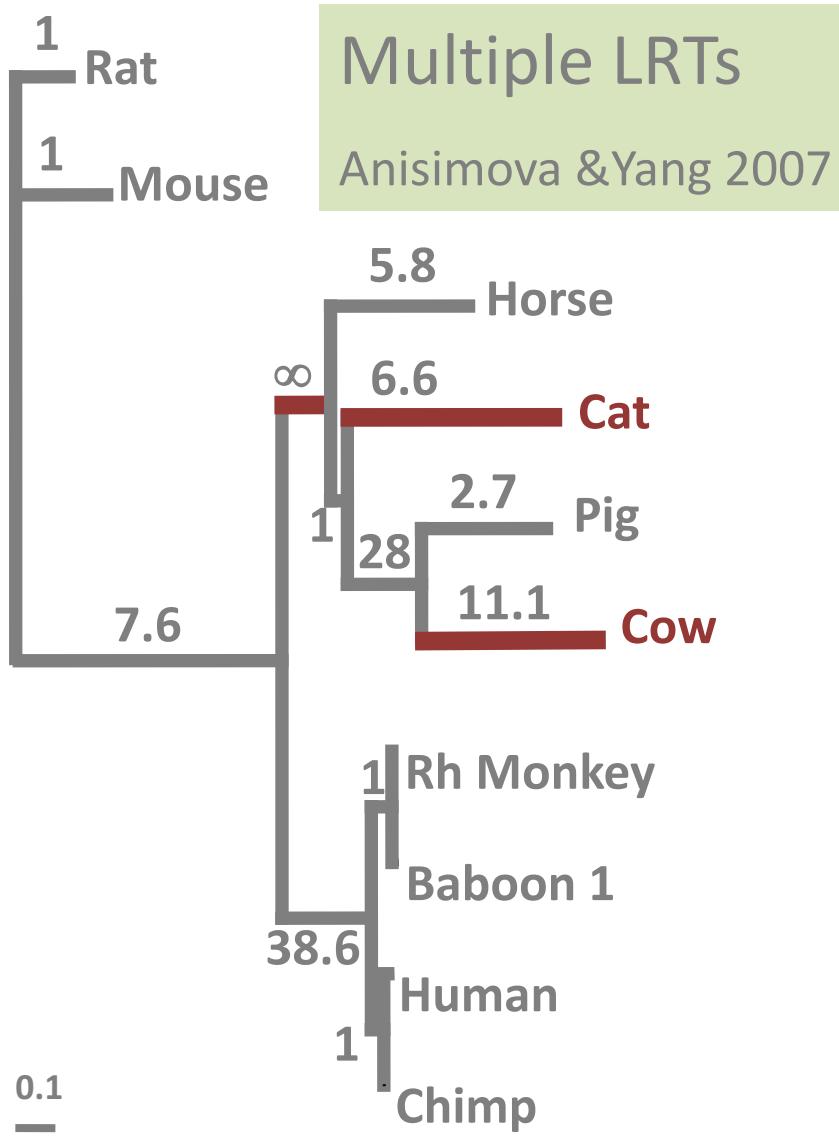
1	0 0 0 0 0 1 1 1 1 1
2	0 1 0 0 0 0 1 1 1 1
3	0 0 0 0 1 0 1 1 1 1
...	
100	0 1 0 0 0 0 1 1 1 1

FDR = 20%

TTTTTTTFFF

FWER = 100%

Multiple branch-site LRTs example: CD2 extra-cellular domain



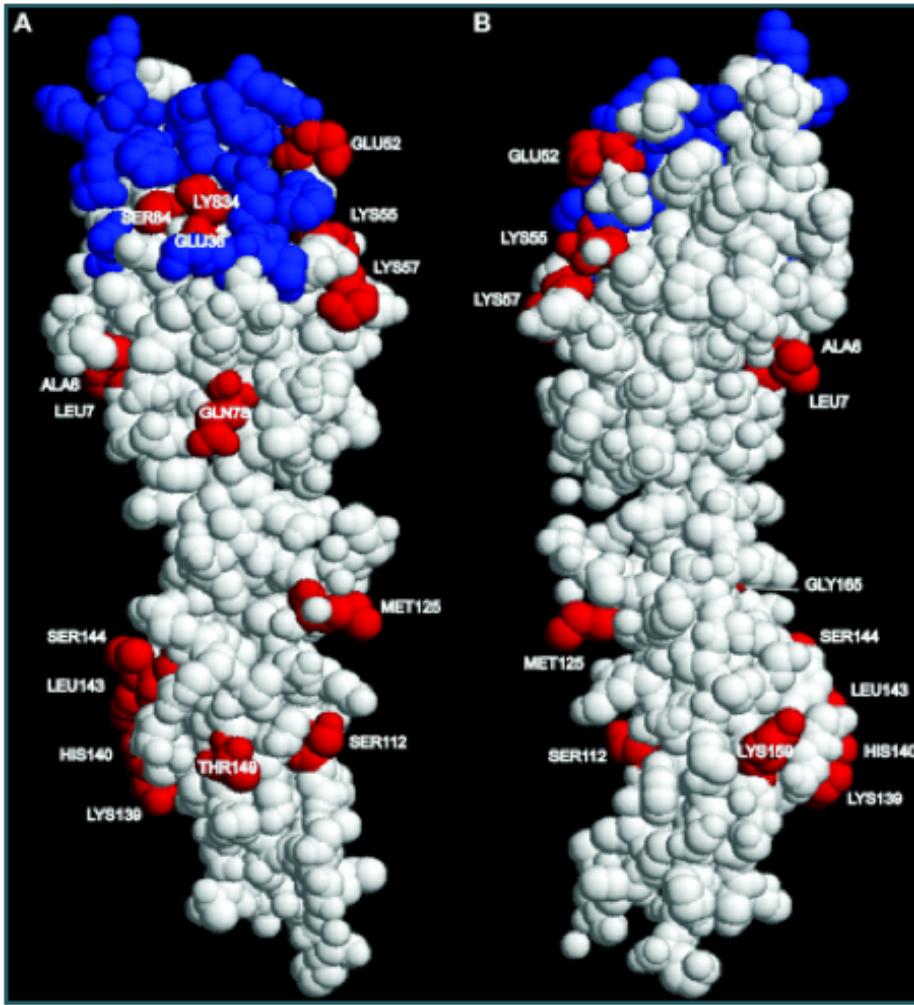
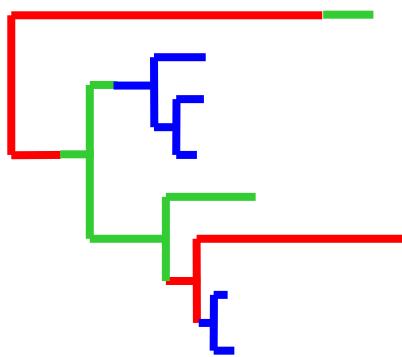


FIGURE 3.—

The three-dimensional structure of human CD2 extracellular domain [Protein Data Bank (PDB) <http://www.rcsb.org/pdb/entry=1HNF>]. Sites shown in red are those sites predicted to be under positive selection (model 8). The sites are labeled according to the numbering scheme used in the PDB file (ALA6 corresponds to site 14 in Table 1). Sites known to be involved in CD58 binding are shown in blue. A and B show two opposite faces of the CD2 molecule. The structure was displayed using RasMol V2.7.2.1.1 (<http://www.openrasmol.org/software/rasmol/>).

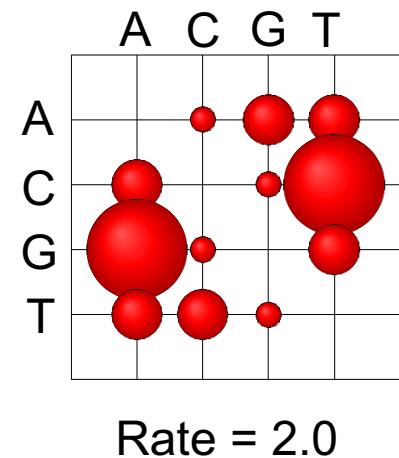
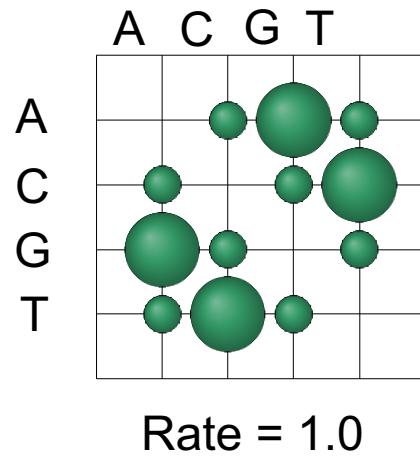
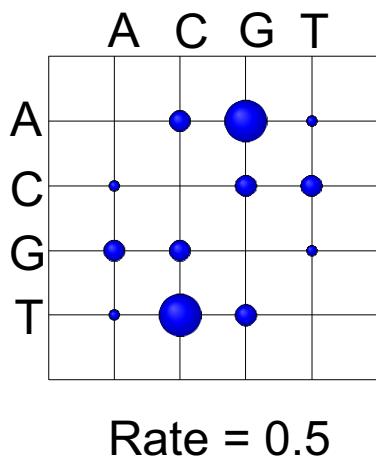
All but two sites under positive selection are found in the extra-cellular domain of CD2

