http://bit.ly/veme-selection

Natural Selection in Coding Sequences.

Sergei L Kosakovsky Pond
Associate Professor
Department of Medicine
University of California San Diego
spond@ucsd.edu
www.hyphy.org/sergei

Preliminaries

- Please download and install HyPhy: http://hyphy.org/wiki/Download
- General user questions and feedback: https://github.com/veg/hyphy/issues
- Datamonkey web-app: http://www.datamonkey.org

A bit of trivia

- The theory of natural selection was first proposed by ...*Patrick Matthew*
- Matthew seemed to regard the idea as more or less self-evident and not in need of further development.
- In a stunning example of how **not** to communicate science, he published his ideas in appendices B and F of his book "*On Naval Timber and Arboriculture*" (1831).
- Unsurprisingly, his peers failed to discover his ideas in such an obscure source, and his work had no impact on the subsequent, more developed, work of Darwin and Wallace (1859).
- Do **not** emulate Patrick Matthew.

Natural Selection

- Mutation, recombination and other processes introduce variation into genomes of organisms
- The fitness of an organism describes how well it can survive/grow/function/replicate in a given environment, or how well it can pass on its genetic material to future generations

- Any particular mutation can be
  - Neutral: no or little change in fitness (the majority of genetic variation falls into this class according to the neutral theory).
  - Deleterious: reduced fitness
  - Adaptive: increased fitness

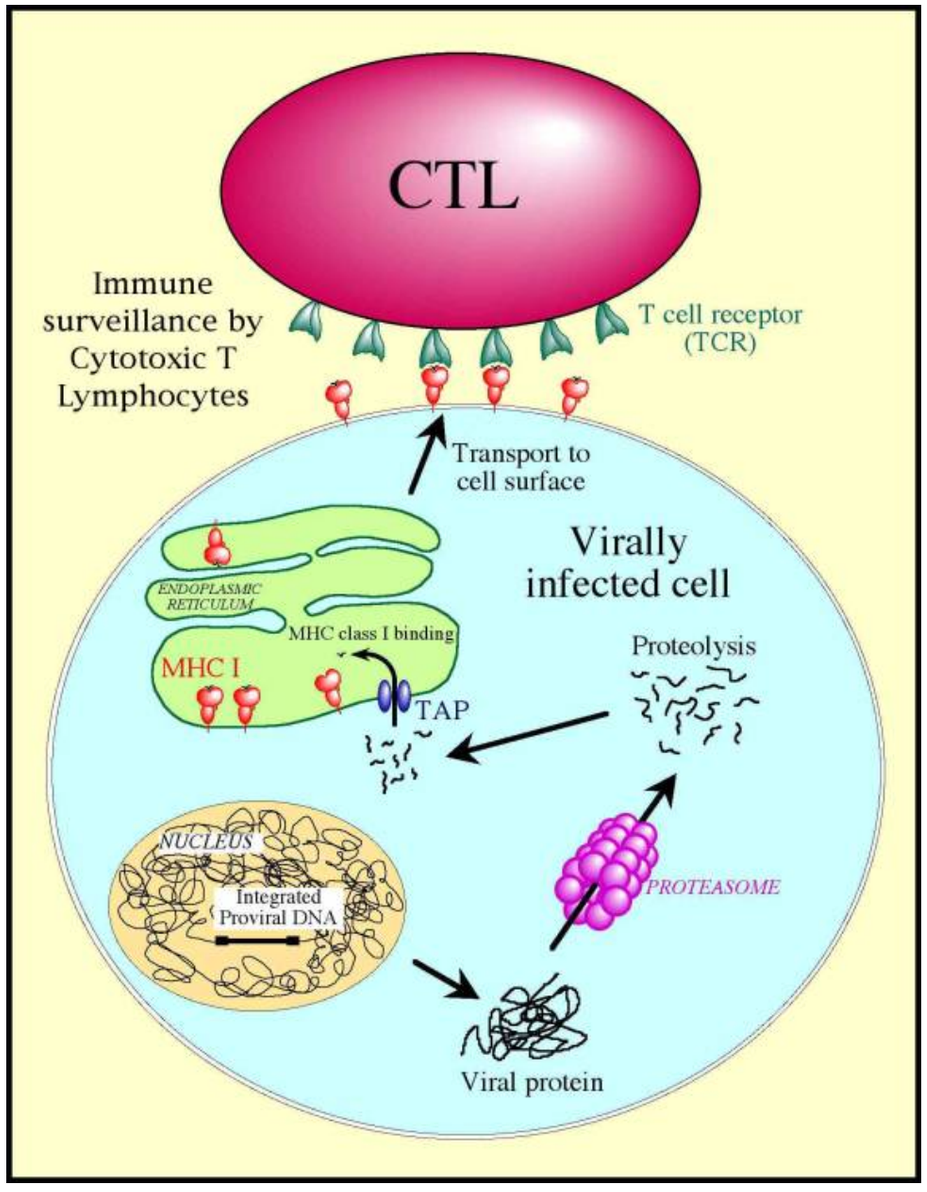
Example in HIV: MHC-restricted CTL killing

- Cytotoxic T-Lympocytes effect cell-mediated immune response

- Viral proteins are cleaved by the proteosome, transported by TAP and loaded onto the MHC Class 1 molecule.

- MHC Class 1 presents a restricted polypeptide (epitope) on the surface of the cell.

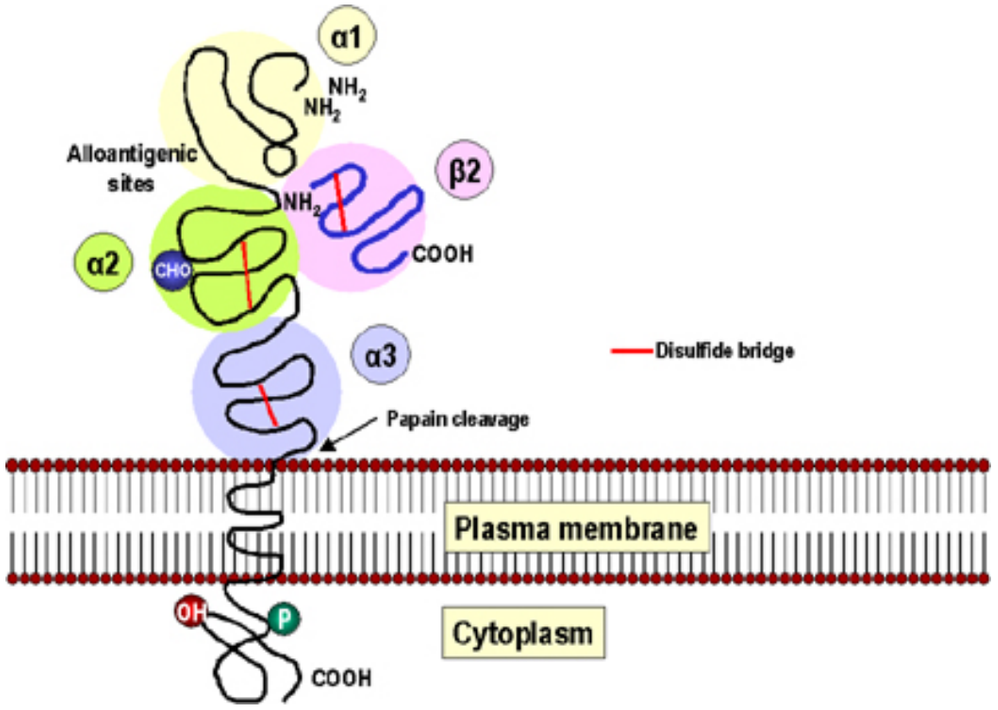
- A CD8+ cell binds to presented foreign peptides via a T cell receptor (TCR) and initiates cell apoptosis.



MHC Class 1 Molecules

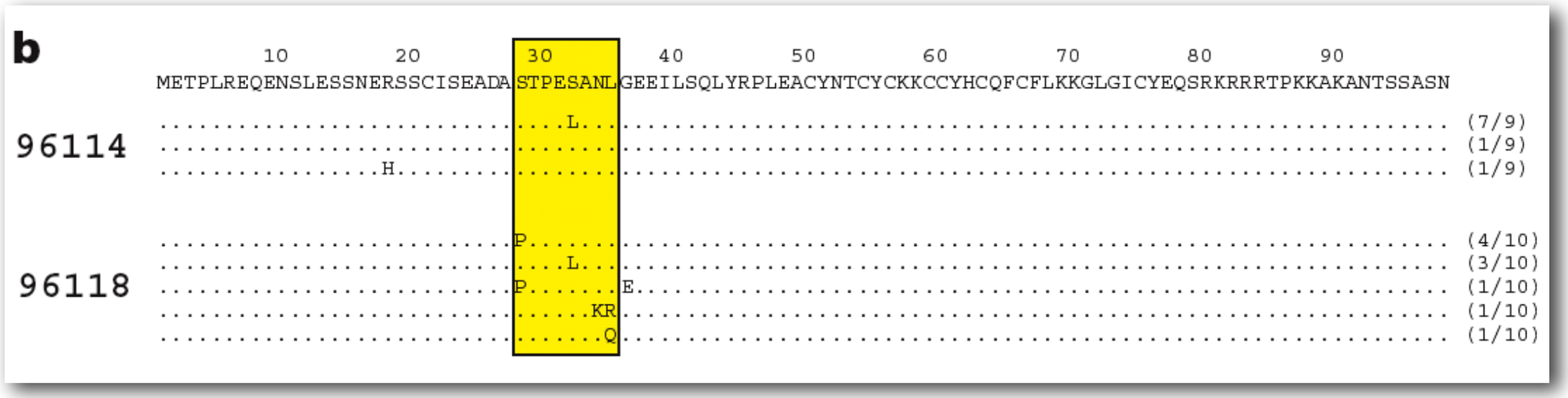
- Present foreign peptides which are 9-10 amino-acid long
- Anchor sites (2 and 9) are usually important for binding and recognition
- Mutations which alter the peptide can hinder or prevent CTL response activation

Antigen Binding Site

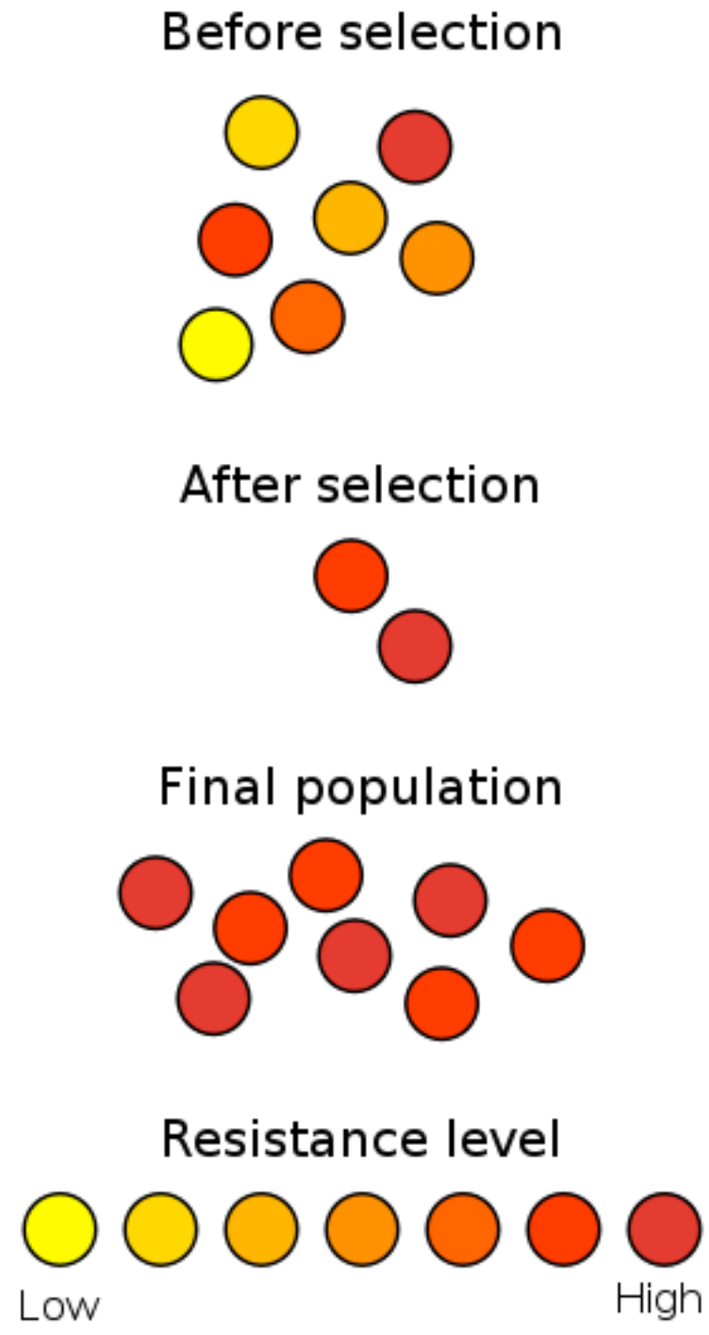


Rapid SIV sequence evolution in macaques

- SIV: the only animal model of HIV (rhesus macaques)
- Experimental infection with MHC-matched strain of SIV
- Virus sequenced from a sample 2 weeks post infection
- Only variation was in an epitope recognized by the MHC
- CTL *escape*

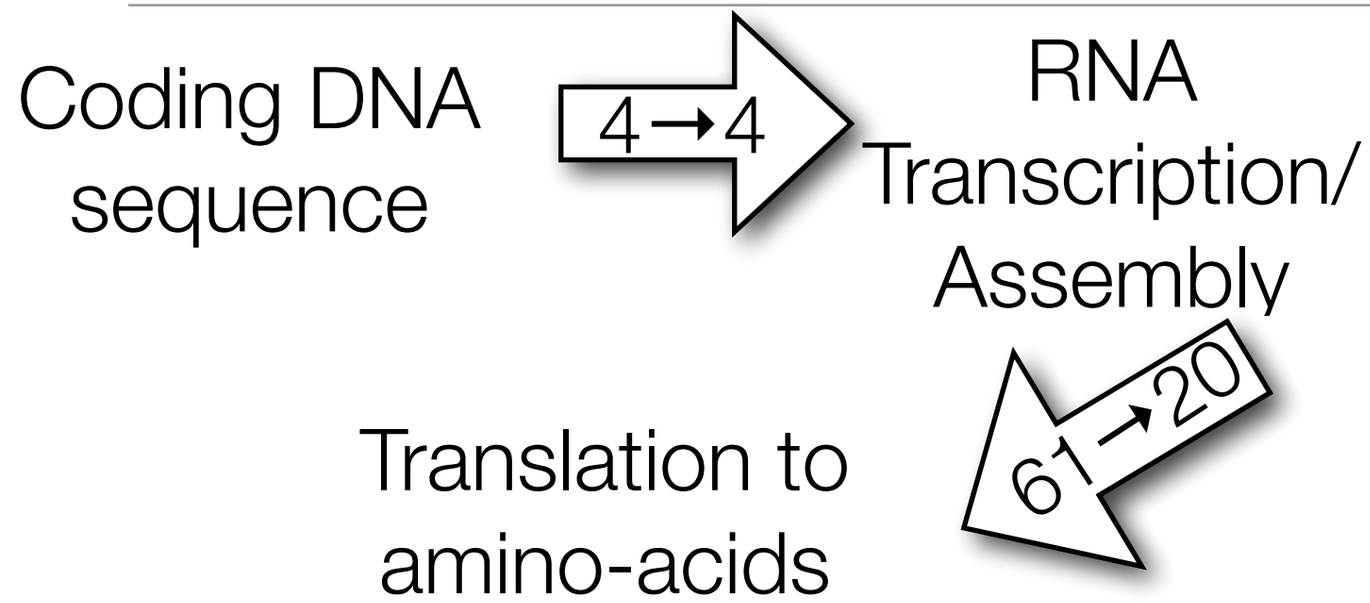


O'Connor et al (2002) Nat Med 8(5):493–499



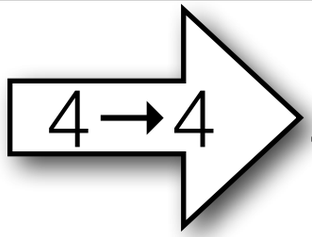
http://en.wikipedia.org/wiki/File:Antibiotic_resistance.svg

Coding Sequences.

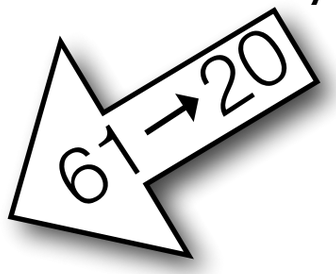


Coding Sequences.

Coding DNA sequence



RNA Transcription/ Assembly

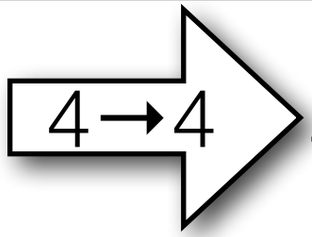


Translation to amino-acids

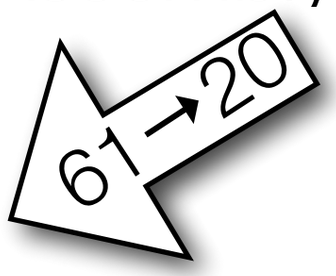
Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

Coding Sequences.

Coding DNA sequence



RNA Transcription/ Assembly



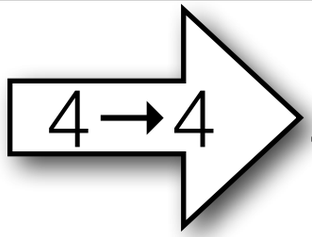
Translation to amino-acids

- Proper unit of evolution is a triplet of nucleotides – a **codon**

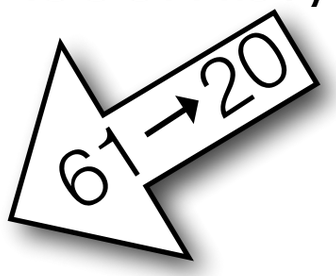
Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

Coding Sequences.

Coding DNA sequence



RNA Transcription/ Assembly



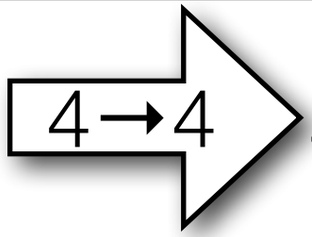
Translation to amino-acids

- Proper unit of evolution is a triplet of nucleotides – a **codon**
- Multiple and differing redundancies in the genetic code

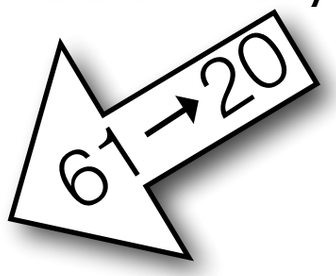
Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

Coding Sequences.

Coding DNA sequence



RNA Transcription/ Assembly



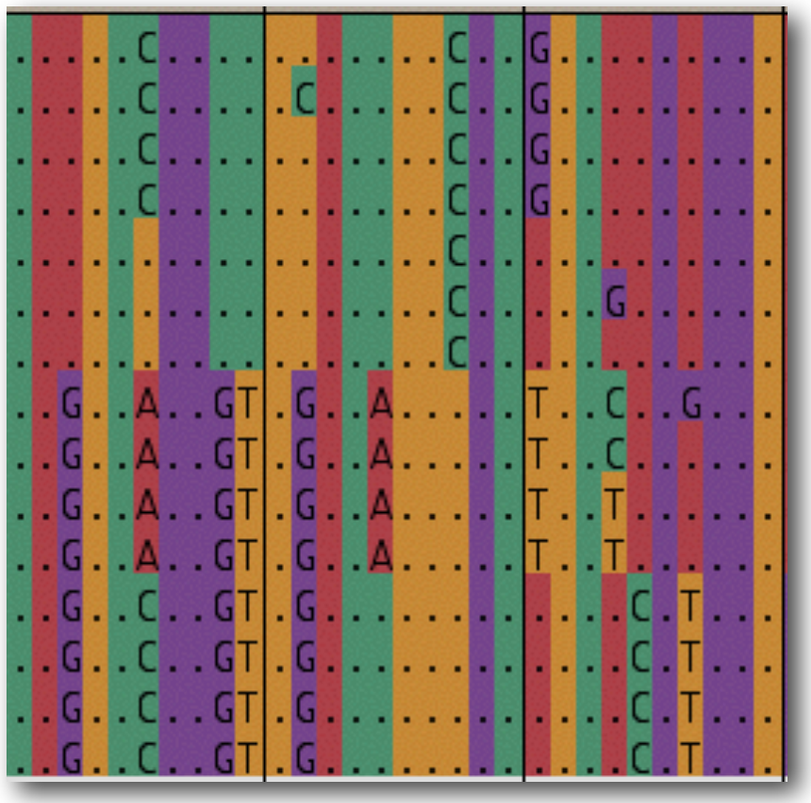
Translation to amino-acids

- Proper unit of evolution is a triplet of nucleotides – a **codon**
- Multiple and differing redundancies in the genetic code
- Synonymous and non-synonymous substitutions are fundamentally different

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

Conservation

- Measles, rinderpest, and peste-de-petite ruminant viruses nucleoprotein.

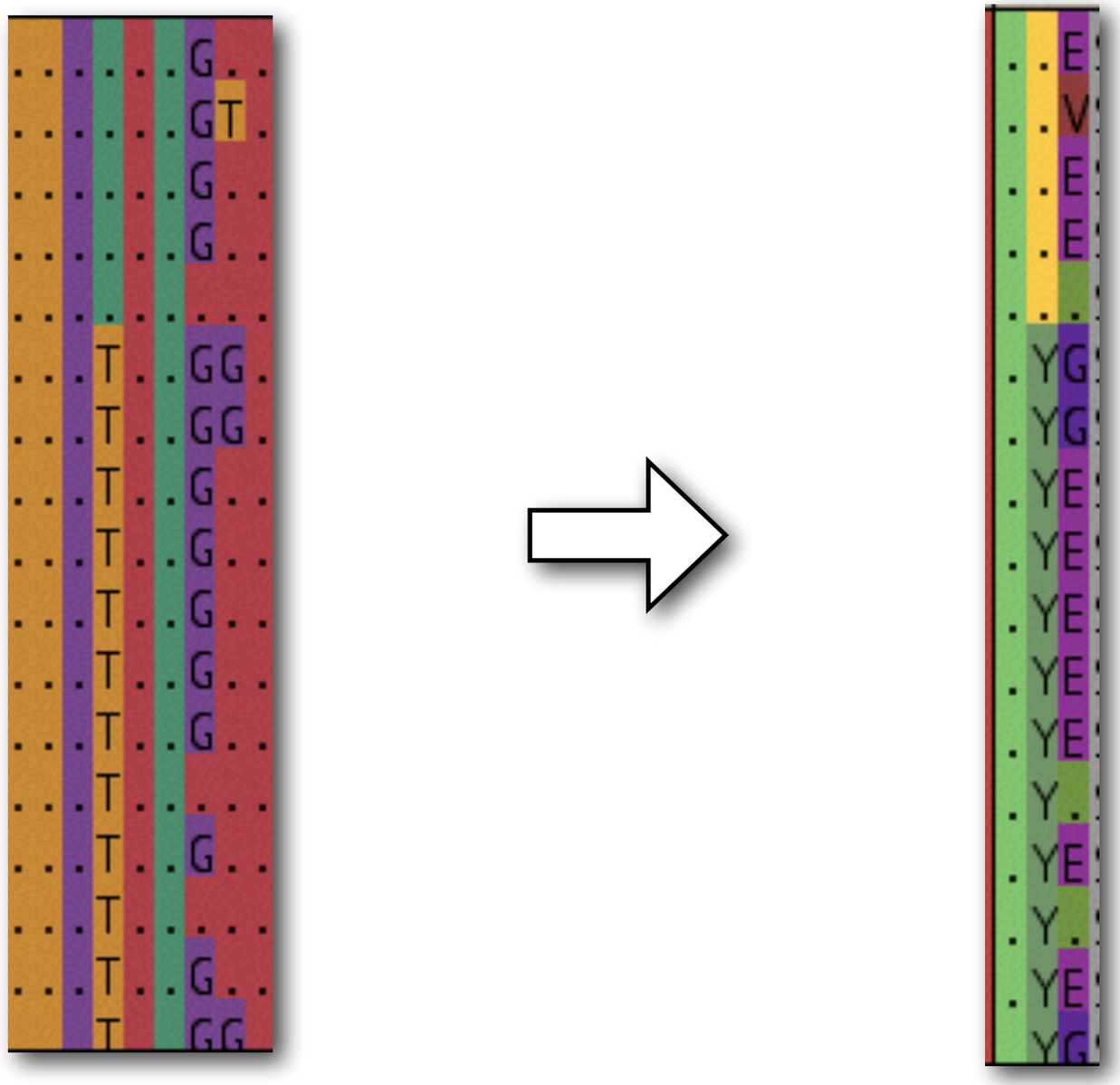


Nucleotides

Aminoacids

Diversification

- An antigenic site in H3N2 IAV hemagglutinin



Nucleotides

Aminoacids

Molecular signatures of selection

- Because synonymous substitutions do not alter the protein, we often posit that they are neutral
- The **rate** of accumulation of synonymous substitutions (**dS**) gives the neutral background
- We can compare the **rate** of accumulation of non-synonymous substitutions (**dN**), which alter the protein sequence, to classify the nature of the evolutionary process

$$dS \sim \frac{\text{number of fixed synonymous mutations}}{\text{proportion of random mutations that are synonymous}}$$

$$dN \sim \frac{\text{number of fixed non-synonymous mutations}}{\text{proportion of random mutations that are non-synonymous}}$$

Evolutionary Modes

Positive Selection (Diversifying)	dS < dN or ω := dN/dS > 1
Negative Selection	dS > dN or ω < 1
Neutral Evolution	dS ≃ dN or ω ≃ 1

Estimating dS and dN

- Consider two aligned sequences

ACA	ATA	A**T**C	TTT	AA**T**	CAA
T	*I*	***I***	*F*	*N*	*Q*

ACA	ATA	A**C**C	TTT	AA**C**	CAA
T	*I*	***T***	*F*	*N*	*Q*

Estimating dS and dN

- Consider two aligned sequences

ACA	ATA	A**T**C	TTT	AA**T**	CAA
T	*I*	***I***	*F*	*N*	*Q*

ACA	ATA	A**C**C	TTT	AA**C**	CAA
T	*I*	***T***	*F*	*N*	*Q*

Can we say that **`dN/dS = 1`**, because there is **one** synonymous and **one** non-synonymous substitution?

This genetic code has 61 sense (non-termination) codons

Substitution types

	Synonymous			Non-synonymous			To a stop codon
	Transitions	Transversions	Total	Transitions	Transversions	Total	Total
1st position:	8	0	8	140	26	166	9
2nd position:	0	0	0	148	28	176	7
3rd position:	58	68	126	2	48	50	7
Total	66	68	134	290	102	392	23

Approximately **3:1 (N/S)** ratio when mutations are fixed at random

Neutral expectation

- A random mutation is **~3 times more likely to be non-synonymous that synonymous**, depending on the variety of factors, such as codon composition, transition/transversion ratios, etc.

- We need to estimate the proportion of random mutations that are synonymous, and use it as the reference to compute **dS**.

