Quantifying Natural Selection in Coding Sequences

Sergei L Kosakovsky Pond
Professor, Department of Biology
Institute for Genomics and Evolutionary Medicine (iGEM)
Temple University
www.hyphy.org/sergei

(slightly modified by Erick)

Sergei Kosakovsky Pond (Temple)

www.hyphy.org/sergei

Preliminaries

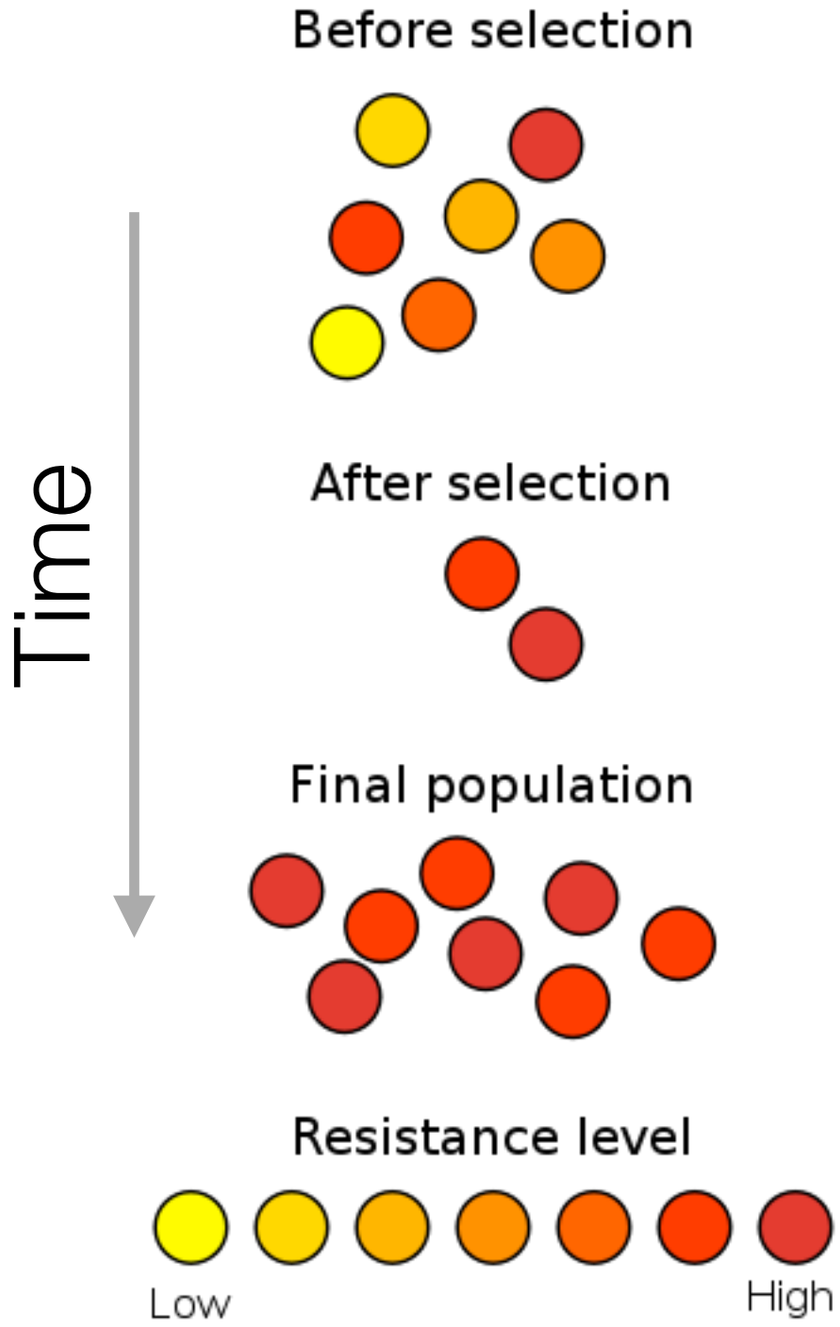
- **Datamonkey** web-app:
    - http://www.datamonkey.org
- Test datasets and practical instructions: bit.ly/hyphy-selection-tutorial

Outline

- The different types of selection analyses enabled by **dN/dS**, told by examples from West Nile virus and HIV and analogies from image analysis
    - Gene-wide selection (BUSTED)
    - Lineage-specific selection (aBSREL)
    - Site-level **episodic** selection (MEME)
    - Site-level **pervasive** selection (FUBAR)
    - Relaxed or intensified selection (RELAX)
- Confounding processes (synonymous rate variation, recombination)
- On the suitability of dN/dS for within-species inference

Natural Selection

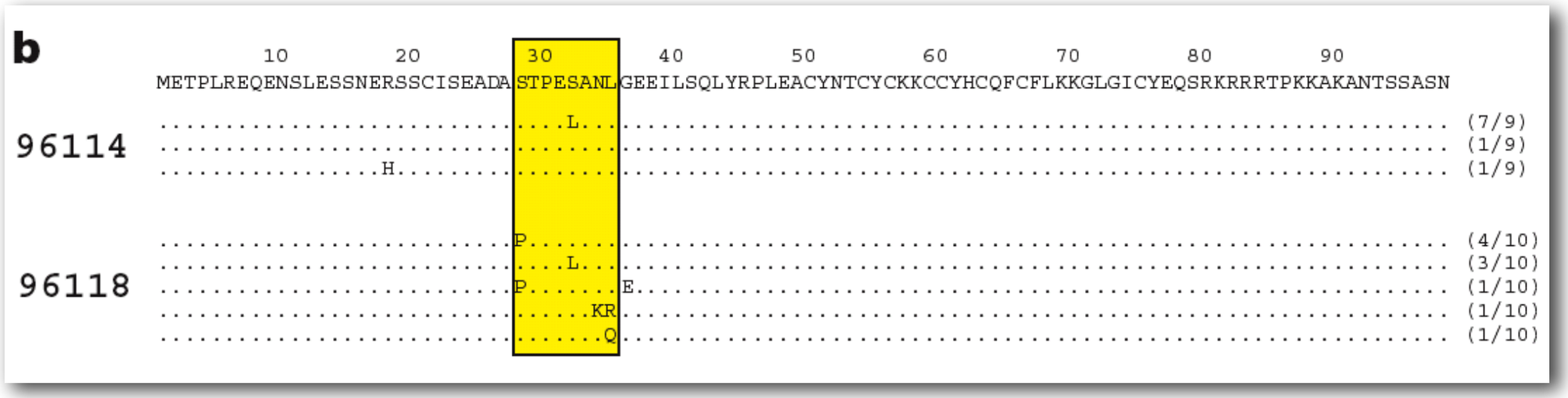
- Any particular mutation can be
  - Neutral: no or little change in fitness (the majority of genetic variation falls into this class according to the neutral theory)
  - Deleterious: reduced fitness
  - Adaptive: increased fitness
- The same mutation can have different fitness costs in different environments (fitness landscape), and different genetic backgrounds (epistasis)



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

Rapid SIV sequence evolution in macaques in response to T-cell driven selection

- 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
- T cell *escape*



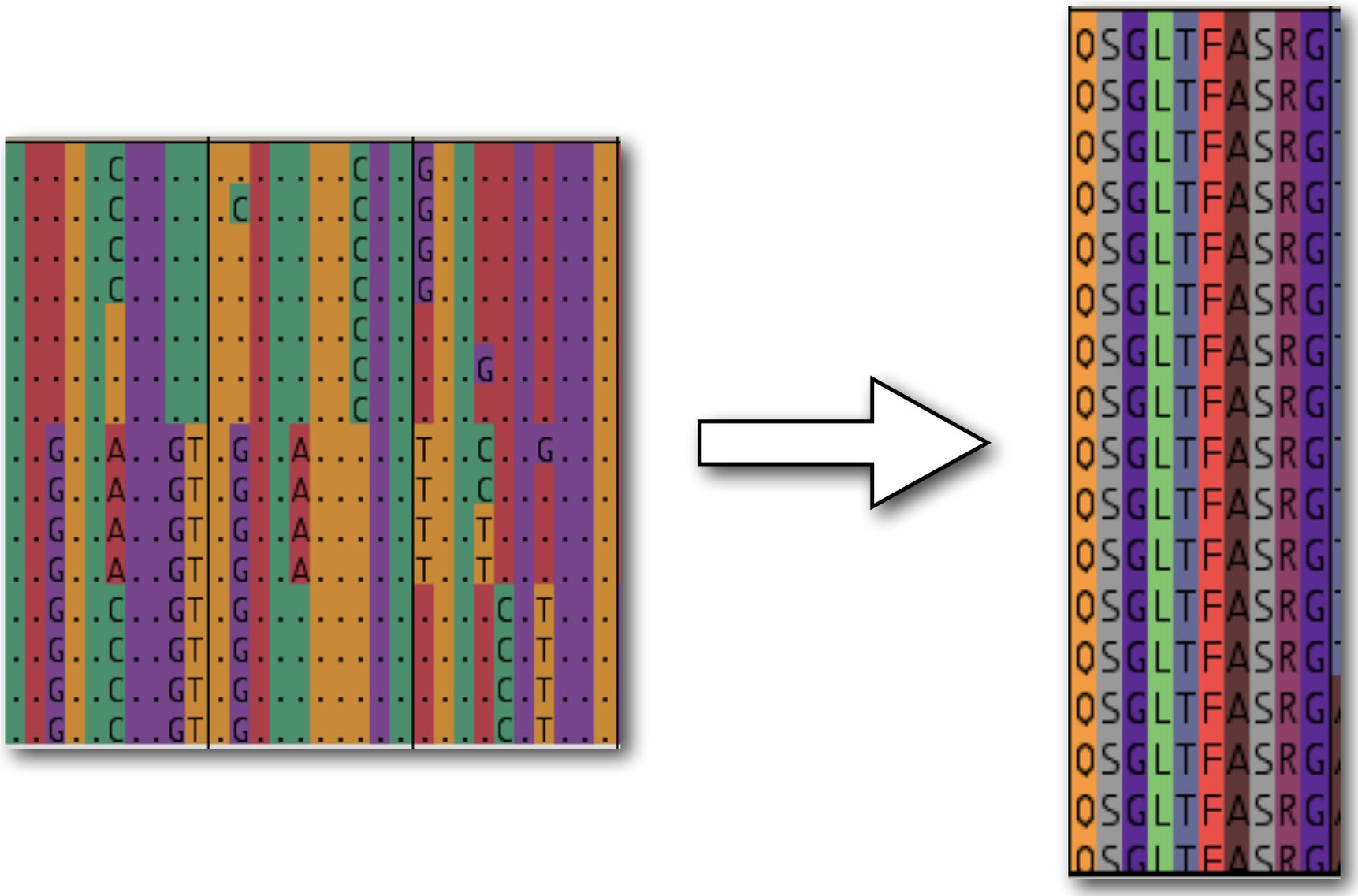
O'Connor et al (2002) Nat Med 8(5):493–499

Evolution of Coding Sequences

- Proper unit of evolution is a triplet of nucleotides — a **codon**
  - **Mutation** happens at the DNA level
  - **Selection** happens (by and large) at the protein level
- **Synonymous** (protein sequence unchanged) and **non-synonymous** (protein sequence changed) substitutions are fundamentally different

Conservation

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

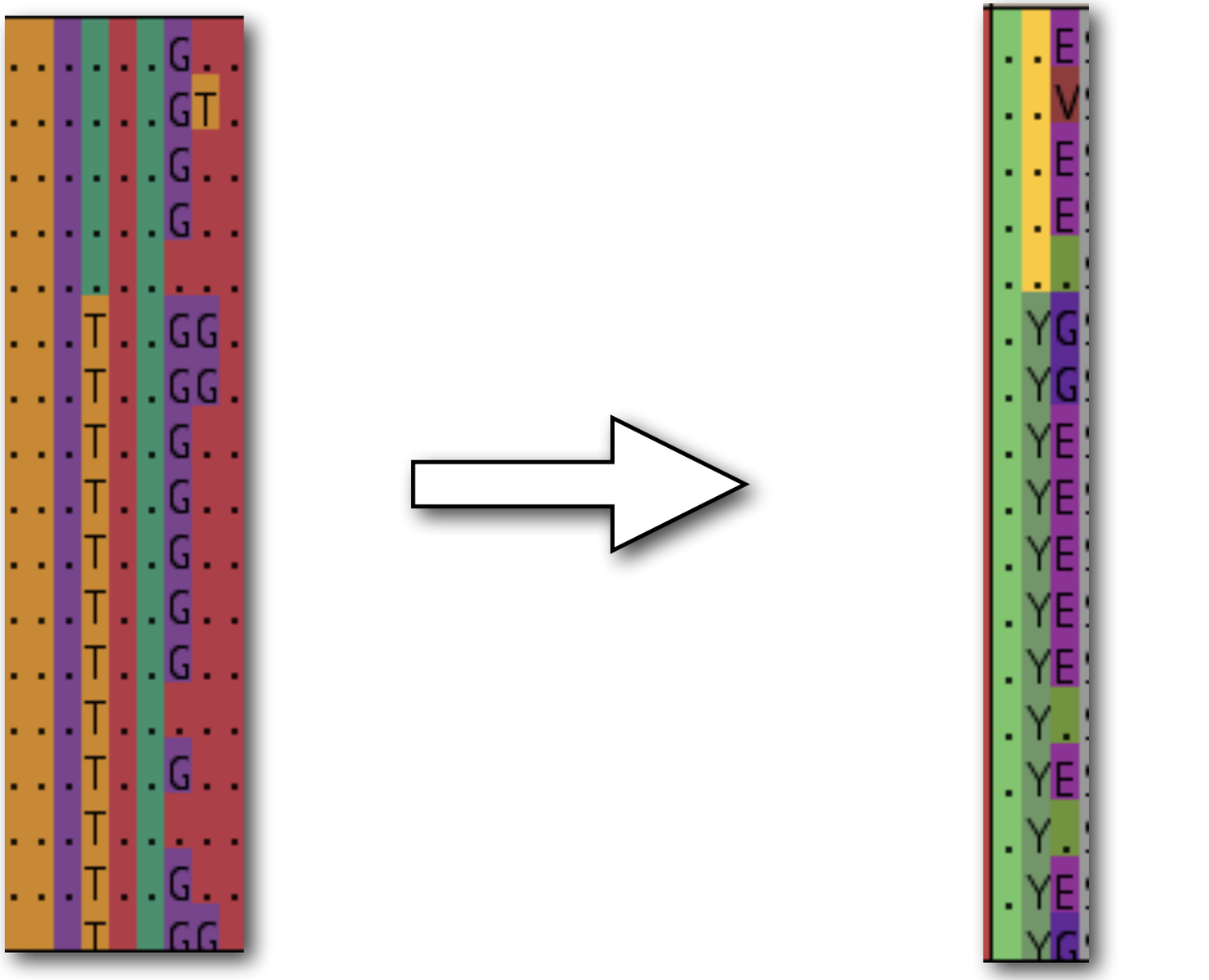


Nucleotides

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 $\omega := dN/dS > 1$
Negative Selection	$dS > dN$ or $\omega < 1$
Neutral Evolution	$dS \simeq dN$ or $\omega \simeq 1$

Estimating dS and dN

Consider two aligned homologous 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 one claim that **`dN/dS = 1`**, because there is **one** synonymous and **one** non-synonymous substitution?

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 a reference to compute **dS**.

- In early literature, these quantities were codified as synonymous and non-synonymous “sites” and/or mutational opportunity.

- As a very crude approximation (assuming that third positions ~ synonymous), each codon has 1 synonymous and 2 non-synonymous sites.

Computing synonymous and non-synonymous sites for GAA (Glutamic Acid)

Start codon:	G	A	A
Site/Change to	1	2	3
A	AAA **Lysine**	*	*
C	CAA **Glutamine**	GCA **Alanine**	GAC **Aspartic Acid**
G	*	GGA **Glycine**	GAG **Glutamic Acid**
T	TAA **Stop**	GTA **Valine**	GAT **Aspartic Acid**
Synonymous changes	0	0	1
Non-synonymous changes	3	3	2
Synonymous sites	**0**	**0**	**1/3**
Non-synonymous sites	**1**	**1**	**2/3**

Aminoacid	Codons	Redundancy
Alanine	GC*	4
Cysteine	TGC,TGT	2
Aspartic Acid	GAC,GAT	2
Glutamic Acid	GAA,GAG	2
Phenylalanine	TTC,TTT	2
Glycine	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

8 non-synonymous site/base combos

1 synonymous site/base combos

Rate matrix for an MG-style codon model

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

X,Y = AAA...TTT (excluding stop codons),
R_{x,y} = neutral rate of substitution from x to y
π_t - frequency of the target nucleotide.
Example substitutions:
AAC→AAT (one step, synonymous - Aspargine)
CAC→GAC (one step, non-synonymous - Histidine to Aspartic Acid)
AAC→GTC (multi-step).