Alternatively, use covarion models

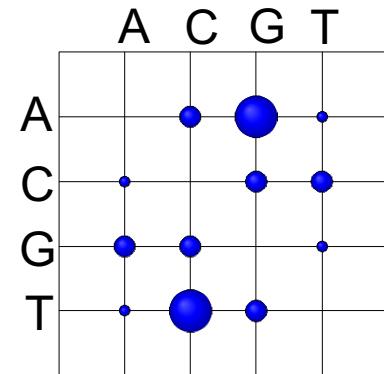
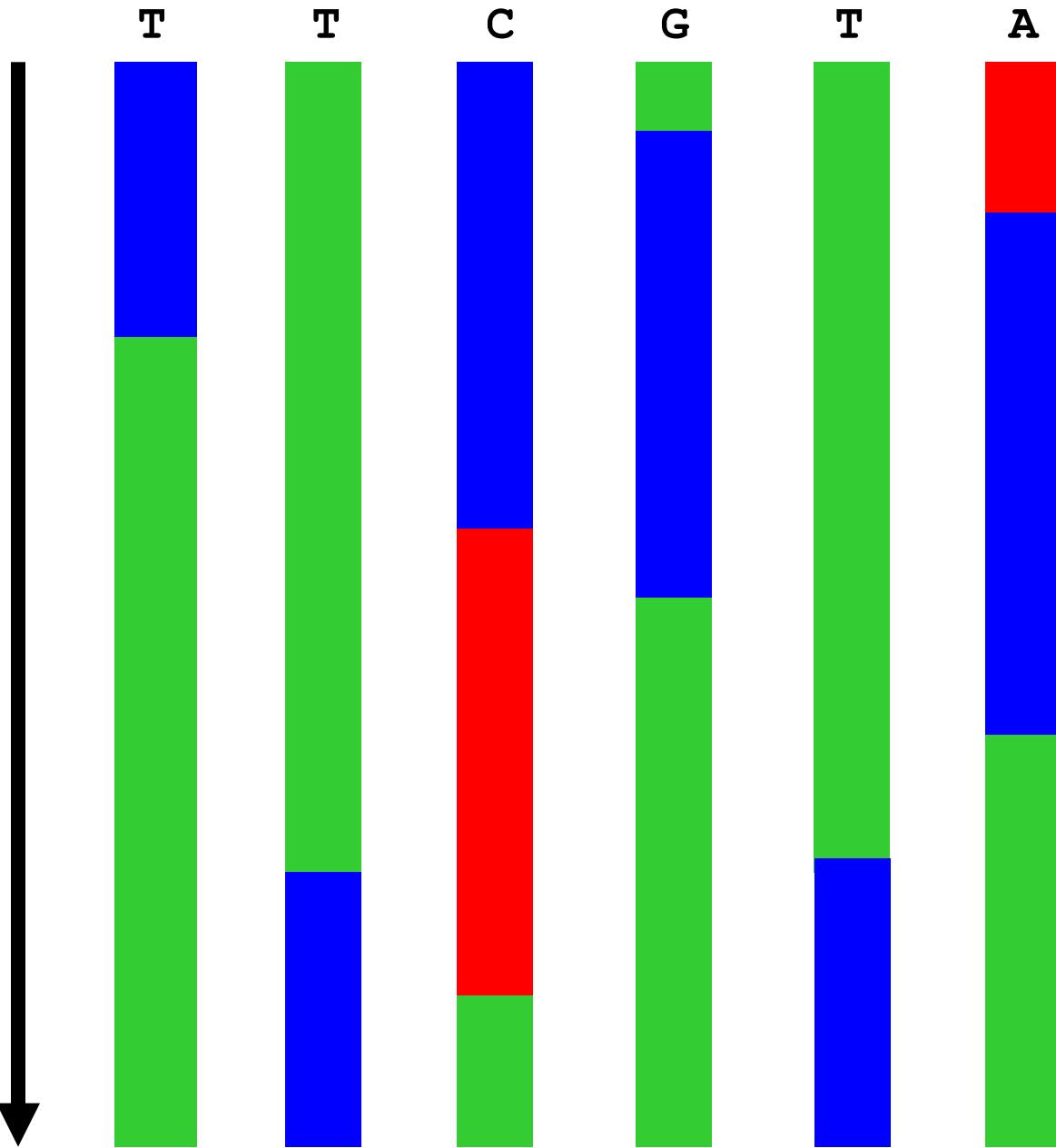


Seq1	TCTTTATTGACGTGTATGGACAATT
Seq2	TCTTTGTTAACGTGCATGGACAATT
Seq3	TCCTTGCTAACATGCATGGACAATT
Seq4	TCTTGCTAACGTGCATGGATAATT
Seq5	TCTT—TAACGTGCATAGATAACTC
Seq6	TCAC—TAACATGTATAGATAACTC
Seq7	TCTCTTCTAACGTGCATTGTGAAGTC
Seq8	TCTCTTTGACATGTATTGAAAAATC

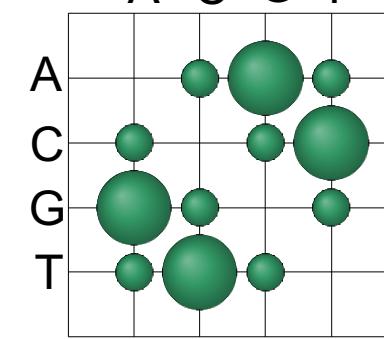
Figure by Simon Whelan



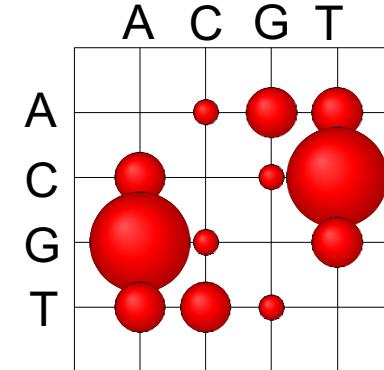
Time



Rate = 0.5



Rate = 1.0



Rate = 2.0

Markov Modulated Codon Model

$$Q_x(ij) = \begin{cases} 0: \text{if codons } i \text{ and } j \text{ differ at more than one nucleotide position} \\ \omega_x \pi_j: \text{nonsynonymous transversion} \\ \pi_j: \text{synonymous transversion} \\ \kappa \omega_x \pi_j: \text{nonsynonymous transition} \\ \kappa \pi_j: \text{synonymous transition} \end{cases}$$

Q_x describes instantaneous rates for sites from selection regime x

Codon models M2 and M3 are considered (each has 3 classes of sites)

Guindon et al. 2004 PNAS

$$\mathbf{R} = \delta \begin{pmatrix} -(p_2 + p_3\alpha) & p_2 & p_3\alpha \\ p_1 & -(p_1 + p_3\beta) & p_3\beta \\ p_1\alpha & p_2\beta & -(p_1\alpha + p_2\beta) \end{pmatrix}$$

\mathbf{R} describes rate switches between selection regimes 1, 2 and 3 ($\omega_1 < \omega_2 < \omega_3$)

p_1, p_2, p_3 are equilibrium frequencies of sites in each selection regime (add up to 1)

α is relative rate of changes between 1 and 3

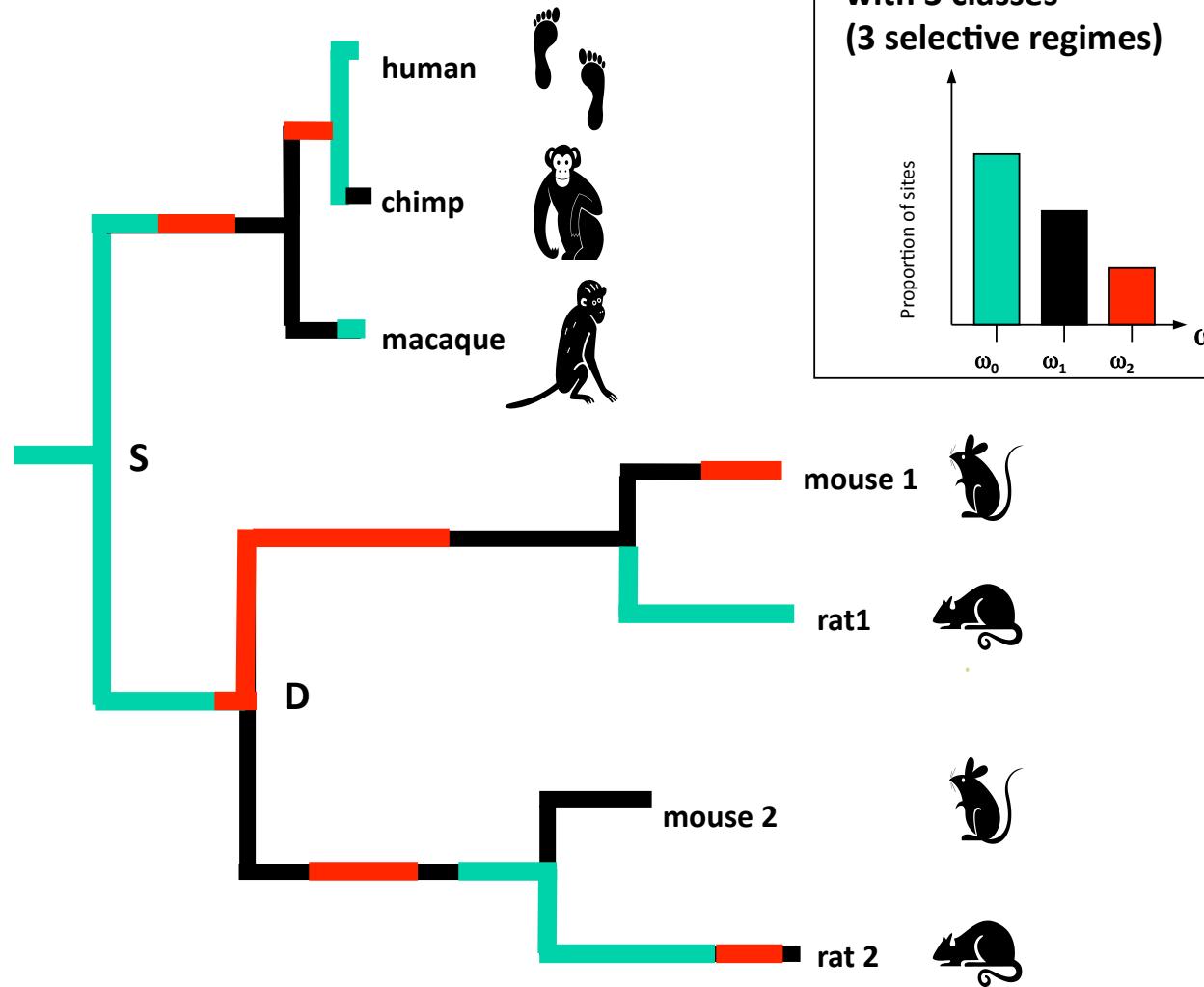
β is relative rate of changes between 2 and 3