- Synonymous and non-synonymous “sites”, mutational opportunity

GAA (Glutamic Acid)

GAA (Glutamic Acid)

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

GAA (Glutamic Acid)

Synonymous

GAG (Glutamic Acid)

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

GAA (Glutamic Acid)

Synonymous

GAG (Glutamic Acid)

Non-Synonymous

AAA (Lysine)

CAA (Glutamine)

TAA (Stop)

GCA (Alanine)

GGA (Glycin)

GTA (Valine)

GAC (Aspartic Acid)

GAT (Aspartic Acid)

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

GAA (Glutamic Acid)

Synonymous

GAG (Glutamic Acid)

Non-Synonymous

AAA (Lysine)

CAA (Glutamine)

TAA (Stop)

GCA (Alanine)

GGA (Glycin)

GTA (Valine)

GAC (Aspartic Acid)

GAT (Aspartic Acid)

8/3 non-synonymous sites

1/3 synonymous sites

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycin	GG*	4
Histidine	CAC,CAT	2
Isoleucine	ATA,ATC,ATT	3
Lysine	AAA,AAG	2
Leucine	CT*,TTA,TTG	6
Methionine	ATG	1
Aspargine	AAC,AAT	2
Proline	CC*	4
Glutamine	CAA,CAG	2
Arginine	AGA,AGG,CG*	6
Serine	AGC,AGT,TC*	6
Threonine	AC*	4
Valine	GT*	4
Tryptophan	TGG	1
Tyrosine	TAC,TAT	2
Stop	TAA,TAG,TGA	3

Nei-Gojobori dN/dS estimate

- For each codon **`C`** we define **`ES(C)`** and **`EN(C)`** - the fractions of synonymous and non-synonymous *neighbors* of a codon
  - E.g., **`ES(GAA) = 1/9, EN(GAA) = 8/9.`**
  - Can also define them as fractions of substitutions that do not lead to stop codons, e.g. **`ES(GAA) = 1/9, EN(GAA) = 7/9.`**

- The sum of **`ES`** and **`EN`** over all codons in a sequence gives an estimate of expected synonymous and non-synonymous sites in a sequence. For two sequences, we average **`ES(C)`** and **`EN(C)`** at each site.

- **`EN/ES`** is thus the expected ratio of non-synonymous to synonymous substitutions under neutral evolution

Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions
M. Nei and T. Gojobori
Mol. Biol. Evol. **3** 418--426 (1986)

Compute ES and EN

Seq1	***ACA***	***ATA***	***ATC***	***TTT***	***AAT***	***CAA***
Syn	1	2/3	2/3	1/3	1/3	1/3
NonSyn	2	7/3	7/3	8/3	8/3	7/3

Seq2	***ACA***	***ATA***	***ACC***	***TTT***	***AAC***	***CAA***
Syn	1	2/3	1	1/3	1/3	1/3
NonSyn	2	7/3	2	8/3	8/3	7/3

Mean						
Syn	1	2/3	5/6	1/3	1/3	1/3
NonSyn	2	7/3	13/6	8/3	8/3	7/3

`ES = 3½, EN = 14⅙`: under neutrality, would expect the ratio of non-synonymous to synonymous substitutions of **`EN/ES ~ 4`**

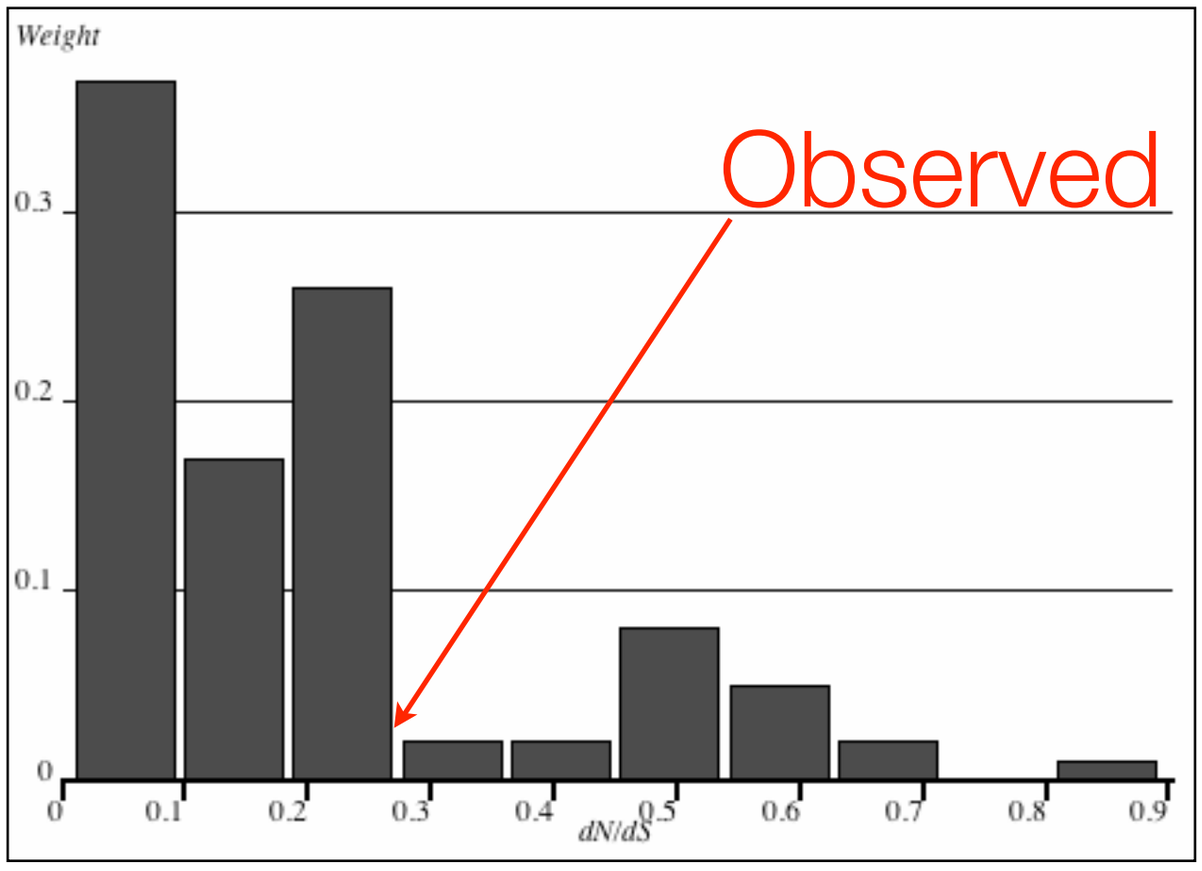
NG example

- The observed **`N/S`** ratio (**`1.0`**) is lower than the expected **`EN/ES`** ratio (**`4.05`**)
- The ratio of the ratios **`(N:S)/(EN:ES)`** yields **`dN/dS=1/4.05~0.25`**
- This ratio quantifies the excess or paucity of non-synonymous substitutions and is near one for neutrally evolving sequences/sites
- Because there are fewer non-synonymous substitutions than expected, we conclude that most non-synonymous mutations are removed by natural selection, i.e. are under **negative selection.**

NG example

- How reliable is the inference based on only 6 codons?
- Obtain sampling variance via bootstrap
- In this case, dN/dS is **significantly** less than `1.0`

Bootstrapped distribution of dN/dS




```
Count    = 100
Mean     = 0.207385
Median   = 0.166687
Variance = 0.0490168
Std.Dev  = 0.221397
COV      = 1.06757
Sum      = 20.7385
Sq. sum  = 9.15351
Skewness = 0.266313
Kurtosis = 33.381
Min      = 0
2.5%     = 0
97.5%    = 0.741176
Max      = 1
```

What about multiple substitutions?

- How many synonymous and how many non-synonymous substitutions does it take to replace `CCA` with `CAG`?

- **Assume** the shortest path
  - Option 1: `CCA (Proline)` ⟹ `CAA (Histidine)` ⟹ `CAG (Glutamine)`
  - Option 2: `CCA (Proline)` ⟹ `CCG (Proline)` ⟹ `CAG (Glutamine)`

- Average over the paths: `0.5` synonymous and `1.5` non-synonymous substitutions

- Intuitively, paths should **not** be equiprobable

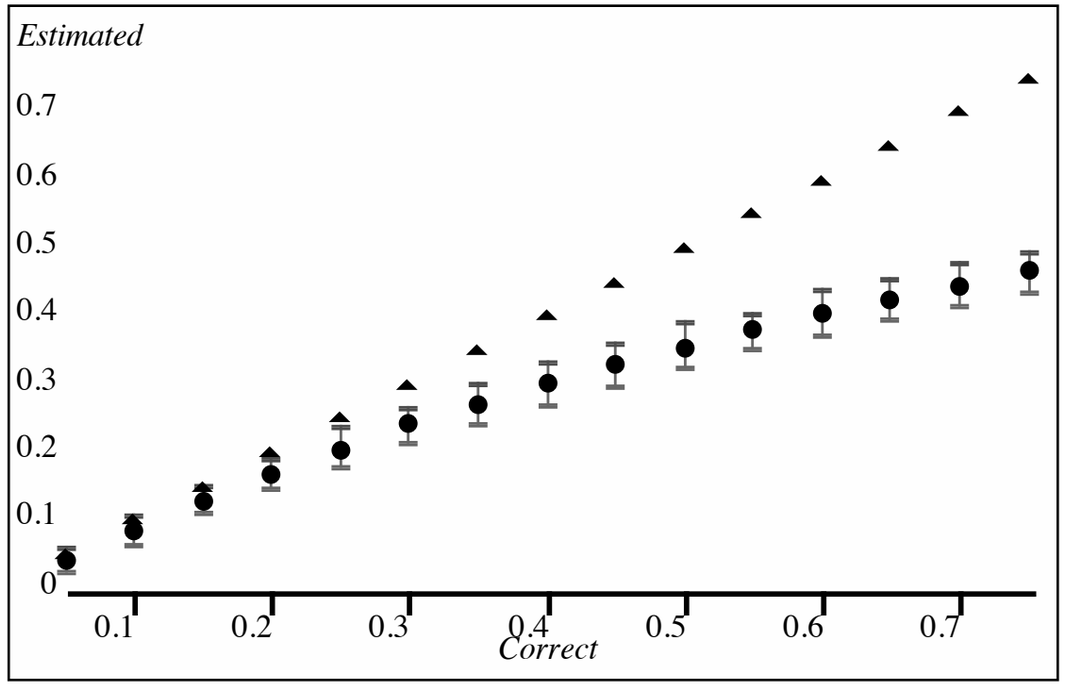
Problems with NG dN/dS: High Divergence

Substitutions = 7
p = 0.4

A T G A A A G C G A
T C
A G T A G A G T G A

Multiple hits

Reversion



- Analogous to how p-distance underestimates true divergence due to multiple hits.
- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- Plotted ‘true’ divergence vs that estimated by p-distance.
- Even for divergence of **`0.25`** (1/4 sites have mutation on average), p-distance already significantly underestimates the true level: **`0.2125 (0.19-0.241 95% range)`**
- Underestimation becomes progressively worse for larger divergence levels.

The effect of phylogenies

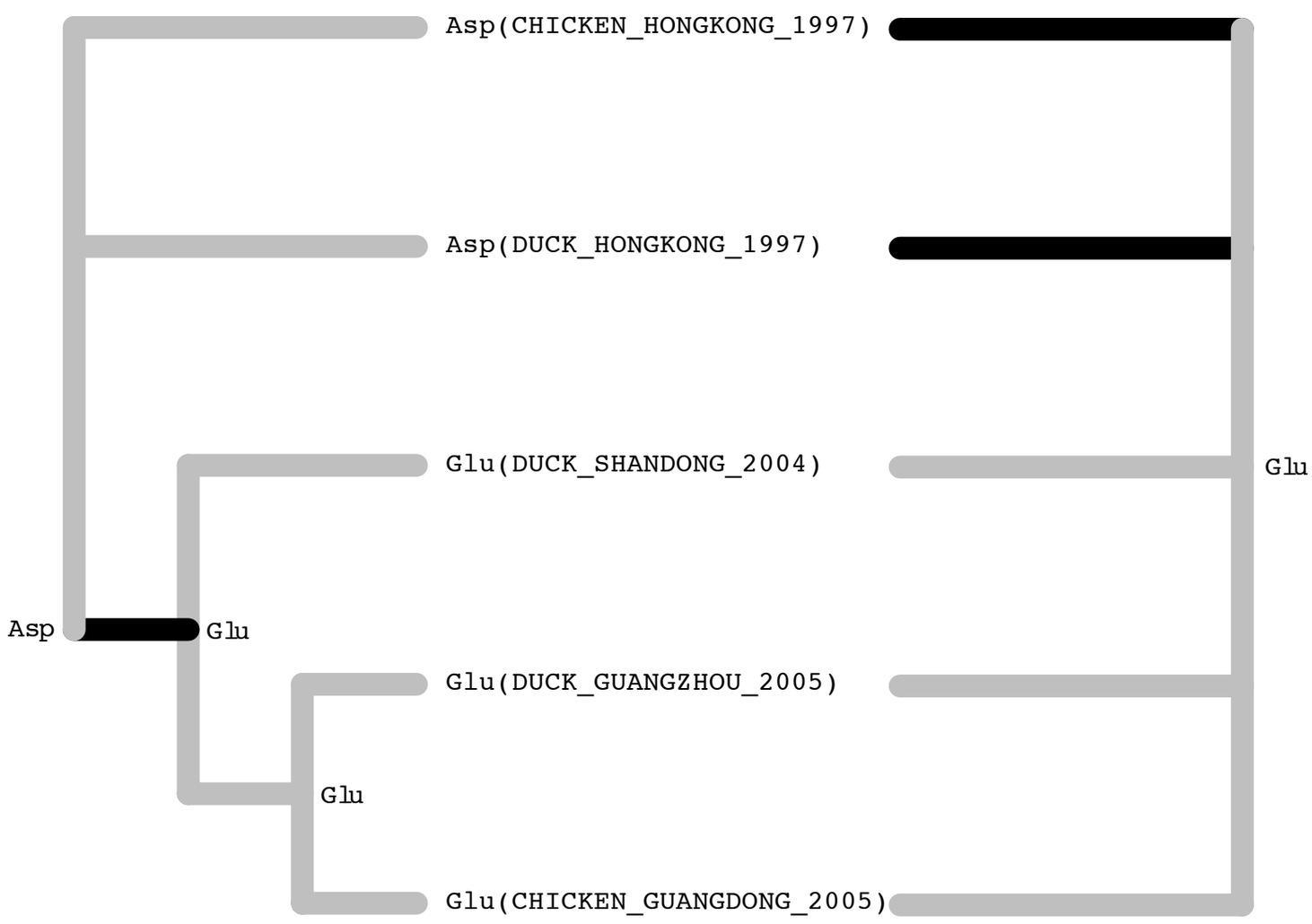


Fig. 1.1. Effect of phylogeny on estimating synonymous and nonsynonymous substitution counts in a dataset of Influenza A/H5N1 haemagglutinin sequences. Using the maximum likelihood tree on the left, the observed variation can be parsimoniously explained with one nonsynonymous substitution along the darker branch, whereas the star tree on the right involves at least two.

Selection is variable across a gene

- Different sites in a gene will be subject to different selective forces
- A *gene-wide* measure of selection is going to average these effects
- Most sites in most genes will be maintained by purifying selection
- Positively selected sites are of great biological interest, because they point to how a particular gene can respond to selective pressures
- Must develop methods that are able to disentangle the contributions of individual sites

Suzuki-Gojobori (SG99)

- Uses a tree to compute **dN/dS** at a given site

    1. Reconstruct ancestral sequences by nucleotide-level parsimony

    2. Compute **EN** and **ES** using labeled branches; define $\mathbf{p_e = ES/EN}$

    3. Compute **S** and **NS** for each site (minimum evolution)

    4. Estimate the probability that the number of synonymous substitutions **S** is unusually low (positive selection) or unusually high (negative selection), using the binomial distribution given $\mathbf{p_e}$ from step 2.

A method for detecting positive selection at single amino acid sites
Y. Suzuki and T. Gojobori
Mol Biol Evol **16** 1315-1328 (1999)

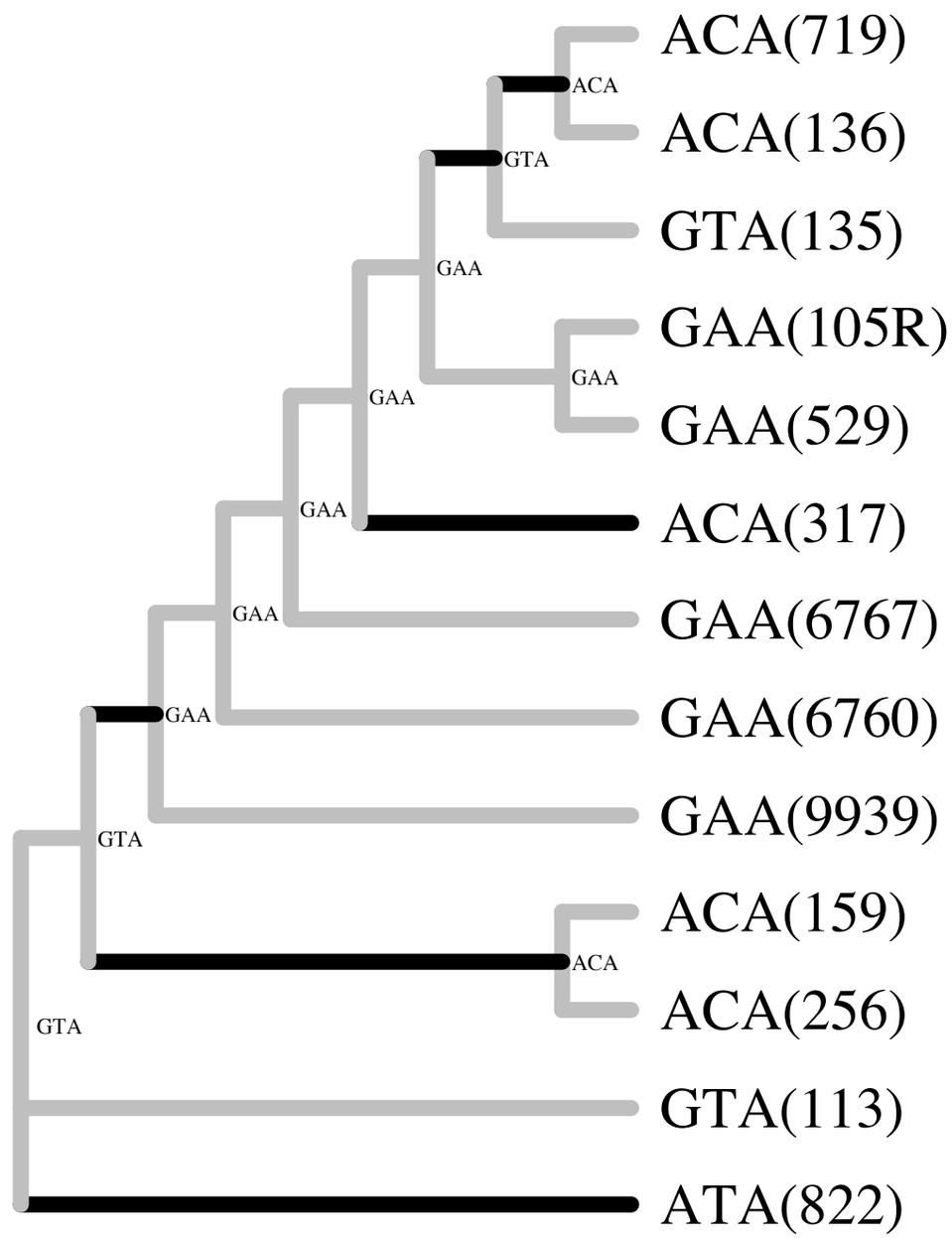
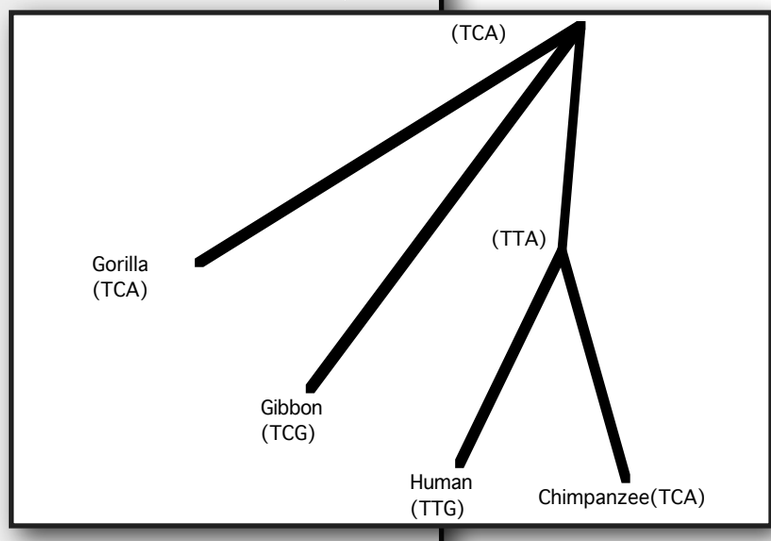


Fig. 1.6. An illustration of SLAC method, applied to a small HIV-1 envelope V3 loop alignment. Sequence names are shown in parentheses. Likelihood state ancestral reconstruction is shown at internal nodes. The parsimonious count yields 0 synonymous and 9 non-synonymous substitutions (highlighted with a dark shade) at that site. Based on the codon composition of the site and branch lengths (not shown), the expected proportion of synonymous substitutions is $p_e = 0.25$. An extended binomial distribution on 9 substitutions with the probability of success of 0.25, the probability of observing 0 synonymous substitutions is 0.07, hence the site is borderline significant for positive selection.

Models of codon evolution

- In 1994, first tractable mechanistic evolutionary models for codon sequences were proposed by Muse and Gaut, and, independently, by Goldman and Yang [in the same issue of MBE, back to back]

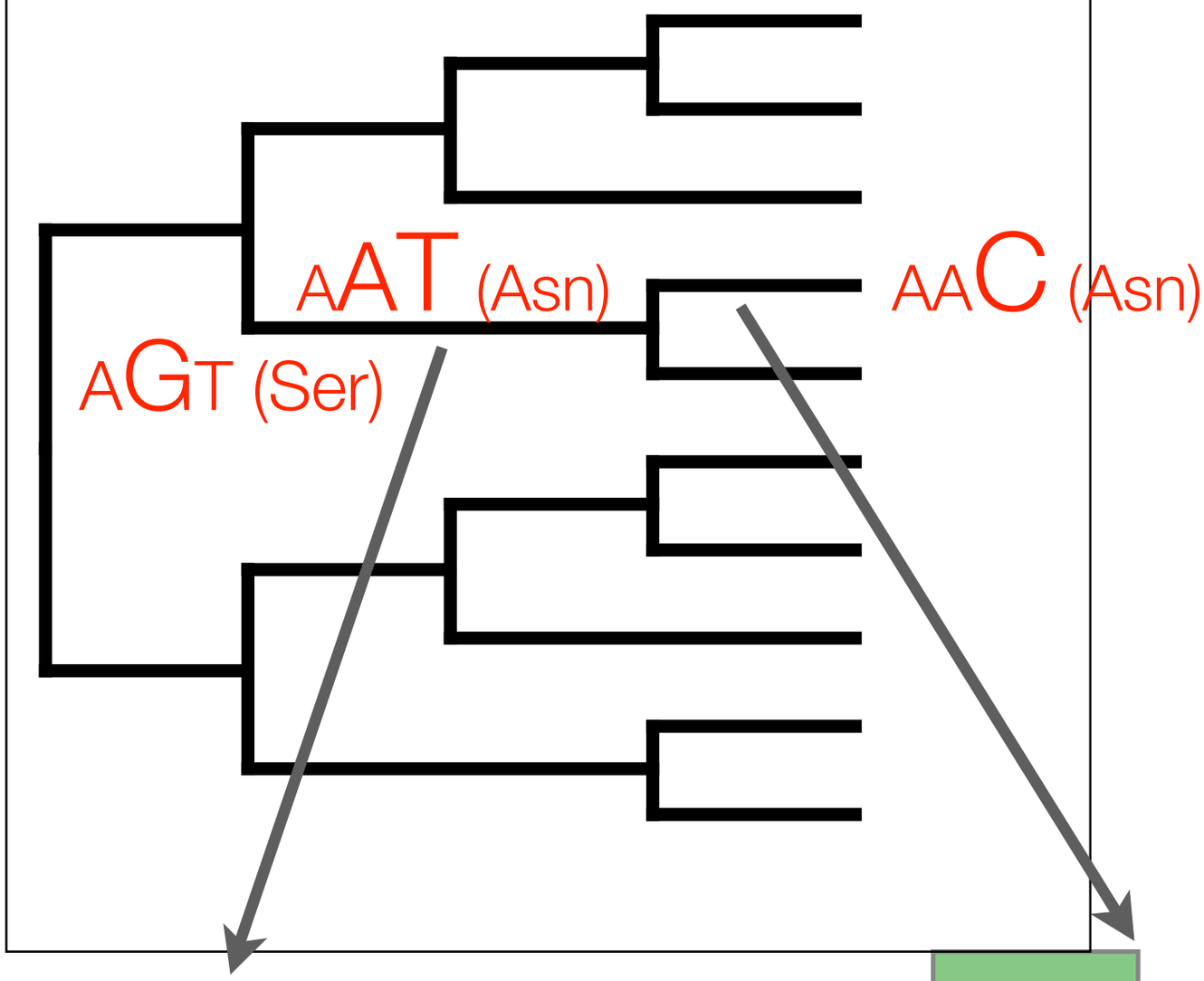
$$(\text{Rate})_{X,Y}(dt) = \begin{cases} \alpha R_{xy}\,\pi_t dt & , \quad \text{one-step, synonymous substitution,} \\ \beta R_{xy}\,\pi_t dt & , \quad \text{one-step, non-synonymous substitution,} \\ 0 & , \quad \text{multi-step.} \end{cases}$$



X,Y = AAA...TTT (excluding stop codons),
π_t - frequency of the target nucleotide.
Example substitutions:
AAC→AAT (one step, synonymous - Aspargine) αR_{CT}
CAC→GAC (one step, non-synonymous - Histidine to Aspartic Acid) βR_{CG}
AAC→GTC (multi-step).

A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome
S. V. Muse and B. S. Gaut
Mol Biol Evol **11** 715-724 (1994)

A codon-based model of nucleotide substitution for protein-coding DNA sequences.
N. Goldman and Z. Yang
Mol Biol Evol **11** 725--736 (1994)



$$\beta^{B_2}_{Ser \rightarrow Asn} \theta_{AG} \pi_A dt \qquad \alpha^{B_1} \theta_{CT} \pi_C dt$$

Nucleotide substitution biases

Synonymous rate at a branch B_1

Nonsynonymous rate for at branch B_2

Stationary character frequency

Computing the transition probabilities

- In order to recover transition probabilities $T(t)$ from the rate matrix Q, one computes the matrix exponential $T(t) = \exp(Qt)$

- Compare this to standard nucleotide models, e.g. HKY85

- Because the computational complexity of matrix exponentiation scales as the cube of the matrix dimension, codon based models require roughly $(61/4)^3 \approx 3500$ more operations than nucleotide models

- This explains why codon probabilistic models were not introduced until the 1990s.

Multiple substitutions

- The model assumes that point mutations alter one nucleotide at a time, hence most of the instantaneous rates (**`3134/3761`** or **`84.2%`** in the case of the universal genetic code) are **`0`**.
- This restriction, however, does not mean that the model disallows any substitutions that involve multiple nucleotides (e.g., **`ACT`**⟹**`AGG`**).
- Such substitutions must simply be realized via several single nucleotide steps, e.g **`ACT`**⟹**`AGT`**⟹**`AGG`**
- In fact the **`(i,j)`** element of **`T(t) = exp(Qt)`** sums the probabilities of all such possible pathways of duration **`t`**, including reversions
- Compare this to the rather naive NG parsimony approach.