αR_{CT}
βR_{CG}

α (syn. rate) and **β** (non-syn. rate)
are the key quantities for all selection analyses

Goldman-Yang (GY) type substitution model

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

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`**.
- Multiple 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

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)

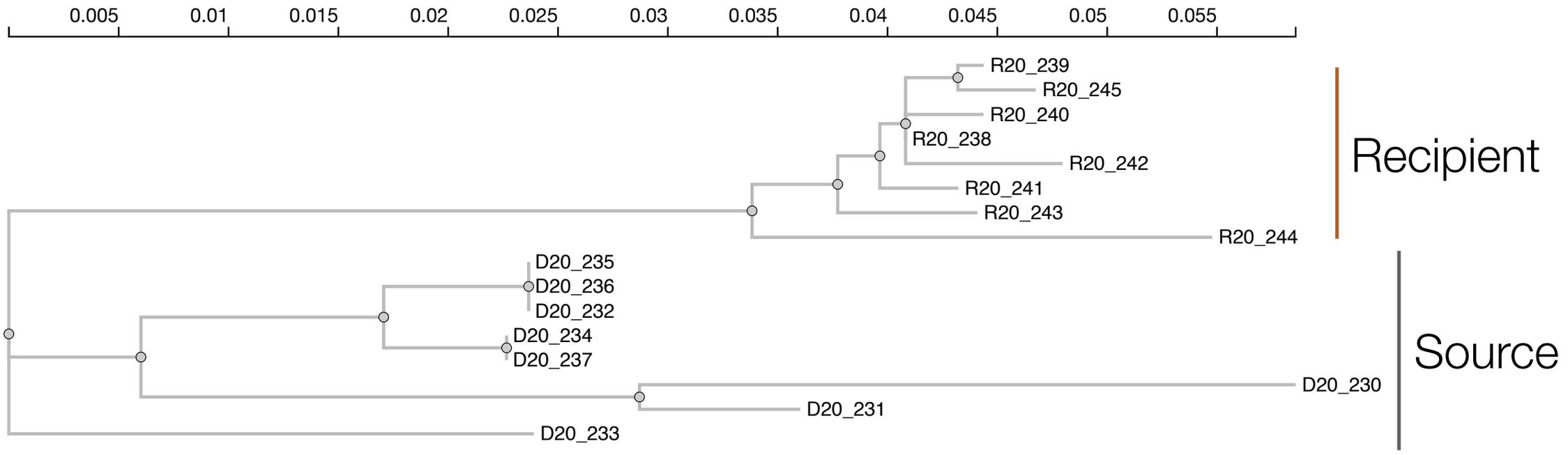
- Codon models also support the concepts of synonymous and non-synonymous distances between sequences using standard properties of Markov processes (exponentially distributed waiting times)

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

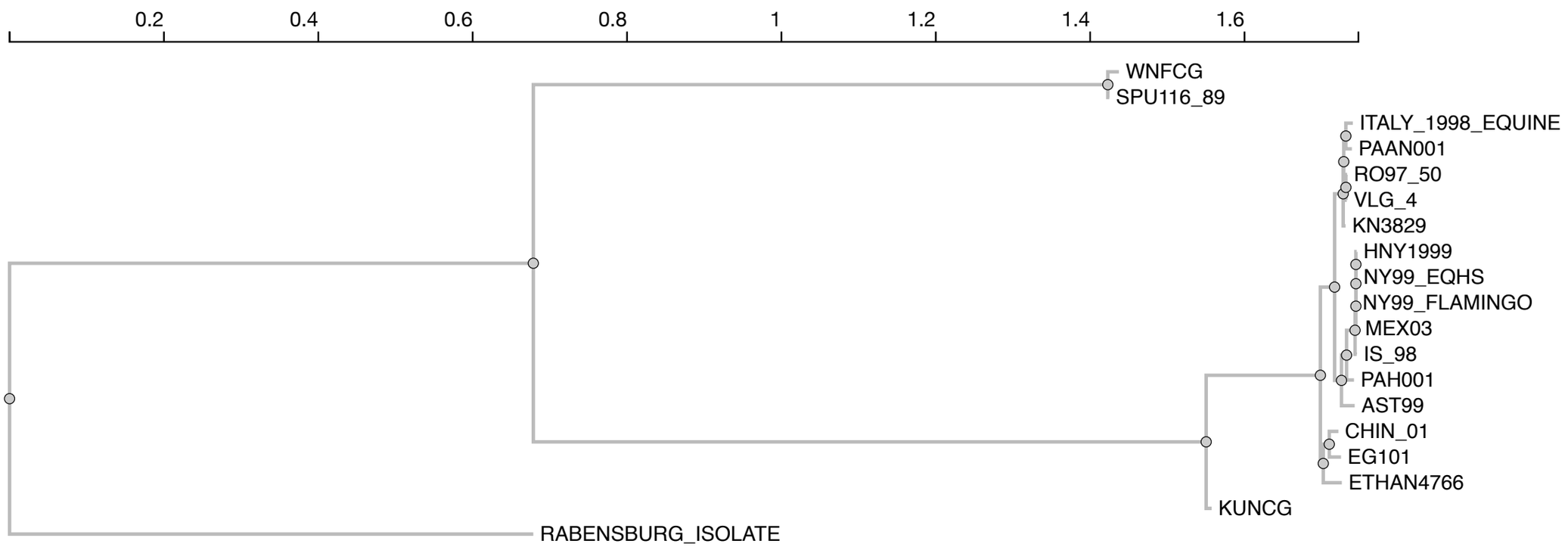
Two example datasets

- **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)

HIV-1 *env*



WN NS3



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

How do you expect these measures to correlate with the ability to detect selection?

WNV NS3

Model	Log L	# p	dN/dS	LRT	p-value
Null	-7668.7	49	1		
Alternative	-6413.5	50	0.009	2510.4	~0

Very strongly conserved

HIV-1 env

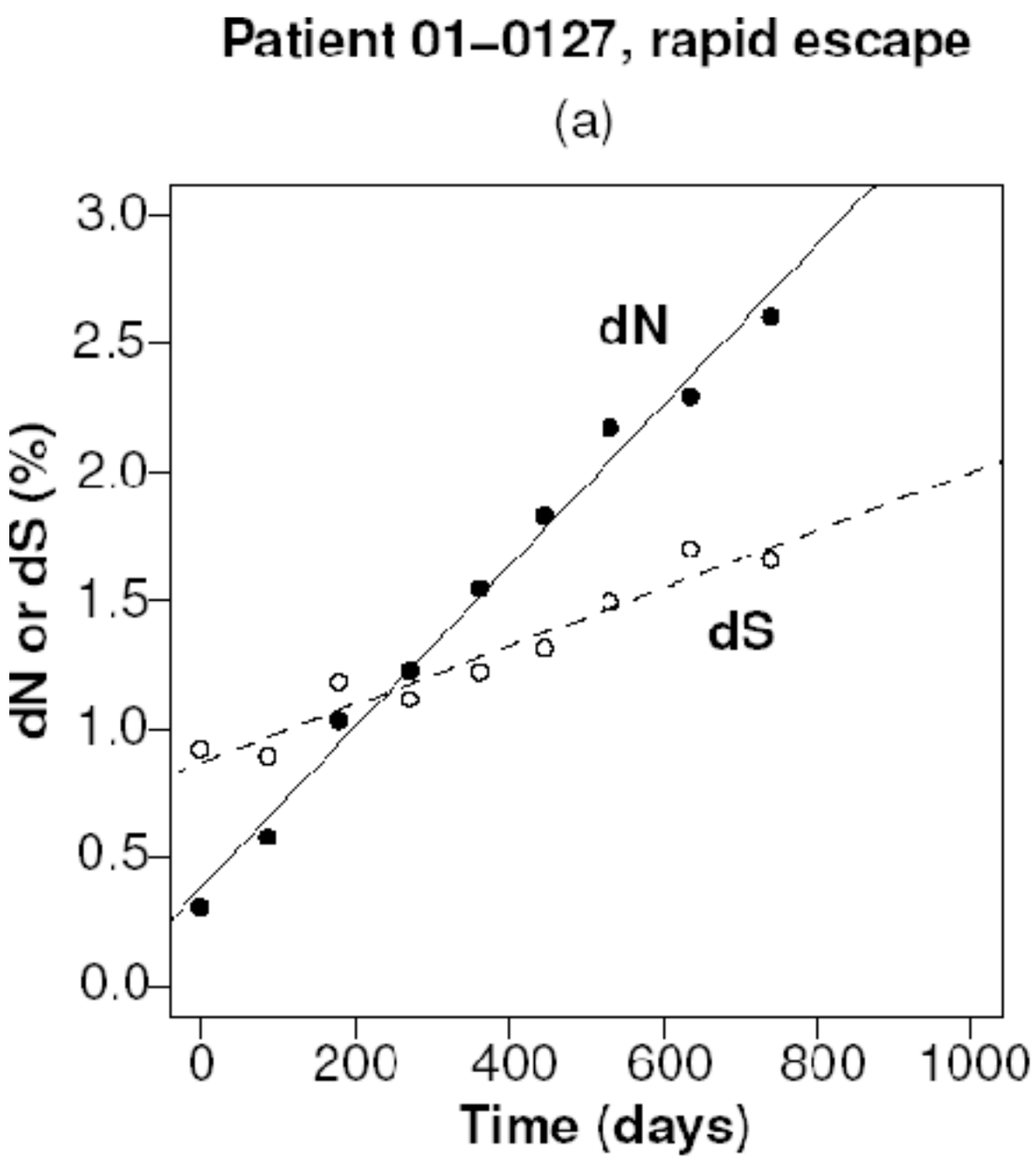
Model	Log L	# p	dN/dS	LRT	p-value
Null	-2078.3	40	1		
Alternative	-2078.2	41	1.128	0.2	~0.6

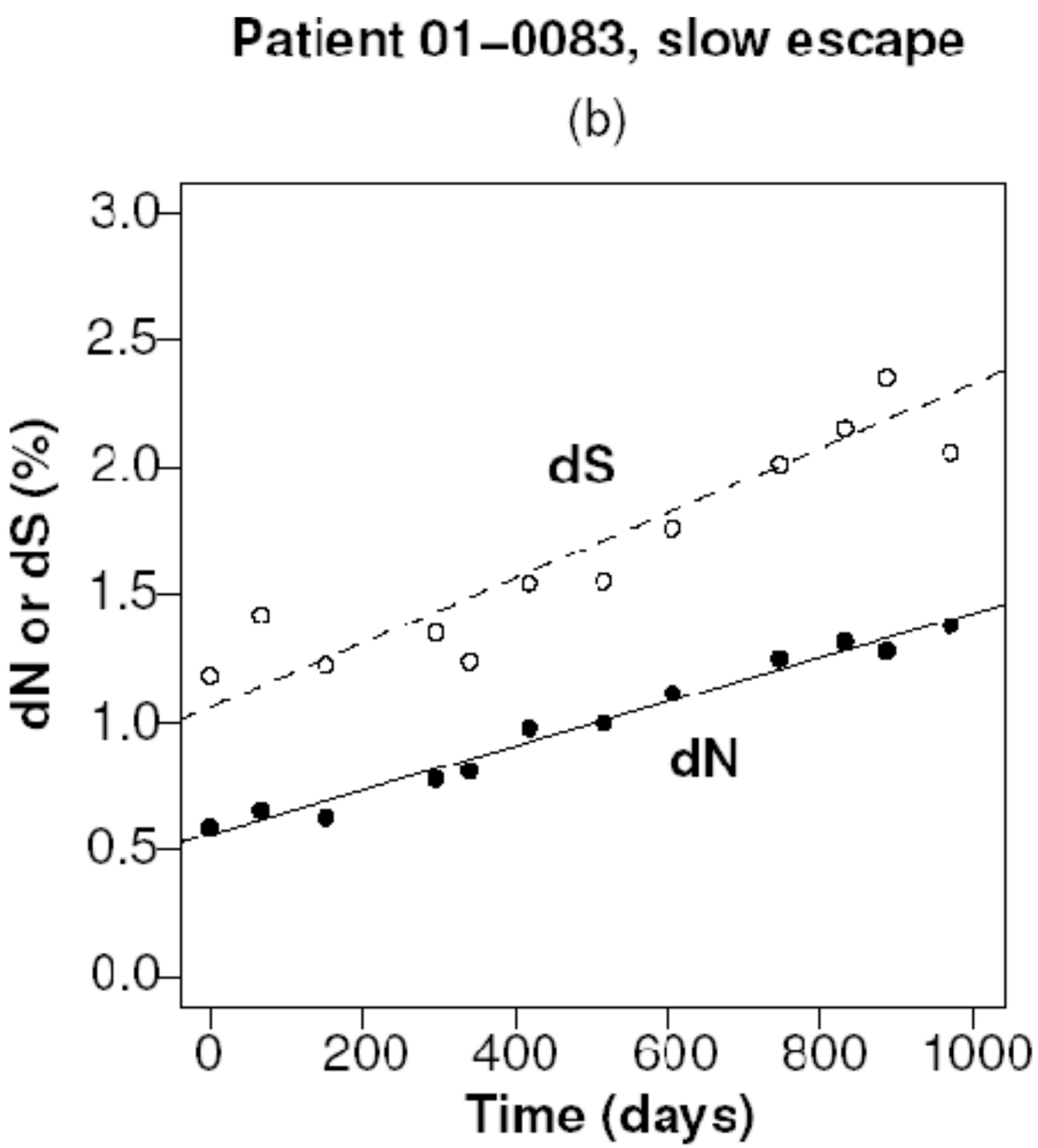
Not significantly different from neutral

Mean gene-wide dN/dS estimates

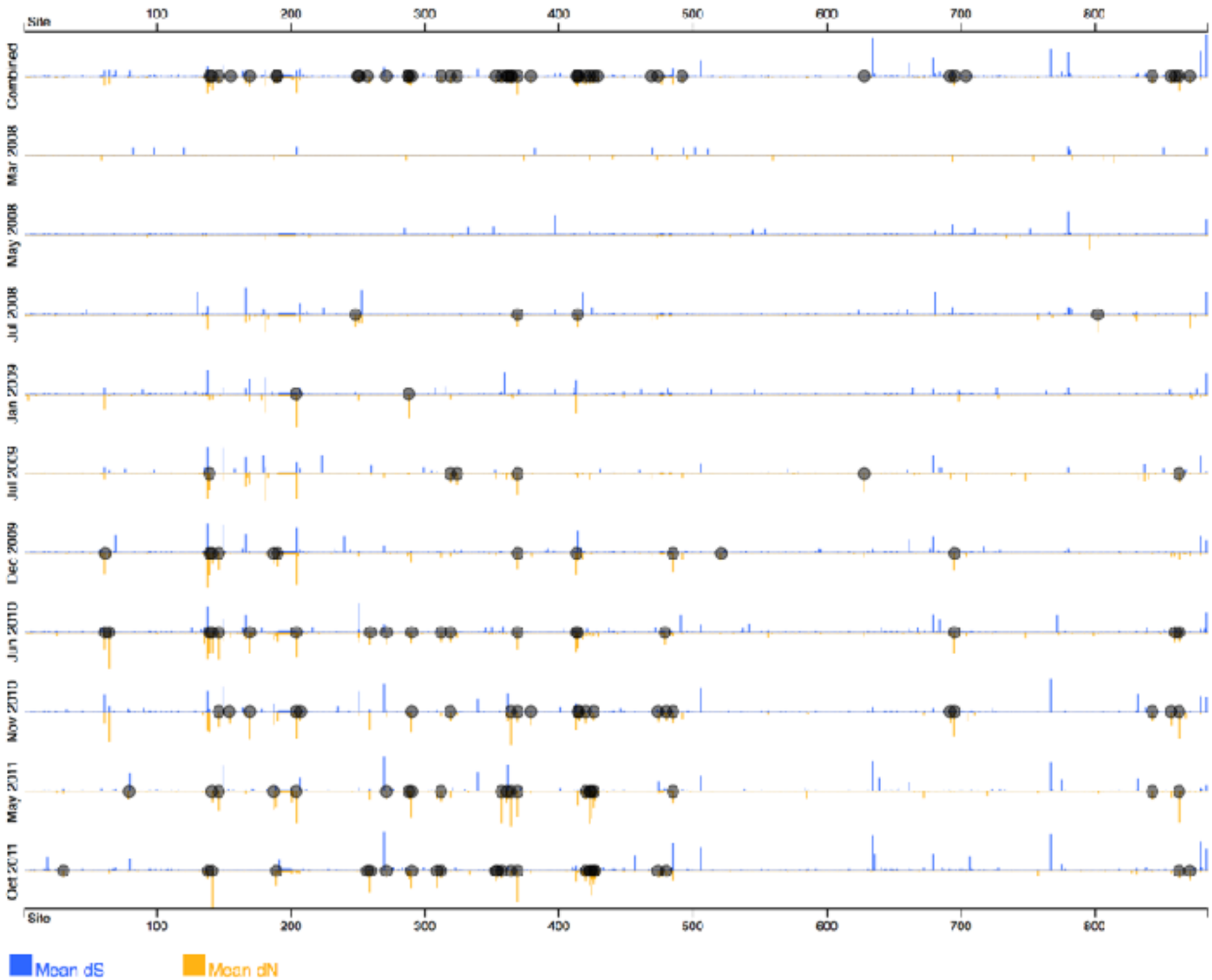
- Are not the way to go, **except** when you have very small (2-3 sequence) datasets
- For example:
  - The humoral arm of the immune system mounts a potent defense against viral infections
  - Existing successful vaccines are based on raising a neutralizing antibody (nAb) response to the pathogen
  - No simple host genetic basis (epitopes) of the specificity of neutralizing antibody responses is known
  - Need to measure these responses

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



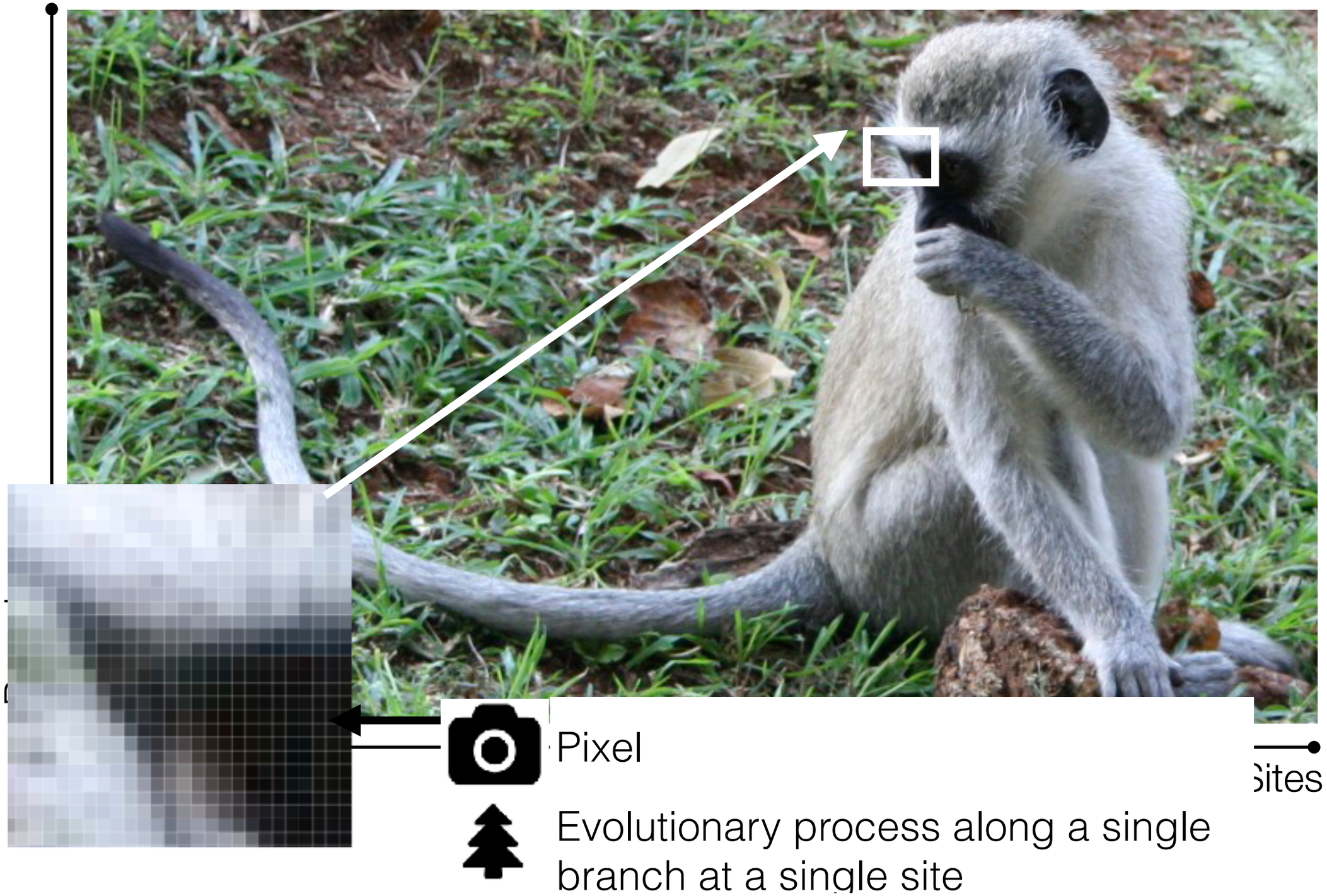


But upon closer look, this pattern is highly variable both across a gene and through time.