Combined process:

$$\mathbf{S} = \begin{pmatrix} \mathbf{Q}_1 & 0 & 0 \\ 0 & \mathbf{Q}_2 & 0 \\ 0 & 0 & \mathbf{Q}_3 \end{pmatrix} + \delta \begin{pmatrix} -(p_2 + p_3\alpha)\mathbf{I} & p_2\mathbf{I} & p_3\alpha\mathbf{I} \\ p_1\mathbf{I} & -(p_1 + p_3\beta)\mathbf{I} & p_3\beta\mathbf{I} \\ p_1\alpha\mathbf{I} & p_2\beta\mathbf{I} & -(p_1\alpha + p_2\beta)\mathbf{I} \end{pmatrix}$$

δ is the rate of switch between selection regimes

Markov Modulated Codon Model



LRTs of temporal variation in selection

$H_0: \delta = 0$ (no switches btw regimes or M3)

$H_1: \delta \neq 0$

$H_0: \delta = 0$ (no switches btw regimes)

$H_1: \beta = \alpha = 1$ (switching but no bias in switching pattern)

$H_0: \beta = \alpha = 1$ (no bias in switching pattern)

$H_1: \beta \neq \alpha$

Model notations: +S1 ($\beta = \alpha = 1$)

+S2 ($\beta = \alpha$ are free)

LRTs of temporal variation in selection

Guindon et el. 2004 PNAS

Table 1. Likelihood analysis of eight HIV-1 env gene sequence data sets

Significant at 5%

	M2	M2+S1	M2+S2	M3	M3+S1	M3+S2
P1						
lnL	-3,050.46	-3,021.78	-3,019.93	-3,036.87	-3,021.15	-3,019.13
$\omega_1 \omega_2 \omega_3$	0.00 1.00 8.31	0.00 1.00 9.40	0.00 1.00 10.01	0.15 1.22 7.50	0.04 0.91 8.62	0.04 0.71 9.43
$p_1 p_2 p_3$	0.39 0.56 0.04	0.67 0.29 0.05	0.64 0.32 0.05	0.70 0.26 0.03	0.69 0.26 0.05	0.60 0.35 0.05
P2						
lnL	-3,672.49	-3,652.61	-3,651.67	-3,658.85	-3,652.30	-3,651.23
$\omega_1 \omega_2 \omega_3$	0.00 1.00 4.39	0.00 1.00 3.86	0.00 1.00 4.47	0.15 1.14 3.85	0.06 1.36 4.23	0.03 0.49 3.98
$p_1 p_2 p_3$	0.30 0.62 0.07	0.57 0.33 0.10	0.55 0.38 0.08	0.58 0.37 0.06	0.65 0.28 0.07	0.46 0.42 0.13
P3						
lnL	-3,205.90	-3,171.99	-3,169.07	-3,184.05	-3,165.13	-3,162.90
$\omega_1 \omega_2 \omega_3$	0.00 1.00 5.20	0.00 1.00 5.07	0.00 1.00 14.17	0.19 2.10 5.95	0.00 2.92 9.99	0.00 2.83 13.82
$p_1 p_2 p_3$	0.36 0.49 0.15	0.71 0.15 0.14	0.75 0.20 0.05	0.73 0.22 0.05	0.78 0.18 0.03	0.79 0.19 0.02
P5						
lnL	-3,889.82	-3,819.30	-3,817.56	-3,838.40	-3,816.79	-3,815.98
$\omega_1 \omega_2 \omega_3$	0.00 1.00 11.88	0.00 1.00 10.01	0.00 1.00 10.44	0.14 1.04 7.34	0.05 1.71 11.51	0.05 1.39 10.80
$p_1 p_2 p_3$	0.35 0.62 0.04	0.73 0.23 0.03	0.71 0.26 0.03	0.77 0.20 0.04	0.84 0.14 0.02	0.79 0.18 0.03
--						
P7						
lnL	-4,121.97	-4,060.46	-4,057.37	-4,084.47	-4,050.26	-4,049.37
$\omega_1 \omega_2 \omega_3$	0.00 1.00 8.40	0.00 1.00 11.61	0.00 1.00 11.81	0.32 2.70 11.84	0.19 3.29 14.56	0.17 3.07 15.09
$p_1 p_2 p_3$	0.25 0.63 0.12	0.61 0.32 0.07	0.58 0.35 0.07	0.79 0.17 0.04	0.83 0.13 0.04	0.81 0.14 0.05
P8						
lnL	-4,174.14	-4,098.80	-4,092.67	-4,136.79	-4,095.89	-4,090.22
$\omega_1 \omega_2 \omega_3$	0.00 1.00 5.34	0.00 1.00 9.20	0.00 1.00 15.05	0.10 1.03 4.17	0.03 1.41 9.93	0.05 1.06 14.85
$p_1 p_2 p_3$	0.38 0.53 0.09	0.68 0.27 0.05	0.68 0.29 0.03	0.64 0.28 0.07	0.74 0.22 0.04	0.71 0.26 0.03
--						

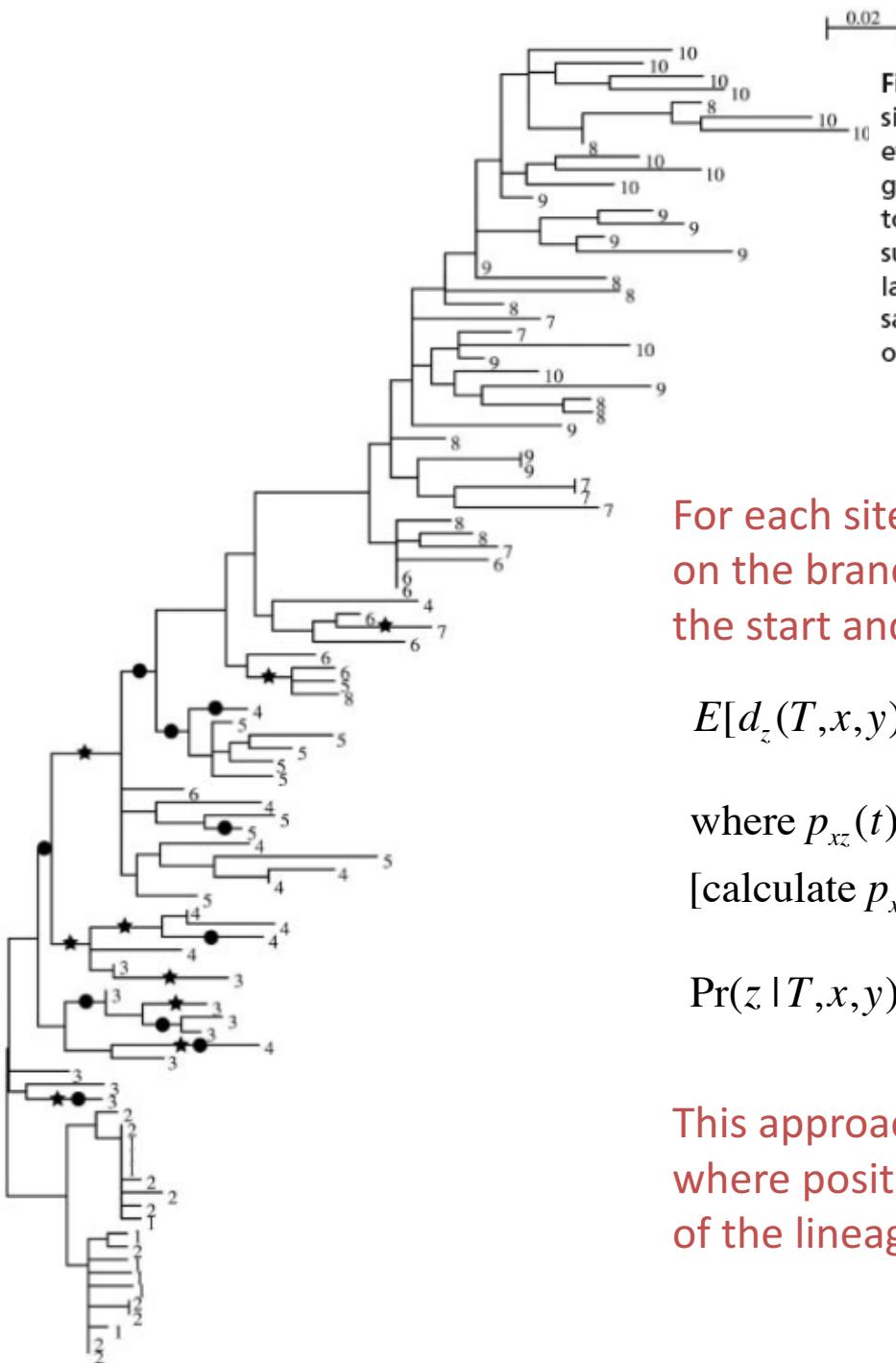


Fig. 1. Phylogenetic positions of substitutions inferred at two amino acid sites of patient 6 data set. M3 strongly supports the hypothesis that sequences evolved under positive selection at these sites, whereas the statistical support given by M3 + S1 to the same hypothesis is less important. ★ and ● correspond to the substitutions inferred at sites 41 and 180, respectively. All of these substitutions are likely to be nonsynonymous. The leaves of the tree are labeled with the rank of the corresponding sample time (1 is the earliest sample and 10 is the latest). The position of the root was determined by using outgroup sequences collected during the earliest stages of the infection.