Alignment-wide estimates

- Using standard MLE approaches it is straightforward to obtain point estimates of **`dN/dS := β/α`**

- Can also easily test whether or not **`dN/dS > 1`**, or **`< 1`** using the likelihood ratio test (LRT)

- Codon models also support the concepts of synonymous and non-synonymous distances between sequences using standard properties of Markov processes (exponentially distributed waiting times)

$$E[subs] = -\sum_i \pi_i \hat{q}_{ii}, \quad E[subs] = E[syn] + E[nonsyn] = -\sum_i \pi_i \hat{q}^s_{ii} - \sum_i \pi_i \hat{q}^{ns}_{ii}.$$

HIV-1 C2-V3 *env* 13 sequences

Model	Log L	# p	ω	LRT	p-value	Mean d
Null	-1121.09	37	1			0.137
Alternative	-1120.83	38	1.18	0.53	0.47	0.136

HIV-1 RT 8 sequences

Model	Log L	# p	ω	LRT	p-value	Mean d
Null	-3258.5	27	1			0.074
Alternative	-3185.88	28	0.23	**145.2**	0	0.077

Diversifying selection in HIV-1 driven by neutralizing antibodies

- The humoral arm of the immune system mounts a potent defense against viral infections

- Existing successful vaccines are based on raising a neutralizing antibody (nAb) response to the pathogen

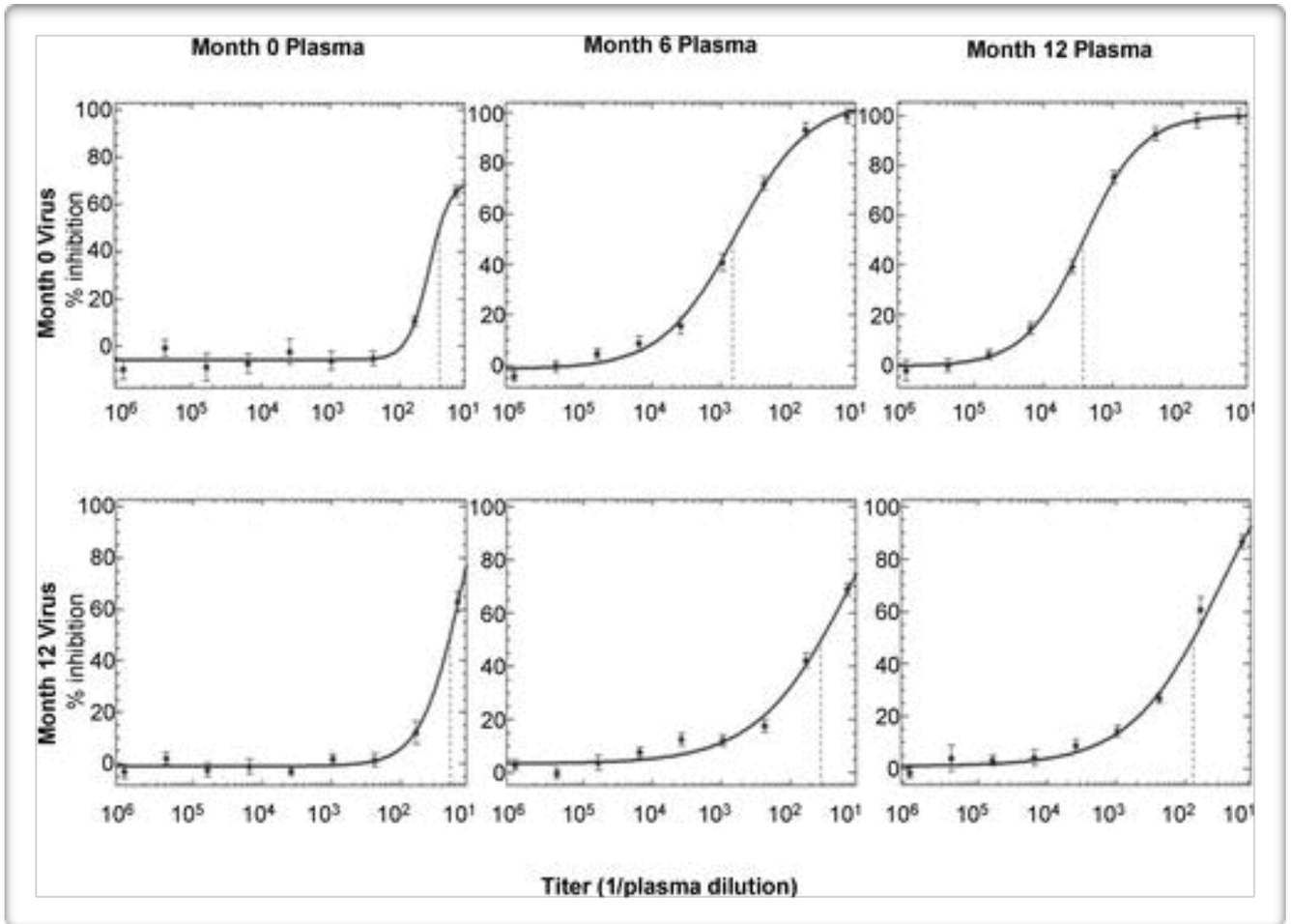
- No simple host genetic basis (epitopes) of the specificity of neutralizing antibody responses is known

- Need to measure these responses

    - Analyzing escape of HIV-1 from neutralizing antibodies

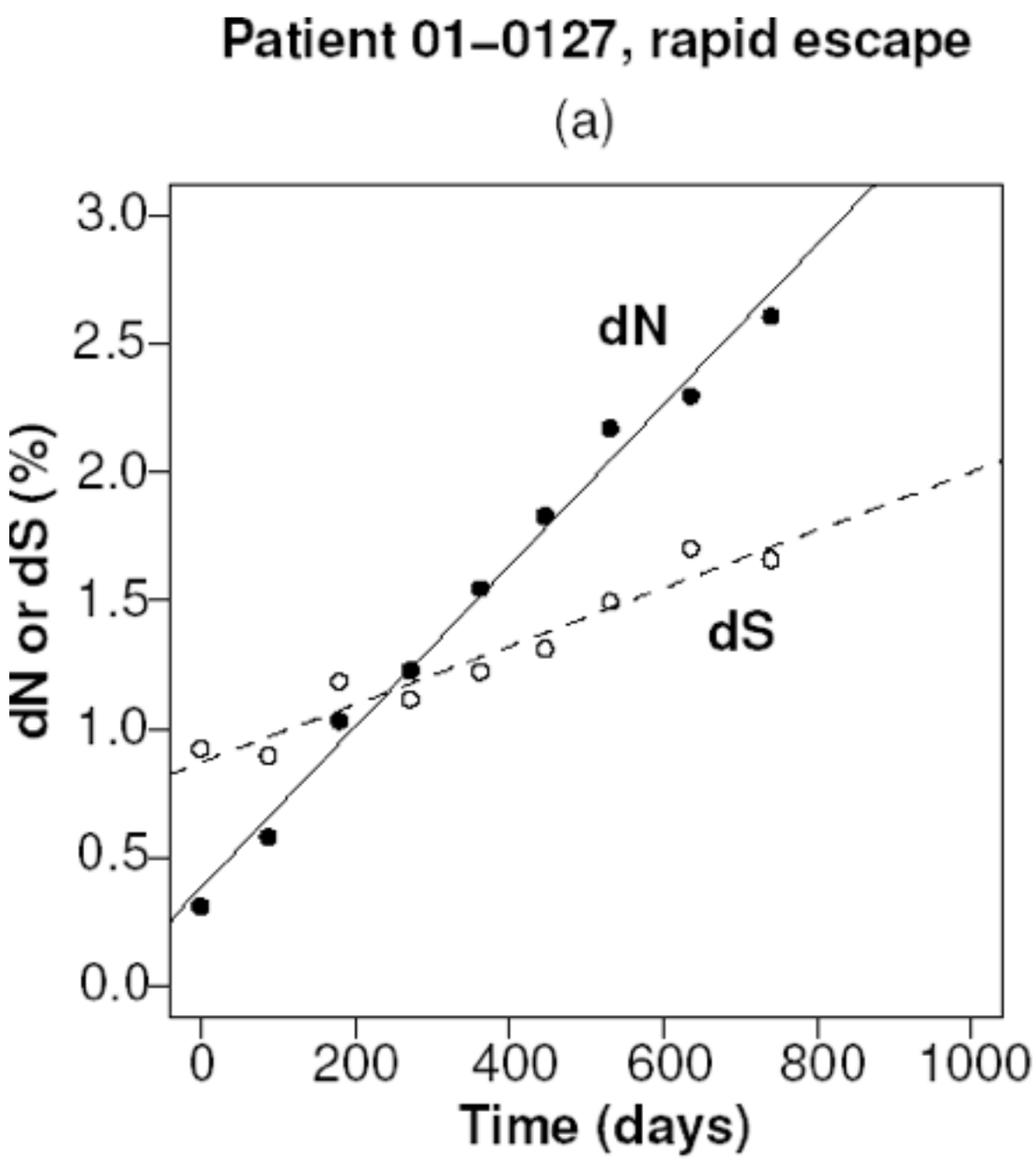
Neutralization curves from an individual with early HIV infection

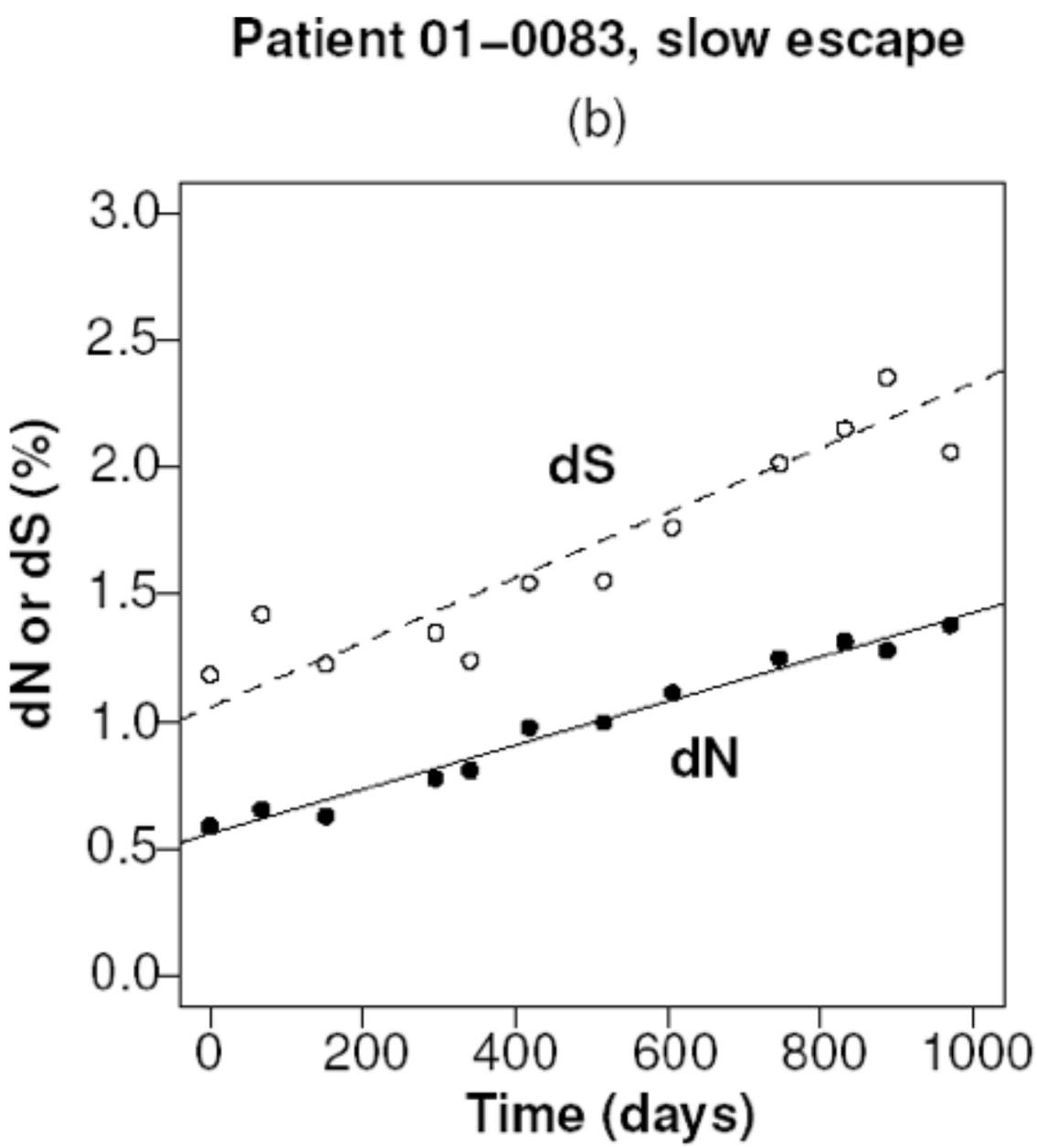
- Neutralization can be measured by the serum dilution needed to reduce viral replication by 50% (typically presented as the inverse of the titer)
- Although variable between individuals, the rate of escape from neutralizing antibodies can be very high during acute/early HIV infection
- Sera are effective at neutralizing earlier viruses, but significantly less effective at neutralizing contemporaneous viruses
- The immune system loses the arms race



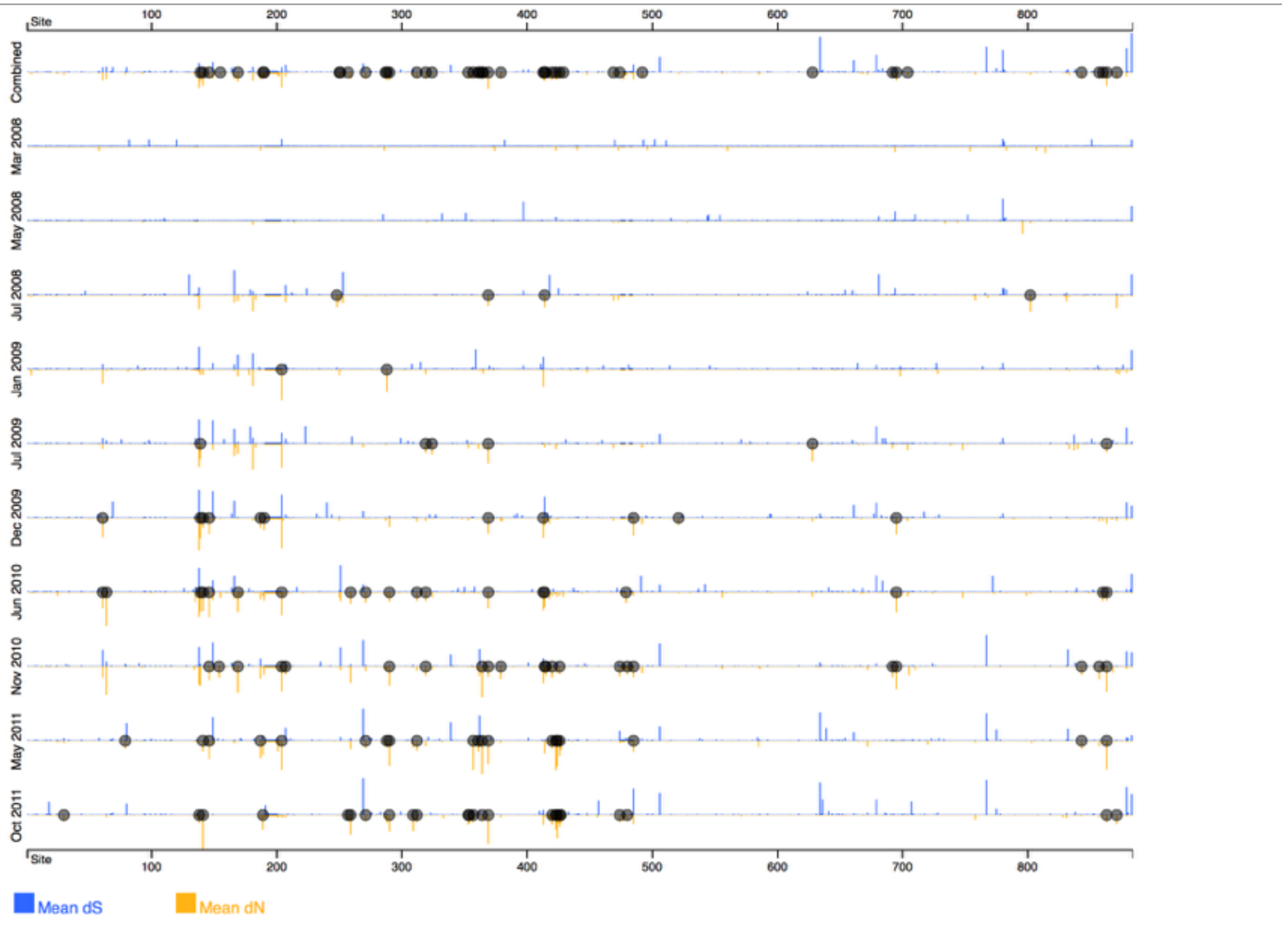
PNAS | **December 20, 2005** | vol. 102 | no. 51 | **18514-18519**

Amino acid substitutions in HIV-1 env accumulate faster during rapid escape

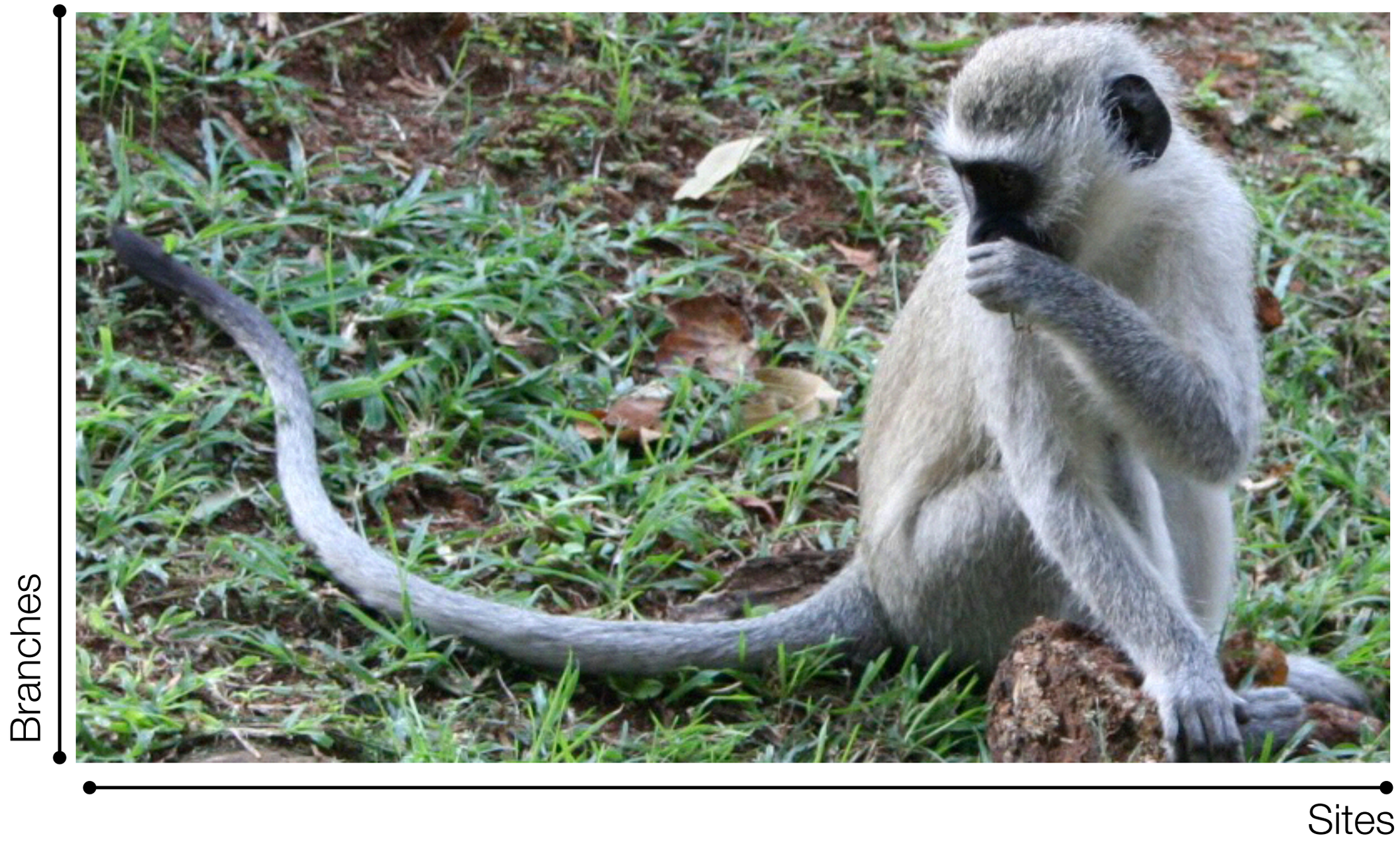




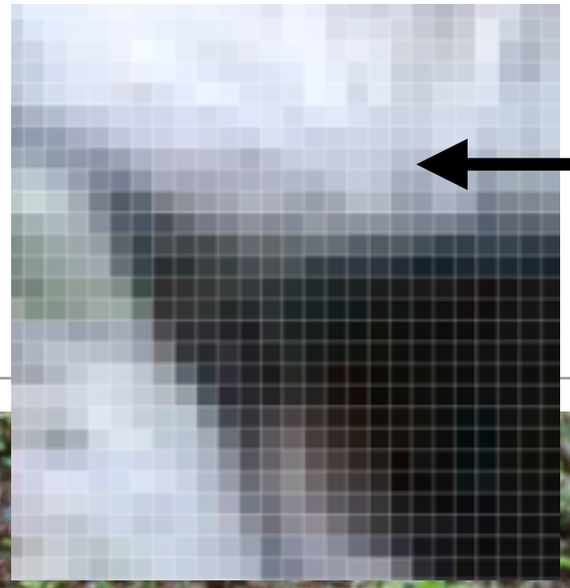
But upon closer look, this pattern is highly variable across a gene and through time.



Reality



Reality



Pixel ~ a single codon site
along a single branch
in the phylogeny.

Branches

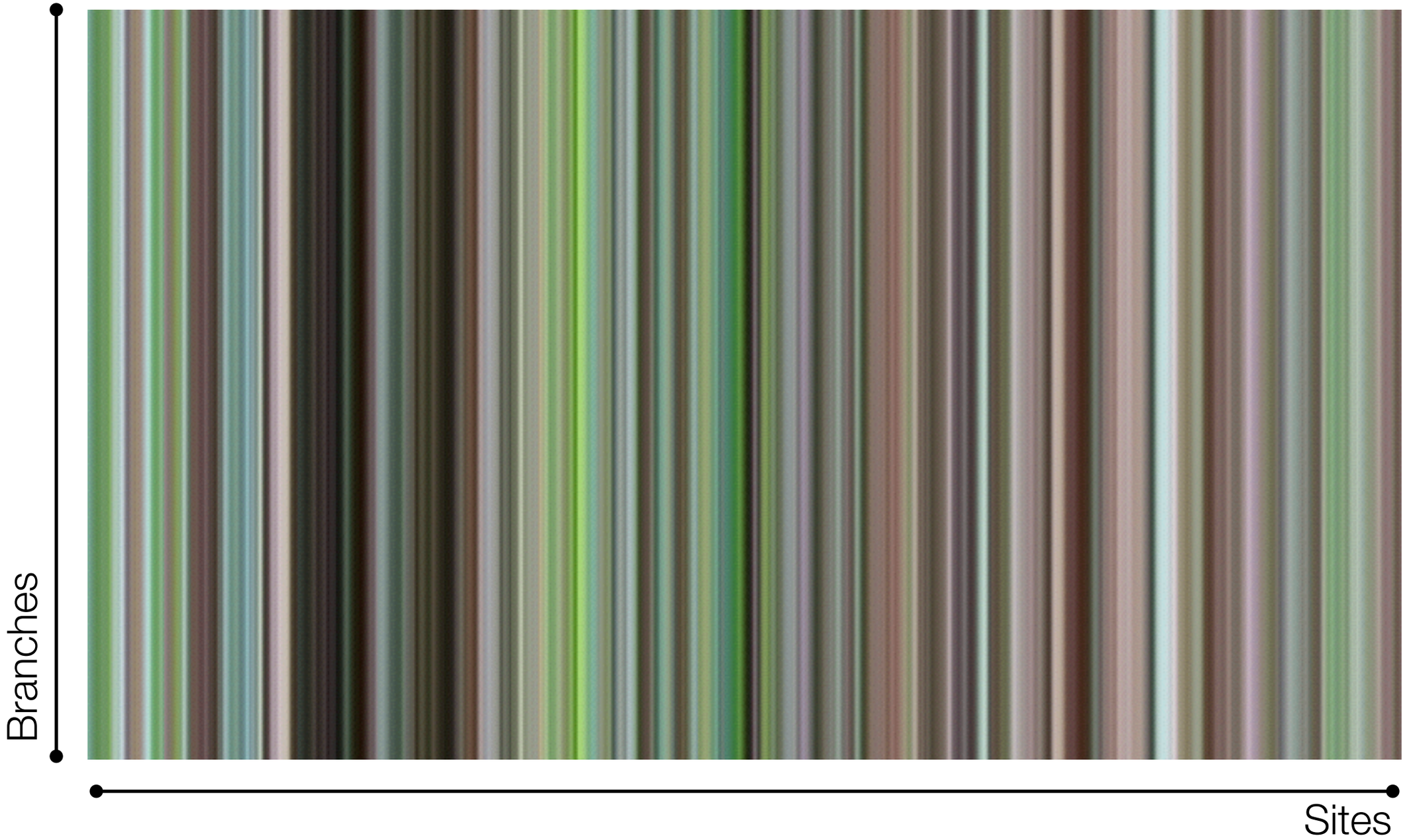
Sites

Gene-wide dN/dS model

Does the mean dN/dS exceed 1?