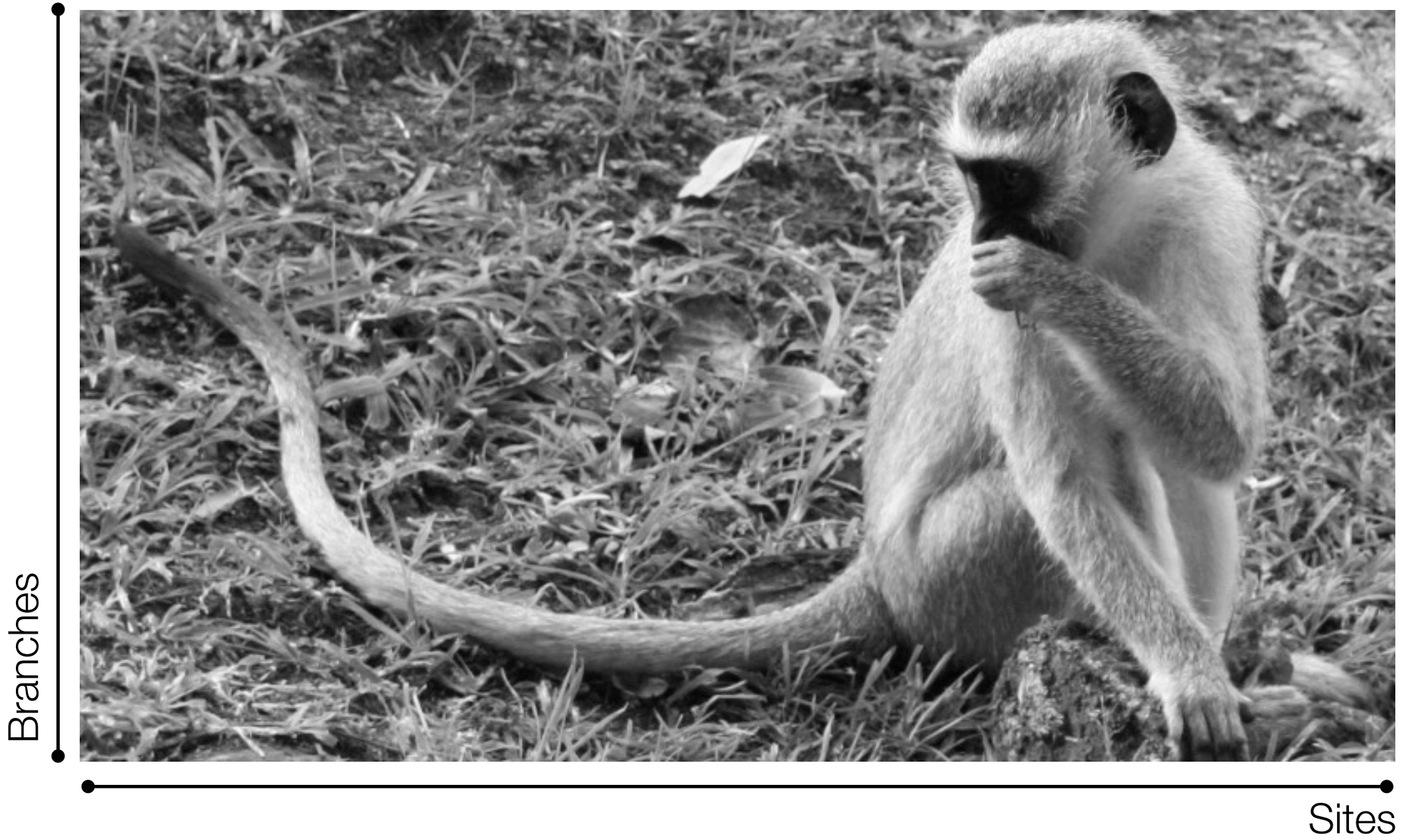


PLoS Pathog 12(1): e1005369. Patient 064

Selection inference as image processing



Forget about the color



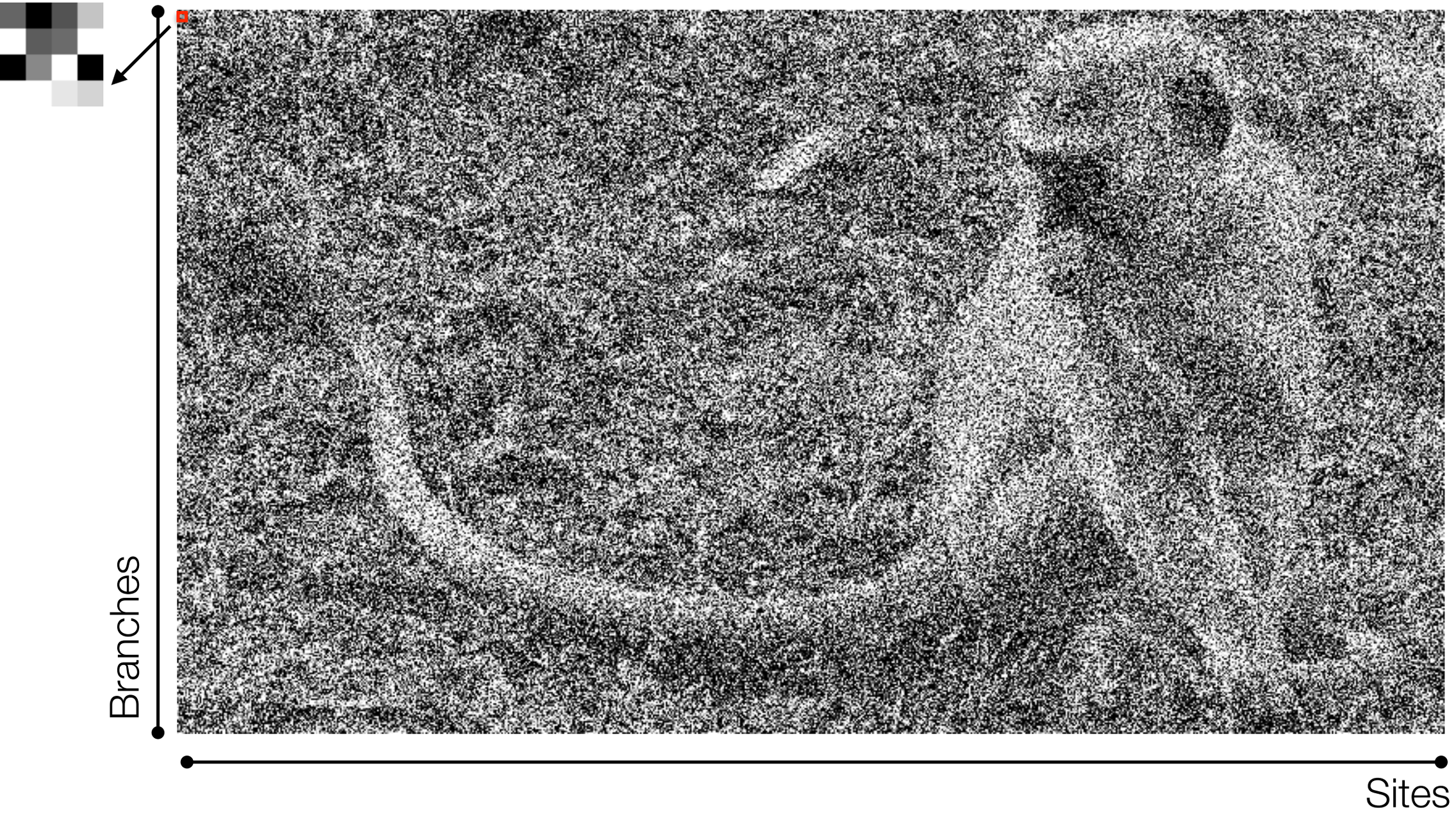
Intensity/brightness

Color

Evolutionary rate (dN/dS)

Type of evolutionary/
function/property change

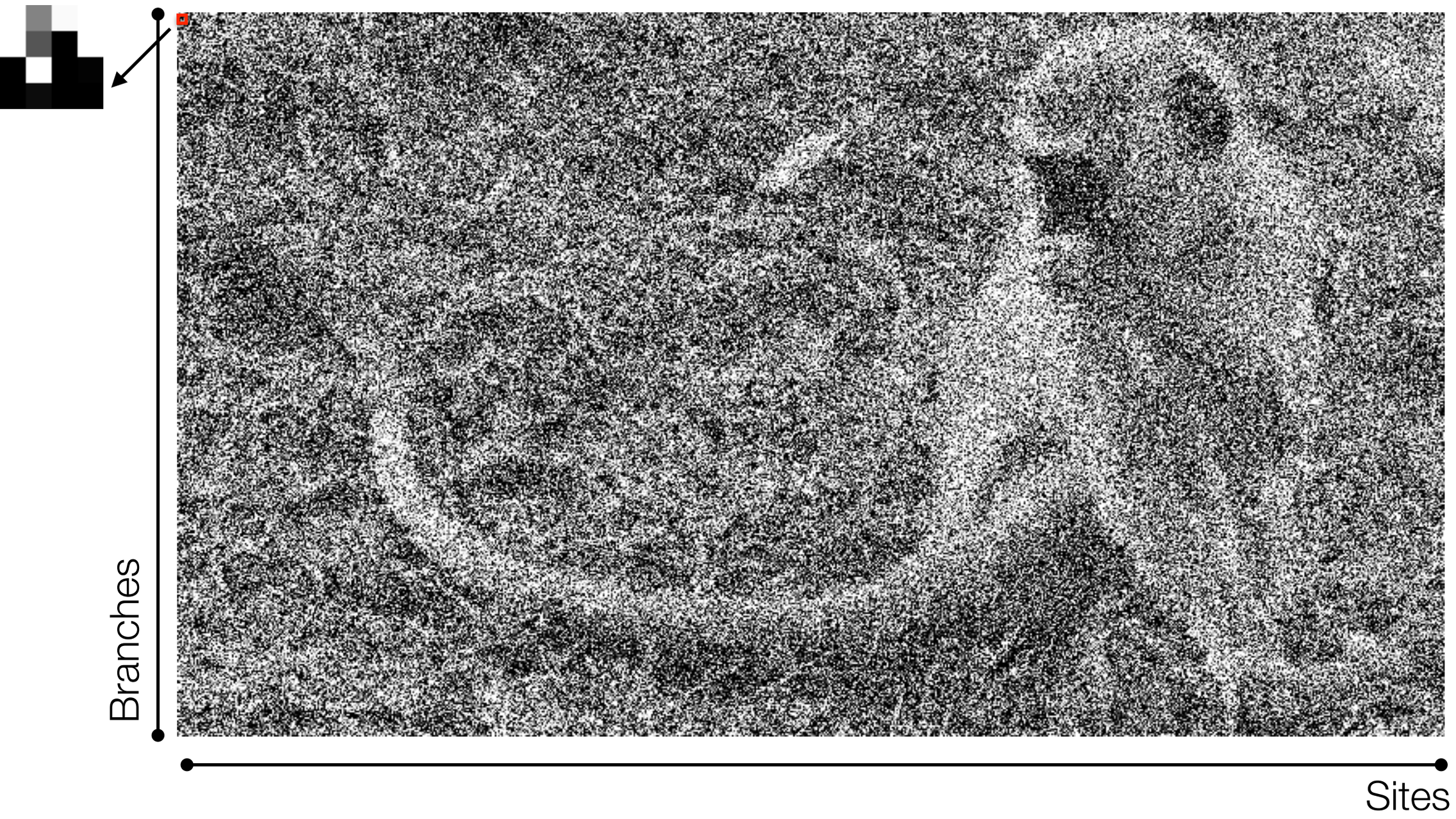
Evolution is largely unobserved and noisy



Visual noise

Saturation, missing data, model misspecification, sampling variation

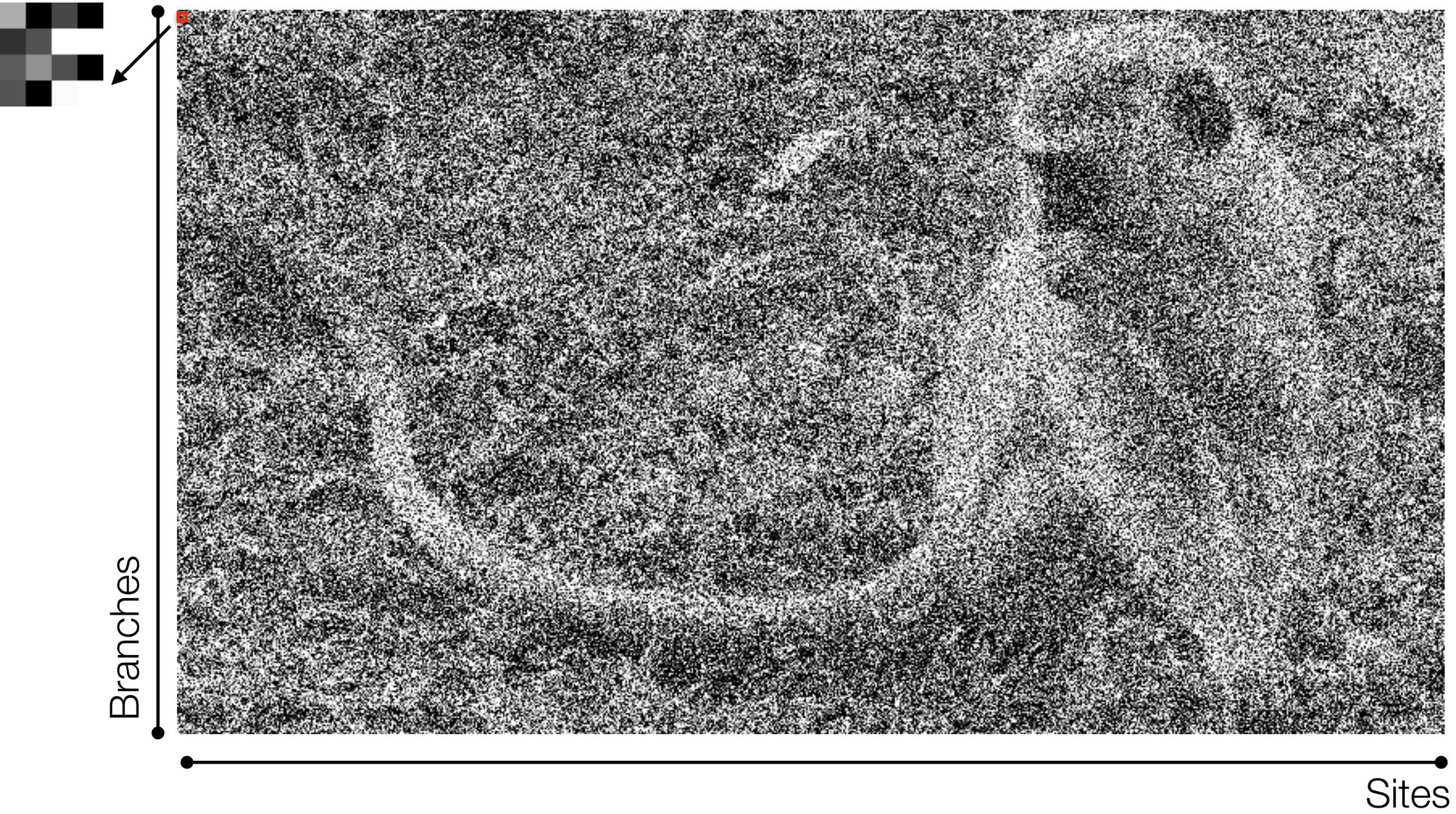
Evolution is largely unobserved and noisy (another replicate)



Visual noise

Saturation, missing data, model misspecification, sampling variation

Evolution is largely unobserved and noisy (another replicate)



Visual noise

Saturation, missing data, model misspecification, sampling variation

- High local variability

- Stable global (monkey) and local (head, tail) patterns, easily discernible

- Desired resolution (branch-site) is not attainable

- Global (and some local) patterns should be inferable and testable

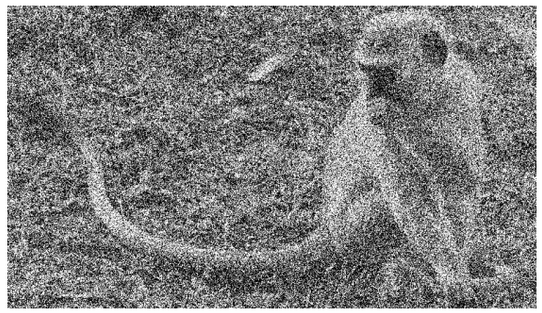
- Statistical inference draws power from sample (and effect) size, need to aggregate data to gain power

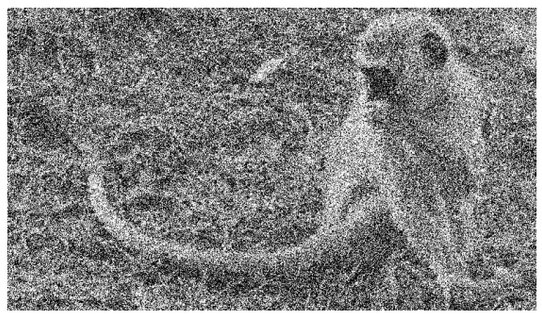
Gene-wide selection (mean dN/dS)

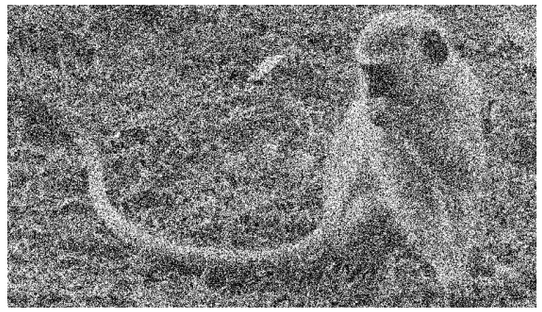
Is the average color sufficiently “bright”

Is there evidence that **gene-wide dN/dS > 1?** Aggregate data over the entire alignment, by inferring a single dN/dS parameter from all sites and branches

- Simple
  - single rate parameter
  - relatively compute-light
- Very robust to local variation
- Sample size ~ sites x branches
- Very low power
  - most genes are **on average** conserved
- No resolution
  - if selection occurred, how much of the gene was involved, and when did it happen
- Rate variation model is definitely misspecified