For each site, the expected time spent in selection class z on the branch of length T , which had selection regime x at the start and y at the end:

$$E[d_z(T,x,y)] = \int_0^T \frac{p_{xz}(t)p_{zy}(T-t)}{p_{xy}(T)} dt$$

where $p_{xz}(t)$ is the probability of change $x \rightarrow z$ over time t
[calculate $p_{xz}(t)$ from $P_R(t) = \exp(tR)$]

$$\Pr(z | T, x, y) = E[d_z(T, x, y)]/T$$

This approach is used to detect sites in the alignment where positive selection is likely to have occurred in most of the lineages

Two decades of large-scale selection scans

Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios

Andrew G. Clark,¹ Stephen Glanowski,³ Rasmus Nielsen,^{1,2*} Paul D. Thomas,⁴ Anish Kejariwal,⁴ Melissa A. Aronowicz,⁵ David M. Tanenbaum,⁵ Daniel Civello,⁶ Fu Lu,⁵ Brian K. O'Connor,⁷ Steve Ferriera,³ Gary Wang,³ Xianqun Zhou,³ Thomas J. White,⁶ John J. Sninsky,⁶ Mark D. Adams,⁵ Michele Cargill^{6,†}

2003, Science

Open access, freely available online

PLOS BIOLOGY

2005

A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees

Rasmus Nielsen^{1,2*}, Carlos Bustamante¹, Andrew G. Clark³, Stephen Glanowski⁴, Timothy B. Sackton³, Melissa J. Hubisz¹, Adi Fledel-Alon¹, David M. Tanenbaum⁵, Daniel Civello⁶, Thomas J. White⁶, John J. Sninsky⁶, Mark D. Adams^{5✉}, Michele Cargill⁶

OPEN ACCESS Freely available online

PLOS GENETICS

2008

Patterns of Positive Selection in Six Mammalian Genomes

Carolin Kosiol¹, Tomáš Vinař¹, Rute R. da Fonseca², Melissa J. Hubisz³, Carlos D. Bustamante¹, Rasmus Nielsen², Adam Siepel^{1*}

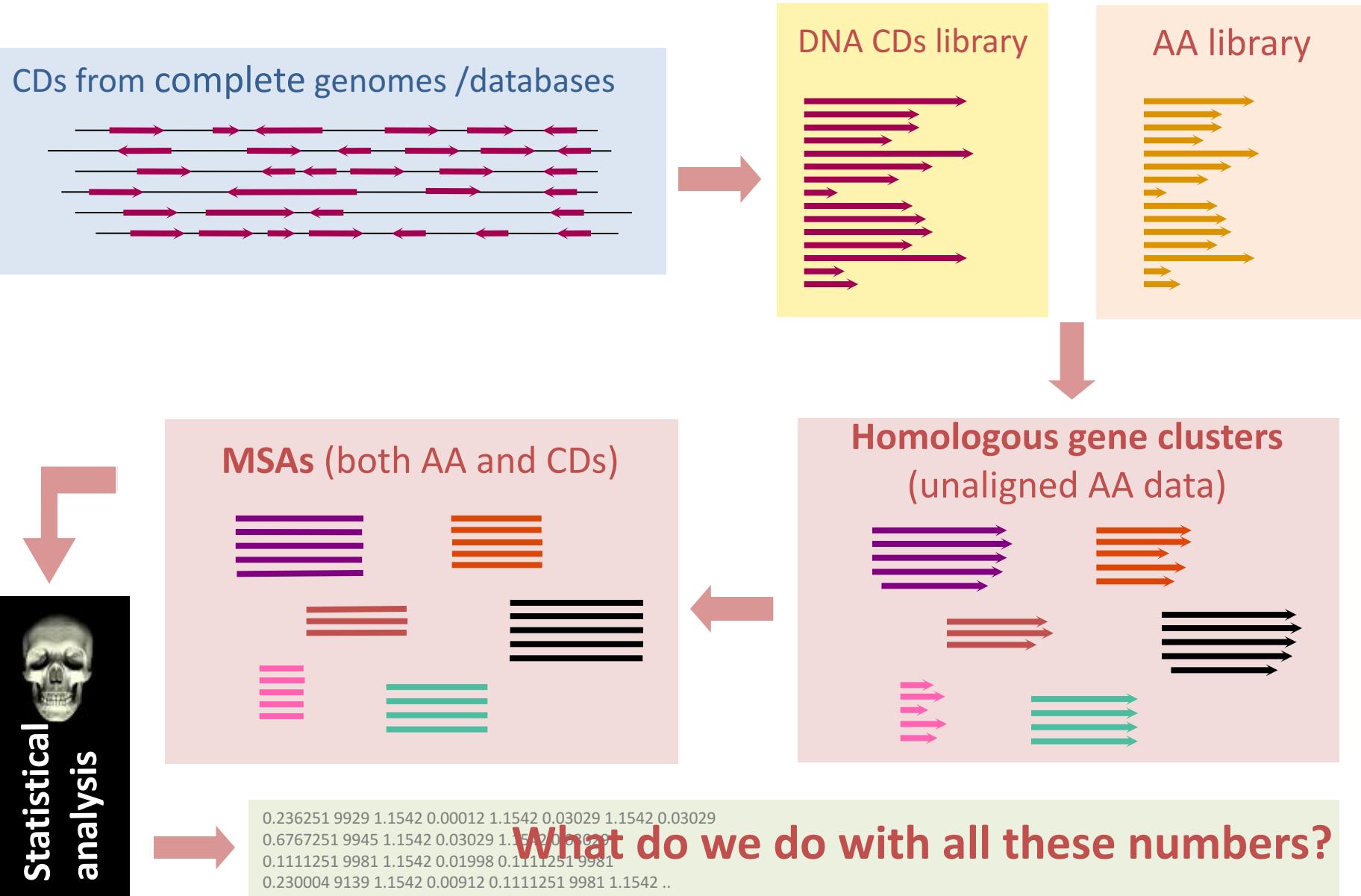
Research article

A systematic search for positive selection in higher plants (Embryophytes)

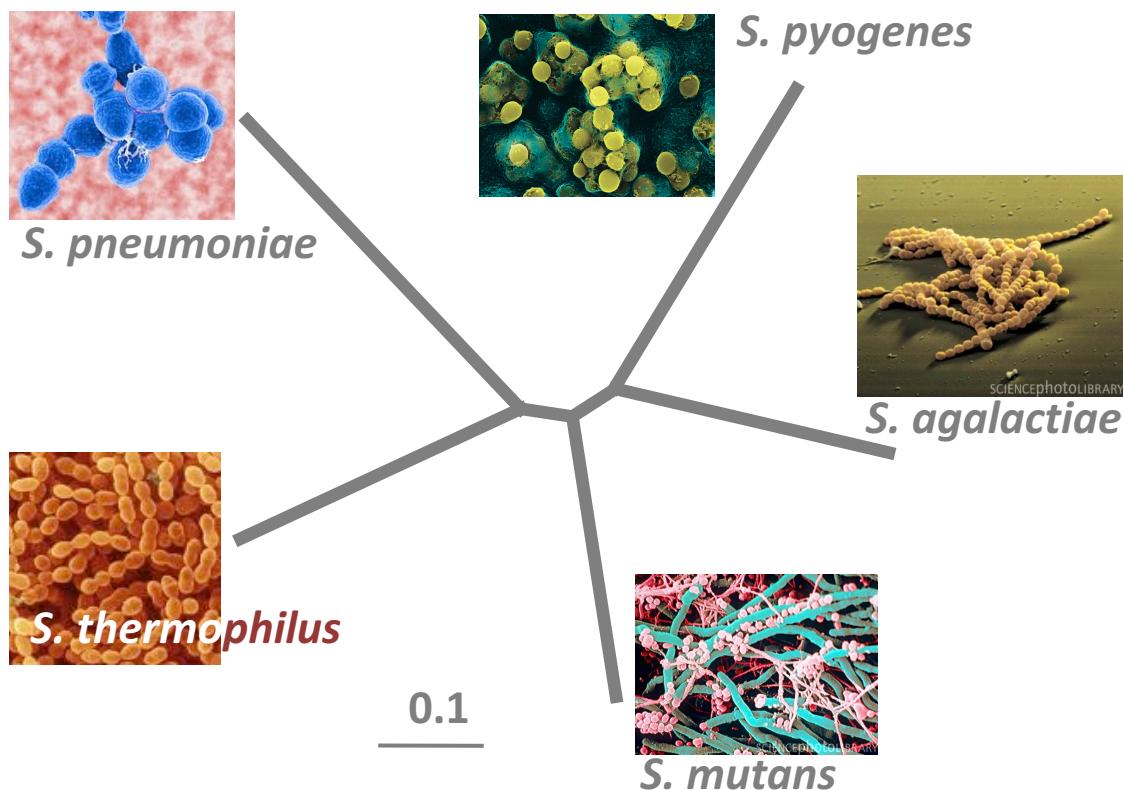
Christian Roth^{1,2,3} and David A Liberles *^{1,3}