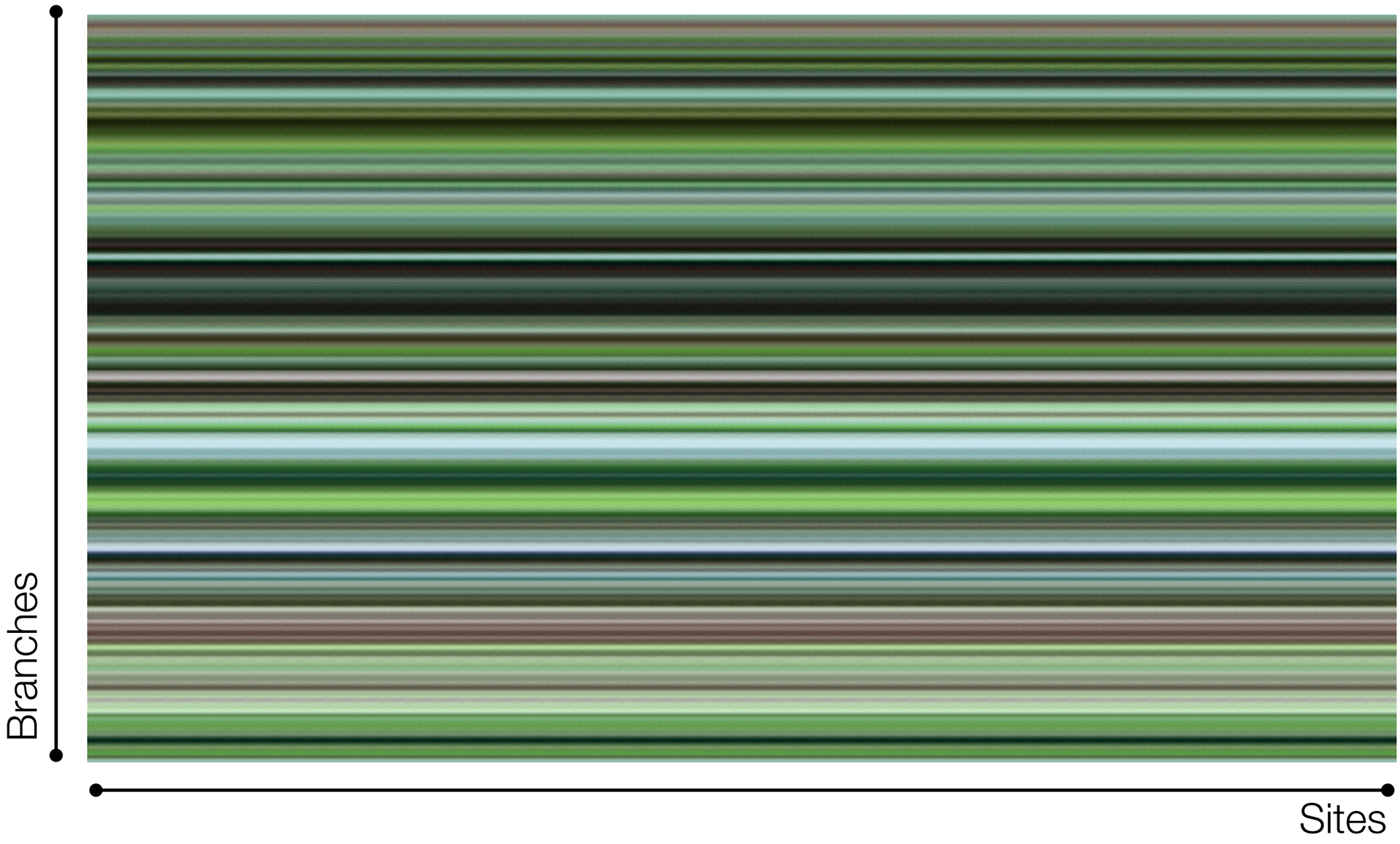
Washes out all the details by averaging over sites and branches

Site-level model



Asks questions about sites by pooling information over branches (separate rate for every site, like FEL or MEME)

Branch-level model



Asks questions about branches by pooling information over sites (separate rate for each branch)

Branch-site model

Asks questions about branches or sites by pooling information over subsets of sites and branches, inferred or given a priori

Evolutionary uncertainty principle

- “However, we caution that despite obvious interest in identifying specific branch-site combinations subject to diversifying selection, such inference is based on very limited data (the evolution of one codon along one branch), and cannot be recommended for purposes other than data exploration and result visualization. This observation could be codified as the **selection inference uncertainty principle** -- one cannot simultaneously infer both the site and the branch subject to diversifying selection.”

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, et al. (2012) **Detecting Individual Sites Subject to Episodic Diversifying Selection**. PLoS Genet 8(7): e1002764.

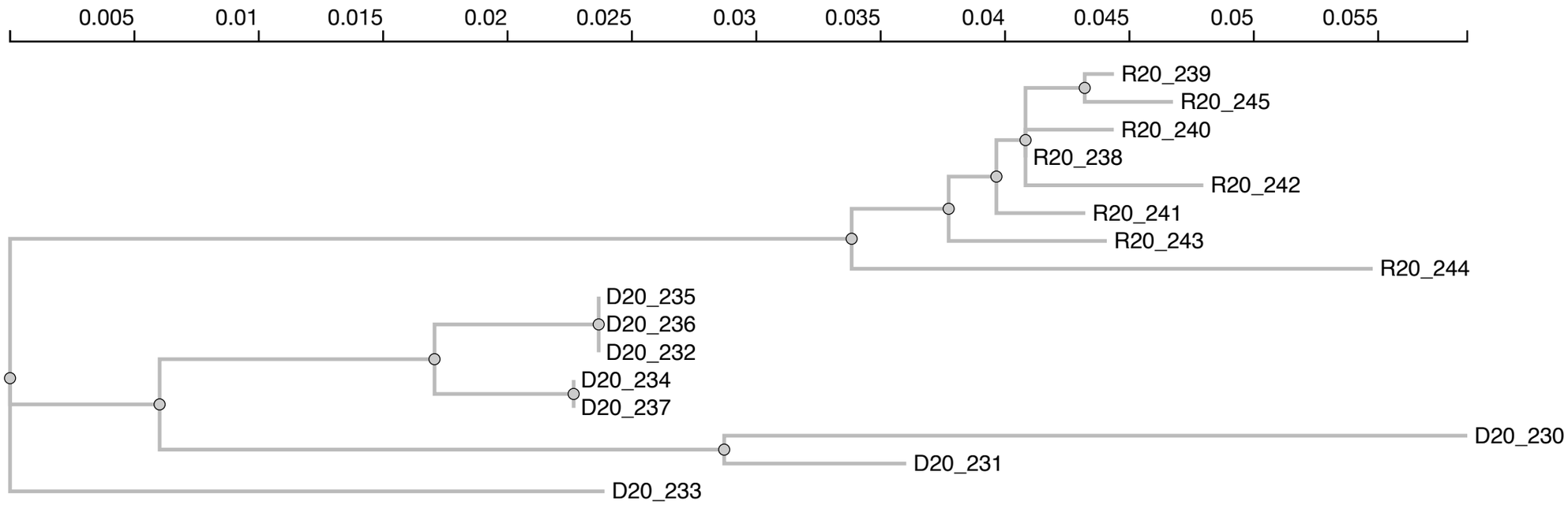
Two example datasets

- **West Nile Virus NS3 protein**
  - An interesting case study of how positive selection detection methods lead to testable hypotheses for function discovery
  - Brault et al 2007, *A single positively selected West Nile viral mutation confers increased virogenesis in American crows*
- **HIV-1 transmission pair**
  - Partial *env* sequences from two epidemiologically linked individuals
  - An example of multiple selective environments (source, recipient, transmission)

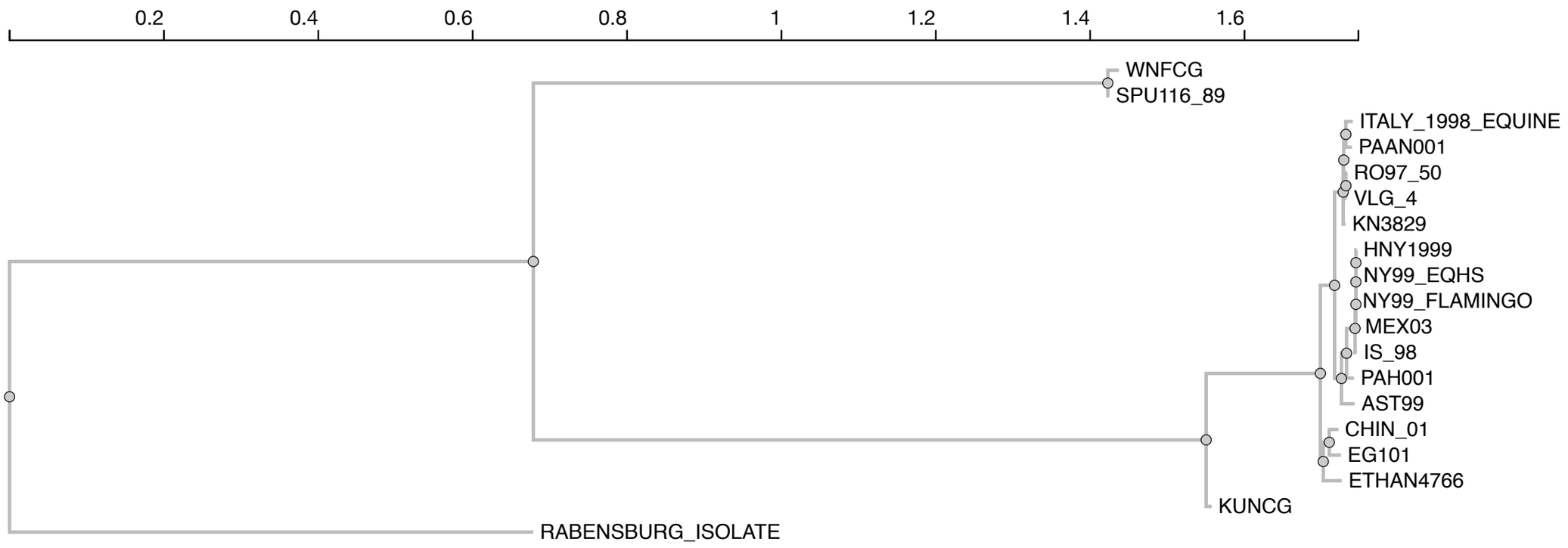
Information content of the alignments

	WNV NS3	HIV-1 *env*
Sequences	19	16
Codons	619	288
Tree Length *MG94 model, subs/site*	3.32	0.20

HIV-1 *env*



WN NS3



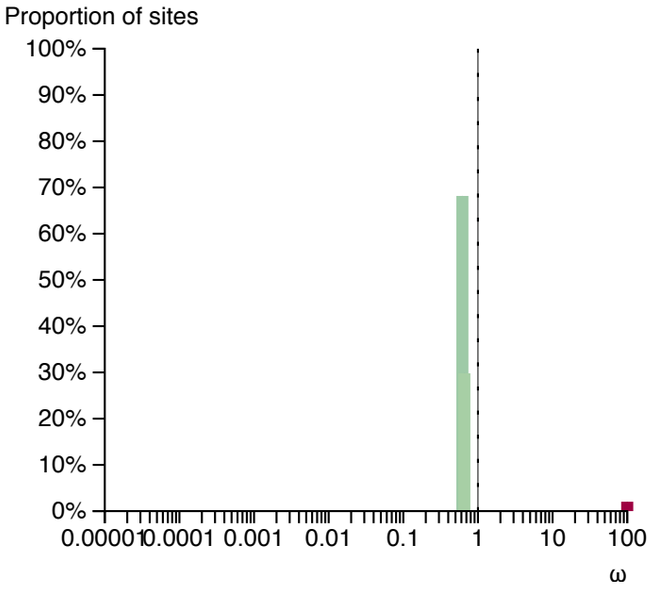
Mean dN/dS analysis

	WNV NS3	HIV-1 *env*
Mean dN/dS	0.009	1.135
95% CI	[0.007-0.011]	[0.95,1.34]
p-value (H_0 : dN/dS = 1)	$<1^{-20}$	0.546
Selection?	Strong purifying.	Indistinguishable from neutral.

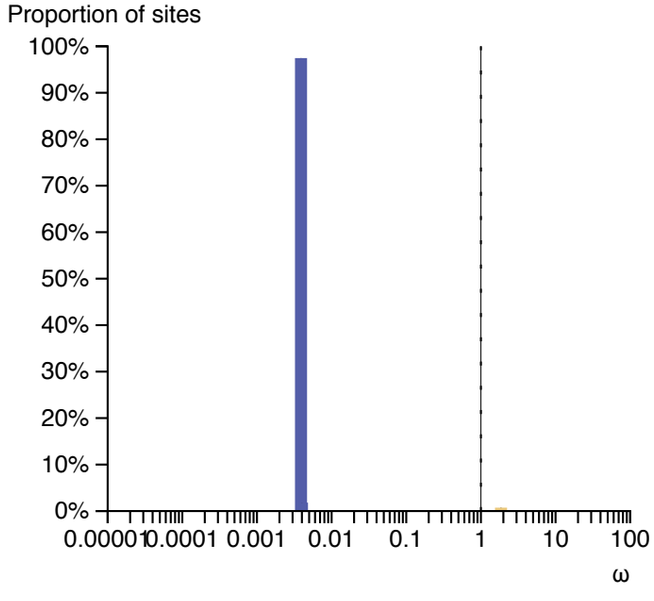
Gene-wide selection analysis using a branch-site method (BUSTED)

	WNV NS3	HIV-1 *env*
Gene-wide **dN/dS** distribution	$\omega_1 = 0.004$ (99%) $\omega_2 = 1.86$ (1%)	$\omega_1 = 0.619$ (68%) $\omega_2 = 0.657$ (30%) $\omega_3 = 105$ (2%)
p-value for selection ($H_0 : \omega_3 = 1$)	0.54	$<1^{-20}$
Log L (no variation)	-6413.50	-2078.13
Log L (branch-site; 4 addt'l parameters)	-6396.18	-2039.99

HIV-1 *env*



WNV NS3

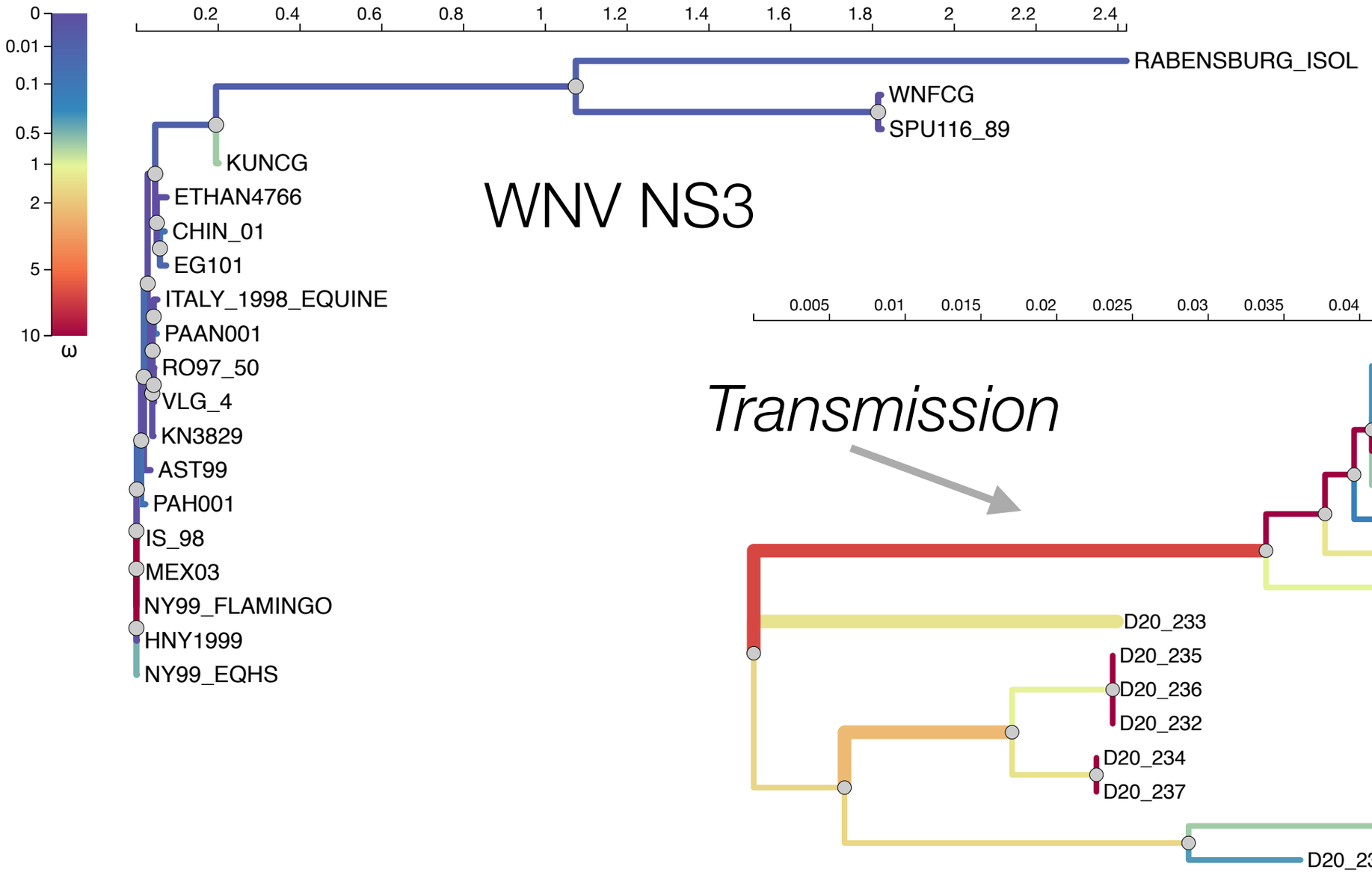


Murrell et al | **Mol. Biol. Evol** | 32(5) | 1365–1371

BUSTED analysis

- **West Nile Virus NS3 protein**
  - Marginal evidence of weak episodic selection (`dN/dS ~ 2`) on a small proportion of sites (`~1%`)
  - The rest of the gene is very strongly conserved (`dN/dS = 0.004`)
- **HIV-1 transmission pair**
  - Very strong evidence of strong episodic diversification (`dN/dS ~ 100`) on a small proportion of sites (`2%`)
  - The rest of the gene evolves with weak purifying selection (`dN/dS = 0.6-0.7`)

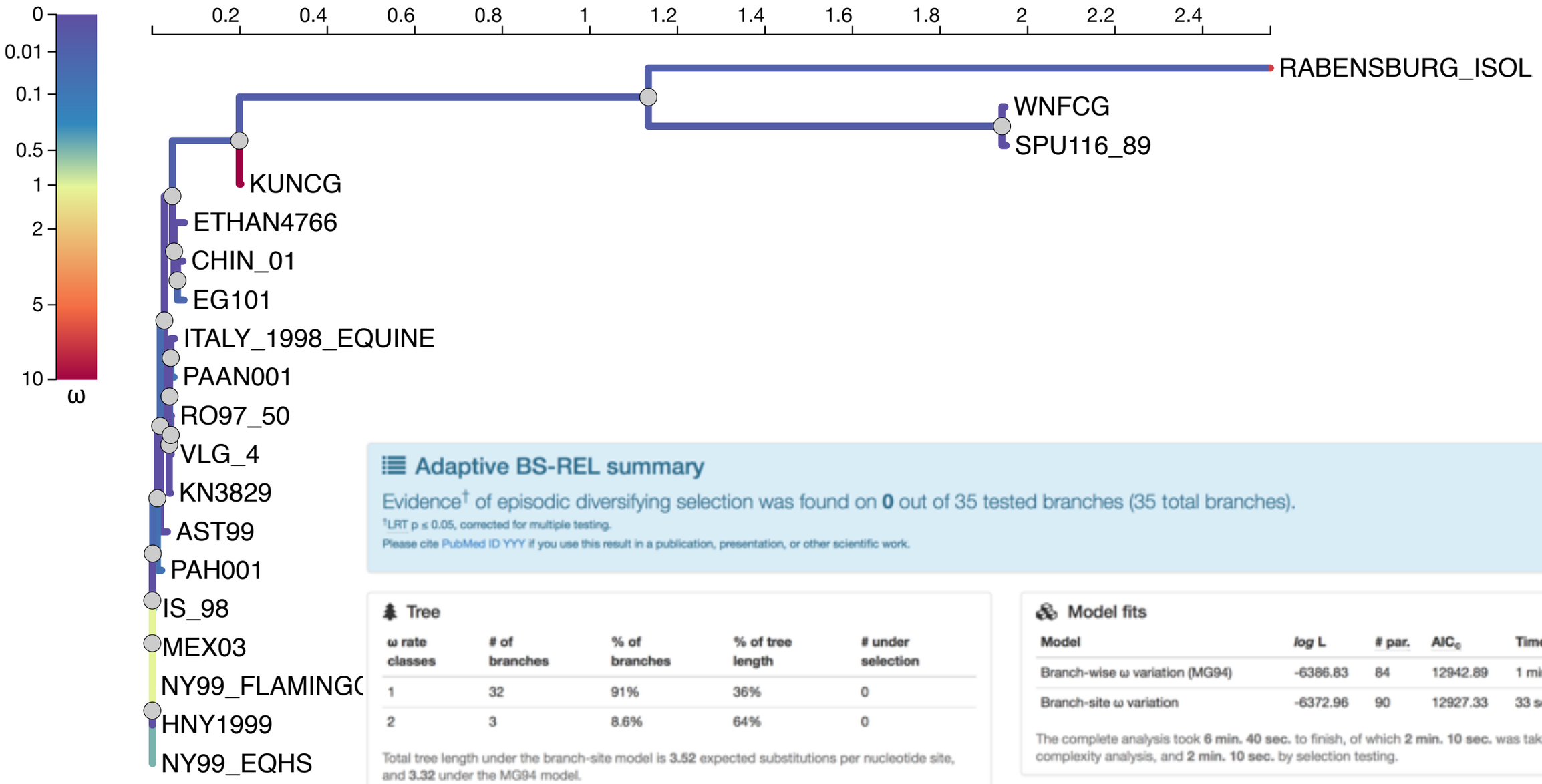
Branch-level variation: separate dN/dS fitted to each branch of the tree (no site-to-site variation)



- Is selection pressure constant through time?
- Are we missing anything by pooling estimates over sites?

aBSREL branch site model. Branch-level analysis. Independent distributions of dN/dS are fitted to branches.

WNV NS3



Adaptive BS-REL summary

Evidence[†] of episodic diversifying selection was found on **0** out of 35 tested branches (35 total branches).

[†]LRT $p \leq 0.05$, corrected for multiple testing.

Please cite PubMed ID YYY if you use this result in a publication, presentation, or other scientific work.

Tree

ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	32	91%	36%	0
2	3	8.6%	64%	0

Total tree length under the branch-site model is **3.52** expected substitutions per nucleotide site, and **3.32** under the MG94 model.

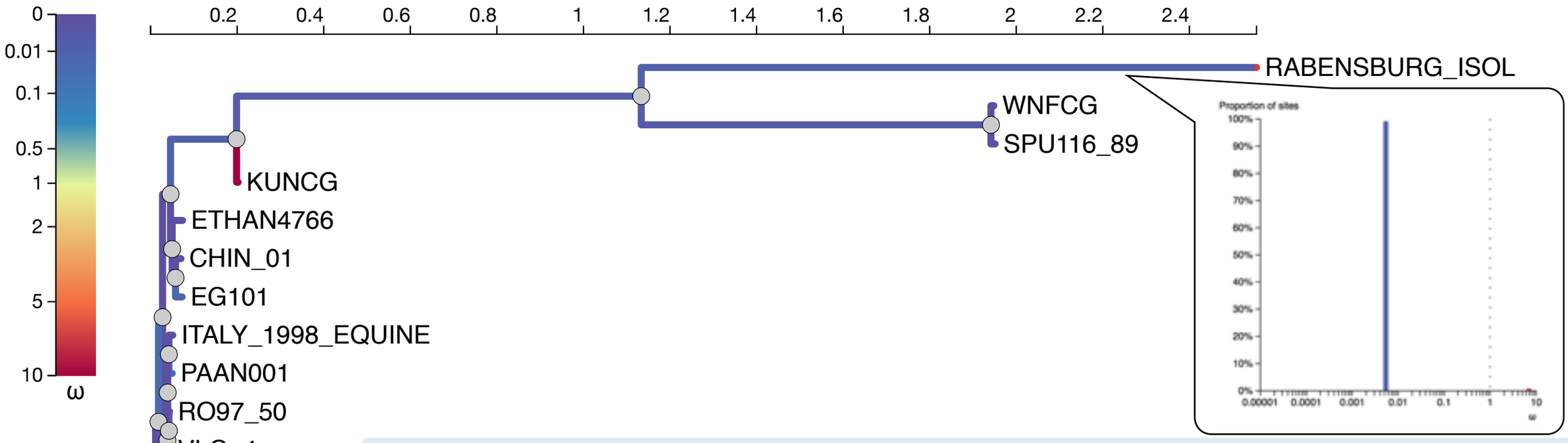
Model fits

Model	*log* L	# par.	AIC_c	Time to fit
Branch-wise ω variation (MG94)	-6386.83	84	12942.89	1 min. 23 sec.
Branch-site ω variation	-6372.96	90	12927.33	33 sec.

The complete analysis took **6 min. 40 sec.** to finish, of which **2 min. 10 sec.** was taken by the complexity analysis, and **2 min. 10 sec.** by selection testing.

aBSREL branch site model. Branch-level analysis. Independent distributions of dN/dS are fitted to branches.

WNV NS3



Adaptive BS-REL summary

Evidence† of episodic diversifying selection was found on **0** out of 35 tested branches (35 total branches).

†LRT $p \leq 0.05$, corrected for multiple testing.

Please cite PubMed ID YYY if you use this result in a publication, presentation, or other scientific work.

Tree

ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	32	91%	36%	0
2	3	8.6%	64%	0

Total tree length under the branch-site model is **3.52** expected substitutions per nucleotide site, and **3.32** under the MG94 model.

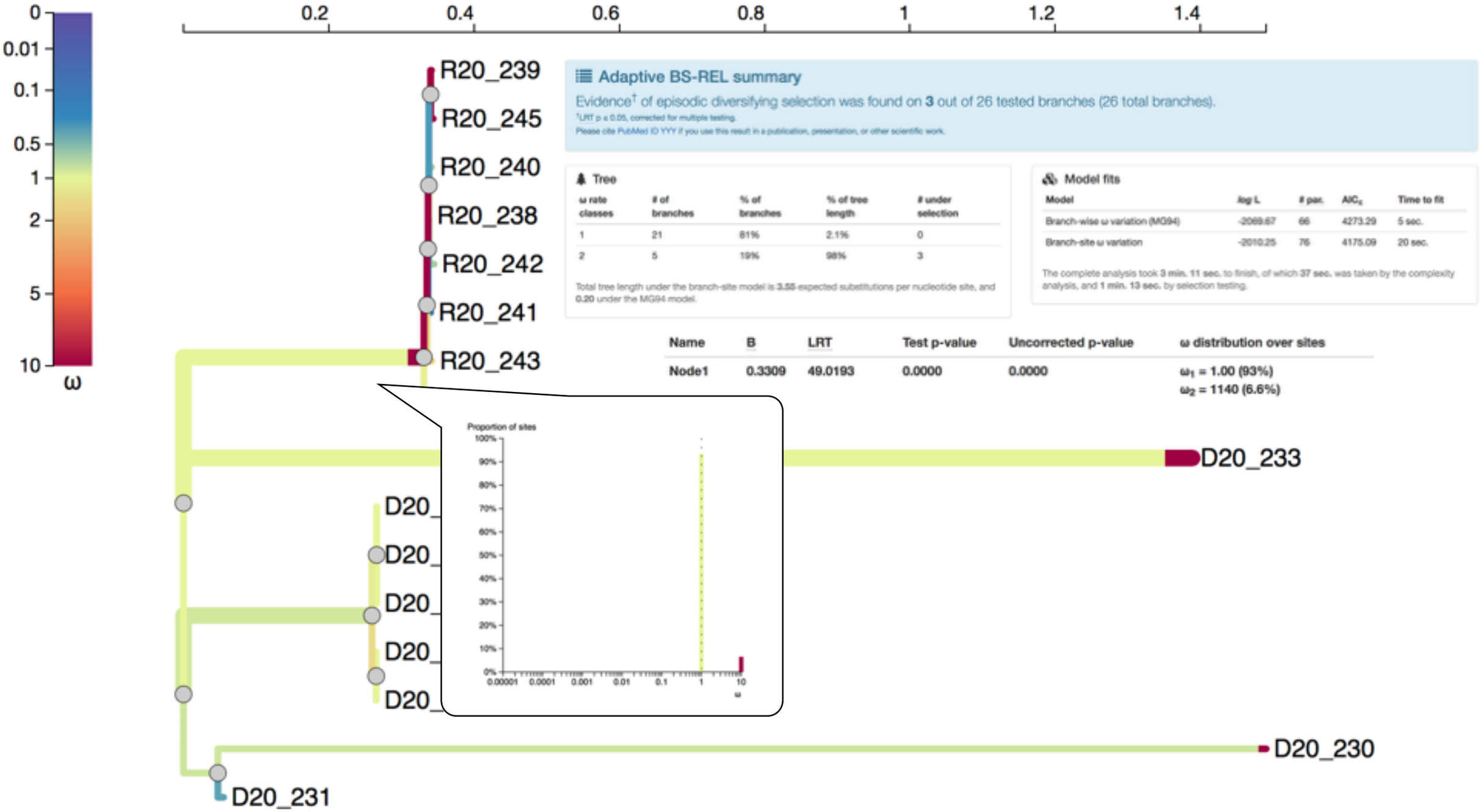
Model fits

Model	*log* L	# par.	AIC_c	Time to fit
Branch-wise ω variation (MG94)	-6386.83	84	12942.89	1 min. 23 sec.
Branch-site ω variation	-6372.96	90	12927.33	33 sec.