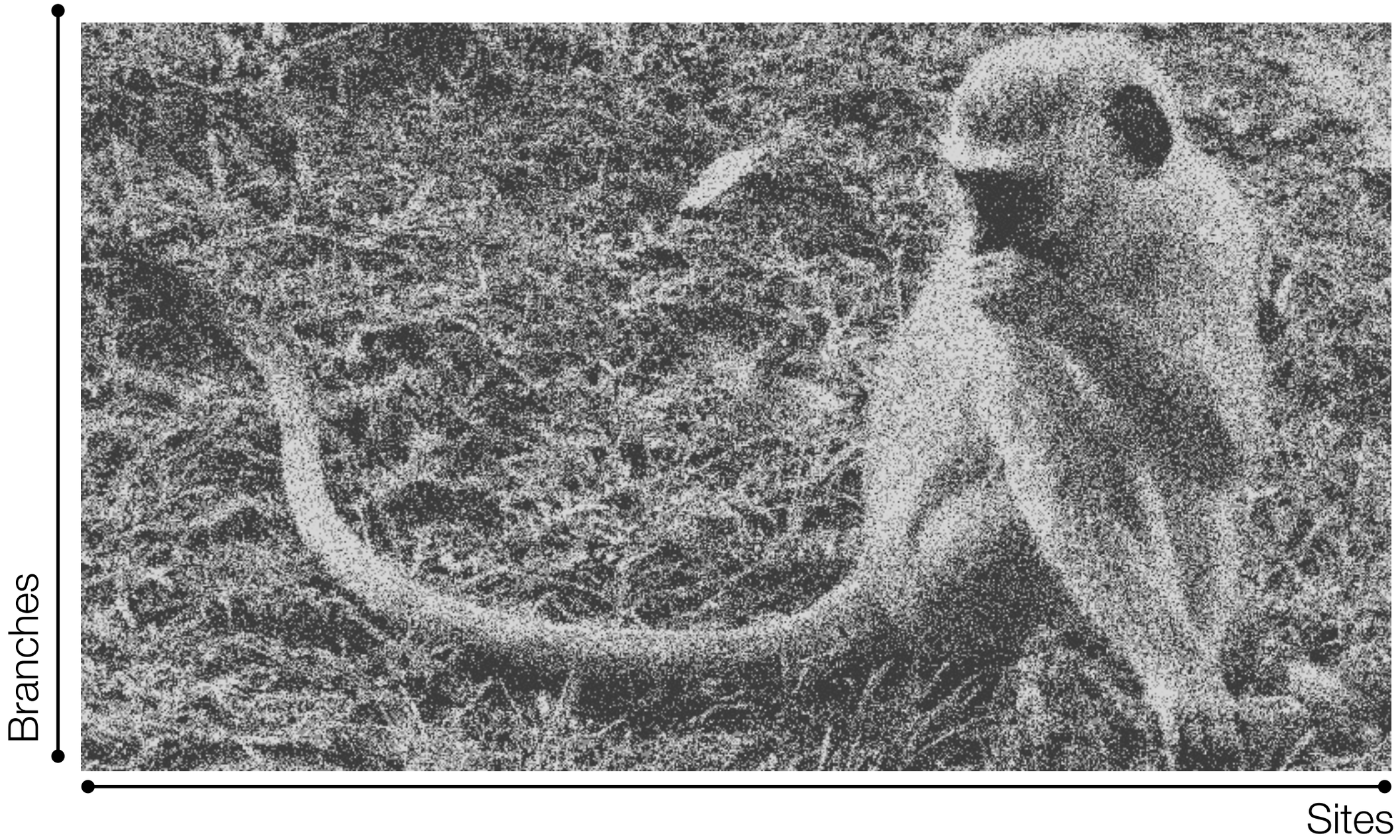






Gene-wide selection

random effects over sites and branches [BUSTED]

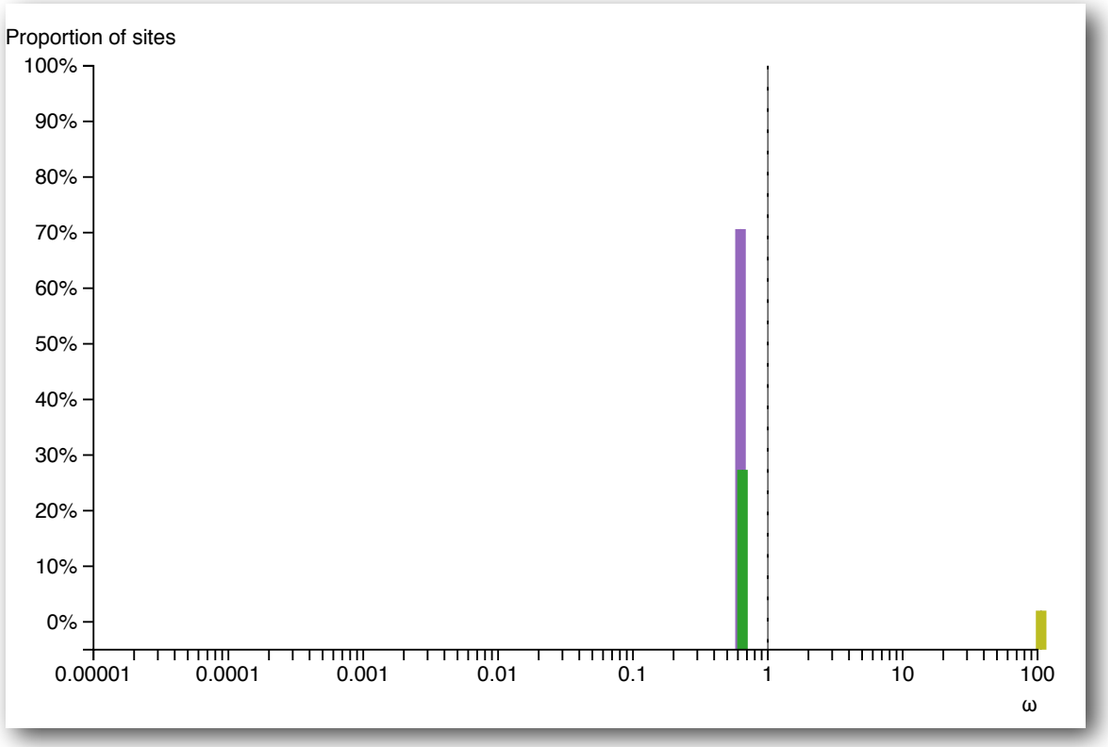


Is there enough **image area** that is sufficiently bright; allow each pixel to be one of 3 colors, chosen adaptively, e.g. to minimize perceptual differences

[BUSTED]: each branch-site combination is a drawn from a 3-bin (dS,dN) distribution. The distribution is estimated from the entire alignment. Tests if dN/dS>1 for some branch/site pairs in the alignment

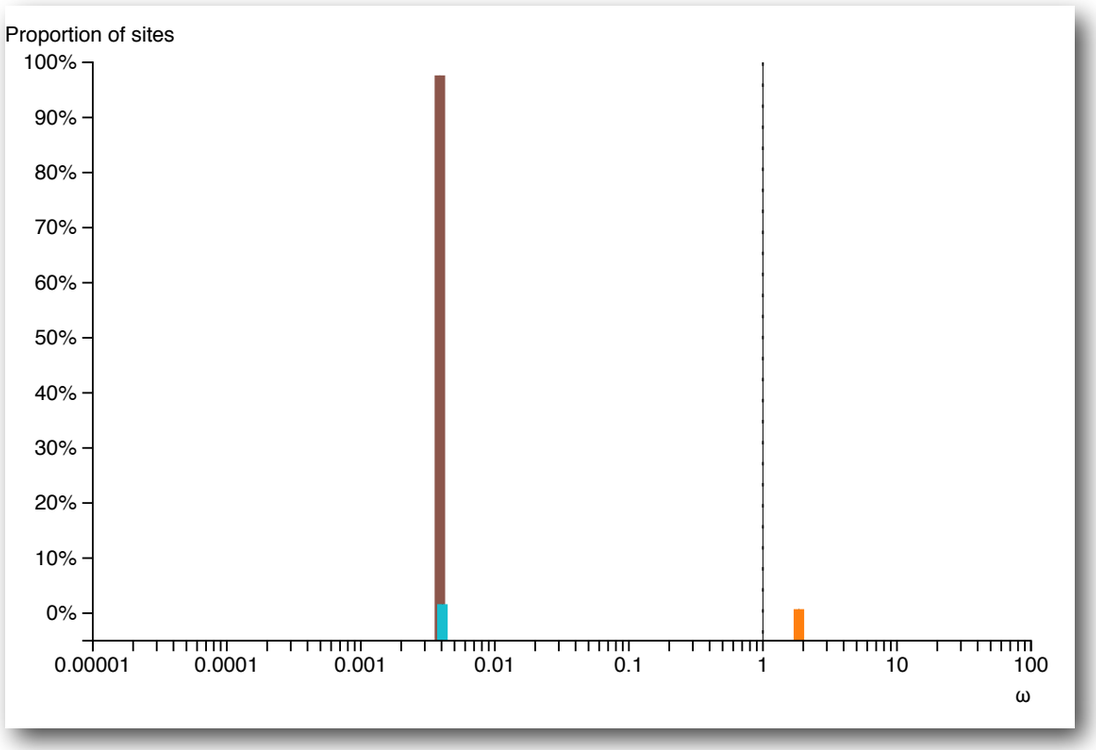
Gene-wide selection analysis using a branch-site method (BUSTED), HIV-1 *env*

Gene-wide **`dN/dS`** distribution	ω_1 = 0.627 (71%) ω_2 = 0.649 (27%) ω_3 = 106 (2%)
`p-value` for selection ($H_0 : \omega_3 = 1$)	$<10^{-15}$
`Log L` (no variation)	-2078.20
`Log L` (branch-site; 4 addt'l parameters)	-2039.99



Gene-wide selection analysis using a branch-site method (BUSTED), WN NS3

Gene-wide **dN/dS** distribution	ω_1 = 0.004 (99.3%) ω_2 = (n/a) ω_3 = 1.86 (0.73%)
p-value for selection ($H_0 : \omega_3 = 1$)	0.54
Log L (no variation)	-6413.50
Log L (branch-site; 4 addt'l parameters)	-6396.18



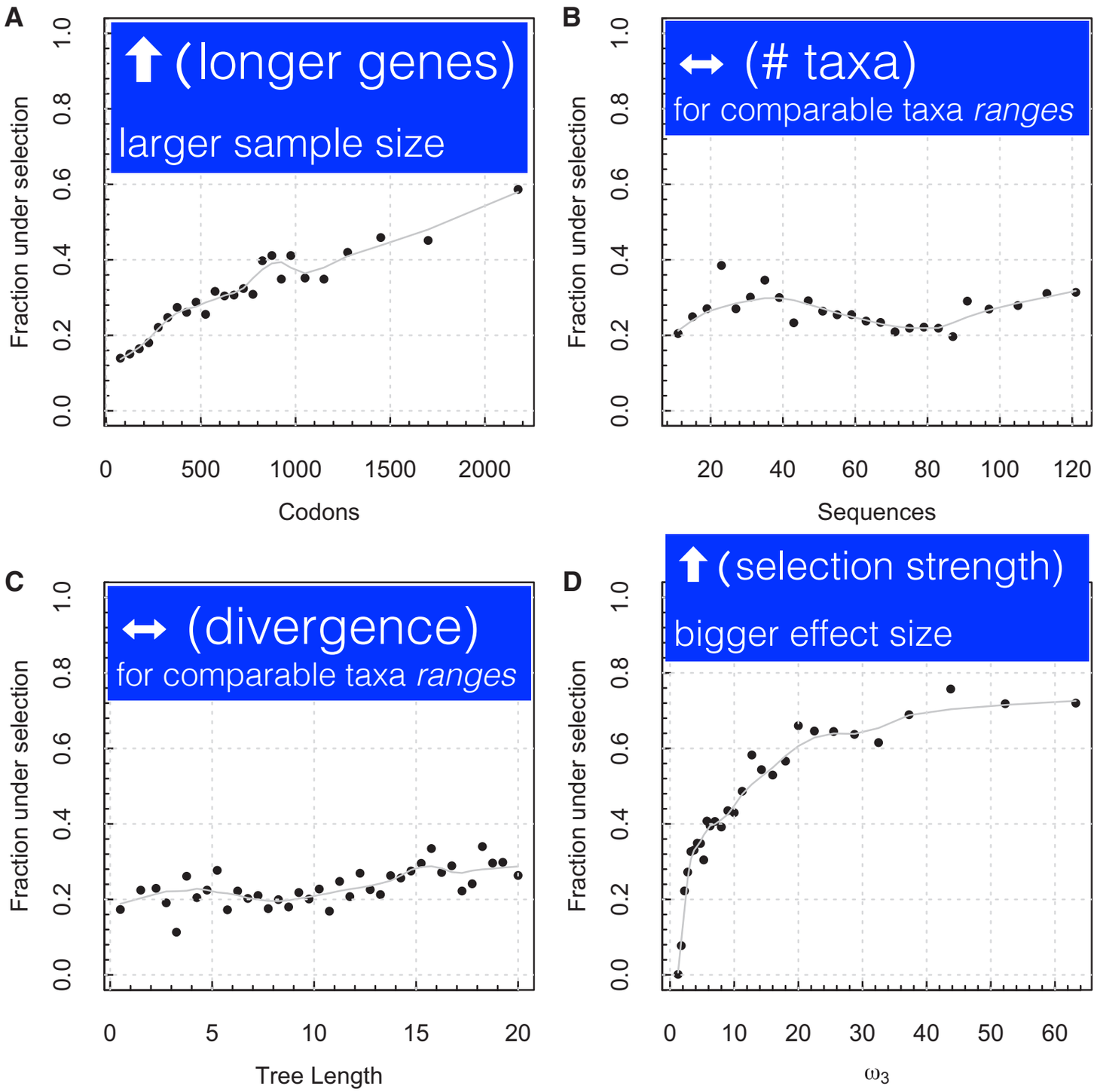
BUSTED analysis

- **West Nile Virus NS3 protein**
  - No statistical support for selection; ML point estimate allocates a small proportion of sites (`~1%`) to the selected group (`dN/dS ~ 2`)
  - 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`)

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

Where does the power come from for BUSTED?

An analysis of ~9,000 curated gene alignments from `selectome.unil.ch`



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

BUSTED site-level inference

- Because BUSTED is a random-effects method, it **pools** information across multiple sites and branches to gain power

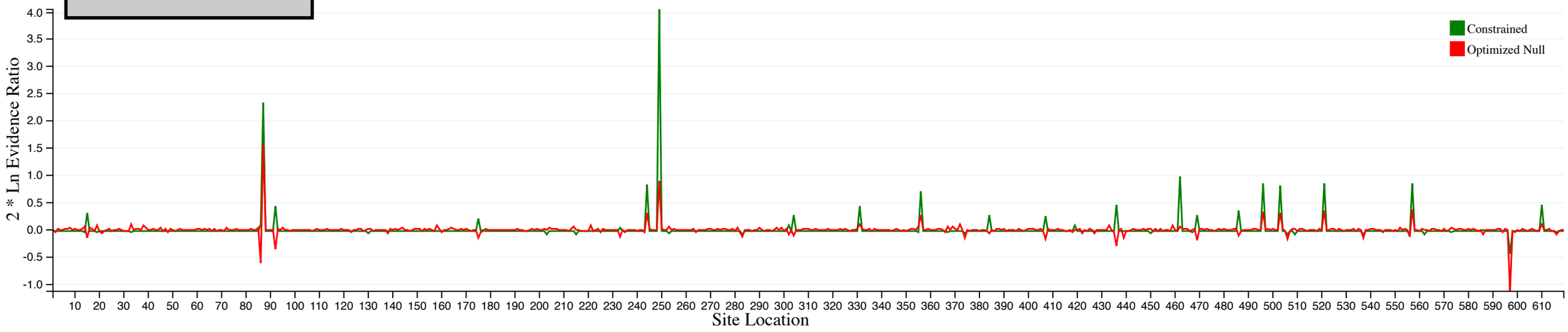
- The cost to this pooling is lack of site-level **resolution**, i.e., it is not immediately obvious which sites and/or branches drive the signal

- Standard ways to extract individual site contributions to the overall signal is to perform a post-hoc analysis, such as empirical Bayes, or “category loading”

- For BUSTED, “category loading” is faster and experimentally better

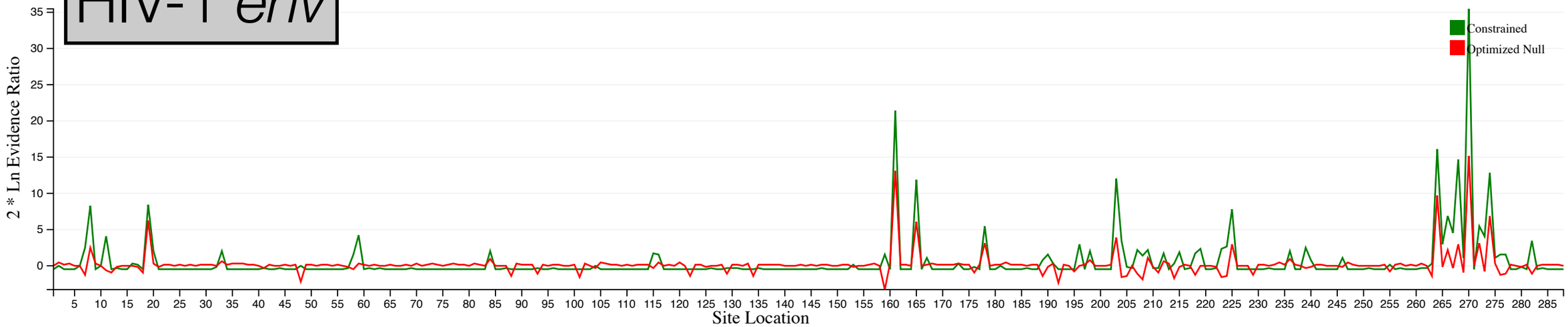
Murrell et al | **Mol. Biol. Evol** | 32(5) | 1365–1371

WN NS3



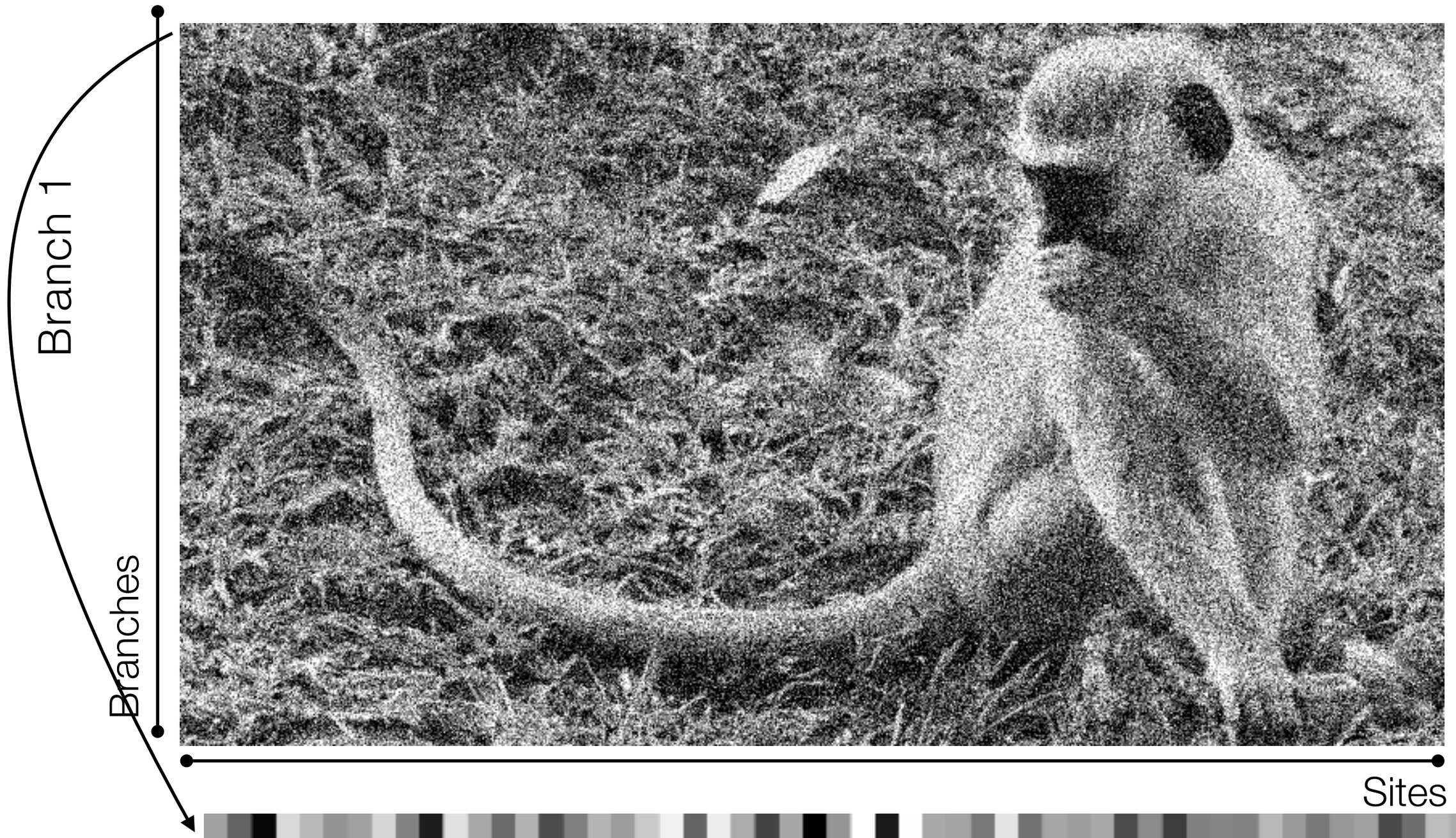
Site Index	Unconstrained Likelihood	Constrained Likelihood	Optimized Null Likelihood	Constrained Evidence Ratio	Optimized Null Evidence Ratio
249 - 249					
249	-65.6860	-67.7124	-66.1326	2.0264	0.4467

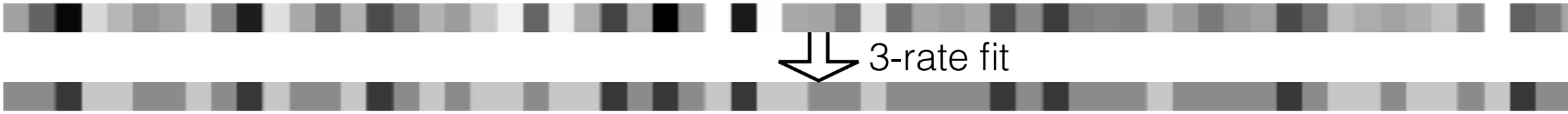
HIV-1 *env*



Site Index	Unconstrained Likelihood	Constrained Likelihood	Optimized Null Likelihood	Constrained Evidence Ratio	Optimized Null Evidence Ratio
161 - 274					
161	-17.4988	-28.2260	-24.1050	10.7272	6.8062
165	-17.4502	-23.4271	-20.5521	5.9769	3.1019
203	-26.0973	-32.1071	-28.0348	6.0098	1.9375
264	-11.6149	-19.6889	-16.4445	8.0741	4.8297
269	-21.5339	-28.9201	-23.0537	7.3862	1.5198

Which branches are under selection?