2006

Large-scale selection scans step-by-step



Natural selection in Streptococcus



Anisimova et al 2007 BMC Evol Biol

12 complete genomes

Positive selection in 136 genes:

29% connected to virulence

10% no ascribable function

7% essential to *S. pneumoniae*

19% with body-site specific patterns of gene expression during invasive disease in *S. pyogenes* (infected blood, cerebrospinal fluid, epithelial cell contact)

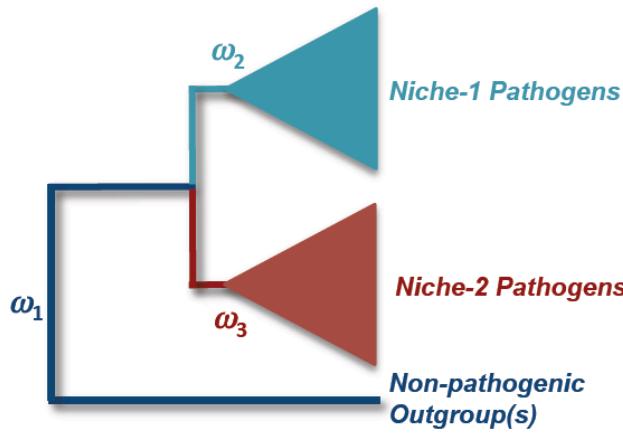
Positive selection affects both core and accessory genes, most likely due to the antagonistic interaction between host and parasite.

Products of both core and auxiliary genes participate in complex networks that comprise the molecular basis of virulence.

Listeria phylogenomics

Mapping selection to phenotype

A: Gene-level data analysis



Null model (1 parameter):

$$H_0 : \omega_1 = \omega_2 = \omega_3$$

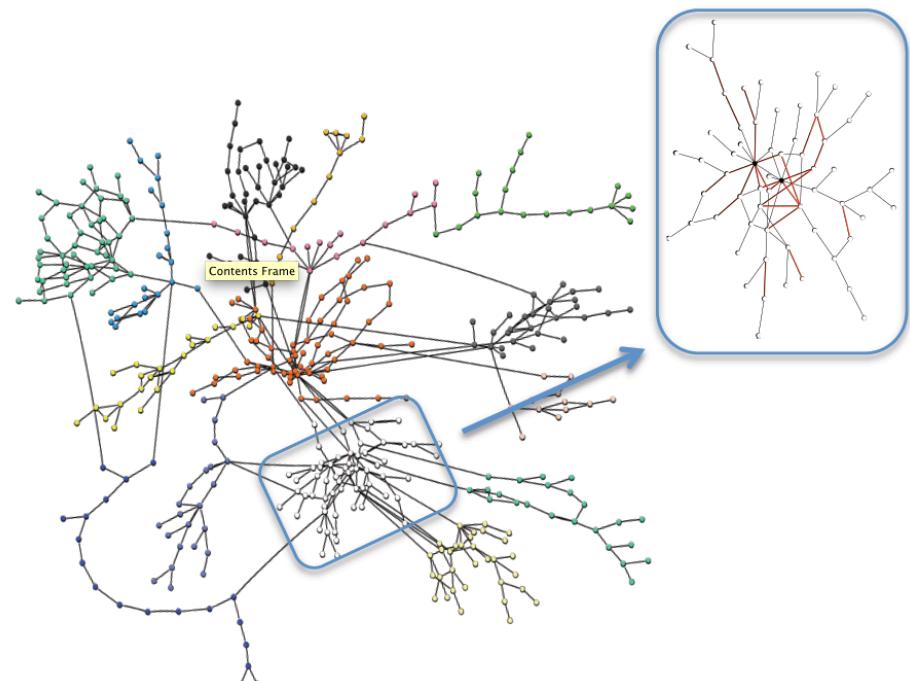
Alternatives (2 parameters):

$$H_1 : \omega_1 \neq \omega_2 = \omega_3$$

$$H_2 : \omega_1 = \omega_2 \neq \omega_3$$

$$H_3 : \omega_1 = \omega_3 \neq \omega_2$$

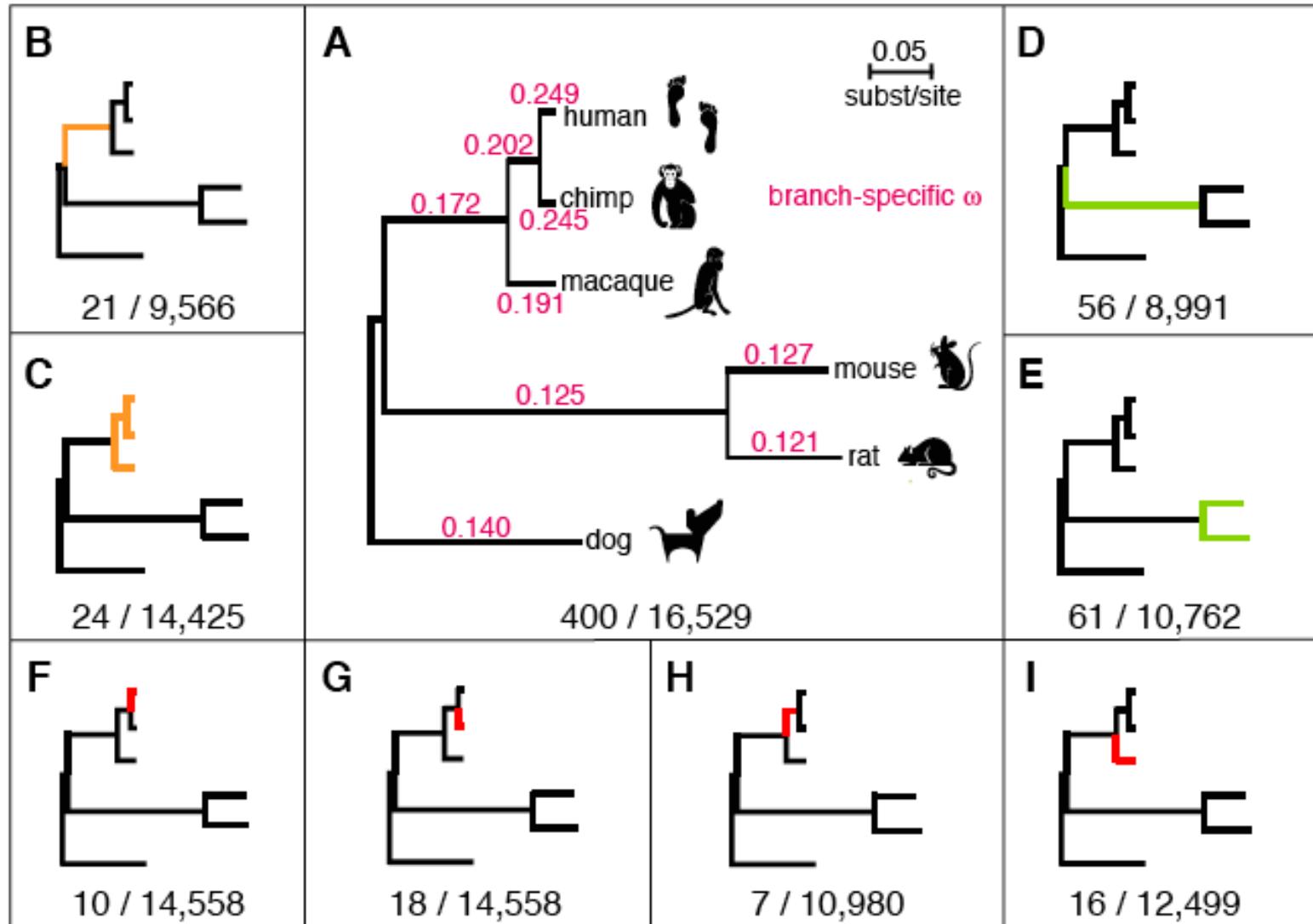
B: Phenotype-level data analysis



Blue box identifies a module in the metabolic network. Red links in the expanded view of this module indicate a significant cluster of genes subject to niche specific selection in “lineage I” of *L. monocytogenes*.