The complete analysis took **6 min. 40 sec.** to finish, of which **2 min. 10 sec.** was taken by the complexity analysis, and **2 min. 10 sec.** by selection testing.

aBSREL branch site model. Branch-level analysis. Independent distributions of dN/dS are fitted to branches.

HIV-1 *env*

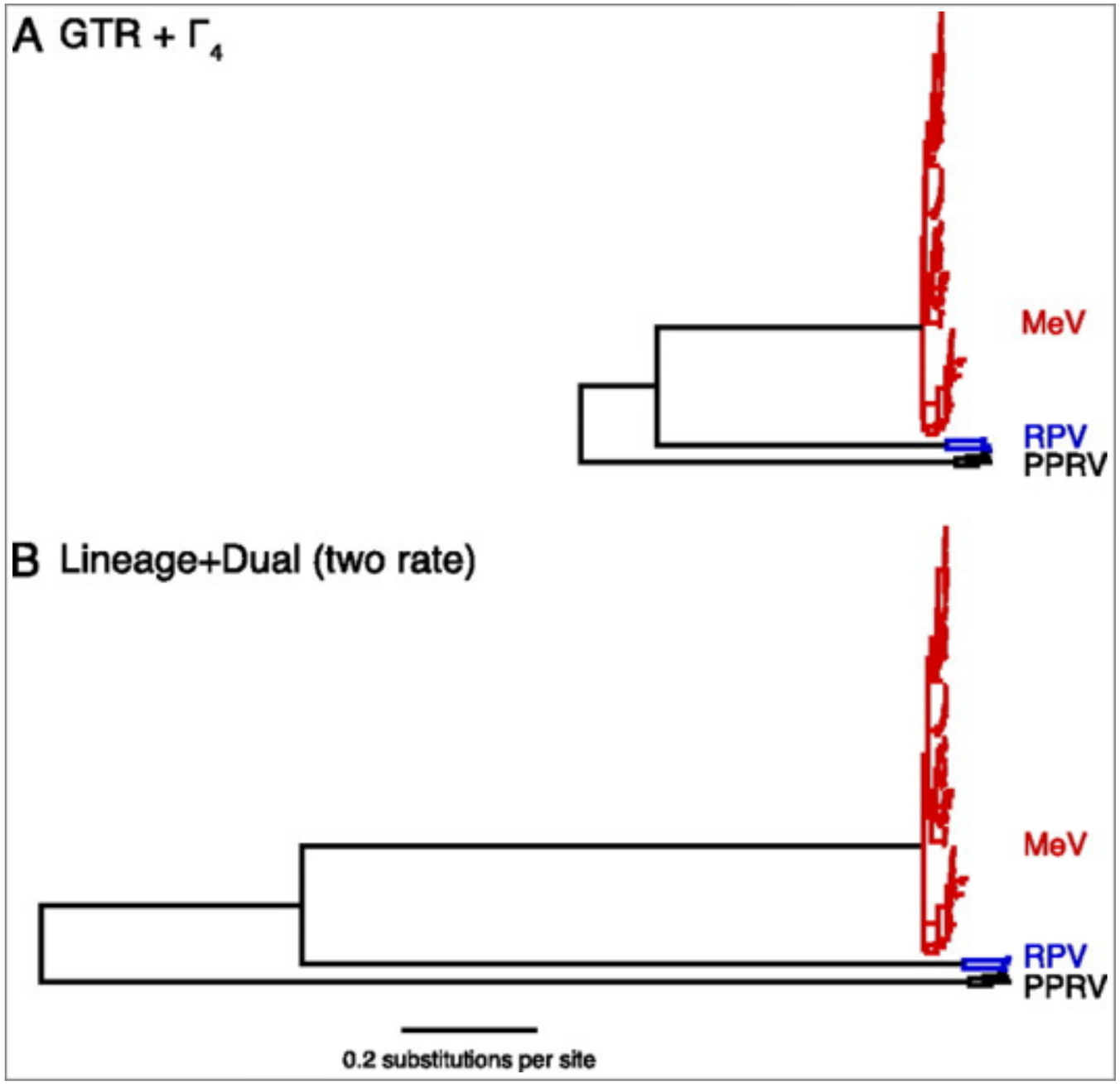


aBSREL analysis

- **West Nile Virus NS3 protein**
  - `91%` branches can be explained with simple (single **`dN/dS`**) models
  - `3` branches (`9%`) have evidence of multiple **`dN/dS`** rate classes over sites, but **none** with significant proportions of sites with **`dN/dS > 1`**
- **HIV-1 transmission pair**
  - `81%` branches can be explained with simple (single **`dN/dS`**) models
  - `5` branches (`19%`) have evidence of multiple **`dN/dS`** rate classes over sites,
  - `3` branches have small (`1-7%`), but statistically significant (**`p<0.05`**, multiple testing corrected) proportions of sites with **`dN/dS > 1`**.

Aside: implications for molecular dating

- Natural selection (especially time-varied and strong purifying selection) often leads to **underestimation** of branch lengths and “young” estimates, inconsistent with other evidence and literature (see Mol Biol Evol. 28(12):3355-65 *Purifying selection can obscure the ancient age of viral lineages*) *e.g.* Measles, EBOV, IAV, Coronaviruses, HSV



Fast site-level analysis (FUBAR): no branch to branch variation; pervasive diversifying selection; random effects

WNV NS3

THE EXPECTED NUMBER OF FALSE POSITIVES IS 0.01 (95% CI: [0-0]).

Codon	α	β	β-α	Posterior Prob β>α	Emp. Bayes Factor	PSRF	N_{eff}	3D rate plot?
249	0.138179	1.51208	1.3739	0.988732	622.977	1.03135	144.342	[SVG] [PNG]

HIV-1 *env*

THE EXPECTED NUMBER OF FALSE POSITIVES IS 0.20 (95% CI: [0-1]).

Codon	α	β	β-α	Posterior Prob β>α	Emp. Bayes Factor	PSRF	N_{eff}	3D rate plot?
161	0.401387	5.64609	5.2447	0.974565	42.7026	1.00373	576.851	[SVG] [PNG]
165	0.311605	2.85981	2.54821	0.963007	29.0123	1.00897	363.228	[SVG] [PNG]
203	0.399165	4.27264	3.87348	0.968713	34.5068	1.00481	514.136	[SVG] [PNG]
204	0.31539	2.80044	2.48505	0.951136	21.6933	1.00947	350.933	[SVG] [PNG]
225	0.292592	2.75072	2.45813	0.955732	24.0613	1.00959	348.105	[SVG] [PNG]
282	0.186706	2.31058	2.12387	0.982502	62.578	1.0103	332.362	[SVG] [PNG]

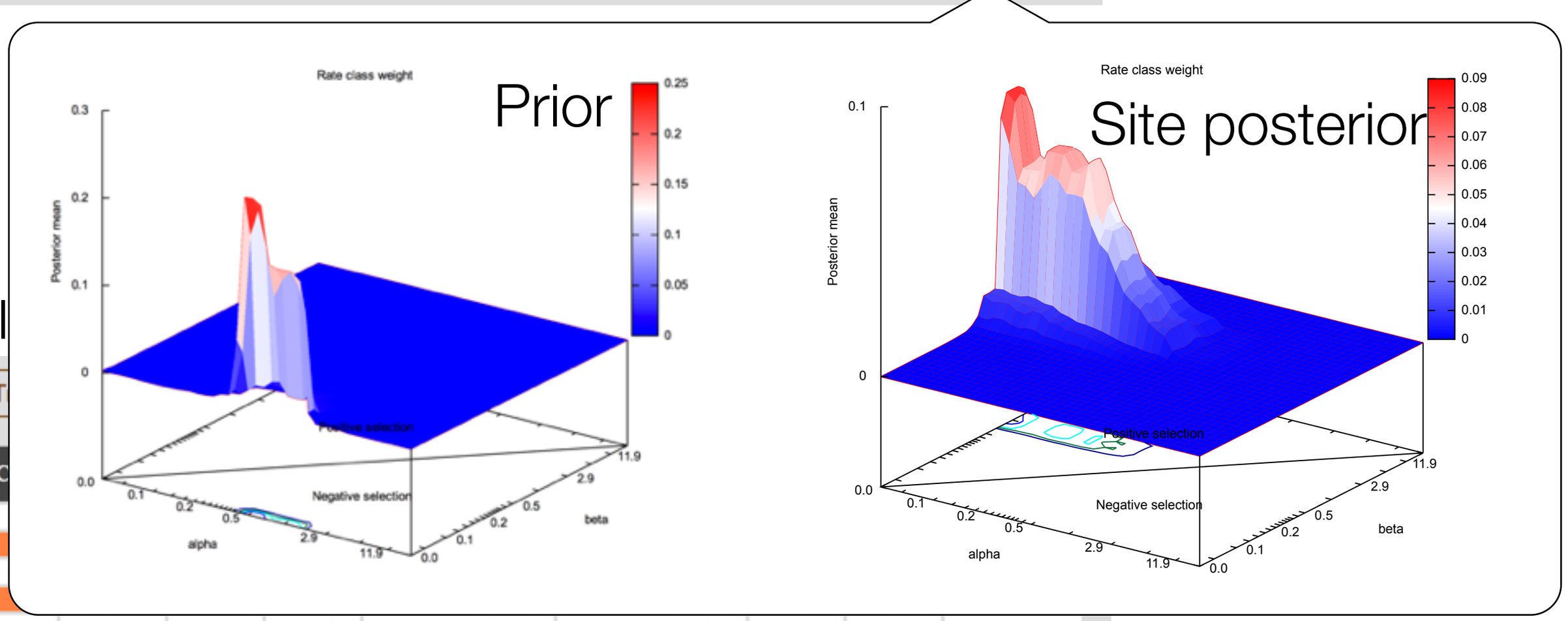
Murrell et al | **Mol. Biol. Evol** | 30(5) | 1196–1205

Fast site-level analysis (FUBAR): no branch to branch variation; pervasive diversifying selection; random effects

WNV NS3

THE EXPECTED NUMBER OF FALSE POSITIVES IS 0.01 (95% CI: [0-0]).

Codon	α	β	β-α	Posterior Prob β>α	Emp. Bayes Factor	PSRF	N_{eff}	3D rate plot?
249	0.138179	1.51208	1.3739	0.988732	622.977	1.03135	144.342	[SVG][PNG]



225	0.292592	2.75072	2.45813	0.955732	24.0613	1.00959	348.105	[SVG][PNG]
282	0.186706	2.31058	2.12387	0.982502	62.578	1.0103	332.362	[SVG][PNG]

Murrell et al | **Mol. Biol. Evol** | 30(5) | 1196–1205

FUBAR results

- **West Nile Virus NS3 protein**
    - A single site (**249**, same as in Brault *et al*) with significant evidence of pervasive diversifying selection.
- **HIV-1 transmission pair**
    - 6 sites with significant evidence of **pervasive** diversifying selection.

Site-level analysis (MEME): branch-level variation; episodic diversifying selection; mixed effects.

WNV NS3

Codon	α	β^-	Pr[$\beta=\beta^-$]	β^+	Pr[$\beta=\beta^+$]	p-value	q-value	Branch-site information
249	0	0	1.00001e-09	2.44107	1	0.0166355	1	[Display]
496	0.947581	0	0.955206	74.0706	0.0447939	0.0838444	1	[Display]
557	0.275214	0	0.96329	171.086	0.0367096	0.0261796	1	[Display]

HIV-1 *env*

Codon	α	β^-	Pr[$\beta=\beta^-$]	β^+	Pr[$\beta=\beta^+$]	p-value	q-value	Branch-site information
161	0	0	0.843437	123.559	0.156563	0.00849354	1	[Display]
165	0.00452376	0.00414013	0.78137	60.3154	0.21863	0.0689752	1	[Display]
178	0	0	0.905795	61.7127	0.0942054	0.0359584	1	[Display]
225	0.00411257	0.00376381	0.784537	72.3307	0.215463	0.0458646	1	[Display]
264	0.000305218	0.000279507	0.913632	881.408	0.0863684	0.0895929	1	[Display]
268	3.66399	0	0.952277	7197.03	0.0477228	0.0882642	1	[Display]
270	1.4879	0	0.673836	108.028	0.326164	0.0819668	1	[Display]
272	0.00301027	0	0.875727	59.7289	0.124273	0.0525552	1	[Display]
274	4.98998	0	0.764384	10000	0.235616	0.0943018	1	[Display]
282	0	0	1.00003e-09	10.2599	1	0.0886042	1	[Display]

Murrell et al | **PLoS Genetics** | 8(7) | e1002764.

MEME results

- **West Nile Virus NS3 protein**
  - **`Three`** sites, (including **`249`**) with significant evidence of **episodic** (or pervasive) diversifying selection.
- **HIV-1 transmission pair**
  - **`10`** sites with significant evidence of **episodic** (or pervasive) diversifying selection; this includes **`4/6`** sites found by FUBAR.

Analysis summary

	WNV NS3	HIV-1 *env*
Gene-wide episodic selection (BUSTED)	No (marginal)	Yes
Branch-level selection (aBSREL)	No	Yes, three branches, including transmission
Site-level pervasive selection (FUBAR)	Yes, `1` site	Yes, `6` sites
Site-level episodic selection (MEME)	Yes, `3` sites (FUBAR + `2` more)	Yes, `10` sites (`4/6` FUBAR + `6` new)

It is **not** unexpected that site-level positive results can occur when a gene-level test does not yield a positive result

---

- **Lack of power for the global test**: if the proportion of sites under selection is very small, a mixture-model test, like BUSTED will miss it

- **Model violations:** MEME and FUBAR supply much more flexible distributions of `dN/dS` over sites

- **False positives at site-level**: our site-level tests have good statistical properties, but each positive site result could be a false positive; FWER correction would make site-level tests too conservative.

- **Summary**: gene-level selection tests need a minimal proportion of sites to be under selection to be powered; site-level tests should not be used to make inferences about gene-level selection.

Result differences between MEME and FUBAR

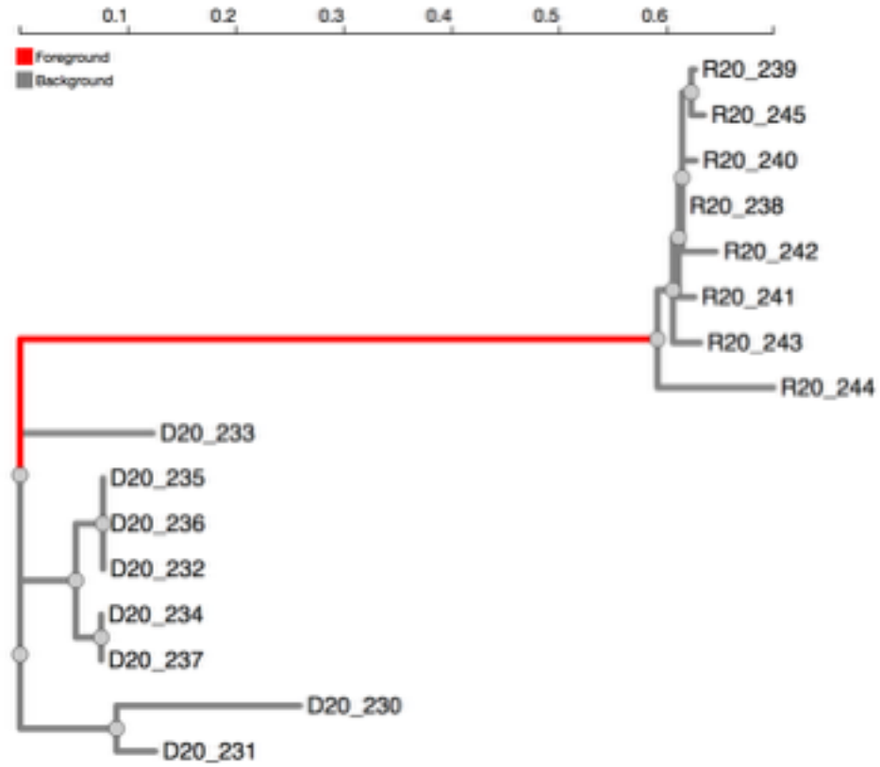
- MEME generally is more powerful, because it does **not** pool dN/dS over all branches, but rather uses a 2-mixture model to include branch-to-branch variation.

- If diversifying selection occurred along a small proportion of branches at a site, averaging over branches will dilute the signal.

Codon	α (dS)	β- (dN<dS)	Pr[β=β-]	β+ (dN>dS)	Pr[β=β+]	p-value
249	0.00	0.00	0.00	**2.44**	**1.00**	0.02
496	0.95	0.00	0.96	**74.07**	**0.04**	0.08
557	0.28	0.00	0.96	**171.09**	**0.04**	0.03

Branch testing; exploratory vs *a priori*

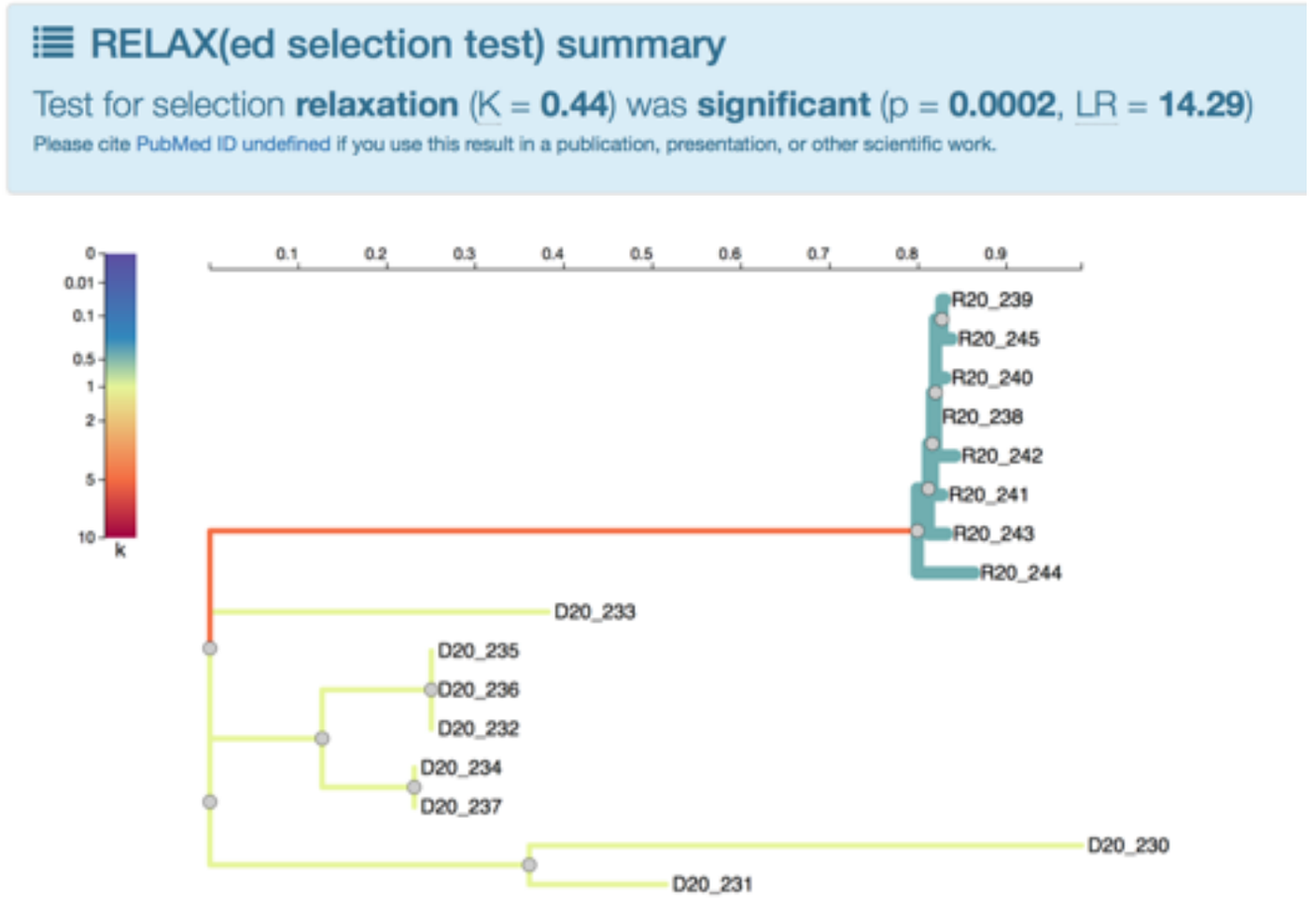
- aBSREL and BUSTED can test all branches for selection (exploratory), or apply the test to a set of branches defined *a priori* (e.g. defining a particular biological hypothesis).
- For BUSTED, *a priori* partitioning of branches can increase power, especially if selective regimes are markedly different on different parts of the tree.
- For example, BUSTED applied to the HIV dataset where the transmission branch is designated as foreground, found a greater proportion sites under stronger selection on this branch that the rest of the tree, and a lower **`p-value`**.

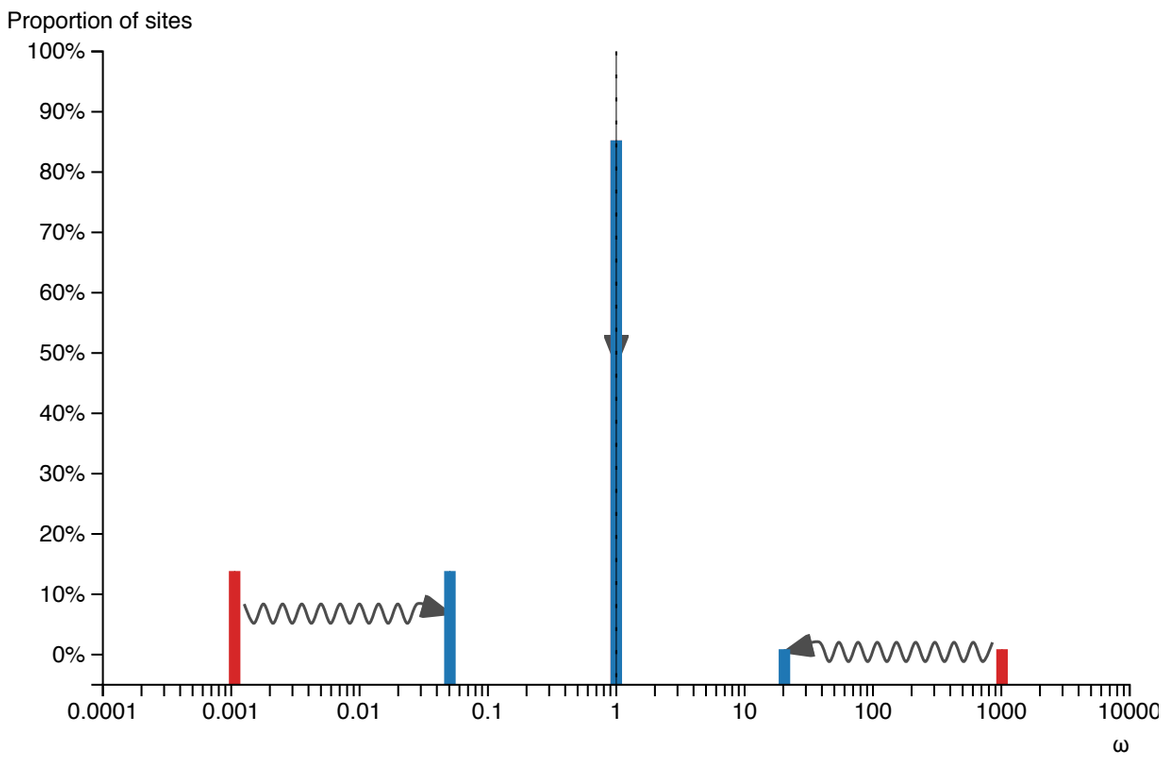


	Background	Foreground
Class 1	ω = 0.51 p = 0.08	ω = 0.00 p = 0.92
Class 2	ω = 0.72 p = 0.91	
Class 3	ω = 116 p = 0.01	ω = 510 p = 0.08

Relaxed/Intensified selection (RELAX)

- Sometimes it is useful to compare **distributions** of `dN/dS` between groups of branches.
- For example, in an HIV-1 transmission pair tree we may want to see if sequences in the recipient are under more or less intense selection compared to the source.
- We partition the tree into test, reference, and nuisance branches, and examine if the **distribution** of `dN/dS` in the test branches is shrunk towards (`K < 1`, relaxed) or expanded away (`K > 1`, intensified) from `dN/dS=1` relative to the reference branches.
- In the HIV env example, selection in the recipient is significantly relaxed compared to that in the source.





Wertheim et al | **Mol Biol Evol** | 32(3) | 820–832 .

Current suggested best practices.

- There are lots of methods you could use to study positive selection, including about 10 developed by our group. The field is still evolving, and this is our current suggestions of what to do with your data, depending on the question you want to answer.

Question	Method	Output
Is there episodic selection anywhere in my gene (or along a set of branches known a priori)?	Branch-site unrestricted statistical test of episodic diversification (BUSTED).	• p-value for gene-wide selection • inferred dN/dS distributions • a “quick and dirty” scan of sites where selection could have operated.
Are there branches in the tree where some sites have been subject to diversifying selection? Also: inferring ancient divergence times.	Adaptive branch site random effects likelihood (aBSREL)	• p-values for each branch • dN/dS distributions for each branch • evolutionary process complexity
Are there sites in the alignment where some of the branches have experienced diversifying selection?	Mixed effects model of evolution (MEME)	• p-values for each site • dN/dS distributions for each site
Are there sites which have experiences diversifying selection **and** my alignment is large?	Fast unconstrained bayesian analysis of selection (FUBAR)	• Posterior probabilities of selection at each site • An estimate of the the gene-wide dN/dS distribution

The role of physico-chemical properties

- In their influential paper "*Evolutionary Divergence and Convergence in Proteins*", written nearly 50 years ago, Zuckerkandl and Pauling (1965) championed the emerging field of chemical paleogenetics – more commonly known today as molecular phylogenetics.

- They discussed what was known about amino acid substitution patterns, how these were affected both by the genetic code and by the physicochemical properties of amino acids, and how the notions of radical and conservative substitutions should be defined.

"*The inadequacy of a priori views on conservatism and non-conservatism is patent. Apparently chemists and protein molecules do not share the same opinions regarding the definition of the most prominent properties of a residue*."

"*A certain new function – say, associated with charge – may be introduced while the former function – say, apolar interaction – is maintained. This simultaneity of conservatism and nonconservatism may well also be one of the basic conditions of protein evolution and organic evolution in general*"

Existing models

- Very few existing evolutionary models for protein-coding sequences consider biochemical residue properties.