For each image **row**, is there a significant proportion of bright pixels, once the column has been reduced to **N** colors only?

[aBSREL]: at a given branch, each site is a draw from an N-bin (dN/dS) distribution, which is inferred from all data for the branch. Test if there is a proportion of sites with dN/dS > 1 (LRT). **N** is derived adaptively from the data.

Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection

Martin D. Smith,[1] Joel O. Wertheim,[2] Steven Weaver,[2] Ben Murrell,[2] Konrad Scheffler,[2,3] and Sergei L. Kosakovsky Pond*[,2]

Mol. Biol. Evol. 32(5):1342–1353

- Best-in-class power
- Able to detect episodes of selection, not just selection on average at a branch
- Does not make unrealistic assumptions for tractability, improves statistical behavior

- Sample size is ~sites, branch level rate estimates could be imprecise
- Cannot reliably estimate which individual sites are subject to selection
- Exploratory testing of all branches leads to loss of power for large data sets (multiple test correction)

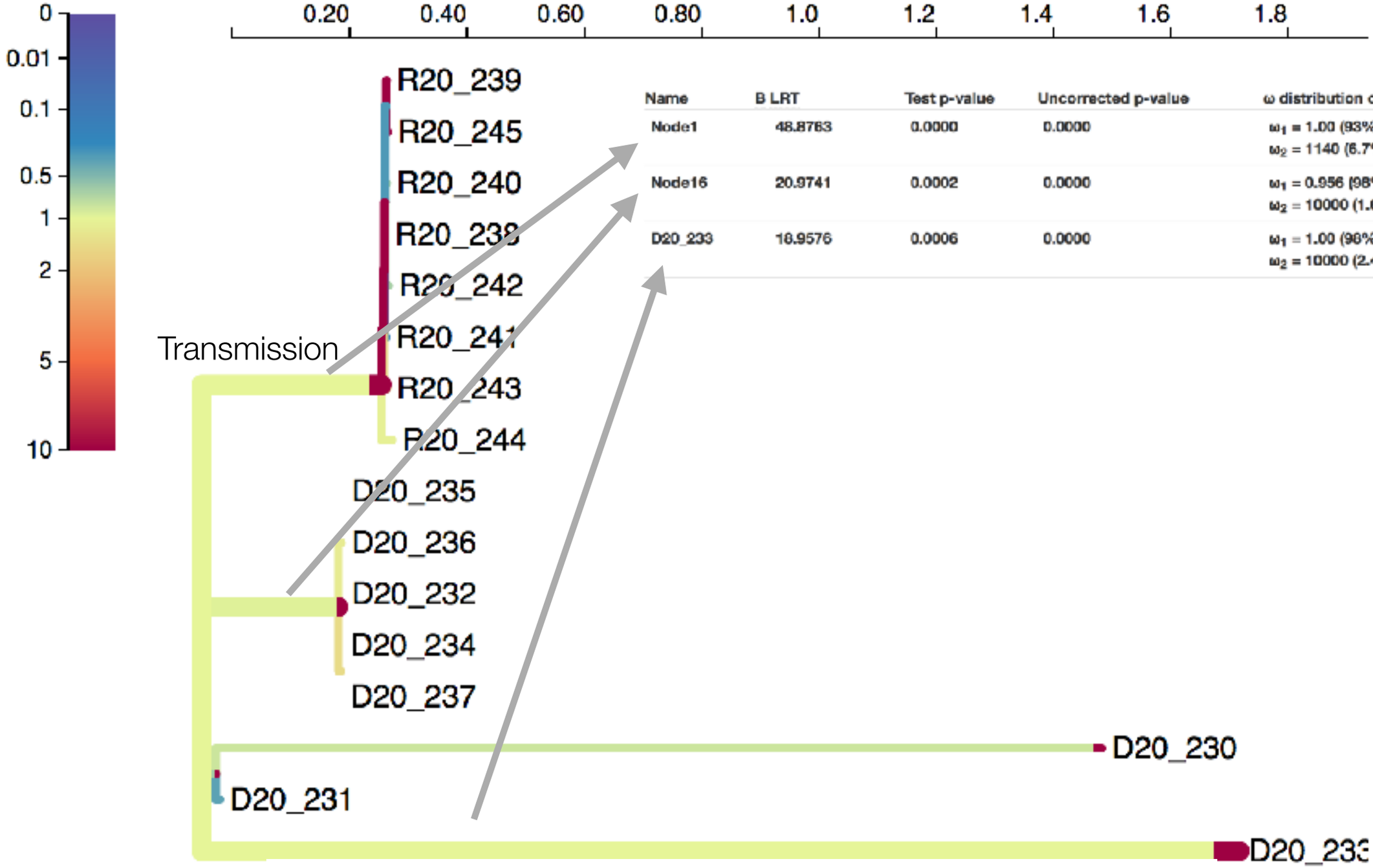
Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection

Martin D. Smith,[1] Joel O. Wertheim,[2] Steven Weaver,[2] Ben Murrell,[2] Konrad Scheffler,[2,3] and Sergei L. Kosakovsky Pond*[,2]

Mol. Biol. Evol. 32(5):1342–1353

- Uses a computationally simple trick to compute the likelihood of data, efficiently summing over all possible assignments of rate classes to branches
  - These cannot be factored into products, unlike sites, because evolution across tree branches is correlated, i.e. a change in the process along one branch affects many others.
- Uses a greedy (but well-performing) step-up procedure to decide how many rate classes to allocate to each branch, prior to testing for selection
  - Perform an evolutionary complexity analysis first (the *adaptive* part), then run selection tests.

HIV-1 *env*



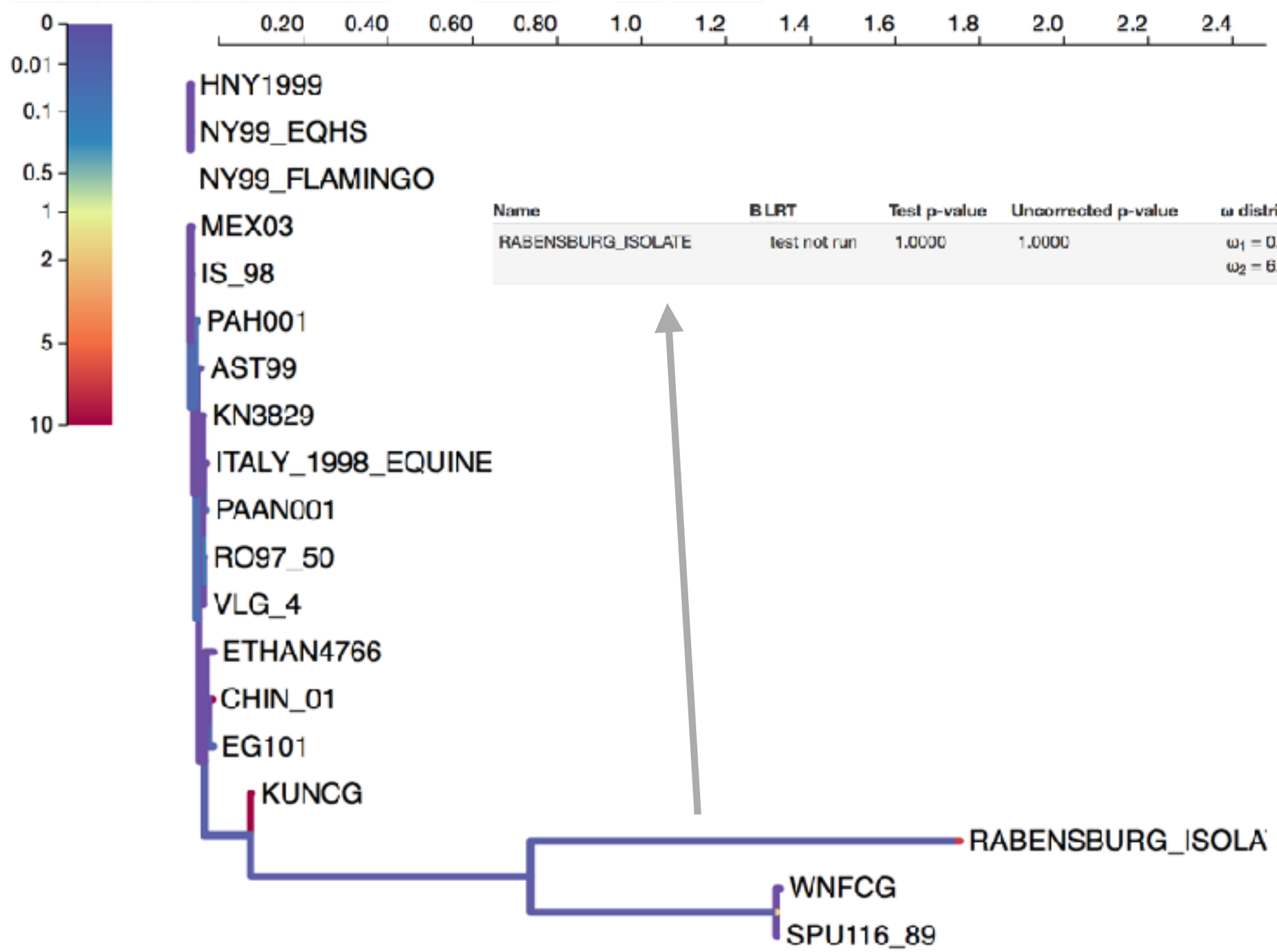
Name	B LRT	Test p-value	Uncorrected p-value	ω distribution over sites
Node1	48.8763	0.0000	0.0000	ω_1 = 1.00 (93%) ω_2 = 1140 (6.7%)
Node16	20.9741	0.0002	0.0000	ω_1 = 0.956 (98%) ω_2 = 10000 (1.6%)
D20_233	16.9576	0.0006	0.0000	ω_1 = 1.00 (96%) ω_2 = 10000 (2.4%)

Tree

ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	21	81%	1.8%	0
2	5	19%	2.6%	3

BRANCH-LEVEL SELECTION [aBSREL] 4

WN *NS3*



Name	BLRT	Test p-value	Uncorrected p-value	ω distribution over sites
RABENSBURG_ISOLATE	test not run	1.0000	1.0000	ω_1 = 0.00536 (99%) ω_2 = 6.95 (0.85%)

Tree

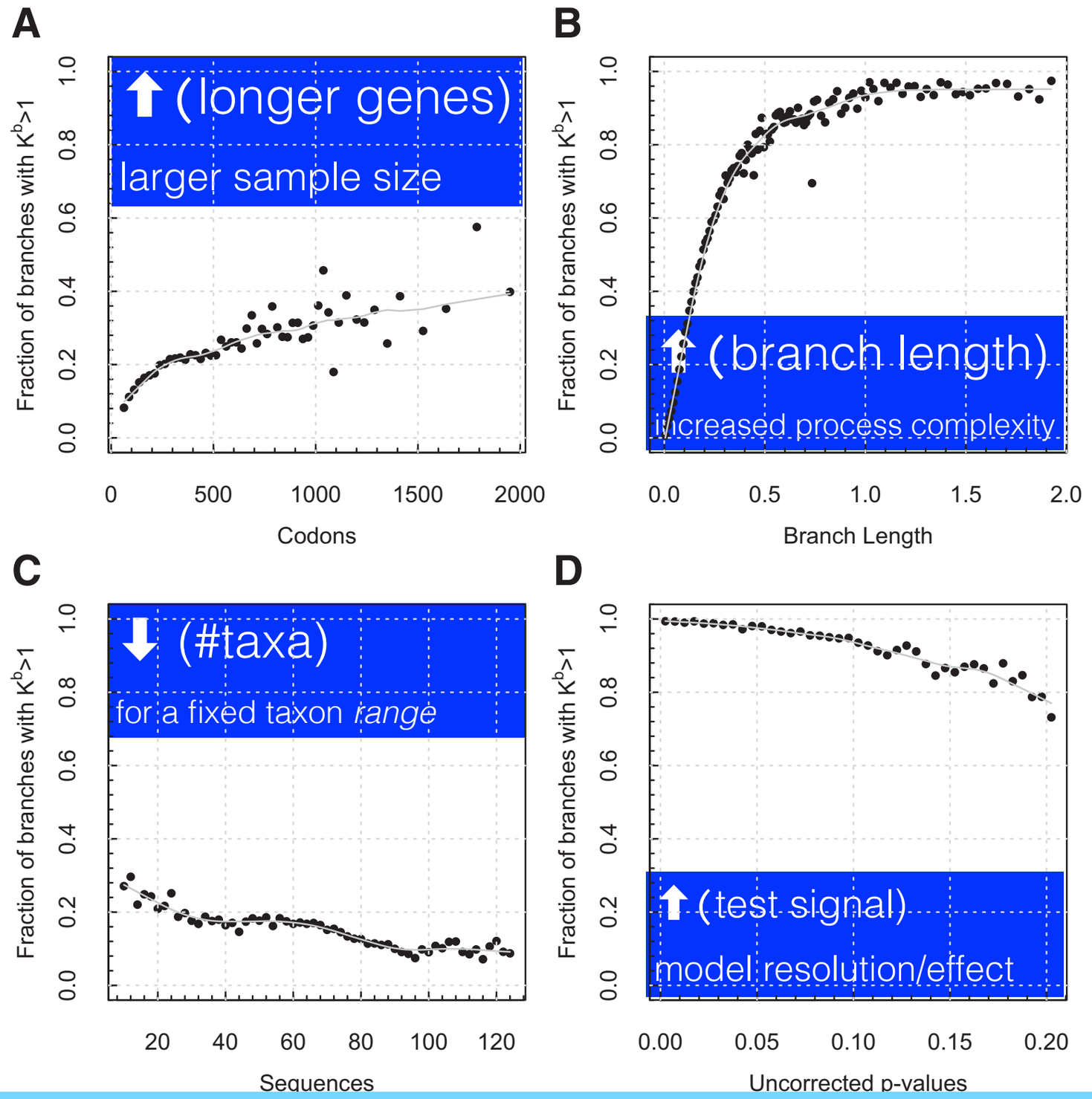
ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	32	91%	35%	0
2	3	8.6%	60%	0

aBSREL analysis

- **West Nile Virus NS3 protein**
  - `91%` branches can be explained with simple (single **`dN/dS`**) models
  - `3` branches (`9%, 60%` of tree length) 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%, 90+%` of tree length) 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`**, including the **transmission** branch

Correlates of evolutionary complexity

An analysis of ~9,000 curated gene alignments from *selectome.unil.ch*



Unanticipated effects of bad modeling assumptions

- Models that fail to account for significant shifts in selective pressures through lineages also significantly underestimate branch lengths

- An instructive example is long-range molecular dating of pathogens, where recent isolates (e.g., 30-50 years of sampling) are used to extrapolate the date when a particular pathogen had emerged

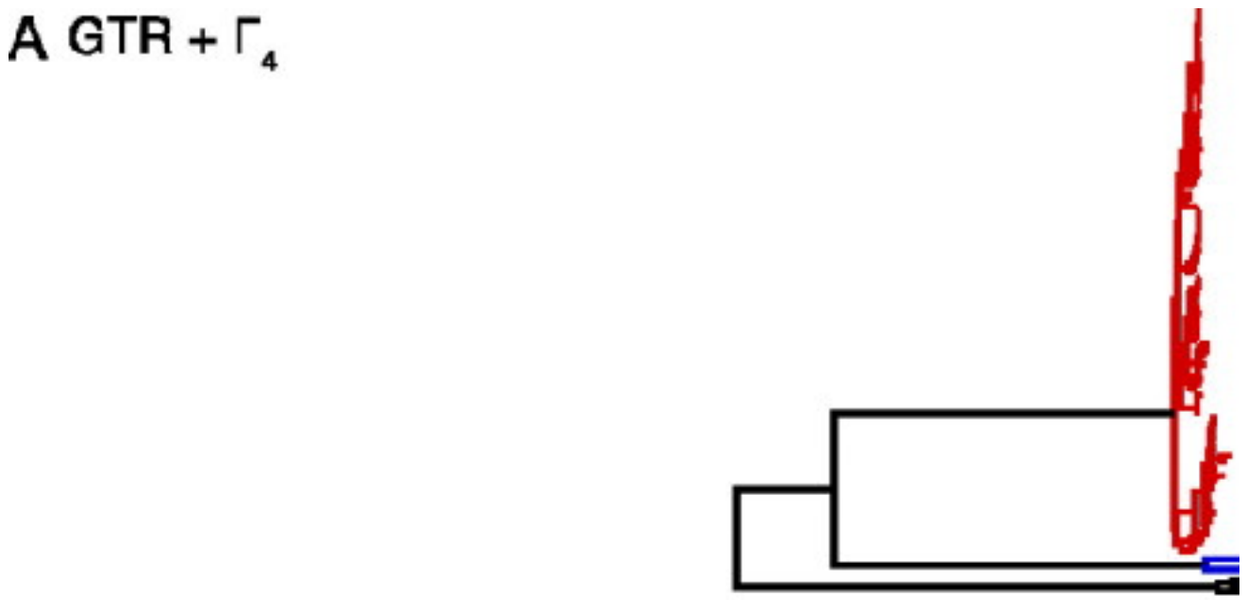
- This creates the situation when terminal branches in the tree have relatively high dN/dS (within-host level evolution), which deep interior branches have very low dN/dS (long term conservation)