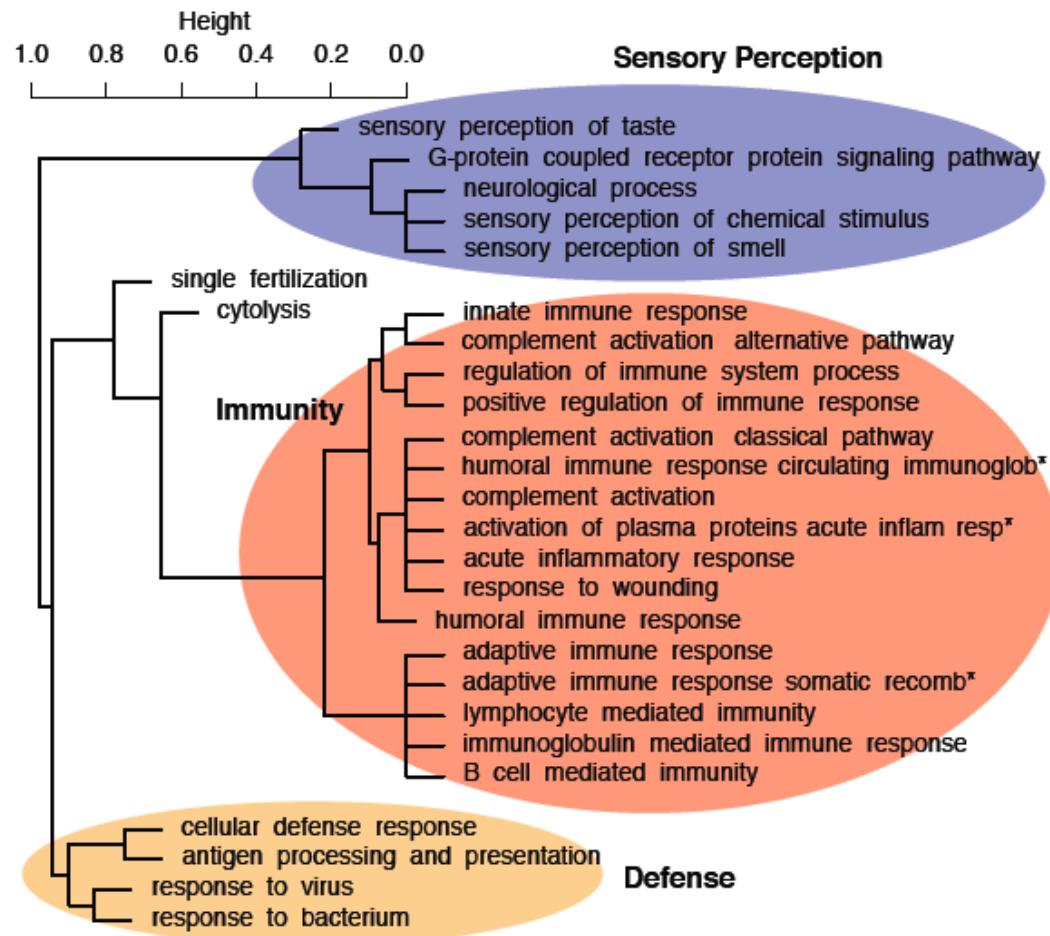
Multiple LRTs: scan of mammalian genomes

Kosiol et al (2008)



Multiple LRTs: scan of mammalian genomes

Kosiol et al (2008)



Dissimilarity measure:

$$d_{AB} = 1 - \frac{|N(A) \cap N(B)|}{\min\{|N(A)|, |N(B)|\}}$$

Hierarchical clustering of GO categories (biological process) over-represented with genes under positive selection

Which proteins are under positive selection?

- Host proteins involved in defence or immunity against viral, bacterial, fungal or parasite attacks (MHC, immunoglobulin VH, class 1 chitinases).
- Viral or pathogen proteins involved in evading host defence (HIV env, nef, gap, pol, etc., capsid in FMD virus, flu virus hemagglutinin gene).
- Proteins or pheromones involved in reproduction (abalone sperm lysin, sea urchin bindin, proteins in mammals)
- Proteins that acquired new functions after gene duplication.
- Miscellaneous (diet, globins,etc.)

Conclusions

Detecting positive selection

- Pairwise methods – very low power
- Branch models allow variation over time but assume one ω for all sites - low power
- Site models allow variation among sites but assume selection pressure does not change over time – have higher power if positive selection is long term
- Branch-site models may be more successful at detecting episodic selection but are more difficult to fit, require more data and often have multiple sub-optimal peaks (caution with genome scans!)

Testing for positive selection

- LRT is accurate even for small datasets
- Power of LRT is better for larger datasets
- Watch out for recombination
- Accurate parameter estimation is more difficult, depends on model assumptions
- Bayesian site prediction is even more difficult than LRTs and parameter estimation
- There is an optimal window of sequence divergence (sequences should be not too similar and not saturated)
- Robustness of results: Use several models & tests
- Check for local optima, especially for complex models

Weaknesses of methods based on codon models

- Model assumptions may be unrealistic (but some assumptions matter more than others)
- The method detects positive selection only if it generates excessive nonsynonymous substitutions. It may lack power in detecting one-off directional selection or when the sequences are highly similar or highly divergent. Little power with population data.
- Do not work for noncoding DNA (but see Wong & Nielsen 2003 Genetics)
- Sensitive to sequence and alignment errors (Fletcher & Yang 2010 Mol Biol Evol 27; Privman et al. 2011 Mol Biol Evol 29; Jordan & Goldman 2012 Mol Biol Evol 29)

Criticisms on codon models

by M. Nei, Y. Suzuki, & A.L. Hughes

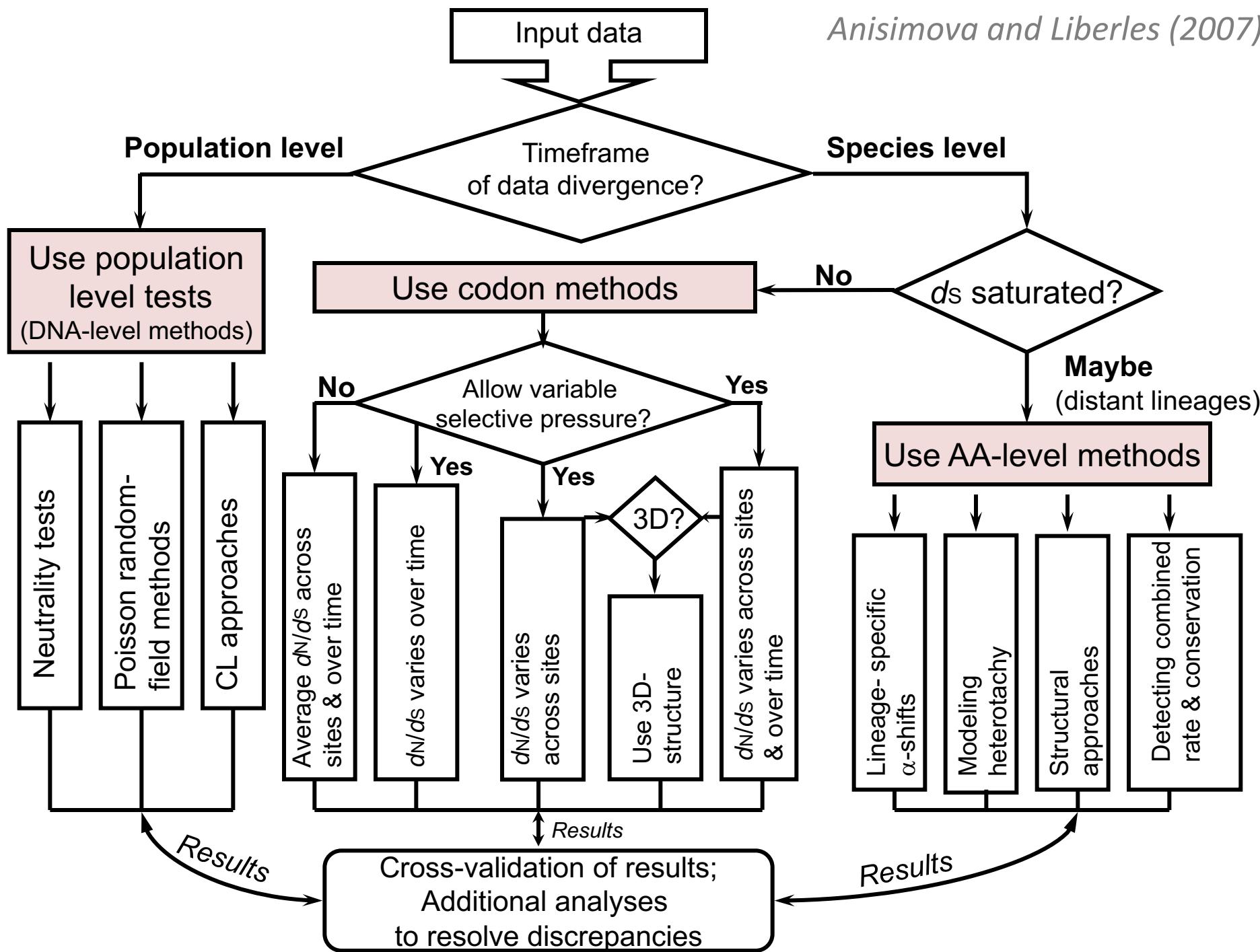
Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99

Nozawa, Suzuki & Nei. 2009. PNAS 106

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol* 28

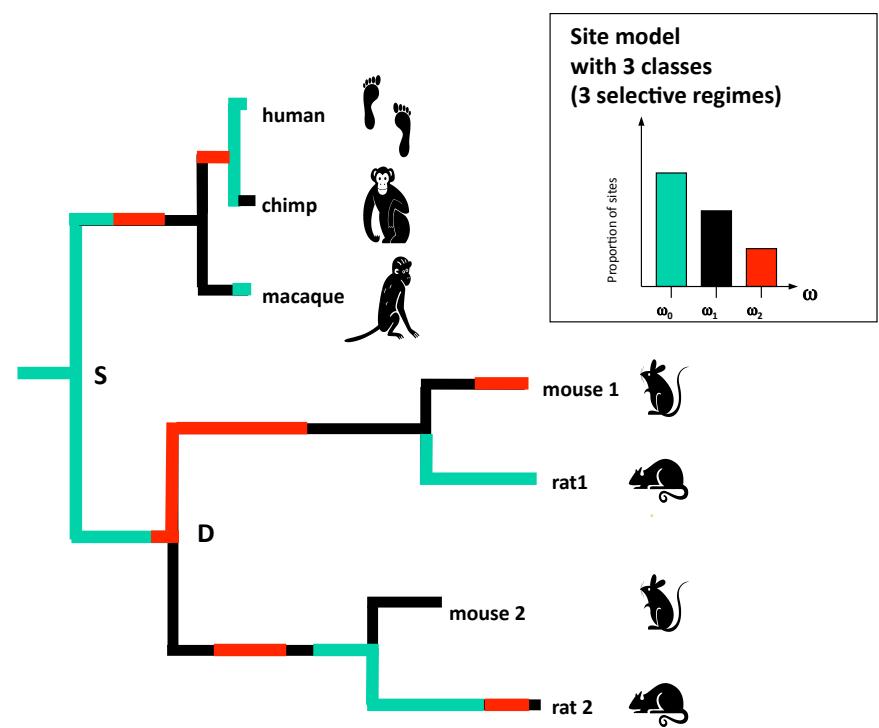
Zhai W, Nielsen R, Goldman N, Yang Z. 2012. Looking for Darwin in genomic sequences - validity and success of statistical methods. *Mol Biol Evol* 29