- None allow these properties to vary from site to site in a tractable fashion.

- Zuckerkandl and Pauling would be disappointed.

Site-to-site variation of property importance

- Some properties are conserved more strongly at some sites than at others.
- When property importance varies from site to site, our notion of radical changes vs conservative changes also varies from site to site.
- Aim: which property changes at which sites are avoided, tolerated, or encouraged?

PRIME: Property Informed Model of Evolution

$$q_{ij} = \begin{cases} \alpha\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = AA(j), \\ \beta_{xy}\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = x \neq AA(j) = y, \\ 0, & \text{for } \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & \text{for } \delta(i,j) = 0, \end{cases}$$

$$\beta_{xy, x \neq y} = f^{(s)}\left(\overrightarrow{d(x,y)}\right)$$

$$f^{(s)}\left(\overrightarrow{d(x,y)}\right) = \alpha^{(s)} \prod_{i=1}^{D} \left[e^{-\lambda_i^{(s)} d_i(x,y)}\right] = \alpha^{(s)} \exp\left[-\sum_{i=1}^{D} \lambda_i^{(s)} |x_i - y_i|\right]$$

PRIME: Property Informed Model of Evolution

- Each site has a set of property importance parameters λ describing how strongly each property is conserved:

$$q_{ij} = \begin{cases} \alpha\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = AA(j), \\ \beta_{xy}\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = x \neq AA(j) = y, \\ 0, & \text{for } \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & \text{for } \delta(i,j) = 0, \end{cases}$$

$$\beta_{xy,x \neq y} = f^{(s)}\left(\overrightarrow{d(x,y)}\right)$$

$$f^{(s)}\left(\overrightarrow{d(x,y)}\right) = \alpha^{(s)} \prod_{i=1}^{D} \left[e^{-\lambda_i^{(s)} d_i(x,y)}\right] = \alpha^{(s)} \exp\left[-\sum_{i=1}^{D} \lambda_i^{(s)} |x_i - y_i|\right]$$

PRIME: Property Informed Model of Evolution

- Each site has a set of property importance parameters λ describing how strongly each property is conserved:
  - Large λ: strongly conserved

$$q_{ij} = \begin{cases} \alpha\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = AA(j), \\ \beta_{xy}\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = x \neq AA(j) = y, \\ 0, & \text{for } \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & \text{for } \delta(i,j) = 0, \end{cases}$$

$$\beta_{xy,x\neq y} = f^{(s)}\left(\overrightarrow{d(x,y)}\right)$$

$$f^{(s)}\left(\overrightarrow{d(x,y)}\right) = \alpha^{(s)} \prod_{i=1}^{D} \left[e^{-\lambda_i^{(s)} d_i(x,y)}\right] = \alpha^{(s)} \exp\left[-\sum_{i=1}^{D} \lambda_i^{(s)} |x_i - y_i|\right]$$

PRIME: Property Informed Model of Evolution

- Each site has a set of property importance parameters λ describing how strongly each property is conserved:
  - Large λ: strongly conserved
  - Small $|\lambda|$: not important

$$q_{ij} = \begin{cases} \alpha\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = AA(j), \\ \beta_{xy}\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = x \neq AA(j) = y, \\ 0, & \text{for } \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & \text{for } \delta(i,j) = 0, \end{cases}$$

$$\beta_{xy,x\neq y} = f^{(s)}\left(\overrightarrow{d(x,y)}\right)$$

$$f^{(s)}\left(\overrightarrow{d(x,y)}\right) = \alpha^{(s)} \prod_{i=1}^{D} \left[e^{-\lambda_i^{(s)} d_i(x,y)}\right] = \alpha^{(s)} \exp\left[-\sum_{i=1}^{D} \lambda_i^{(s)} |x_i - y_i|\right]$$

PRIME: Property Informed Model of Evolution

- Each site has a set of property importance parameters λ describing how strongly each property is conserved:
  - Large λ: strongly conserved
  - Small $|\lambda|$: not important
  - Negative λ: driven to change

$$q_{ij} = \begin{cases} \alpha\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = AA(j), \\ \beta_{xy}\theta_{ij}\pi_j, & \text{for } \delta(i,j) = 1, AA(i) = x \neq AA(j) = y, \\ 0, & \text{for } \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & \text{for } \delta(i,j) = 0, \end{cases}$$

$$\beta_{xy,x\neq y} = f^{(s)}\left(\overrightarrow{d(x,y)}\right)$$

$$f^{(s)}\left(\overrightarrow{d(x,y)}\right) = \alpha^{(s)} \prod_{i=1}^{D} \left[e^{-\lambda_i^{(s)} d_i(x,y)}\right] = \alpha^{(s)} \exp\left[-\sum_{i=1}^{D} \lambda_i^{(s)} |x_i - y_i|\right]$$

Which properties?

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

- Highly correlated – cannot ascribe effects uniquely to one property.

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

- Highly correlated – cannot ascribe effects uniquely to one property.

- Compound properties can be defined arbitrarily.

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

- Highly correlated – cannot ascribe effects uniquely to one property.

- Compound properties can be defined arbitrarily.

- Can specify any set of properties.

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

- Highly correlated – cannot ascribe effects uniquely to one property.

- Compound properties can be defined arbitrarily.

- Can specify any set of properties.

- No set of properties fits best on all data sets.

Which properties?

- More than 500 physico-chemical amino acid properties have been measured.

- Highly correlated – cannot ascribe effects uniquely to one property.

- Compound properties can be defined arbitrarily.

- Can specify any set of properties.

- No set of properties fits best on all data sets.

- Here we use the Conant-Stadler properties (easy to interpret).

The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

SLYNTVATL

The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**

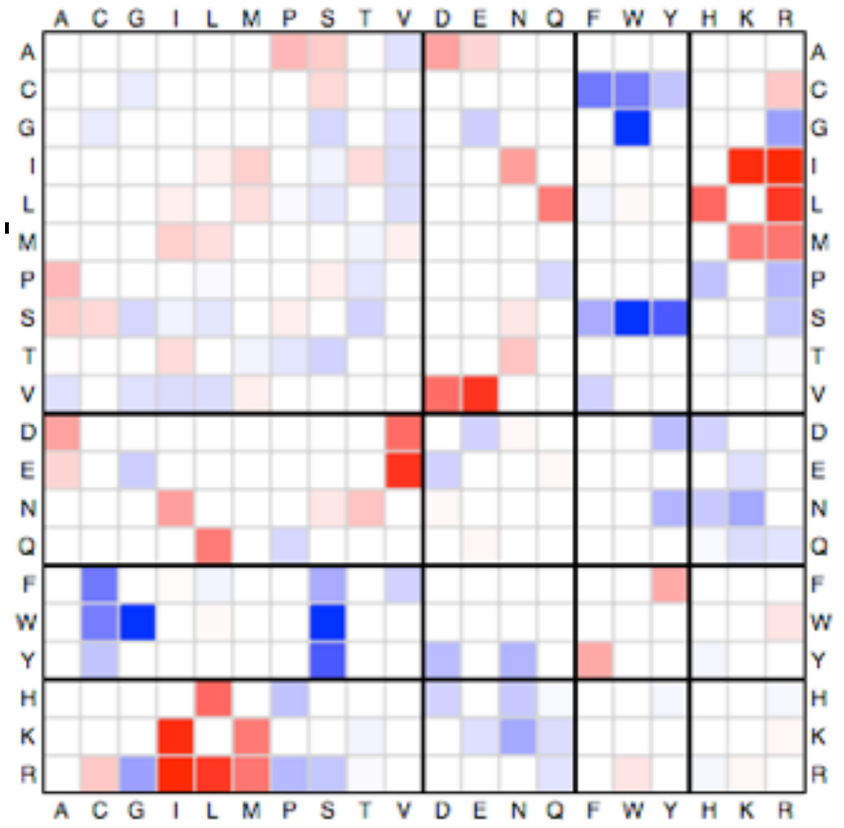
The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**

Site 79: volume is conserved.



The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

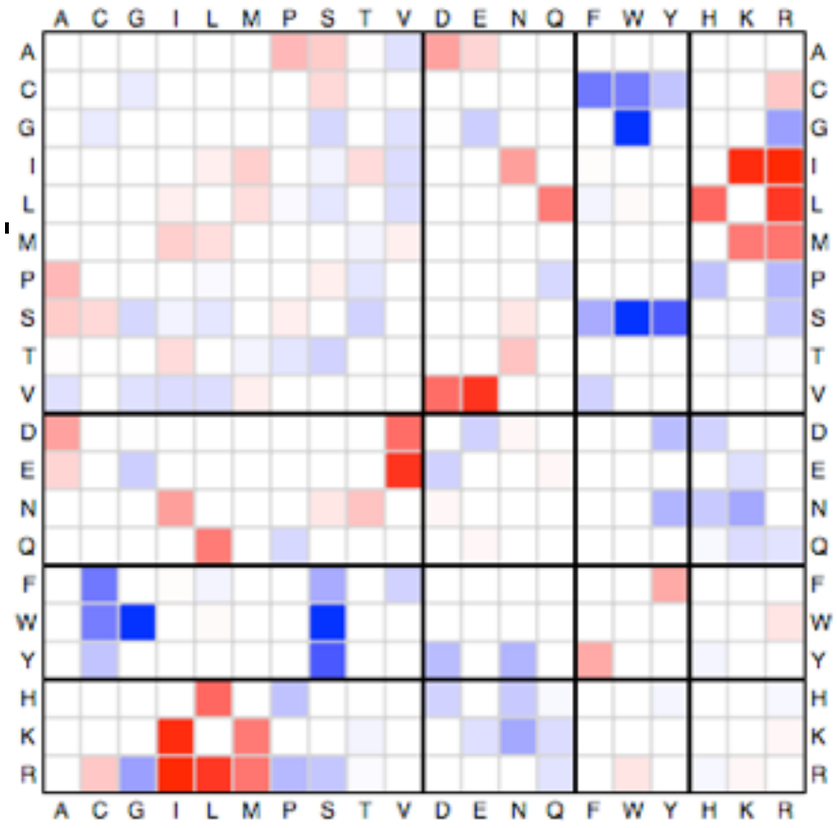
These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 84

Site 79: volume is conserved.

Evolution under positive selection in the presence of a specific biochemical constraint.



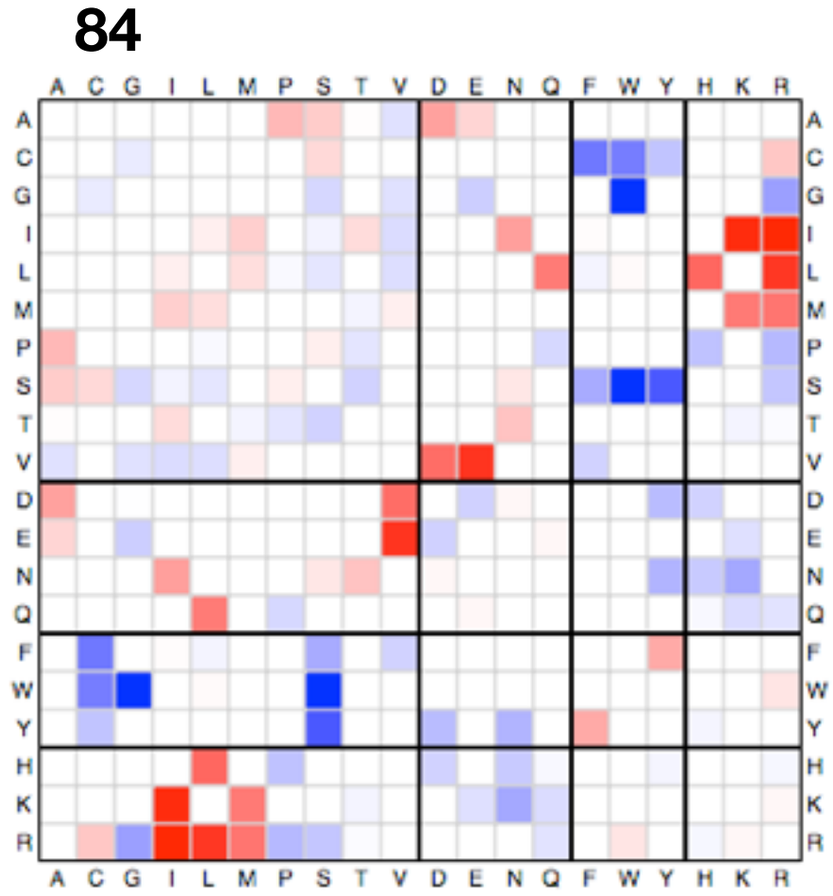
The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**

Site 79: hydropathy is selected to change.



The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

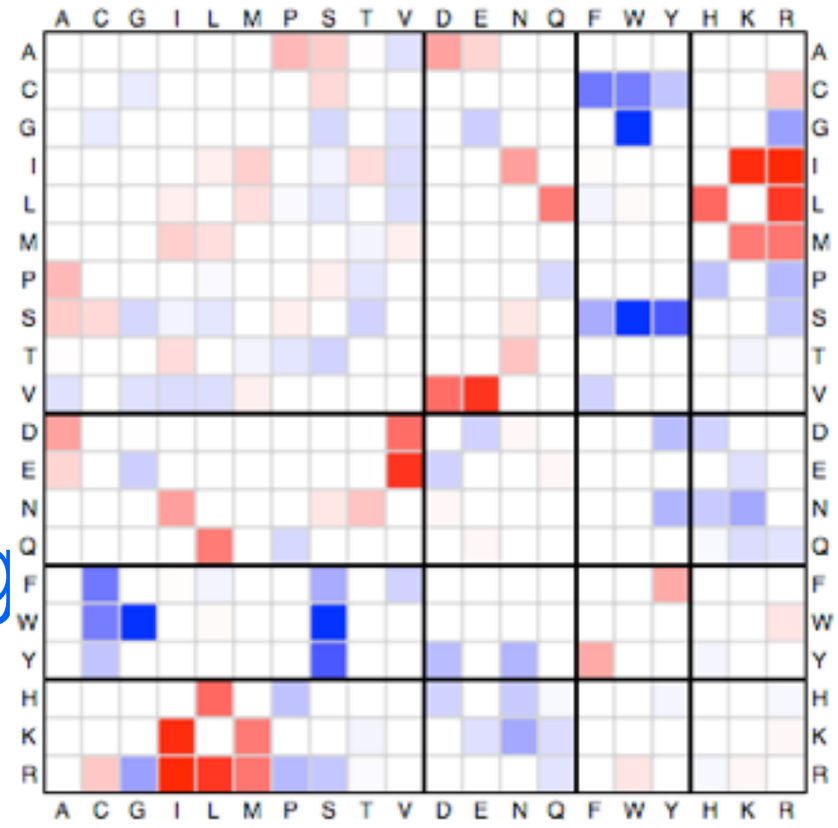
These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**

Site 79: hydropathy is selected to change.

A mechanism through which the epitope is broken while maintaining function.

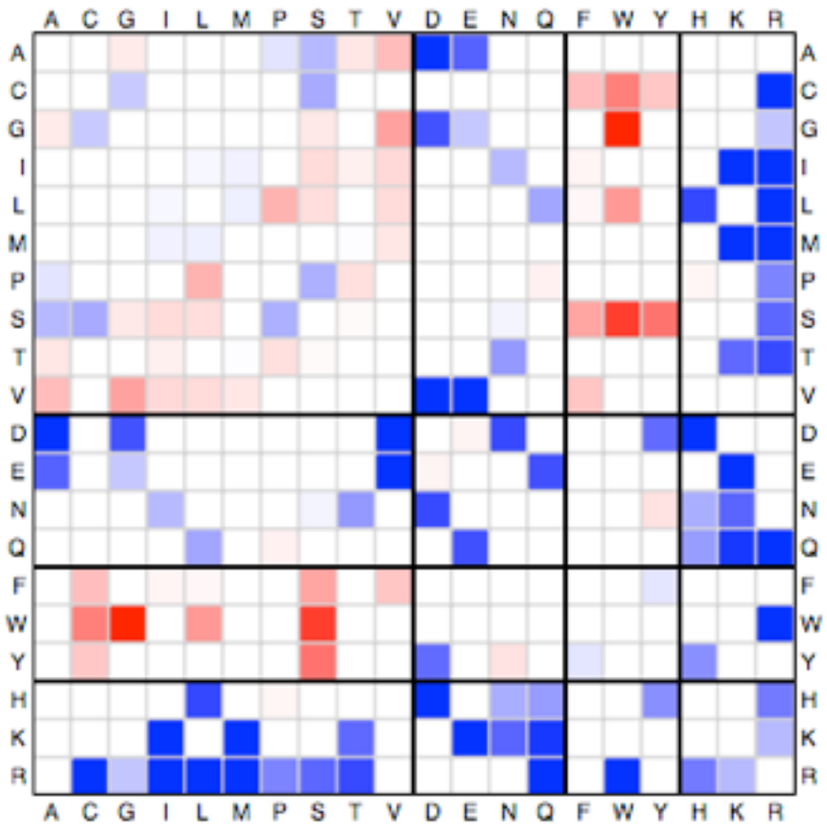


The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**



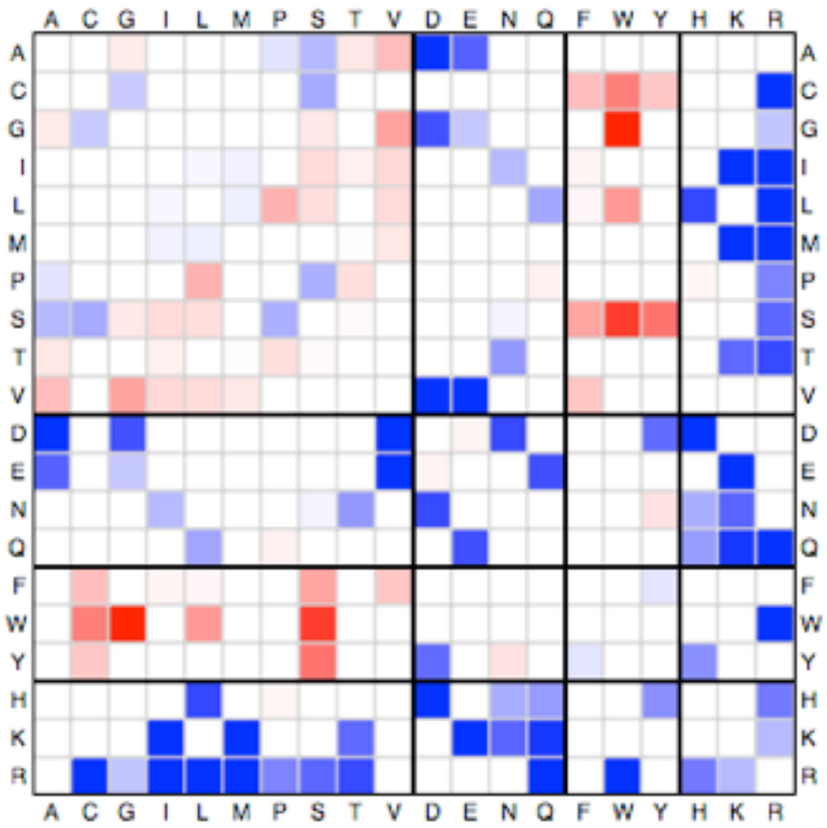
Site 84: polarity is conserved.

The SLYNTVATL epitope (SL9: gag p17 sites 77-85) is targeted by A*0201 CTL response.

These anchor sites are known to be under positive diversifying selection [Edwards *et al.*, JVI 2005]:

SLYNTVATL

79 **84**

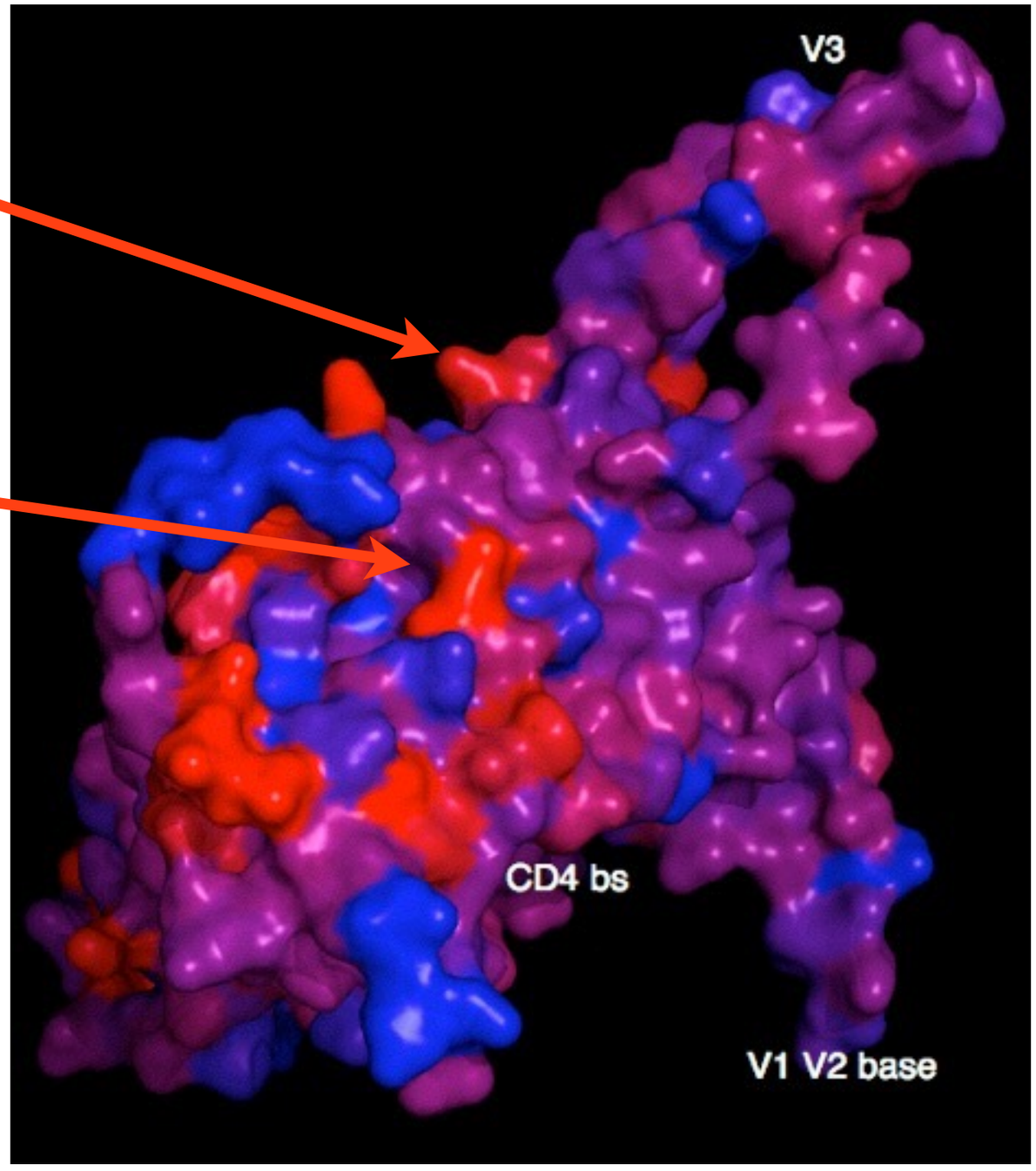


Site 84: polarity is conserved.

Site 84: volume is under selection to change.