- Using models that do not vary selection pressure across lineages yields a patently false "*too young*" estimate for the origin of **measles** (about 600 years ago)

- This estimate is refuted by clear historical records which suggest that measles is at least 1,500-5,000 years old

  - *This includes a treatise by a Persian physician Rhazes about **differential diagnosis of measles and smallpox** published circa 600 AD.*

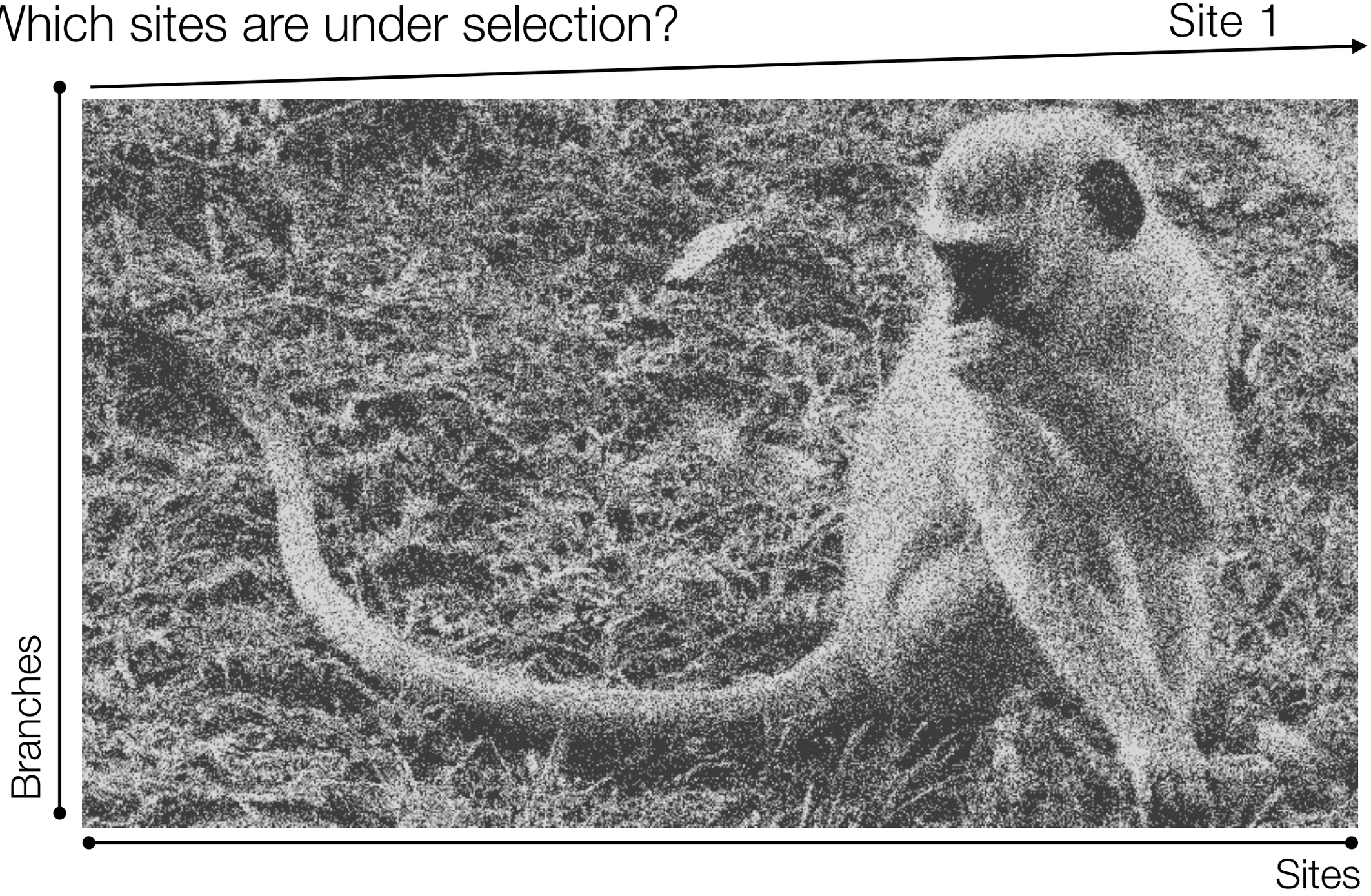
- Same patterns found for corona-viruses, ebola, avian influenza and herpesvirus



Wertheim and Pond (2011) Mol Biol Evol. 28(12):3355-65

Murrell et al 2012

Which sites are under selection?



For each image column, is there a significant proportion of bright pixels, once the column has been reduced to 2 colors only?

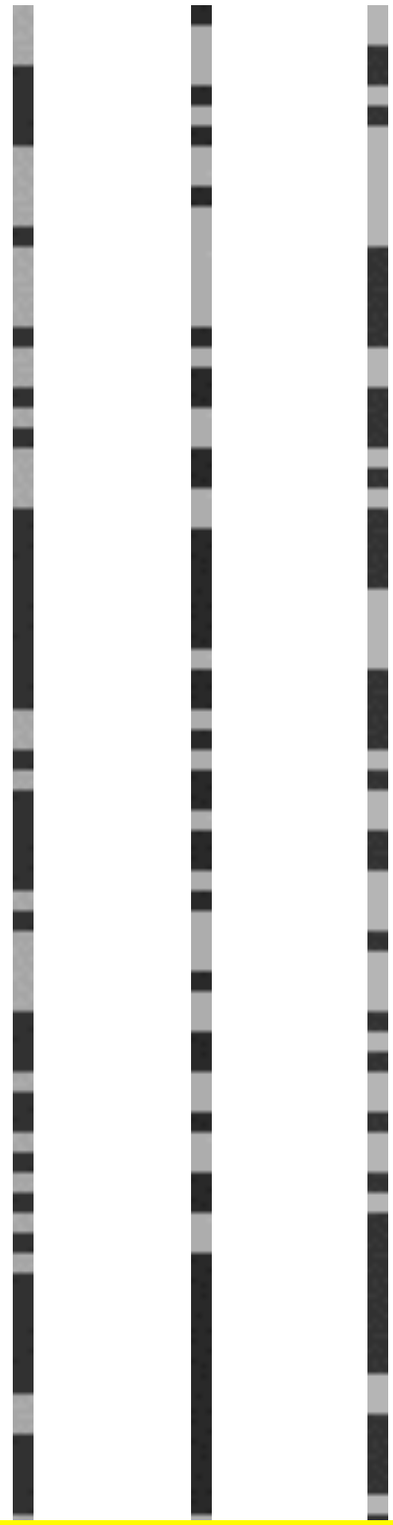
[MEME]: at a given **site**, each branch is a draw from a 2-bin (dS, dN) distribution, which is inferred from that site only. Test if there is a proportion of branches with dN>dS (LRT)

Detecting Individual Sites Subject to Episodic Diversifying Selection

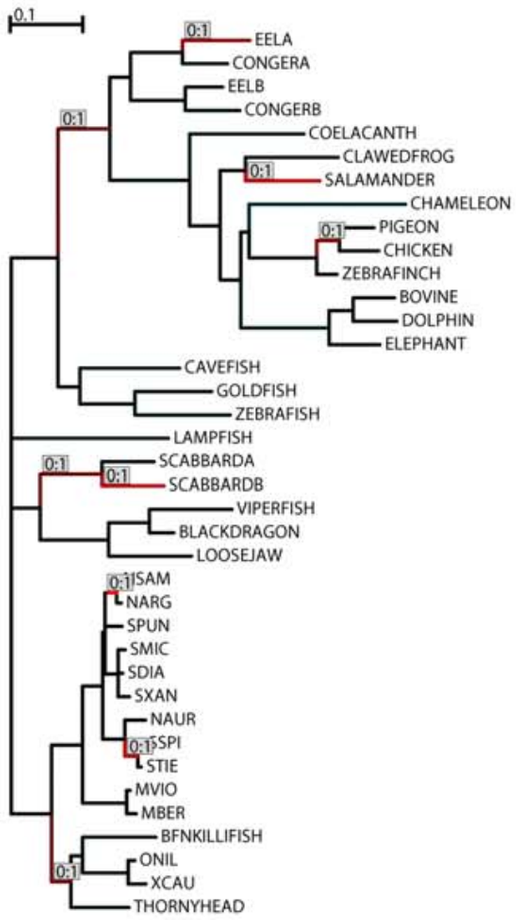
Ben Murrell[1,2], Joel O. Wertheim[3], Sasha Moola[2], Thomas Weighill[2], Konrad Scheffler[2,4], Sergei L. Kosakovsky Pond[4]*

PLoS Genetics | www.plosgenetics.org 1 July 2012 | Volume 8 | Issue 7 | e1002764

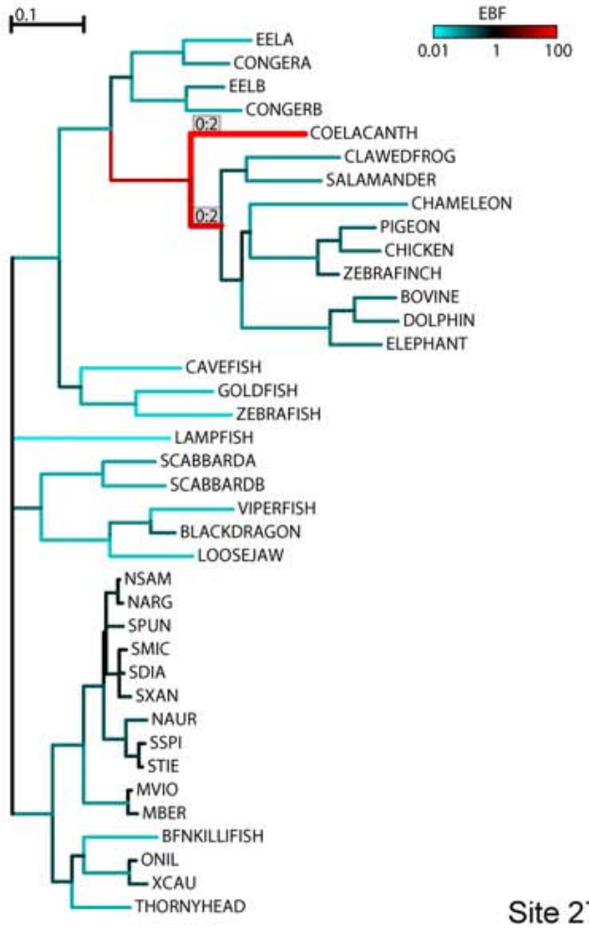
- Best-in-class power
- Able to detect episodes of selection, not just selection on average at a site
- Embarrassingly parallel (farm out each site), so runs reasonably fast



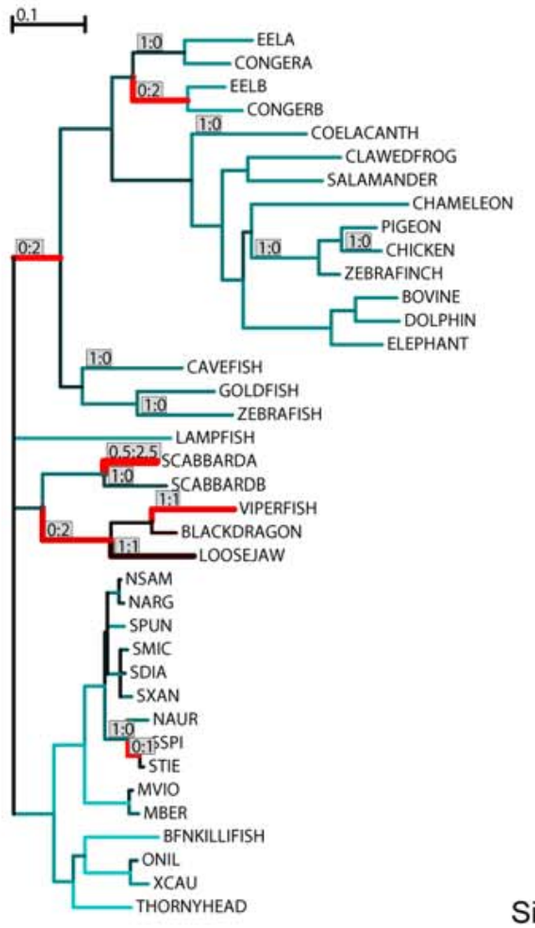
- Sample size is ~sequences, site level rate estimates imprecise
- Cannot estimate which individual branches are subject to selection
- Does not scale especially well with the number of sequences



Site 54



Site 273



Site 210

Pervasive selection, also picked up by older methods

Episodic selection, missed by old methods

Episodic selection, followed by conservation. Miscalled by old methods as purifying selection only

Murrell et al | **PLoS Genet 8(7): e1002764**

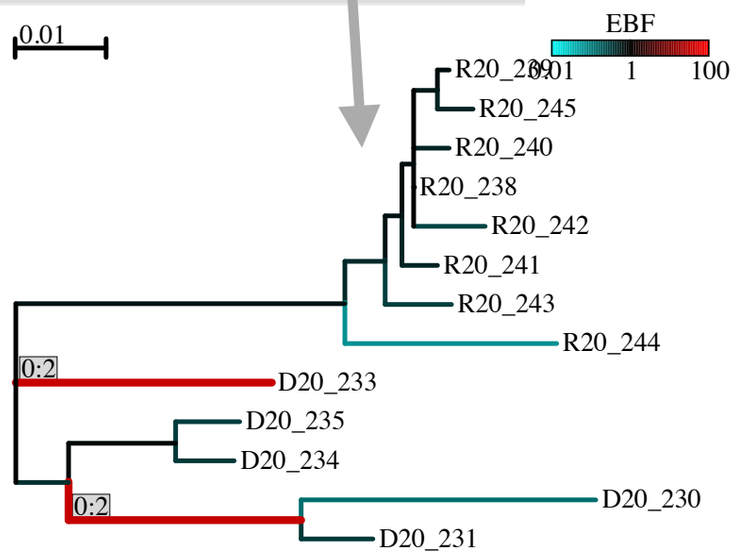
HIV-1 env

Found **11** sites with **evidence** of episodic diversifying selection (0.1 significance level Retabulate)

*This summary table reports the distribution of synonymous (α) and non-synonymous (β) substitution rates over **sites** inferred by the MEME model, where the proportion of branches with β>α is significantly greater than 0. p-value is derived using a mixture of χ^2 distributions, and q-values are obtained using Simes' procedure, which controls the false discovery rate under the strict neutral null (likely to be conservative).*

Codon	α	β^-	Pr[$\beta=\beta^-$]	β^+	Pr[$\beta=\beta^+$]	p-value	q-value	Branch-site information
19	0	0	0.943201	773.518	0.0567992	0.0317833	1	[Display]
161	0	0	0.829382	115.825	0.170618	0.00844037	1	[Display]
165	0	0	0.779355	60.0702	0.220645	0.0688644	1	[Display]
178	0	0	0.902811	67.331	0.0971885	0.0357418	1	[Display]
225	0	0	0.78381	72.3626	0.21619	0.0458125	1	[Display]
261	0.000305384	0.00027966	0.913635	881.437	0.0863648	0.0895985	1	[Display]
268	0	0	0.895791	10000	0.104209	0.0350278	1	[Display]
270	0	0	0.760392	226.778	0.239608	0.057291	1	[Display]
272	0	0	0.874576	59.5745	0.125424	0.0525324	1	[Display]
274	4.99073	0	0.764401	260.143	0.235599	0.062812	1	[Display]
282	0	0	1e-09	10.8217	1	0.0881541	1	[Display]

Site 161
82% of branches with α=β=0
18% of branches with α=0, β=116



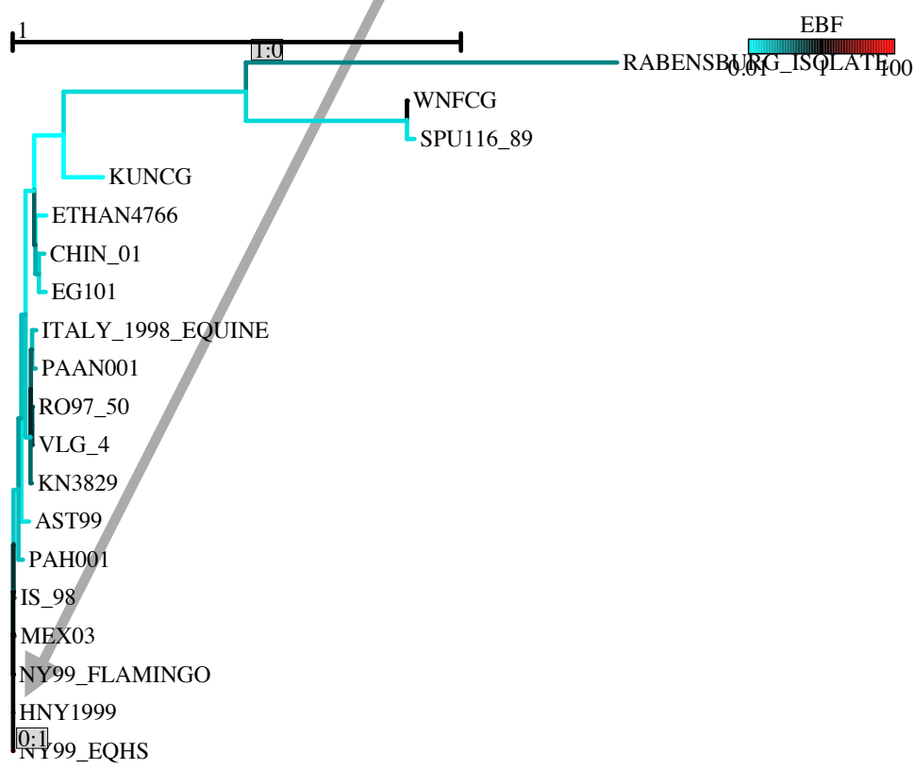
WN NS3

Found **3** sites with evidence of episodic diversifying selection (0.1 significance level Retabulate)

*This summary table reports the distribution of synonymous (α) and non-synonymous (β) substitution rates over **sites** inferred by the MEME model, where the proportion of branches with β>α is significantly greater than 0. p-value is derived using a mixture of χ^2 distributions, and q-values are obtained using Simes' procedure, which controls the false discovery rate under the strict neutral null (likely to be conservative).*

Codon	α	β^-	Pr[β=β^-]	β^+	Pr[β=β^+]	p-value	q-value	Branch-site information
249	0	0	1.00003e-09	2.44107	1	0.0166364	1	[Display]
496	0.947581	0	0.955206	74.0758	0.0447943	0.0838408	1	[Display]
557	0.275201	0	0.963363	171.171	0.036637	0.0261761	1	[Display]

Site 557
96% of branches with α=0.28, β=0
4% of branches with α=0.28, β=171



MEME results

- **West Nile Virus NS3 protein**
  - **Three** sites, (including **249**) with significant evidence of **episodic** (or pervasive) diversifying selection.
- **HIV-1 transmission pair**
  - **Eleven** sites with significant evidence of **episodic** (or pervasive) diversifying selection.

Why MEME?