MacCallum, C. & Hill, E. 2006 Being positive about selection. *PLoS Biol* 4, e87



The many faces of codon models

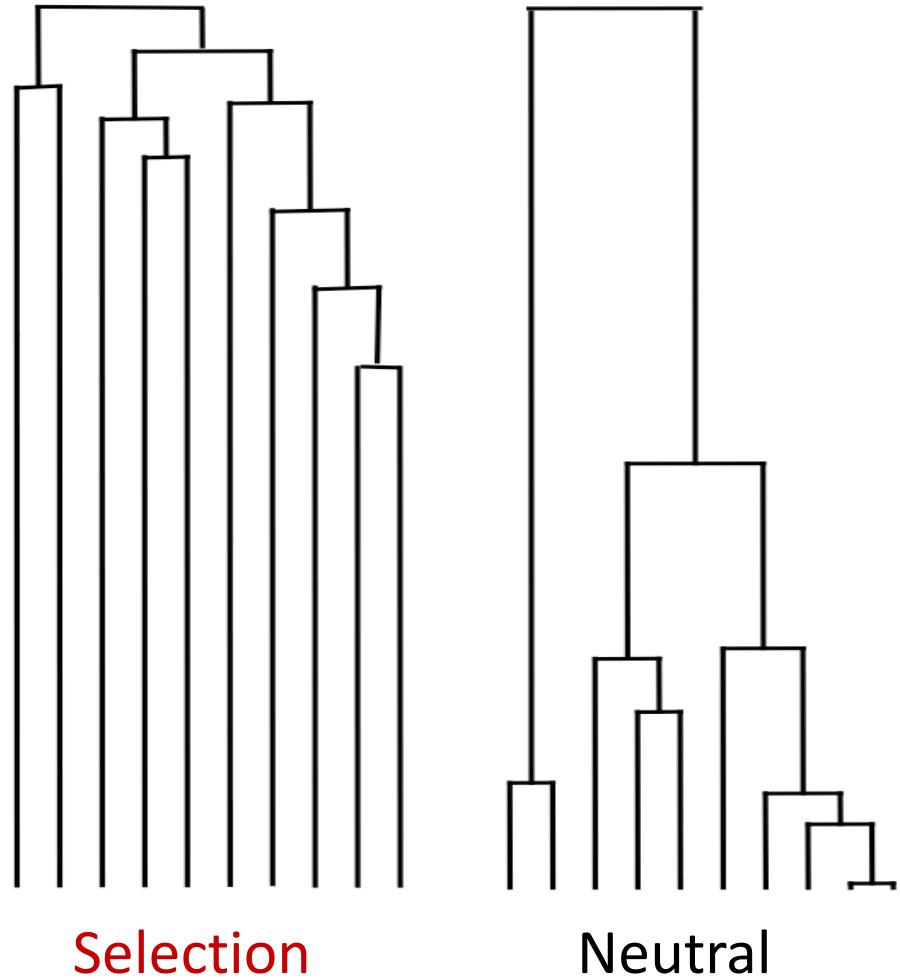
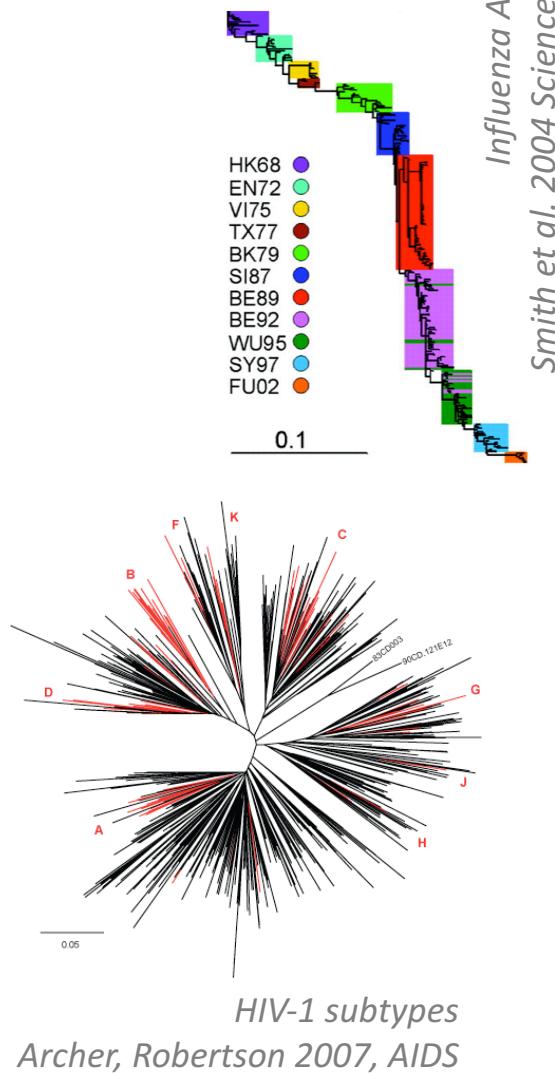
- Detecting selection
- Studying codon bias
- Inferring phylogenies
- Dating speciation events
- Ancestral reconstruction
- Changes in time & space
- Predicting coding regions
- Improved alignment
- Inferring gene features
(phyloHMM, netHMM)
- Simulation of data

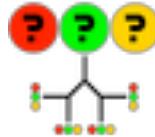


Markov modulated model:
Guindon et al. 2004

Reviews of codon models:
Kosiol and Anisimova 2012
Anisimova and Kosiol 2009

Selection affects the shape of tree





CodonPhyML : maximum likelihood tree inference

Hundreds of codon models

- Parametric, empirical, semi-parametric
- Comparable likelihoods across AA, DNA, codon data

High performance computing

- BLAS, LAPACK, OpenMP
- Heuristic using $\exp(Qt)$ via Taylor
- Blocking heuristic (FixQ)

Anisimova, Gascuel 2006 Syst Biol

Guindon et al. 2010 Syst Biol

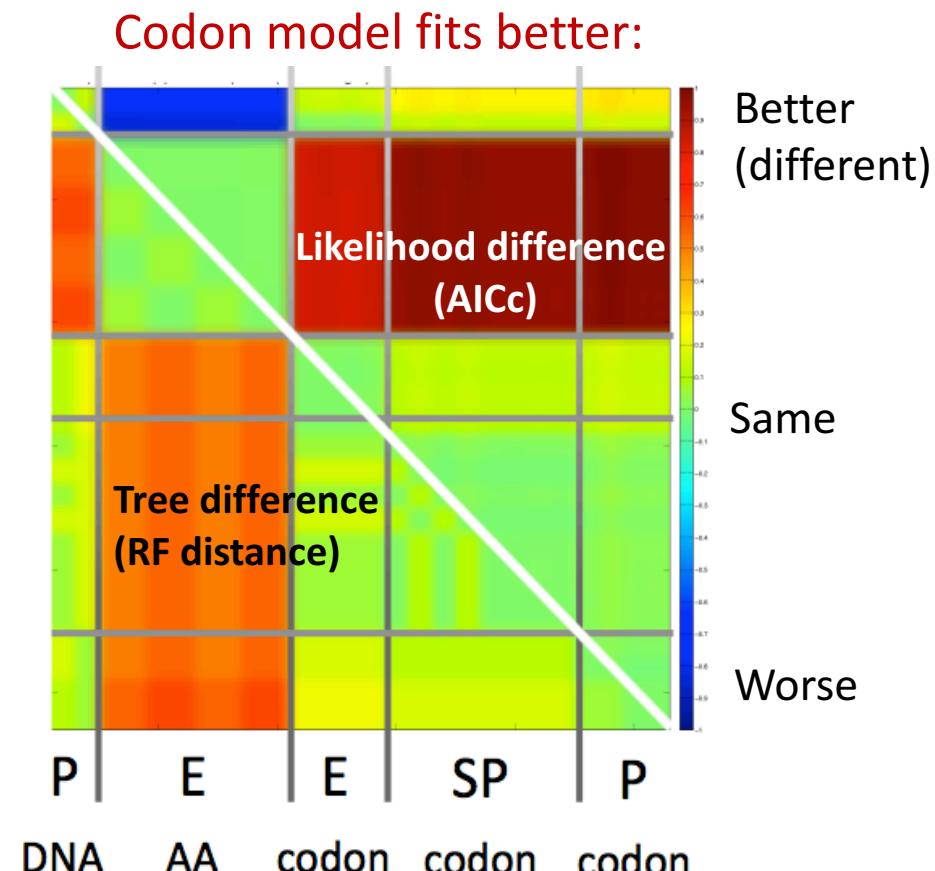
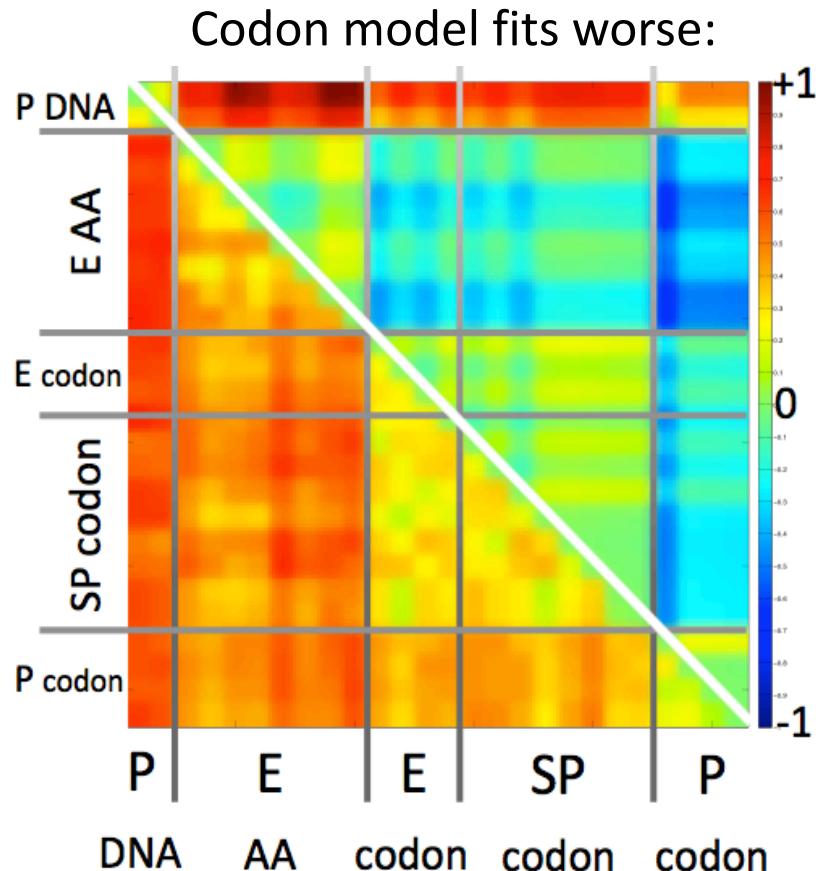
Anisimova et al. 2011 Syst Biol