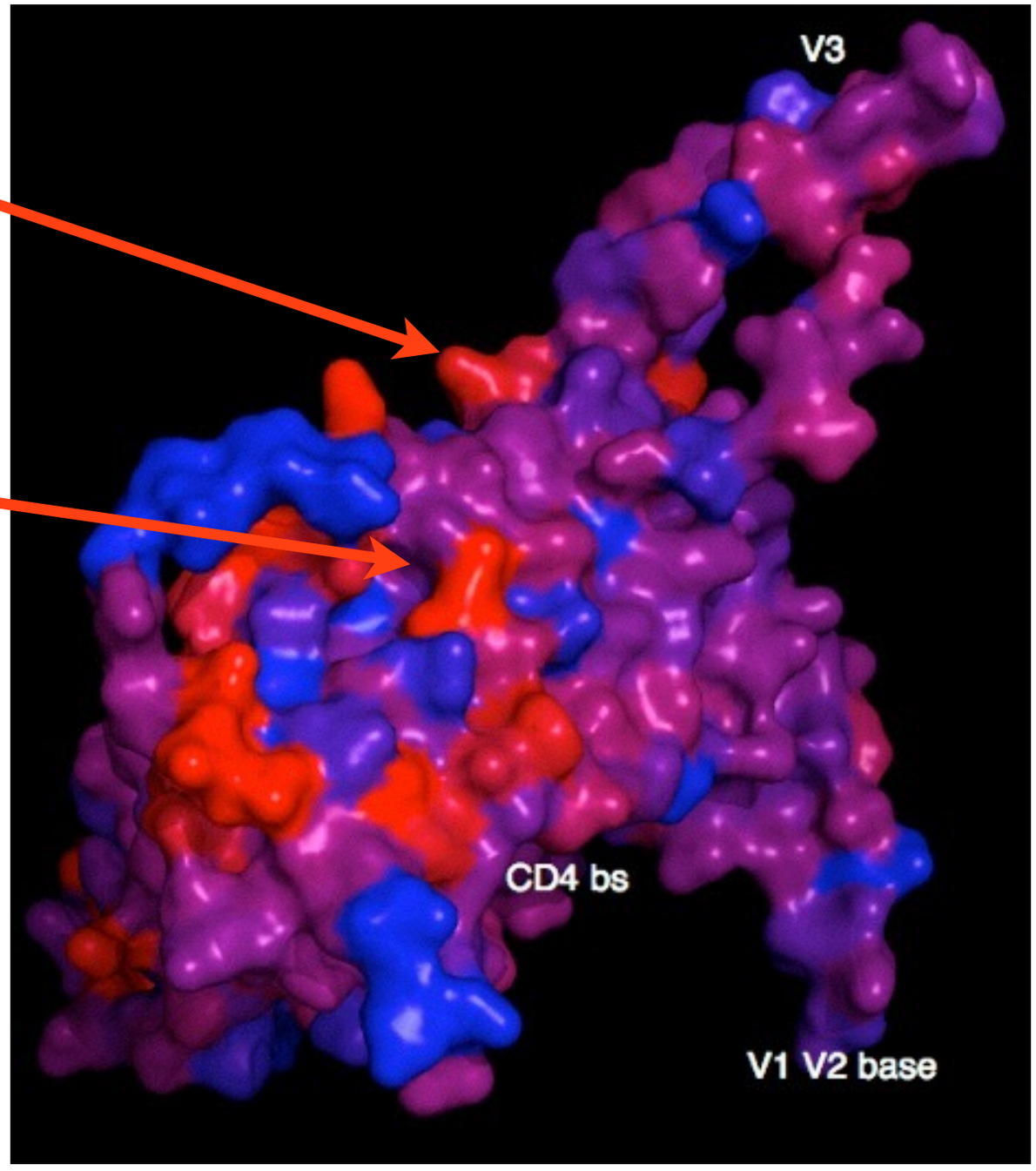
Hydrophilic

Hydrophobic



Hydrophilic

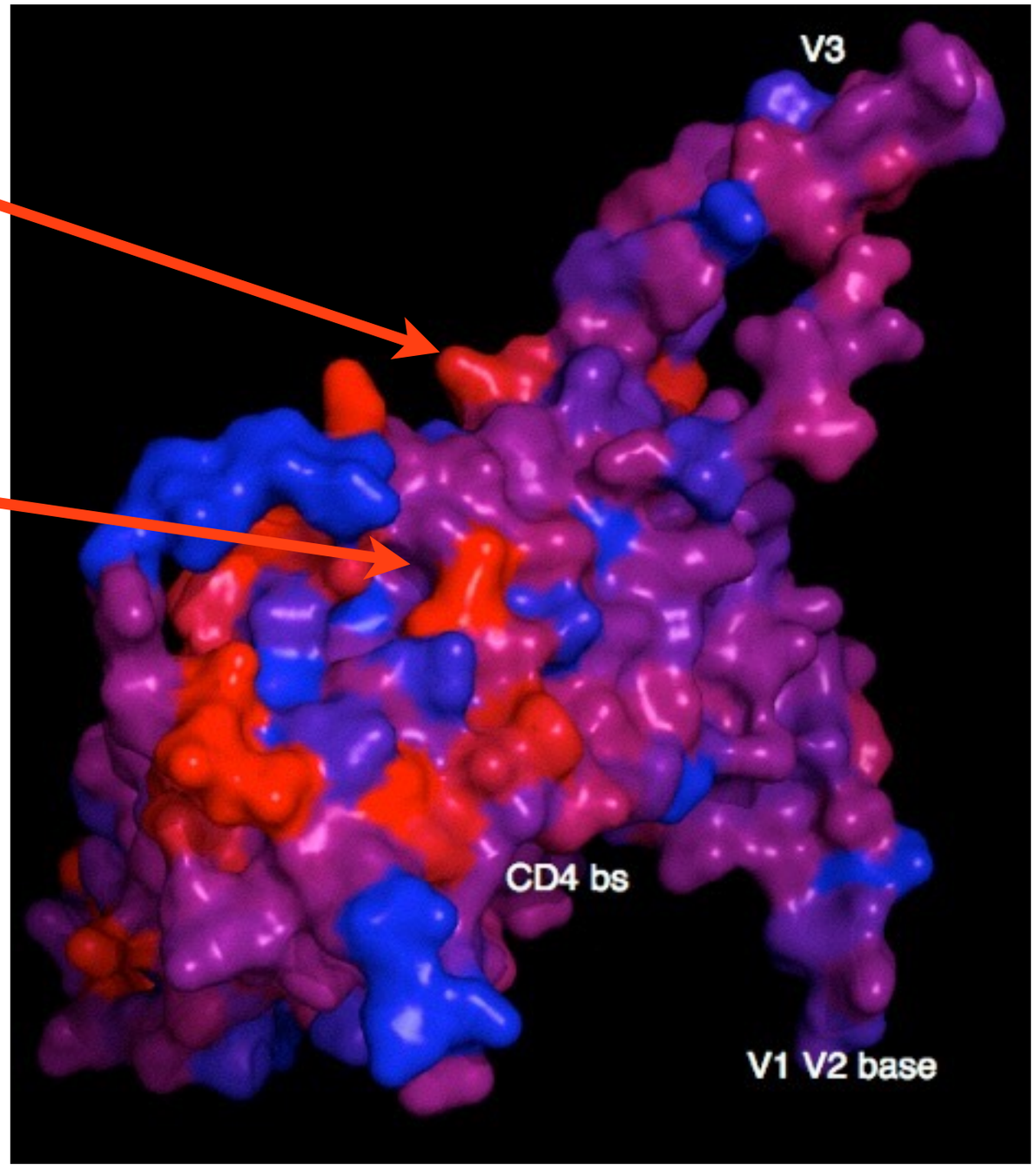
Hydrophobic



Radical vs conservative substitutions

Hydrophilic

Hydrophobic

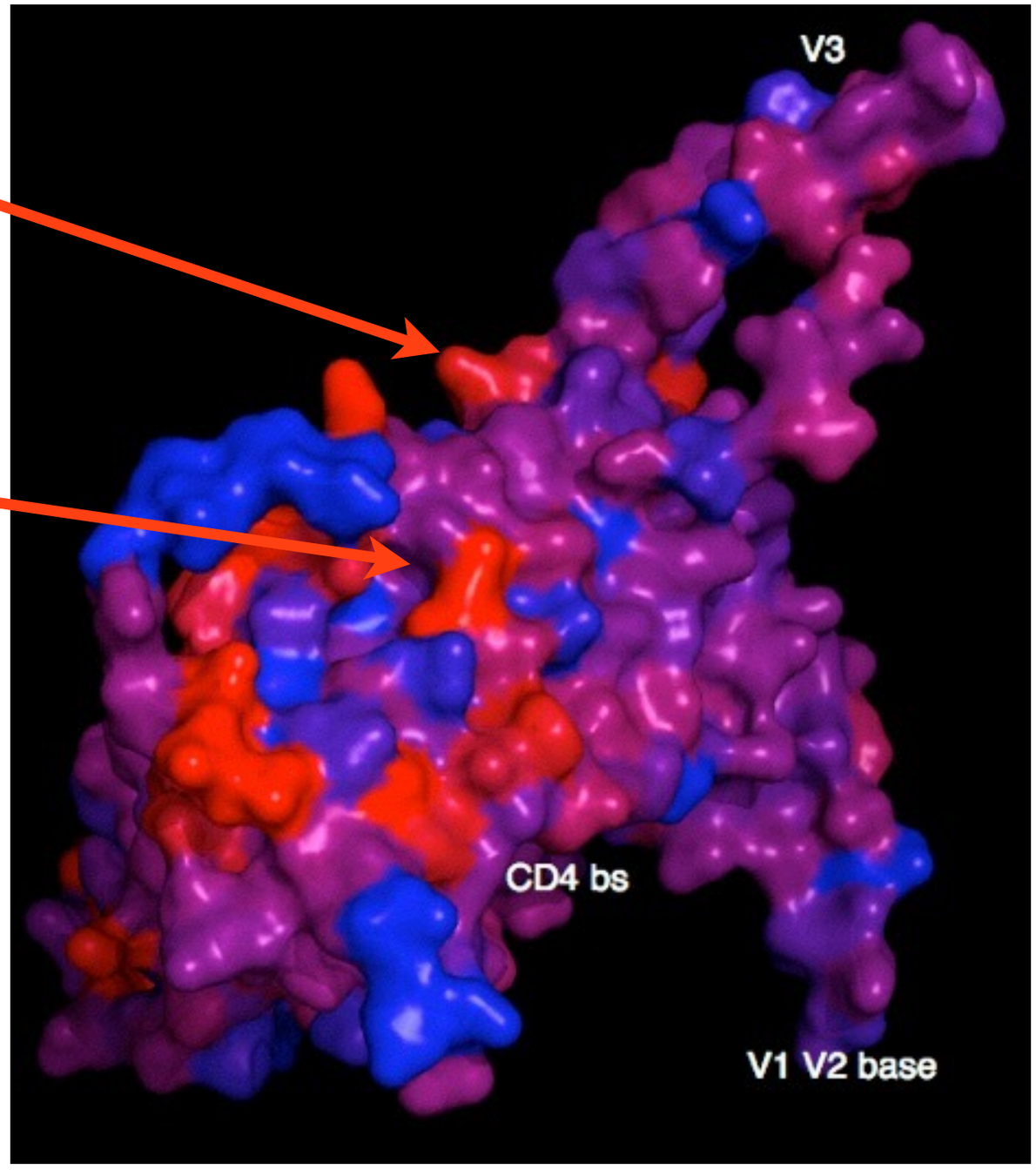


Radical vs conservative substitutions

Radical: large changes in one or more properties

Hydrophilic

Hydrophobic



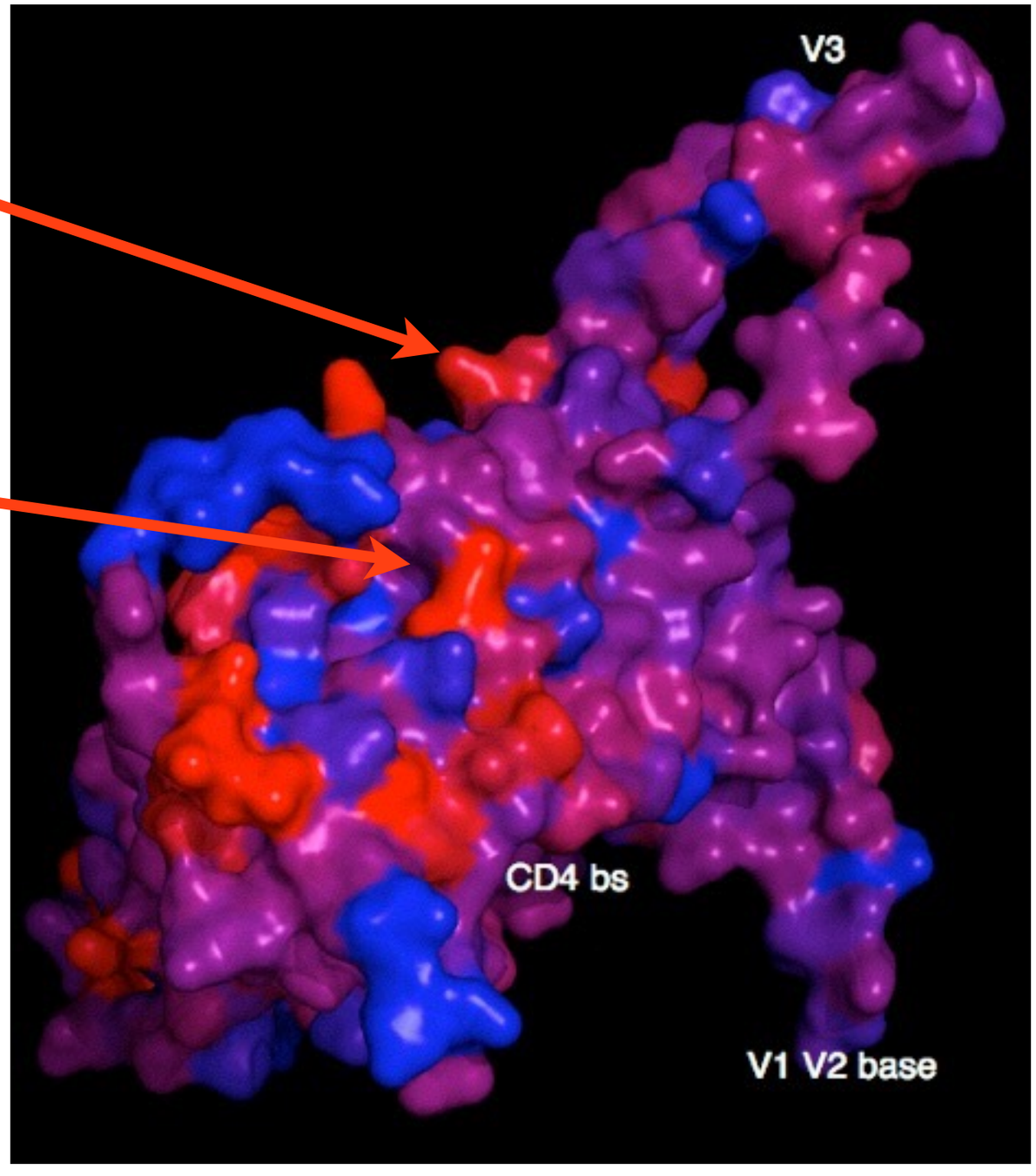
Radical vs conservative substitutions

Radical: large changes in one or more properties

But which properties?

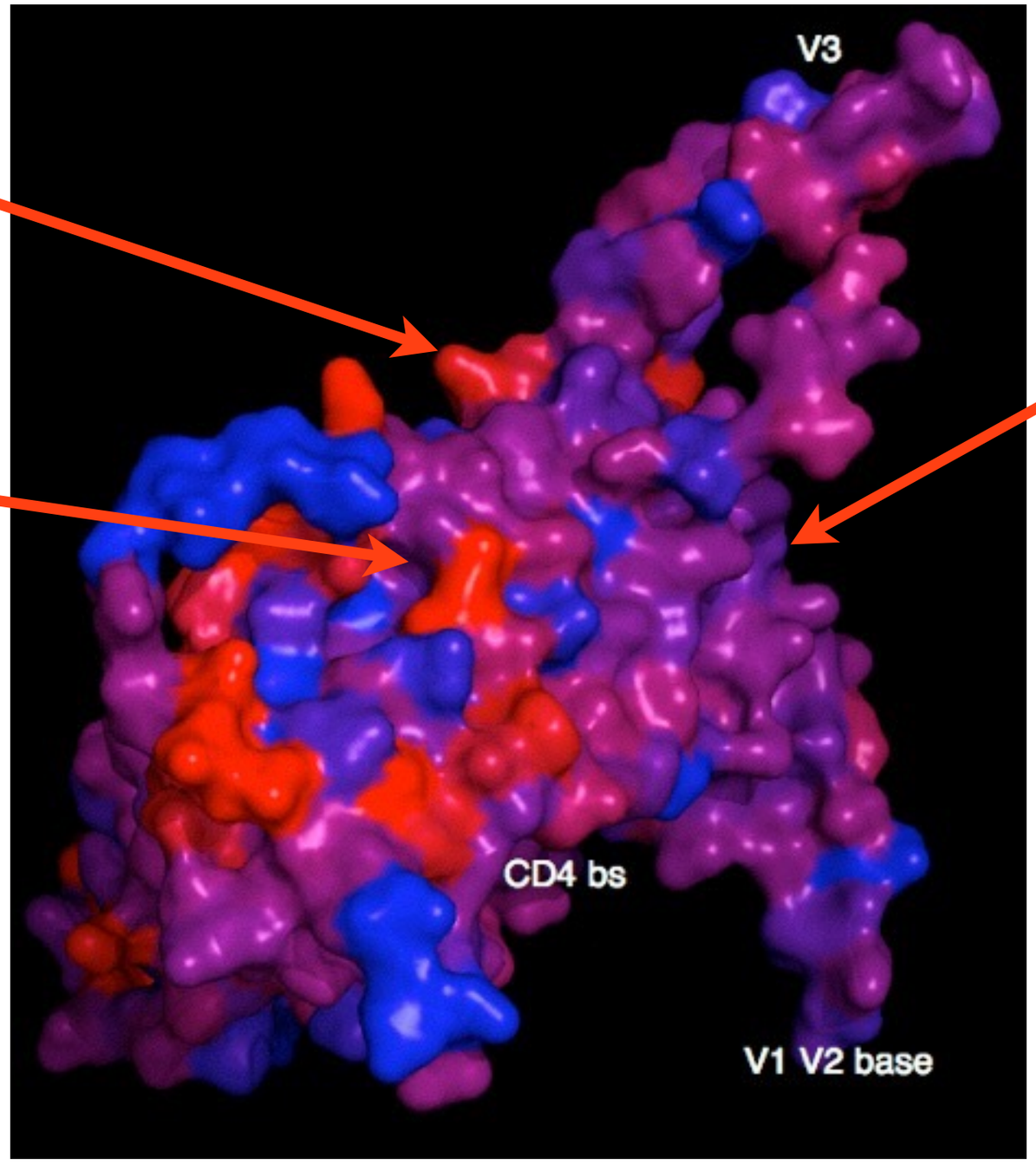
Hydrophobicity matters

Hydrophobicity matters



Hydrophobicity
matters

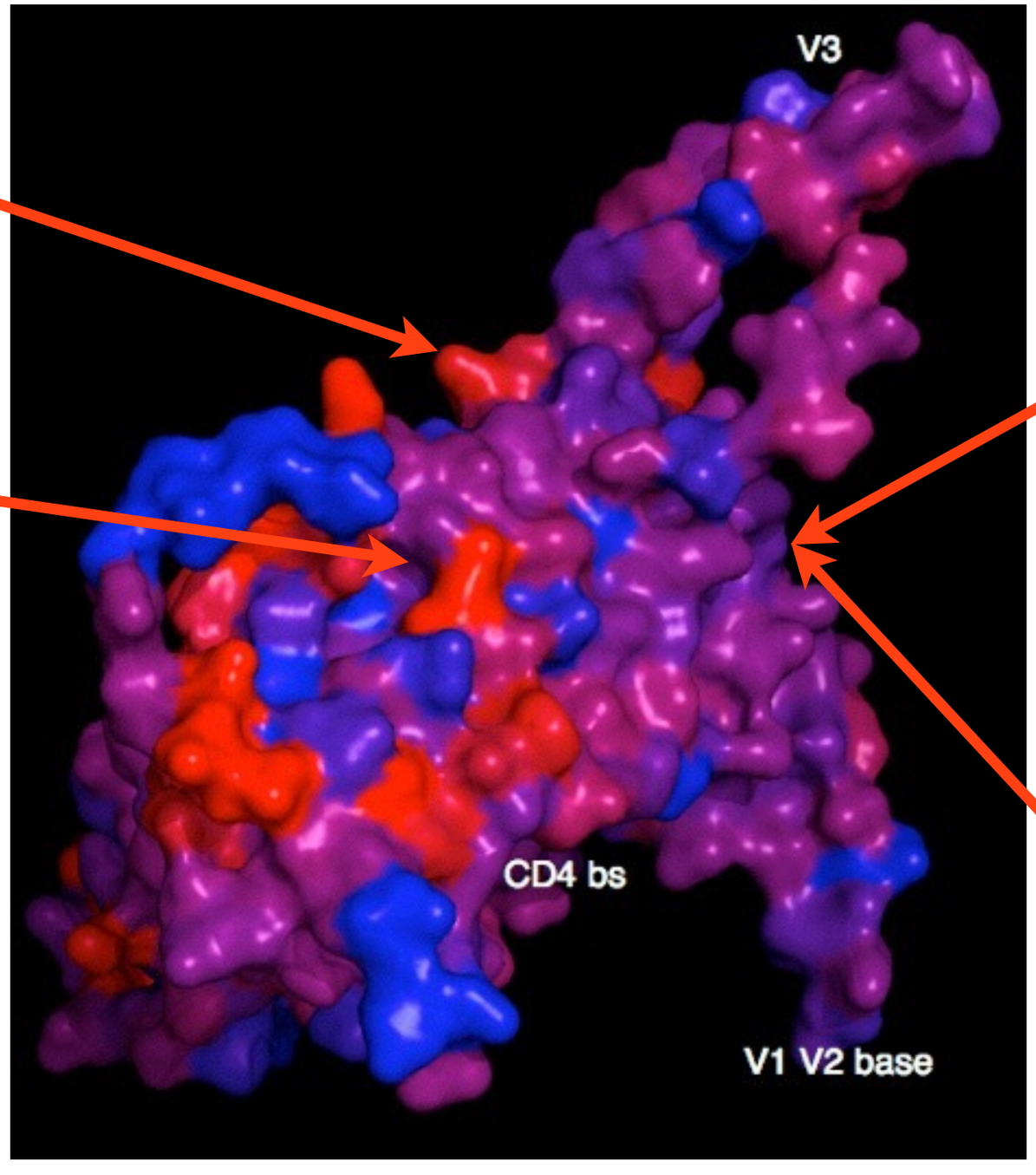
Hydrophobicity
matters



Hydrophobicity
unimportant?

Hydrophobicity
matters

Hydrophobicity
matters



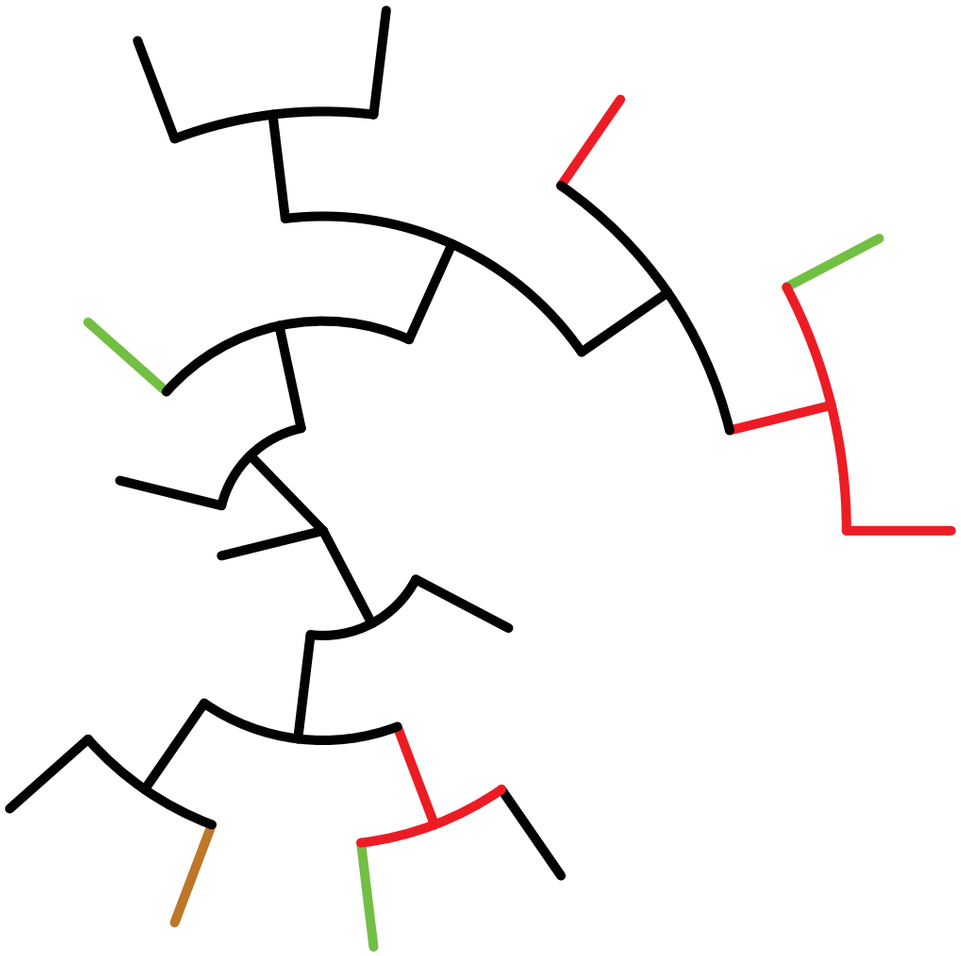
Hydrophobicity
unimportant?

But maybe
charge affects
folding here?

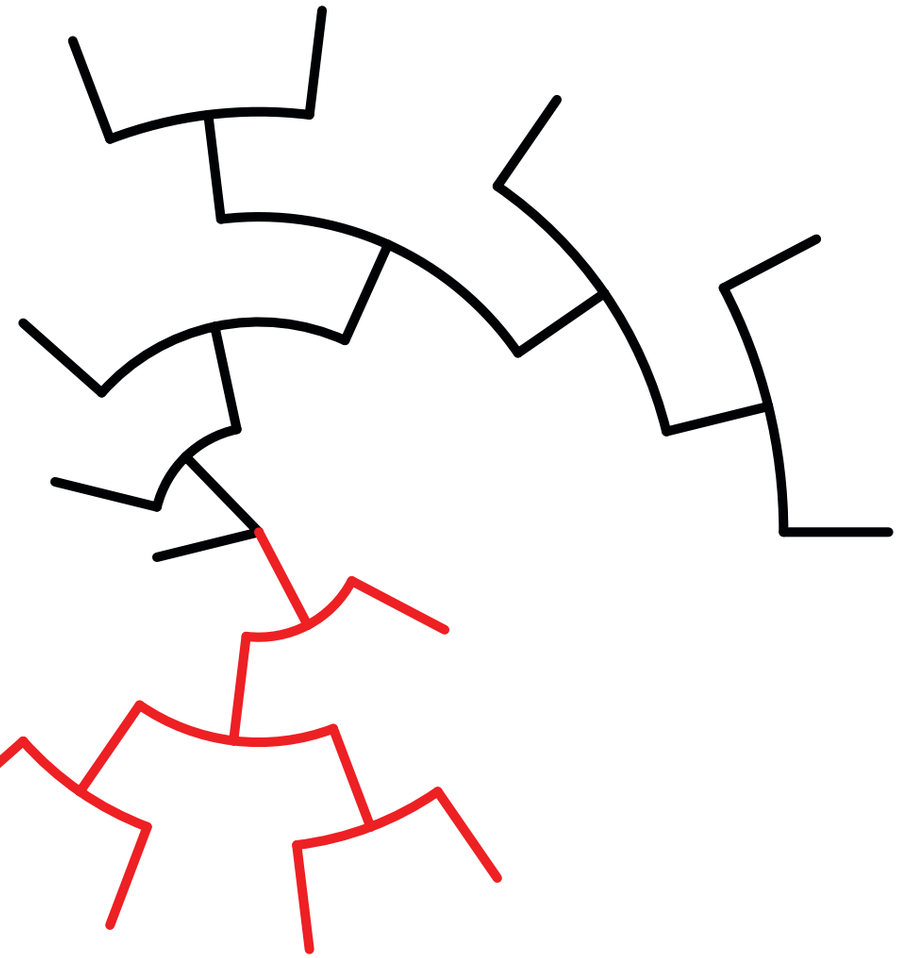
PRIME analysis

Property	WNV NS3	HIV-1 *env*
Chemical Composition	None	Conserved at site 282 (which is overall positively selected according to MEME)
Polarity	None	None
Volume	None	Conserved at site 7
Isoelectric Point	None	Conserved at site 274 (which is overall positively selected according to MEME)
Hydropathy	None	None

Diversifying vs Directional Selection



Diversifying/disruptive
Many non-synonymous substitutions, detected well by dN/dS analyses

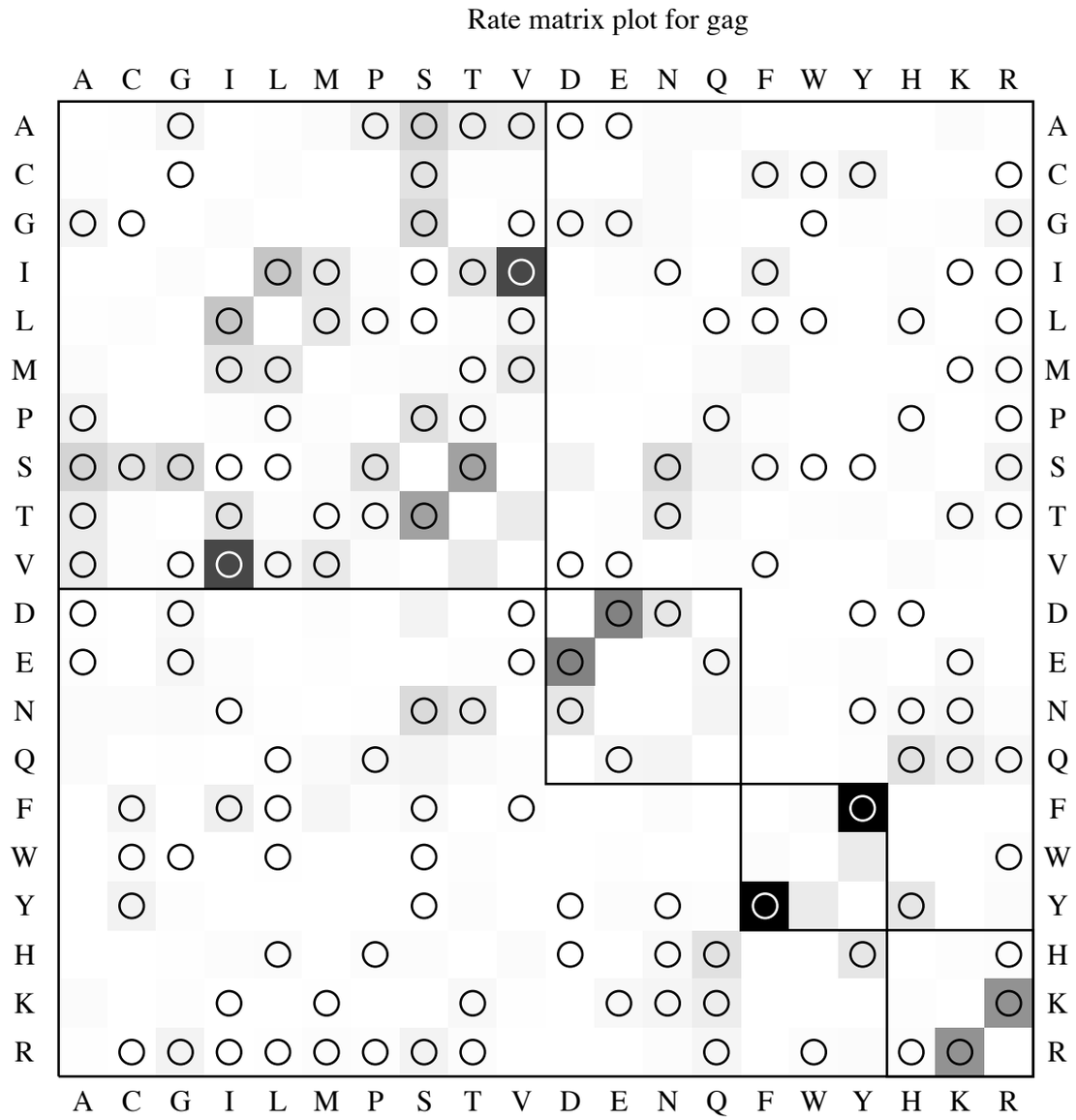


Directional/positive
Very few substitutions, but a significant change in allele frequencies. Confounds dN/dS methods, because their frequency stationary model does not describe the biology well at all