- Affords a much greater power to detect selection
- Mitigates the pathological effect when adding sequences to a sample can reduce, or remove, signal of selection

> "The greater power of MEME indicates that selection acting at individual sites is considerably more widespread than constant **ω** models would suggest. It also suggests that natural selection **is predominantly episodic**, with transient periods of adaptive evolution masked by the prevalence of purifying or neutral selection on other branches. We emphasize that MEME is not just a quantitative improvement over existing models: for 56 sites in our empirical analyses, we obtain qualitatively different conclusions. FEL asserts that these sites evolved under significant **purifying** selection, but MEME is able to identify the signature of **positive selection on some branches**"

Murrell et al | **PLoS Genet 8(7): e1002764**

Why MEME?

- Affords a much greater power to detect selection
- Mitigates the pathological effect when adding sequences to a sample can reduce, or remove, signal of selection

> "Although a previous analysis of 38 vertebrate rhodopsin sequences found no sites under selection at posterior probability >95%, the same authors found 7 selected sites in the subset of 11 squirrelfish sequences, and 2 selected sites when the subset of 28 fish sequences was analyzed. These results run counter to the expectation that more data should provide greater power to detect selection. MEME, on the other hand, [typically] detects more selected sites when more sequences are included."

Murrell et al | **PLoS Genet 8(7): e1002764**

Analysis summary

	WNV NS3	HIV-1 *env*
Gene-wide episodic selection (BUSTED)	No	Yes
Branch-level selection (aBSREL)	No	Yes, three branches, including transmission
Site-level episodic selection (MEME)	Yes, 3 sites	Yes, 11 sites

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.

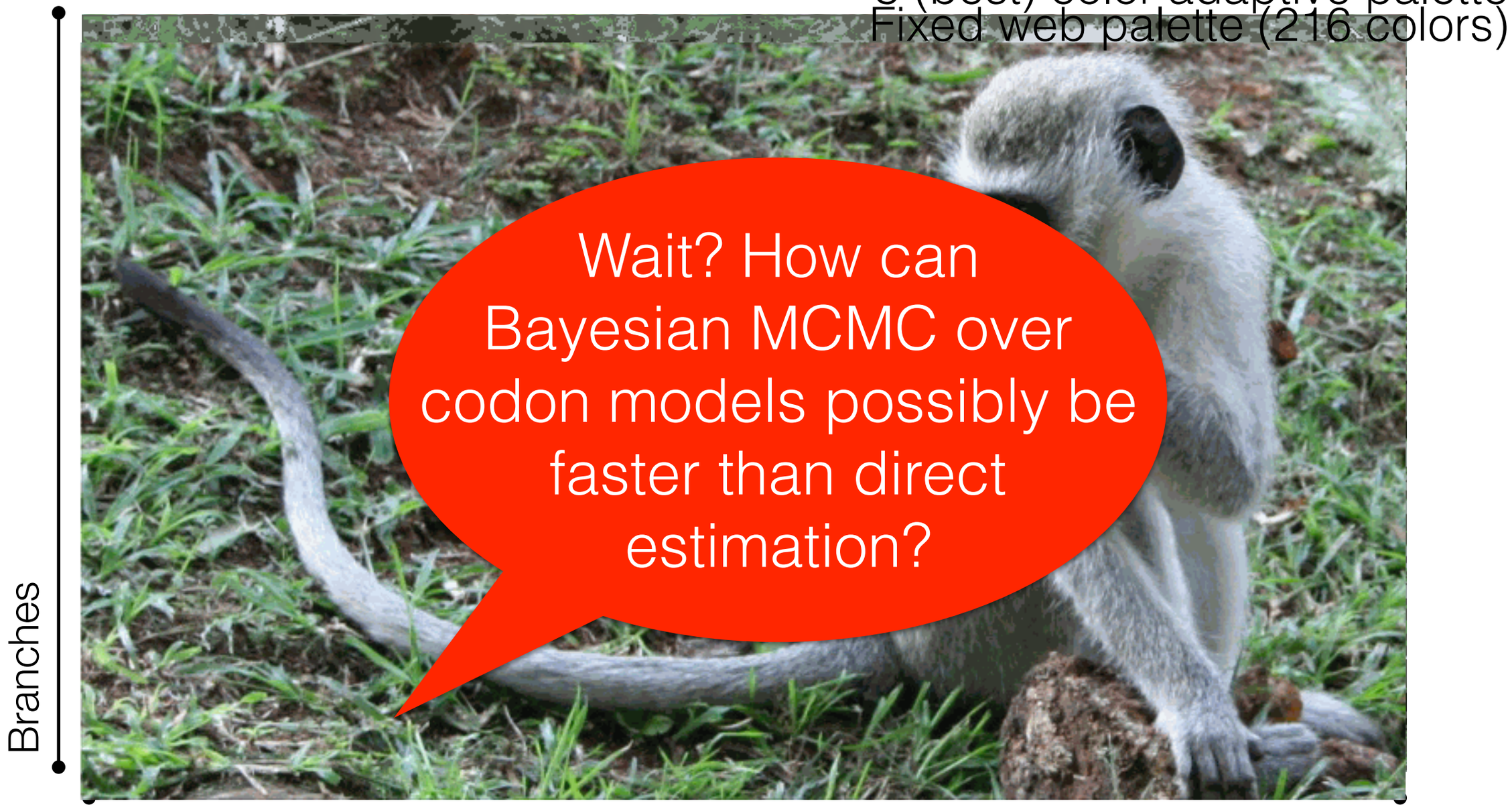
- **Model violations:** MEME supplies much more flexible distributions of **`dN/dS`** over sites; compared to alignment-wide 3-bit BUSTED distribution.

- **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.

> 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. In this manuscript [MEME], we describe how to infer the location of sites, pooling information over branches; previously [aBSREL] we have outlined a complementary approach to find selected branches by pooling information over sites.
>
> *Murrell et al 2012*

FUBAR: selection testing done fast



Average colors over sites; use a relatively large but fixed palette to approximate the image

[FUBAR]: Fix a grid of dS and dN values, use the data to sample (Bayesian MCMC) weights to individual grid points; this forms the prior distribution on rates; use empirical Bayes to obtain site-level estimates of posterior probability that dN > dS

- The time consuming part of traditional random-effects models is the estimation of the aliment-wide dN/dS distribution

- Each hyper-parameter adjustment entails an expensive phylogenetic likelihood calculation

- Larger data sets —> more complex mixtures needed to avoid smoothing, i.e., more parameters, more evaluations, and a non-linear dependance on data-set sizes

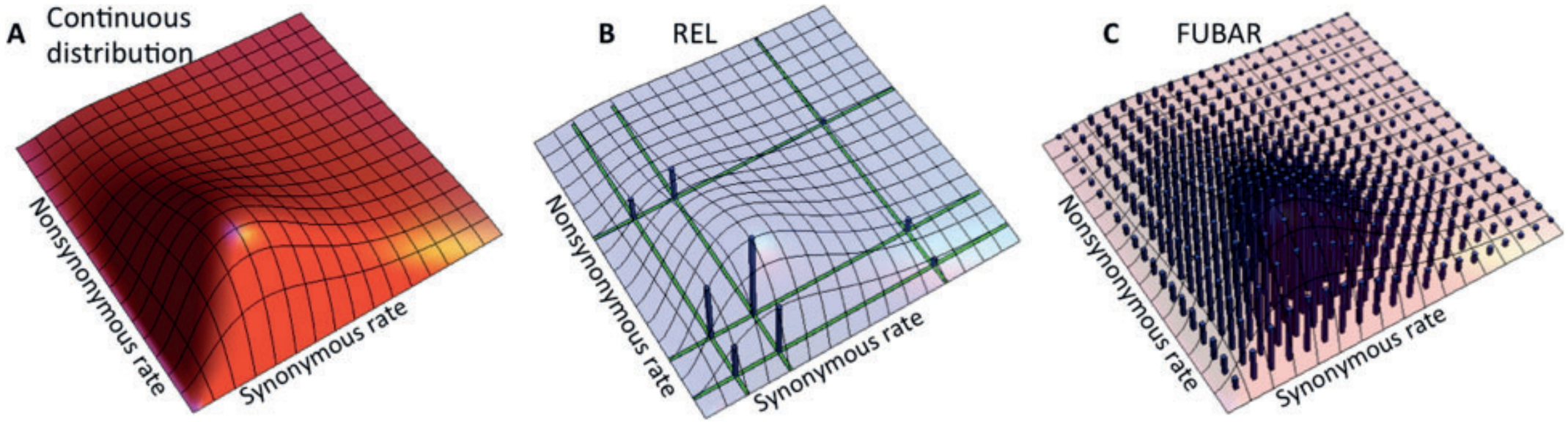
- With FUBAR we make the following approximations:
  - Branch lengths, GTR biases etc, are estimated using simple (nucleotide models) and held fixed
  - We fix a 15x15 or 20x20 grid of (dS,dN) values *a priori*; the data only inform how much weight will be allocated to each point
  - Only need to evaluate the expensive codon-based phylogenetic likelihood once for each grid point: **complexity only increases linearly with the size of the data.** This step is also embarrassingly parallel.
  - Allocating weights to individual points is done using MCMC (or Gibbs sampling, or variational Bayes); this step does not require ANY further evaluations of the phylogenetic likelihood, i.e., **its cost does not depend on the size of the alignment**

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

FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection

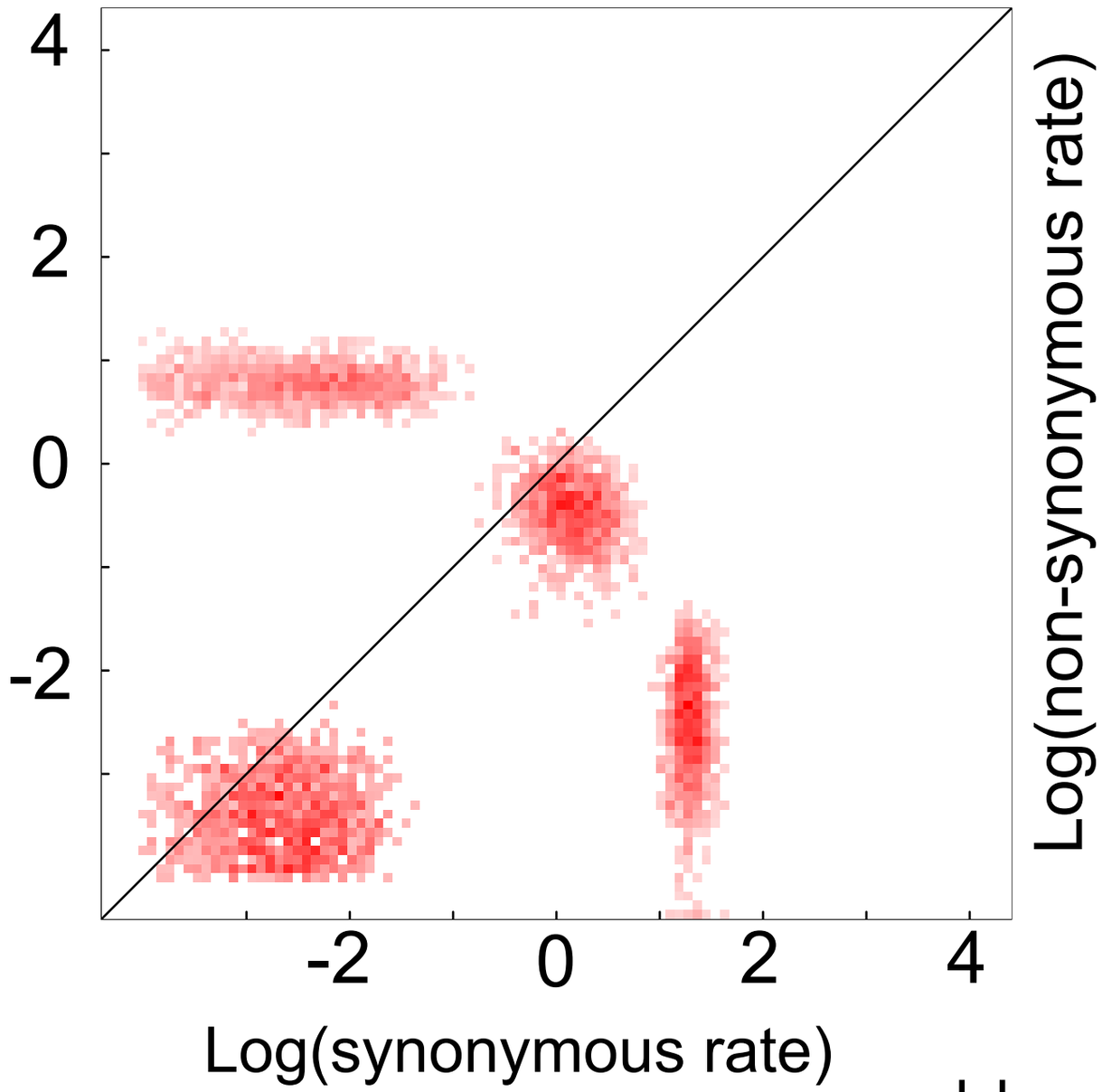
Ben Murrell,[1,2,3] Sasha Moola,[1,3] Amandla Mabona,[1,4] Thomas Weighill,[1] Daniel Sheward,[5] Sergei L. Kosakovsky Pond,[6] and Konrad Scheffler*[,1,6]