Gil et al. 2013 Mol Biol Evol

CodonPhyML: Model & tree comparison on real data

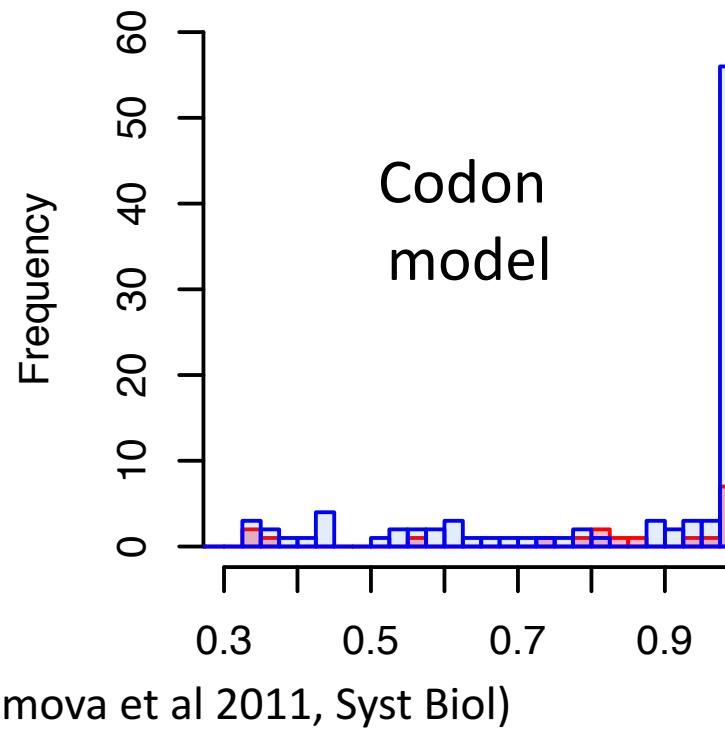
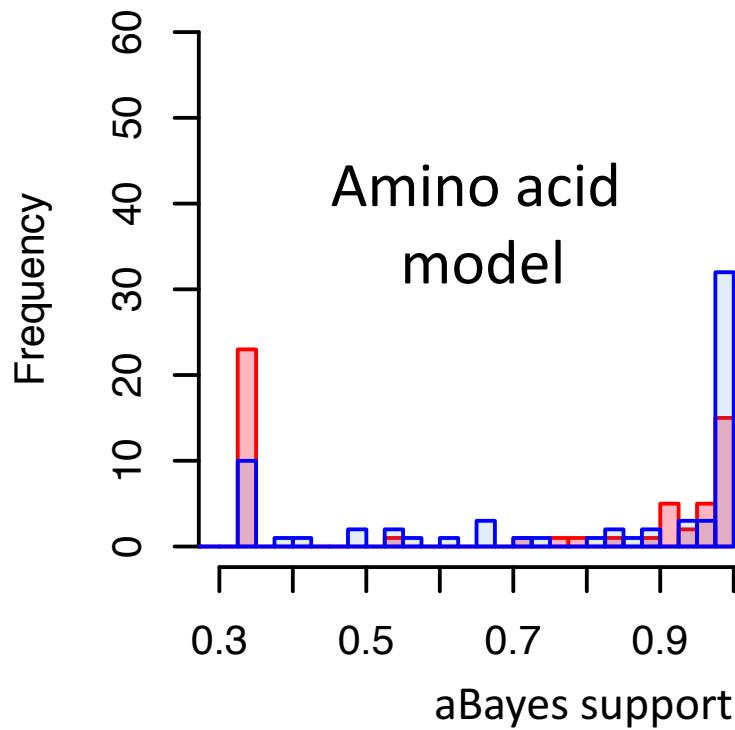
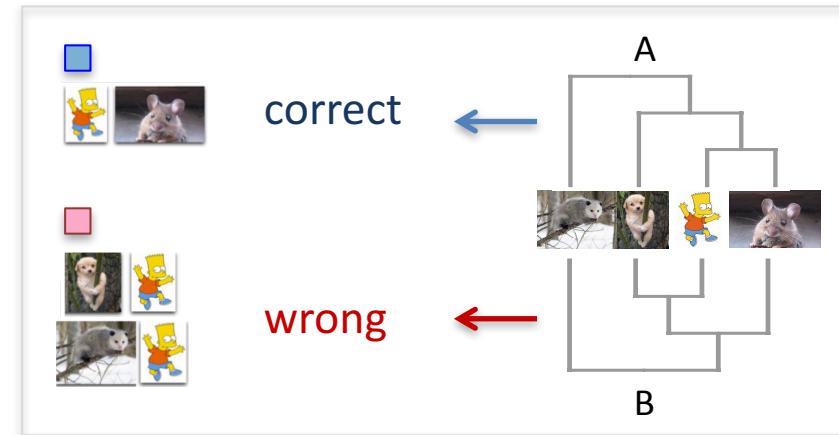
Model types: DNA, AA, codon

E = empirical, SP = semi-parametric, P = parametric



CodonPhyML: evaluating inferred splits

22 mammalian species
72 protein orthologs



<http://sourceforge.net/projects/codonphym>

Summary Files Reviews Support Wiki Discussion Donate Code Bugs

 codonPhyML
anisimova, laduplessis, mgil_, mszanetti, stefanzoller

6 Recommendations
85 Downloads (This Week)
Last Update: 4 hours ago

Download source code

Tweet 0 Like 0 Browse All Files

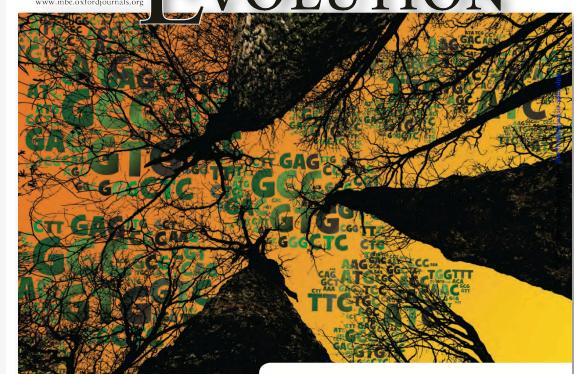
Screenshot of the codonPhyML software interface showing command-line input and graphical user interface windows.

Description

codonPhyML uses Markovian codon models of evolution in phylogeny reconstruction. Given a set of species characterized by their DNA sequences as input, codonPhyML will return the phylogenetic tree that best describes their evolutionary relationship. Our paper describing codonPhyML has been accepted for publication in the journal "Molecular Biology and Evolution". For more details, follow the link:
<http://mbe.oxfordjournals.org/content/early/2013/02/23/molbev.mst034.short>

Volume 30 • Number 8 • August 2013

MOLECULAR BIOLOGY AND EVOLUTION



Society for Molecular Biology and Evolution

Print ISSN 0737-4038
Online ISSN 1537-1719

- Effects of domestication on brain expressions in dogs
- Epistasis among antibiotic resistance variations
- Evolution of duplicated genes
- Experimental method for finding transcription start sites
- High-altitude adaptations in Ethiopians and Tibetans



A black and white photograph of a tree trunk and branches against a background of DNA sequence text. The background is filled with a dense pattern of green and blue DNA codons (e.g., ATG, TCG, GCA) on a yellowish-green gradient. A large, dark, gnarled tree trunk and branches are silhouetted against this background. In the lower right foreground, a hand holds a piece of orange paper with the word "Questions?" written in large, bold, black letters.

Questions?

Remaining exercises

Focus:

ML estimation with branch-site models
Try out with codon tree (CodonPhyML)