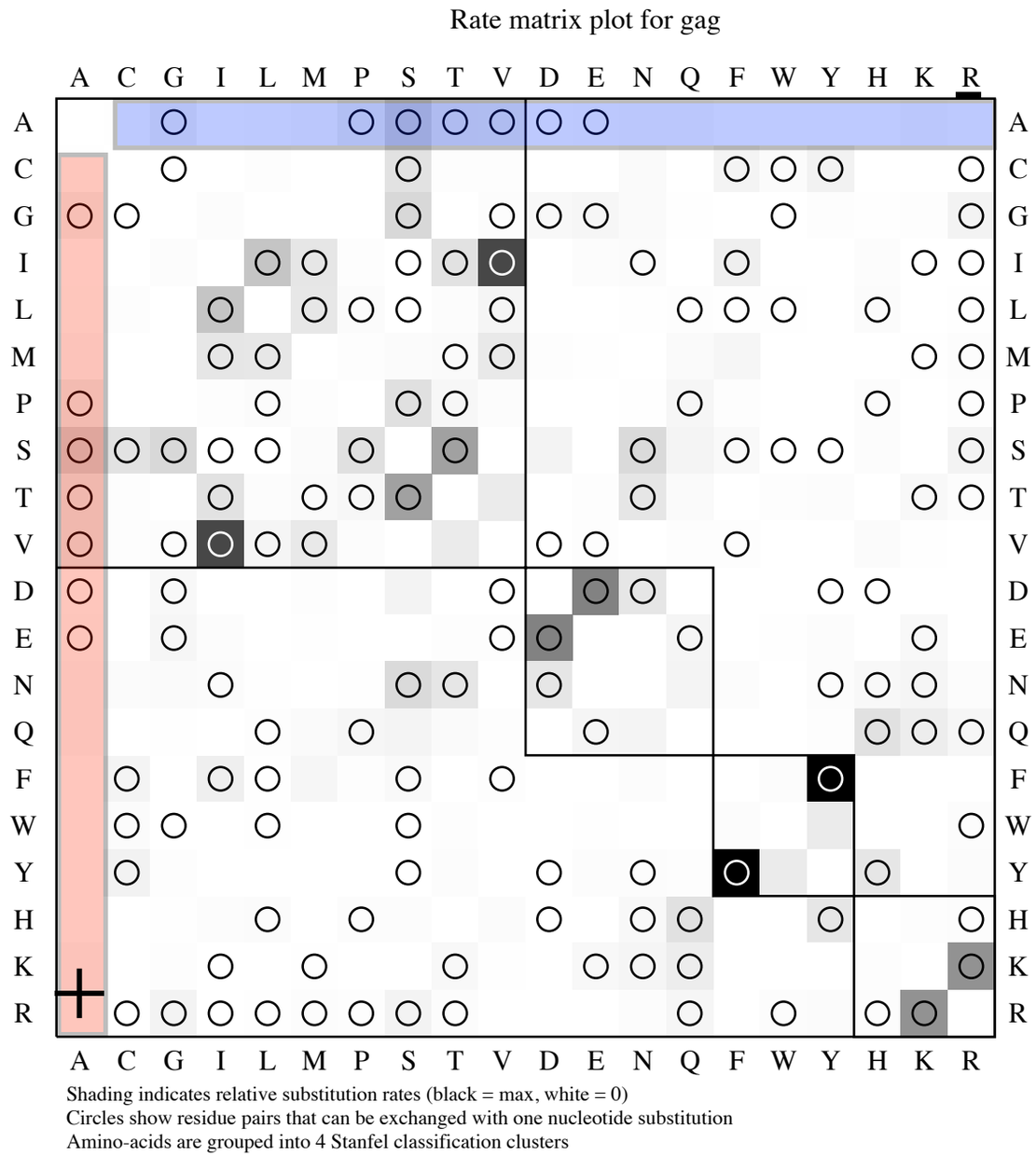
Directional Evolution of Protein Sequences (DEPS)

- **Idea**: Use protein data. Fit an appropriate background null protein model (e.g. WAG, HIV, IAV). Fit a non-reversible alternative model, which allows at a proportion of sites (p) to accumulate substitutions to a given target residue (**T**) faster (factor of **B**>1) than the background model allows. Test if **p**>0 for each **T**.
- **Assumptions**: directional selection applies constant pressure across the site (i.e. it is not episodic); background model is appropriate to describe how amino-acids substitution preferences on average.
- **Best for**: detecting which amino-acids are preferentially substituted for at a subset of sites, e.g. selective sweeps or balancing selection.
- **Powered by**: strength of selection (i.e. the more the frequency of the residue changes, the better) or repeated substitutions towards the same residue.

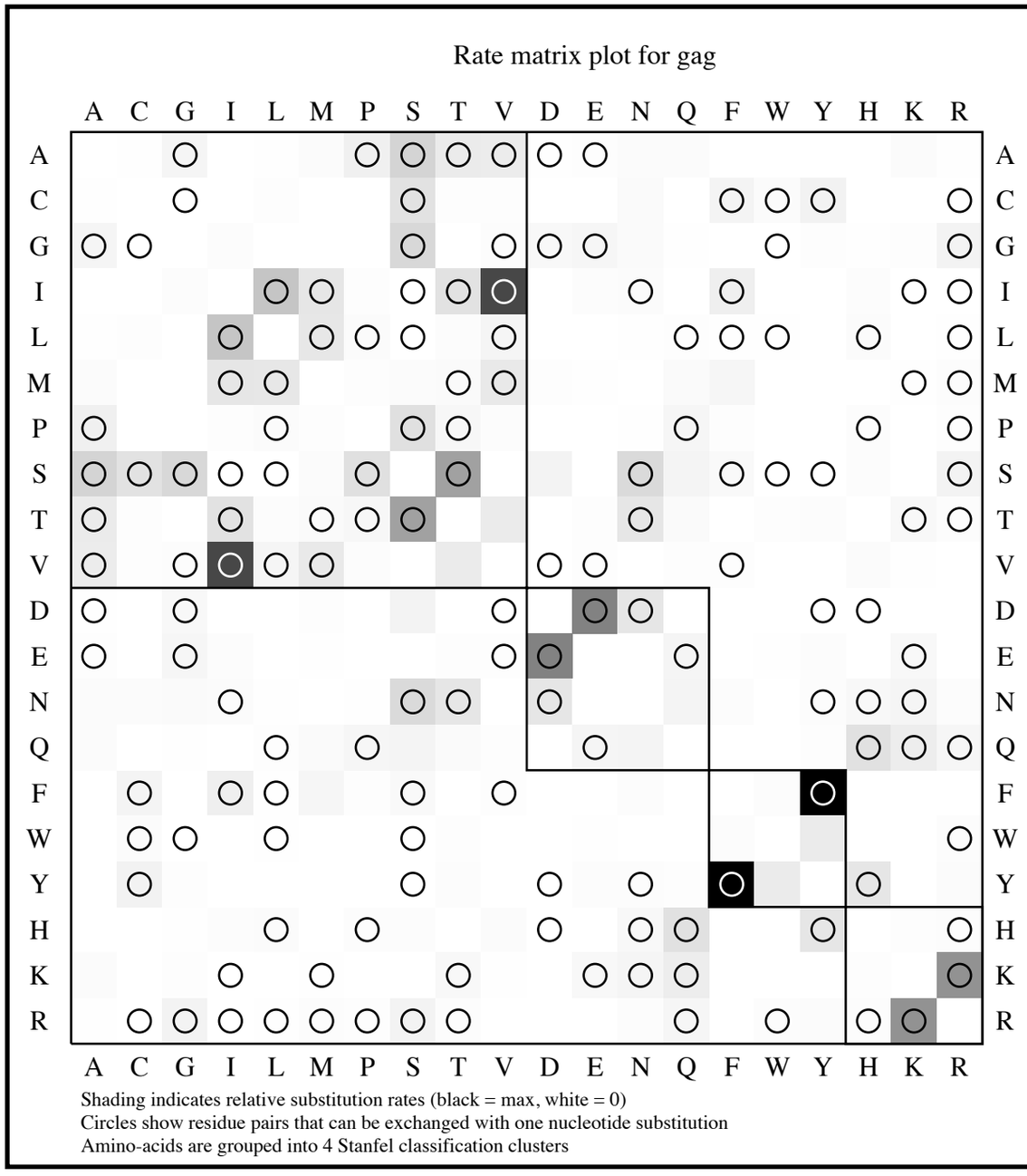
Mol. Biol. Evol. | 25(9):1809-24 | 2008

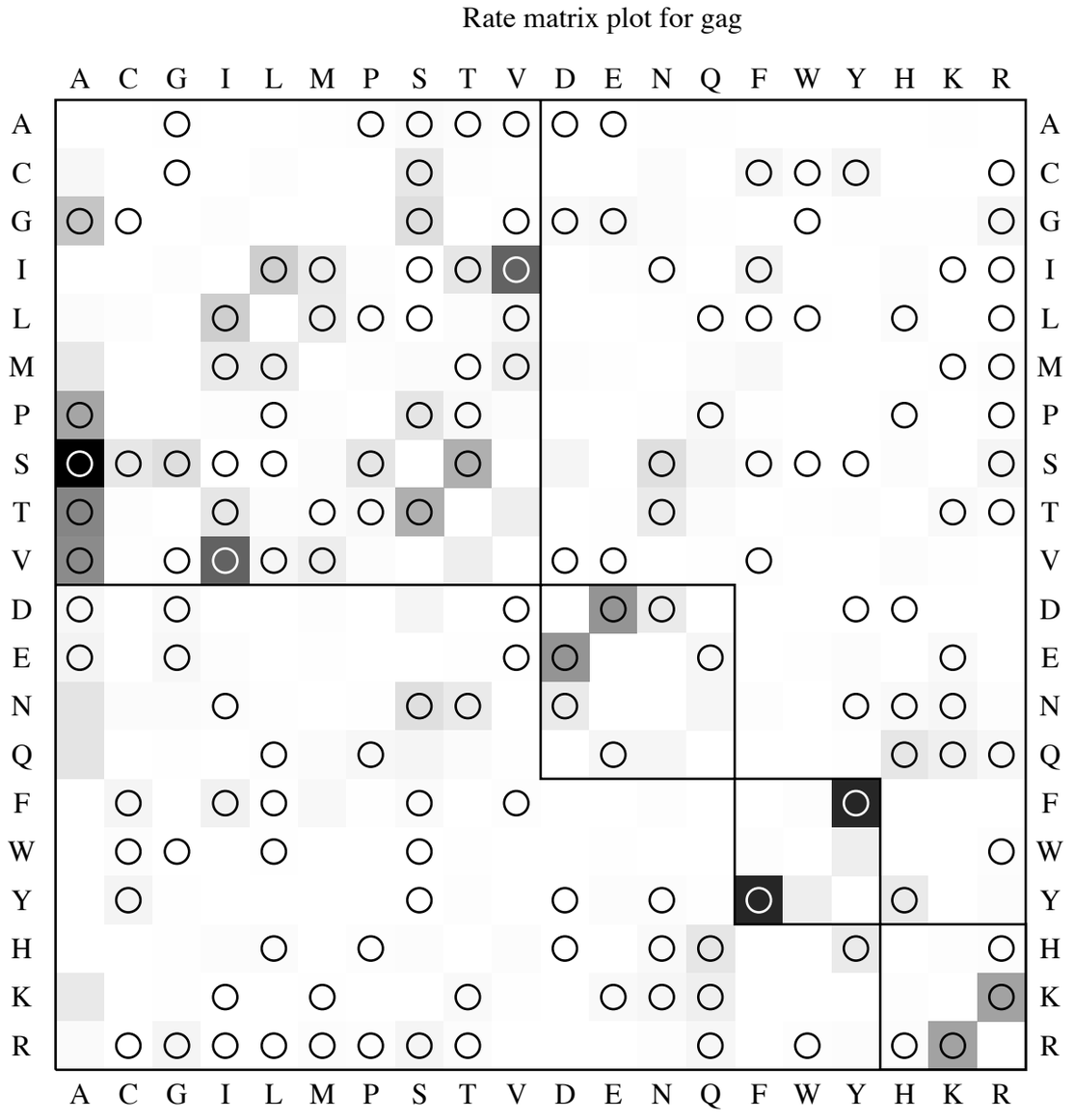


Base model (gag)



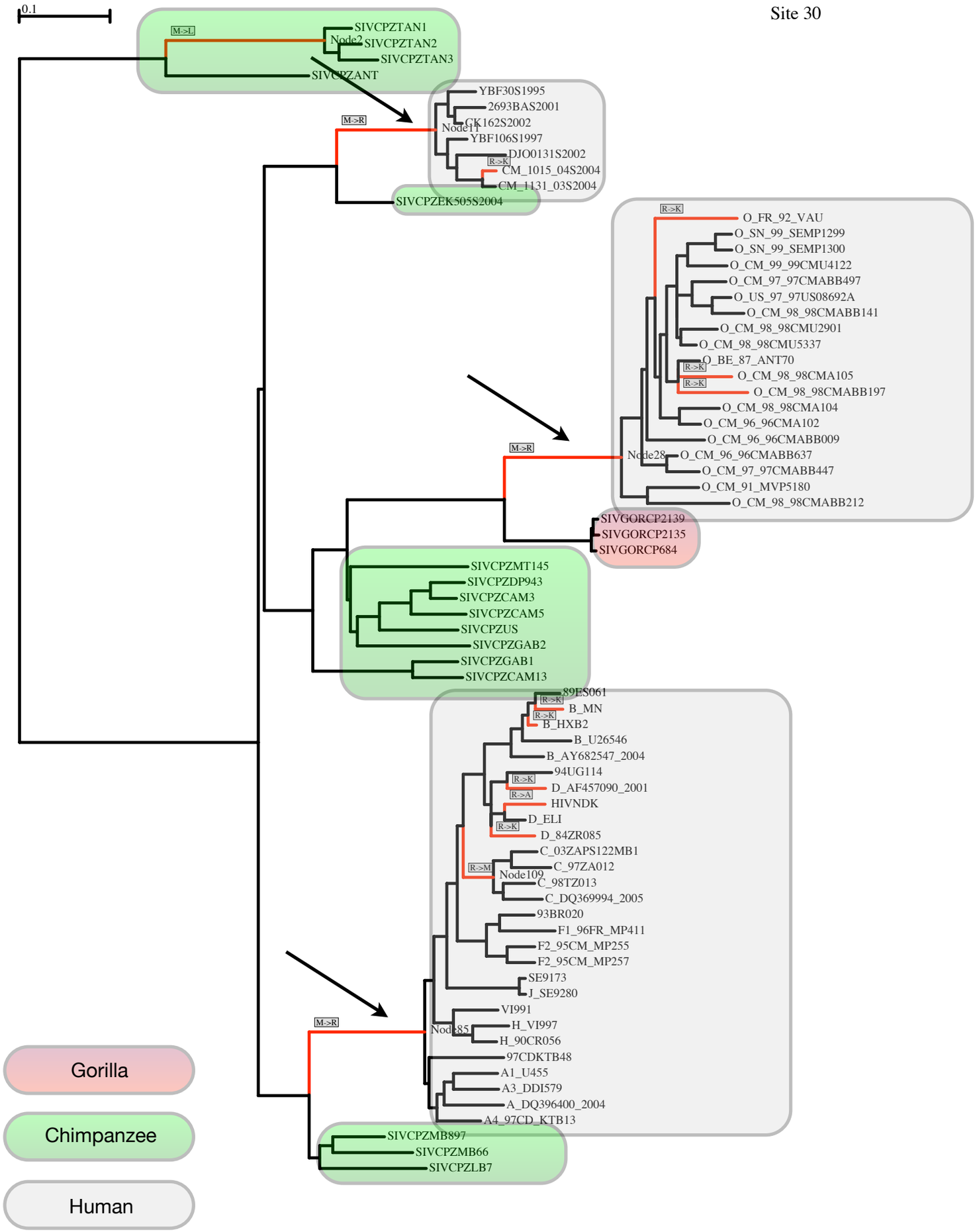
Base model (gag)



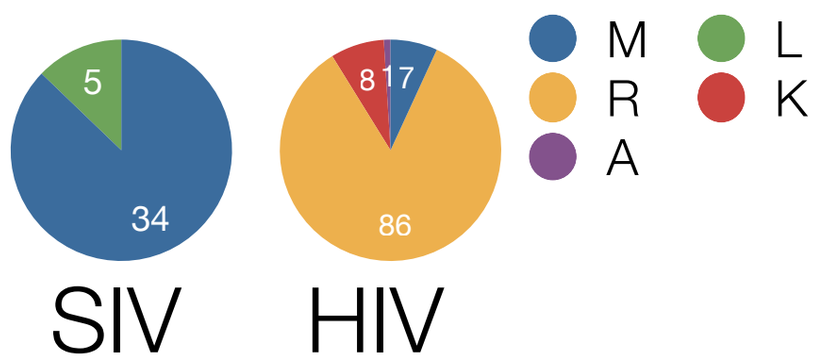


Base model (gag)

‘A’ model (bias factor 7)



- The only residue inferred by L. Wain et al (2007) to be involved in host adaptation (increased replication efficiency in T-cells)
- There are 11 synonymous substitutions and 16 non-synonymous substitutions
- The site has **dN/dS** = 0.2 and is flagged as negatively selected by FEL (**p<0.01**).
- DEPS Bayes Factor for selection on **R** along the three transmission branches is 107



Recombination

- Affects a large variety of organisms, from viruses to mammals (e.g. gene family evolution)
- Manifests itself by incongruent phylogenetic signal
- This can be exploited to detect which sequence regions recombined and which sequences were involved
- Recombination can influence or even mislead selection detection methods.
- Using an incorrect tree to analyze a segment of a recombinant analysis can bias **dS** and **dN** estimation
- The basic intuition is that an incorrect tree will generally break up identity by descent and hence make it appear as if more substitutions took place than did in reality.

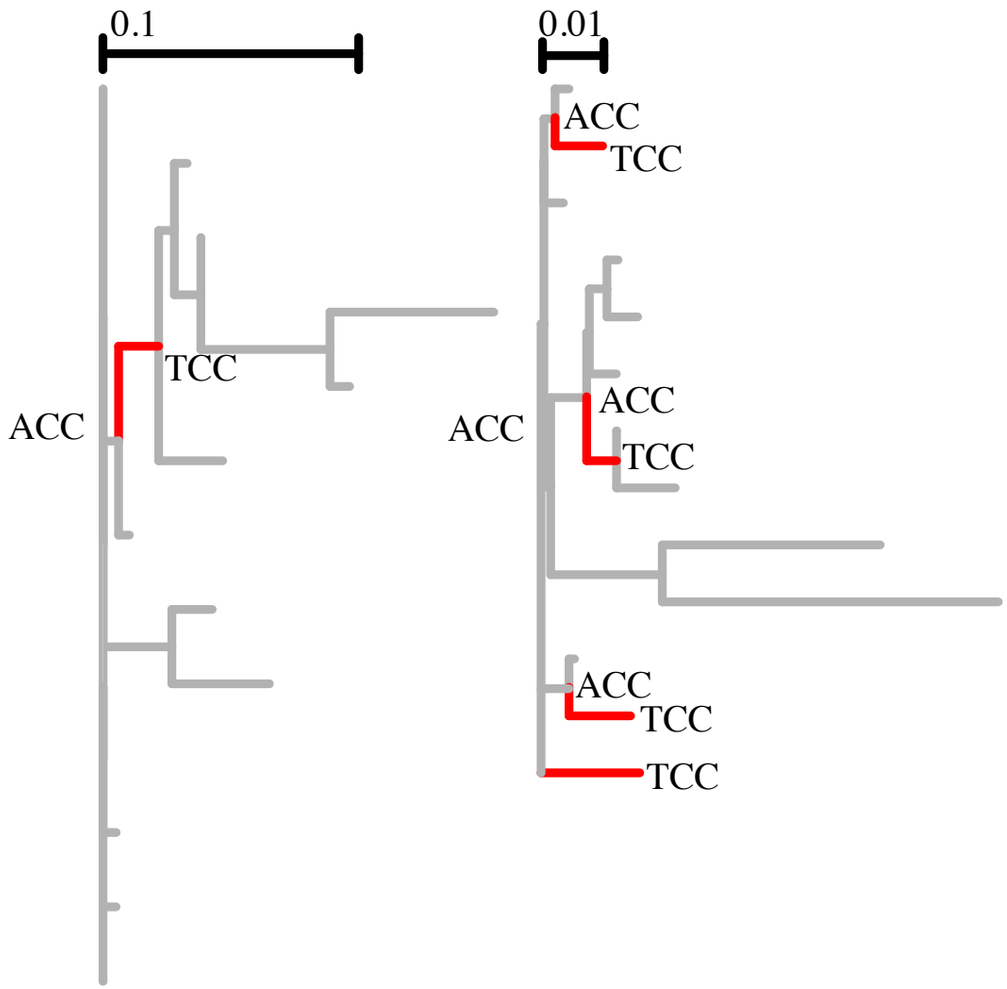


Figure 4.2: The effect of recombination on inferring diversifying selection. Reconstructed evolutionary history of codon 516 of the Cache Valley Fever virus glycoprotein alignment is shown according to GARD inferred segment phylogeny (left) or a single phylogeny inferred from the entire alignment (right). Ignoring the confounding effect of recombination causes the number of nonsynonymous substitutions to be overestimated. A fixed effects likelihood (FEL, Kosakovsky Pond and Frost (2005)) analysis infers codon 516 to be under diversifying selection when recombination is ignored ($p = 0.02$), but not when it is corrected for using a partitioning approach ($p = 0.28$).

Accounting for recombination

- First screen the alignment to find putative non-recombinant fragments (e.g. using GARD)

- Apply a model-based test (MEME, FUBAR, PRIME) using multiple phylogenies (one per fragment), but inferring other parameters (e.g. kappa and base frequencies) from the entire alignment

- This has been shown to work very well on simulated and empirical data

- This is the approach does not work for analyses assuming a single tree (BUSTED, aBSREL).

Mol. Biol. Evol. | 23(10):1891-901 | 2006

Table 4. Effect of correcting for recombination when using fixed effects likelihood to detect positively selected sites.

Virus and gene	Positively Selected Codons	
	Uncorrected FEL	Corrected FEL
Cache Valley G	212,516,546,551	None
Canine Distemper H	158, **179, 264, 444**	**179, 264, 444**, 548
Crimean Congo hemm. fever NP	**195**	9,**195**
Hantaan G2	None	None
Human Parainfluenza (1) HN	37,**91, 358**, 556	**91, 358**
Influenza A (human H2N2) HA	**87**, 166, **252, 358**	**87**, 147,**252, 358**
Influenza B NA	**42,106,345,436**	**42,106,345,436**
Mumps F	**57, 480**	**57, 480**
Mumps HN	399	None
Newcastle disease F	**1**,4,**5,7,16**,18,**108**,516	**1,5,7,16,108**,493,505
Newcastle disease HN	**2**,54,**58,228,262,284,306,471**	**2,58,228,262,284,306,471**
Newcastle disease N	**425, 430, 466**	**425, 430**, 462, **466**
Newcastle disease P	12,**56,65,174,179**,188,**189, 204**, **208, 213**,217,**218**,239,**306,332**	**56, 65**, 146, 153, **174, 179, 189**, 193, **204,208, 213, 218**, 261,**306,332**
Puumala NP	79	None

Test $p < 0.1$ was used to classify sites as selected. Codon sites found under selection by both methods are shown in bold.

Synonymous rate variation

- **dS** = constant for all sites (assumed by many models); this assumption appears to be nearly universally violated in biological data, due to e.g. secondary structure, localized codon usage bias, overlapping reading frames, etc.
- This can lead to, e.g. incorrect identification of relaxed constraint as selection
- FUBAR, MEME, and PRIME fully account for **dS** variation; BUSTED and aBSREL provide experimental support.

Table 1
Data Sets Analyzed for Presence of Synonymous Rate Variation

				MG94 × REV Nonsynonymous GDD 3		MG94 × REV Dual GDD 3 × 3			
Data	Reference	Sequences	Codons	log *L*	Tree Length	log *L*	Tree Length	*P* Value	ΔAIC
Sperm lysin	(Yang and Swanson 2002)	25	135	−4,409	2.85 (0.06)	−4,397.3	2.93 (0.06)	0.0001	15.36
Primate COXI	(Seo, Kishino, and Thorne 2004)	21	506	−12,013.3	8.5 (0.22)	−11,976.6	5.8 (0.15)	<0.0001	65.27
Drosophila *adh*	(Yang et al. 2000)	23	254	−4,586.2	1.41 (0.03)	−4,583.4	1.47 (0.03)	0.23	−2.35
HIV-1 *vif*	(Yang et al. 2000)	29	192	−3,347.2	0.97 (0.02)	−3,334.4	0.99 (0.02)	<0.0001	17.63
β-globin	(Yang et al. 2000)	17	144	−3,659.3	2.6 (0.08)	−3,649.1	3.3 (0.1)	0.0004	12.43
Influenza A*	(Yang 2000)	349	329	−10,916.5	1.42 (0.002)	−10,860.7	1.42 (0.002)	<0.0001	103.7
Camelid VHH*	(Harmsen et al. 2000)	212	96	−16,540.8	14.9 (0.04)	−16,391.2	14.9 (0.04)	<0.0001	291.24
Encephalitis *env*	(Yang et al. 2000)	23	500	−6,774.4	0.85 (0.02)	−6,752.8	0.89 (0.02)	<0.0001	35.15
Flavivirus NS5	(Yang et al. 2000)	18	183	−9,137.8	6.3 (0.19)	−9,110.2	7.8 (0.24)	<0.0001	47.25
Hepatitis D antigen	(Anisimova and Yang 2004)	33	196	−5,137.7	1.9 (0.03)	−5,074.2	2.02 (0.03)	<0.0001	118.98

Mol. Biol. Evol. 22(12):2375–2385. 2005