Mol. Biol. Evol. 30(5):1196–1205

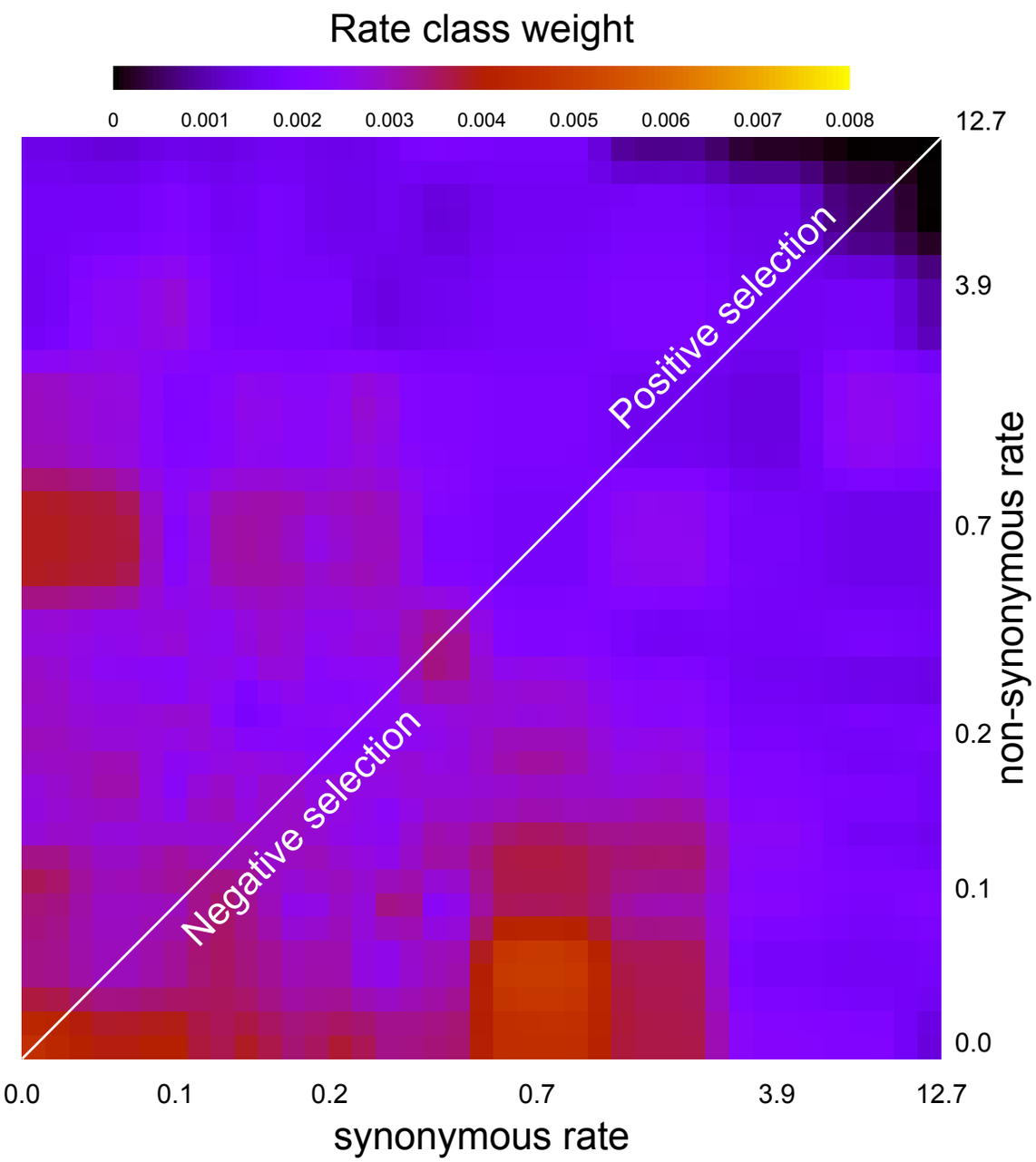


Fitting a small number (4) of dN and dS values directly

with post-hoc error estimates

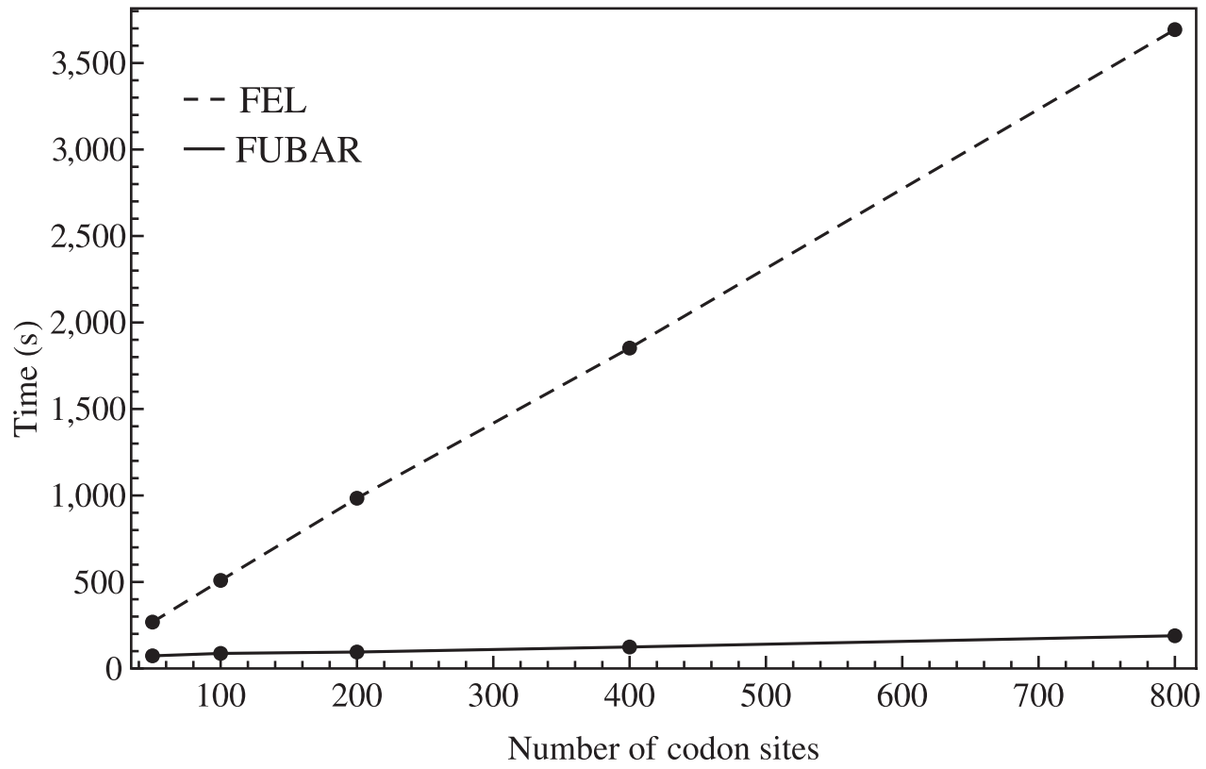


Using a FUBAR grid



Hepatitis E Virus Genotype 4 ORF3

Brayne et al | *J. Virol.* May 2017 91:9 16 e02241-16



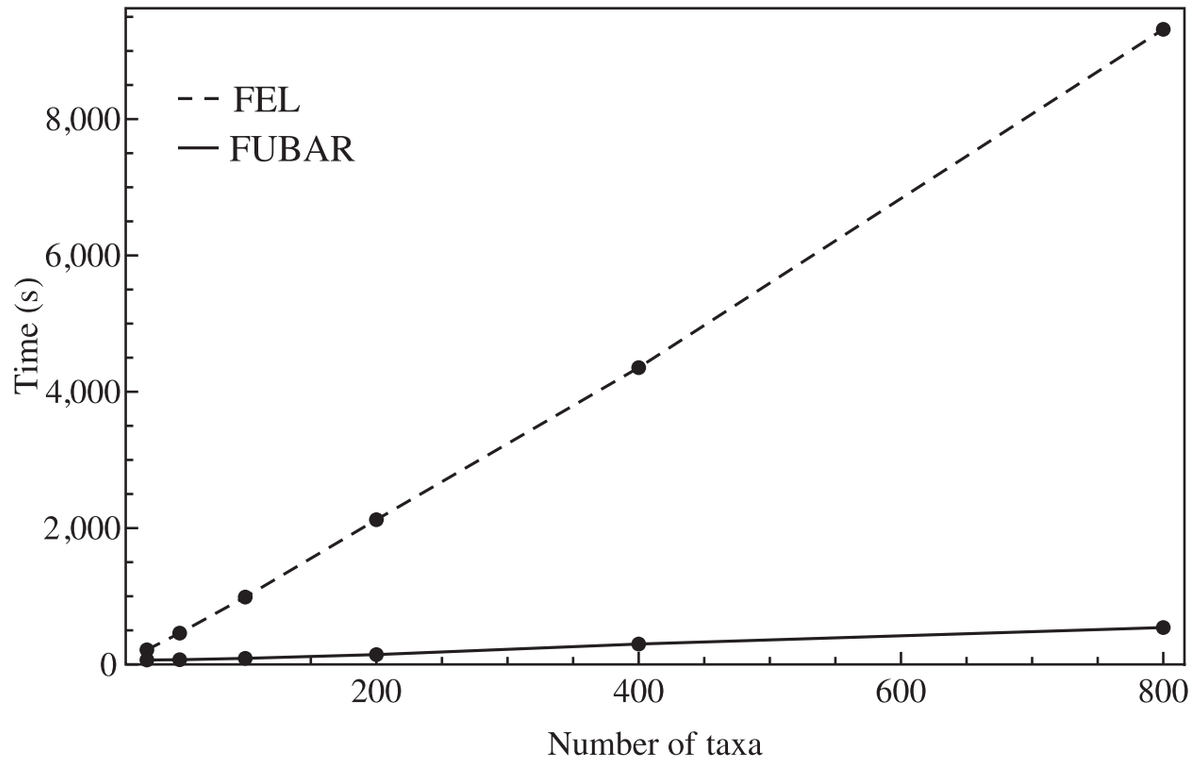


FIG. 2. Execution times for FEL and FUBAR as a function of the number of codon sites (top) and number of taxa (bottom).

FUBAR is dramatically faster (and as good or better)

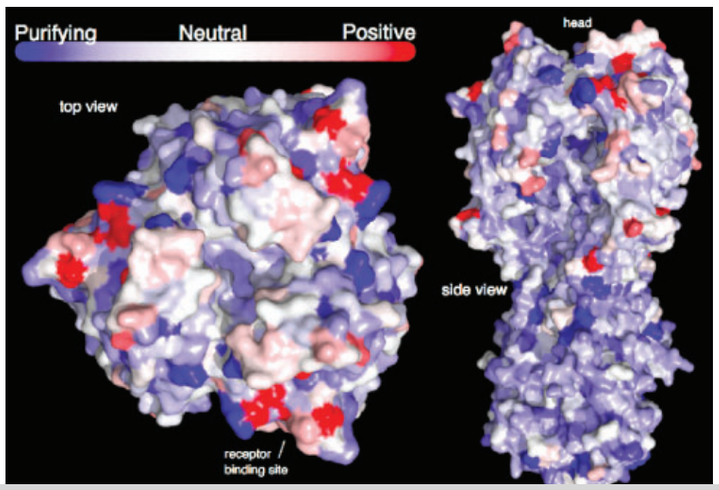
Table 2. Run Time Comparisons between Different Selection Detection Methods on 16 Empirical Data Sets, Sorted on the Duration of the FUBAR Run.

Data Set	Taxa	Codons	Mean Divergence Subs/Site	FUBAR Run Times (s)	Run Times (Times Slower than FUBAR)			
					FEL	REL	PAML M2a	PAML M8
Echinoderm H3	37	111	0.33	40	5.1	*12.0*	7.1	*46.1*
Flavivirus NS5	18	342	0.48	45	8.6	4.5	9.3	*25.5*
Drosophila adh	23	254	0.26	53	3.4	4.0	2.7	4.3
West Nile virus NS3	19	619	0.13	58	6.1	5.9	*37.2*	*105.5*
Hepatitis D virus Ag	33	196	0.29	59	4.0	3.3	*10.1*	*22.4*
Primate lysozyme	19	130	0.08	62	0.5	3.0	0.7	1.8
Vertebrate rhodopsin	38	330	0.34	62	*12.0*	4.9	8.4	*18.2*
Japanese encephalitis virus env	23	500	0.13	68	4.8	8.8	1.6	4.0
Mamallian β-globin	17	144	0.38	74	1.5	8.4	2.3	5.6
Abalone sperm lysin	25	134	0.43	78	1.9	3.9	3.7	9.3
HIV-1 vif	29	192	0.08	84	2.6	3.8	2.3	4.5
Salmonella recA	42	353	0.04	102	2.1	2.9	2.6	*12.3*
Camelid VHH	212	96	0.27	120	6.3	*17.2*	*141.0*	*311.1*
Diatom SIT	97	300	0.54	136	*10.2*	5.1	*21.5*	*19.3*
Influenza A virus H3N2 HA	349	329	0.04	210	*15.0*	*14.4*	*221.1*	*616.4*
HIV-1 rt	476	335	0.08	278	*15.2*	*14.4*	Ø[a]	Ø[a]

NOTE.—Run times that are at least 10 times greater than those of FUBAR are italicized, and those at least 100 times greater are underlined.
[a]PAML reported an error regarding too many ambiguities in the data set.

FUBAR is dramatically faster (and as good or better)

> We reconstructed the phylogeny for **3,142** complete H3 nucleotide sequences isolated from humans using FastTree 2. The FUBAR selection analysis (which we restricted to 10 CPUs, just as for the timing comparisons) took **one and a half hours.**

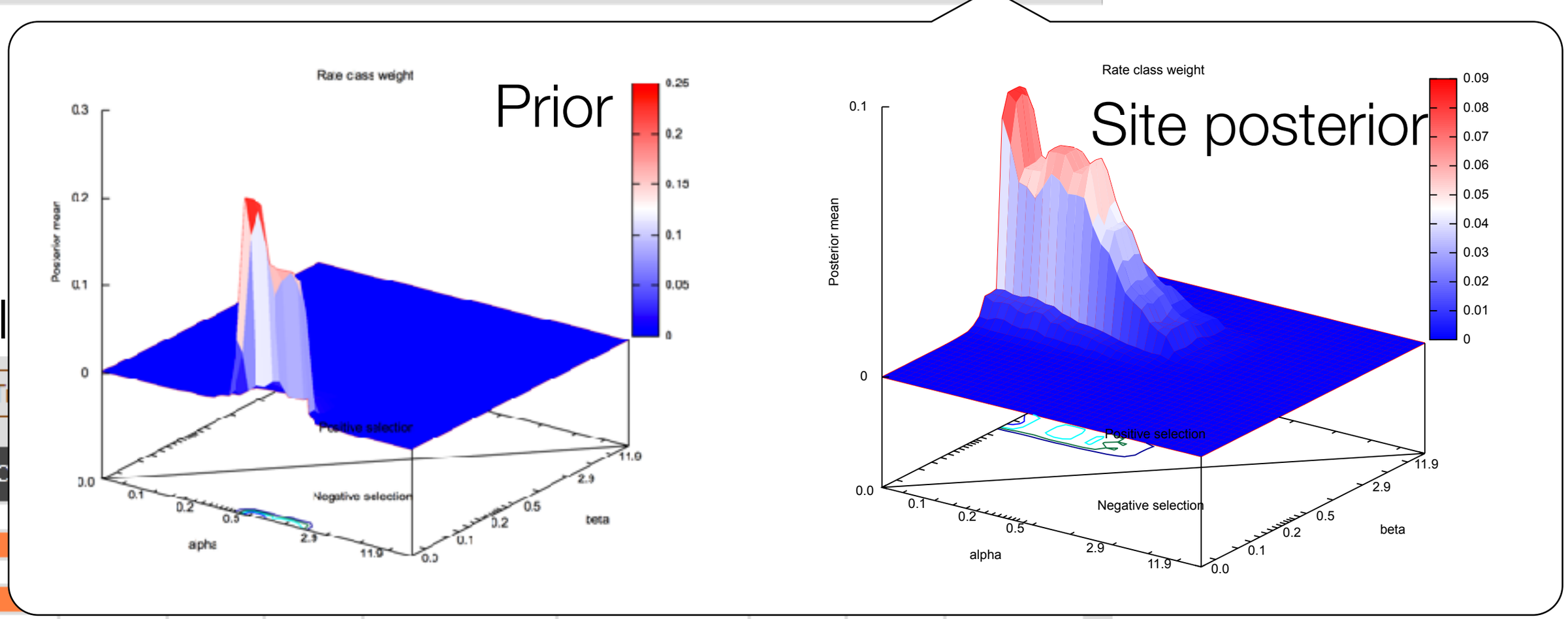


Fast site-level analysis (FUBAR): no branch to branch variation; pervasive diversifying selection; random effects

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]



HI

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.

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
Are parts of the tree evolving with different selective pressures relative to other parts of the tree?	RELAX (a test for relaxed selection)	• p-value for whether or not there is relaxed or intensified selection • inferred dN/dS distributions for different branch sets • more flexible distribution companions possible

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.

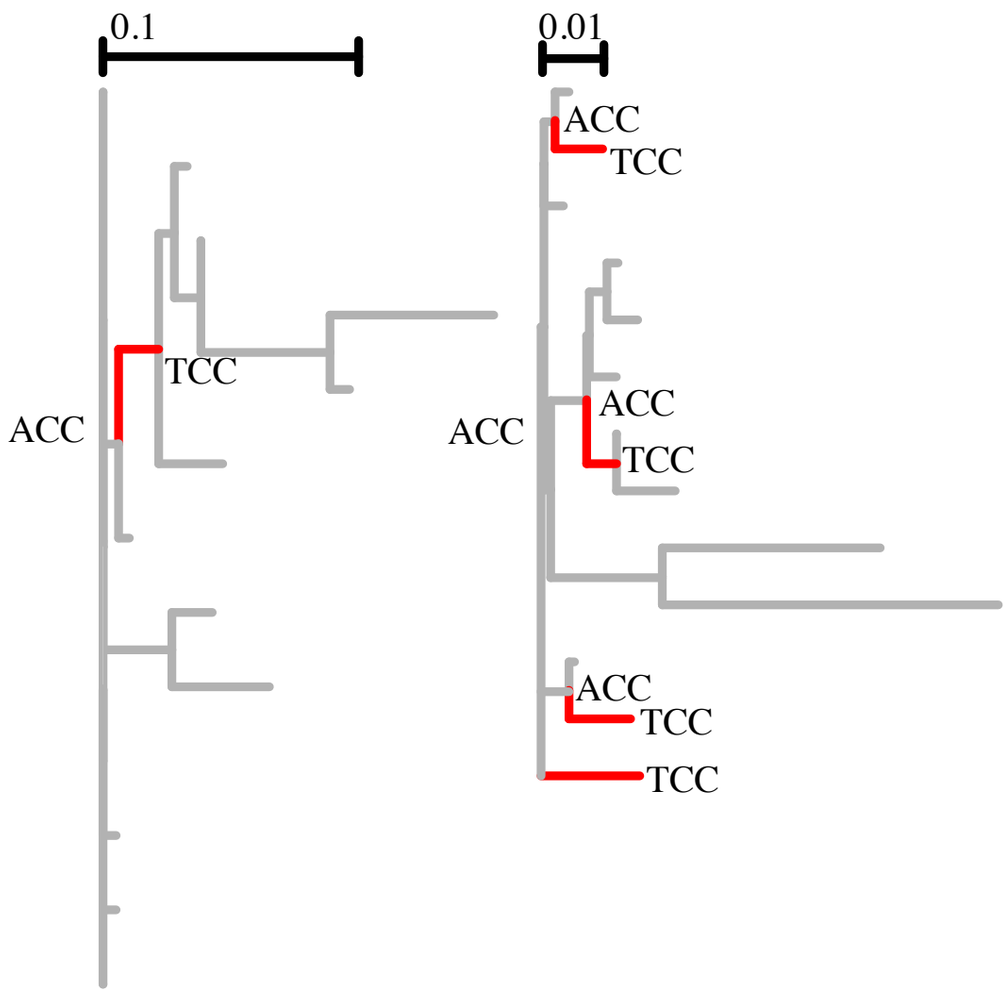


Figure 4.2: The effect of recombination on inferring diversifying selection. Reconstructed evolutionary history of codon 516 of the Cache Valley Fever virus glycoprotein alignment is shown according to GARD inferred segment phylogeny (left) or a single phylogeny inferred from the entire alignment (right). 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) using multiple phylogenies (one per fragment), but inferring other parameters (e.g. nucleotide substitution biases 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).

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

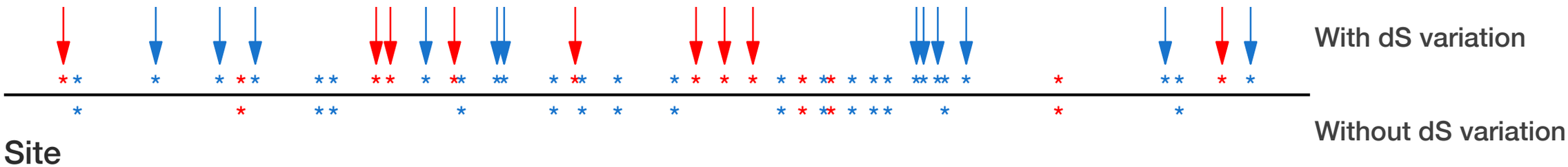
- **dS** = constant for all sites (assumed by many models); this assumption appears to be nearly universally violated in biological data, due to e.g. secondary structure, localized codon usage bias, overlapping reading frames, etc.
- This can lead to, e.g. incorrect identification of relaxed constraint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

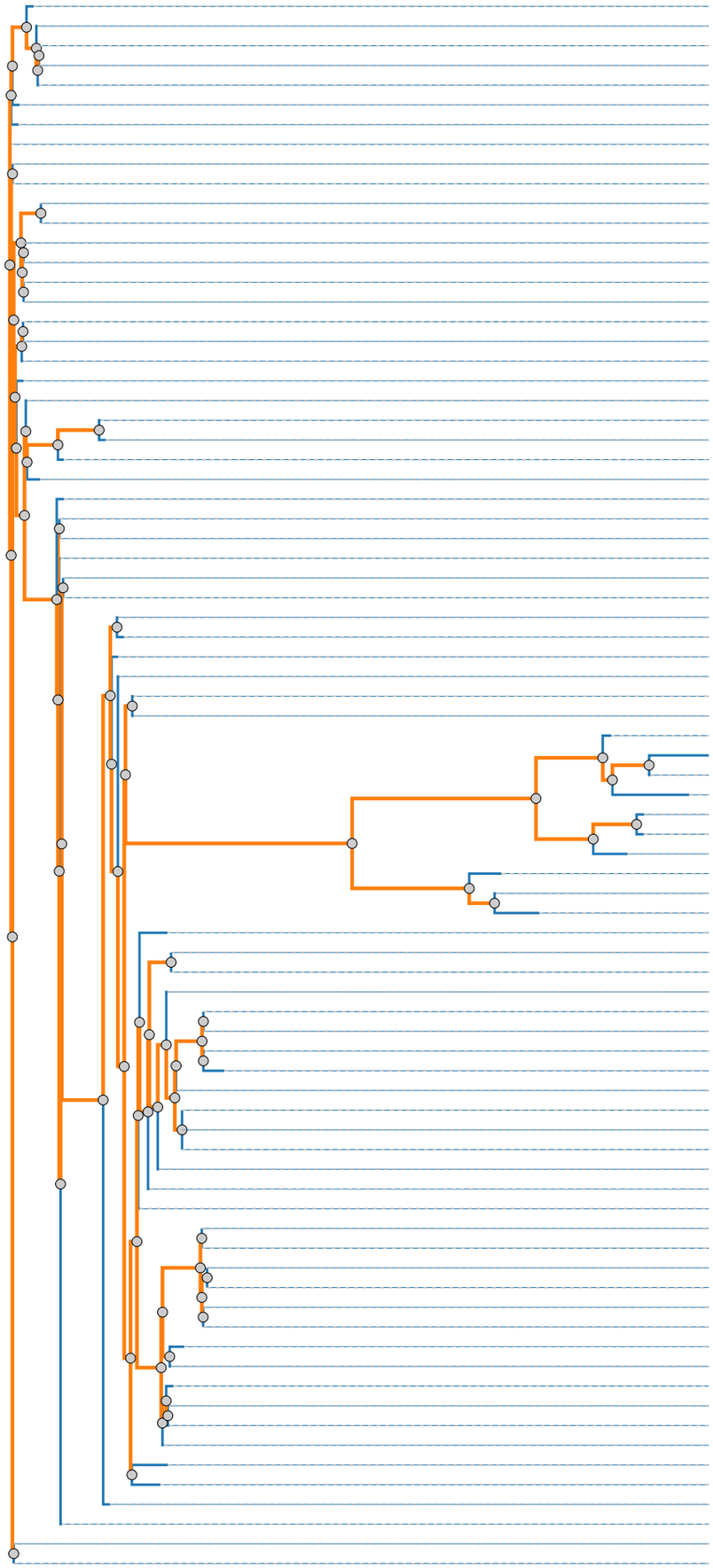
Mol. Biol. Evol. 22(12):2375–2385. 2005

Sites detected by FEL with and without dS variation



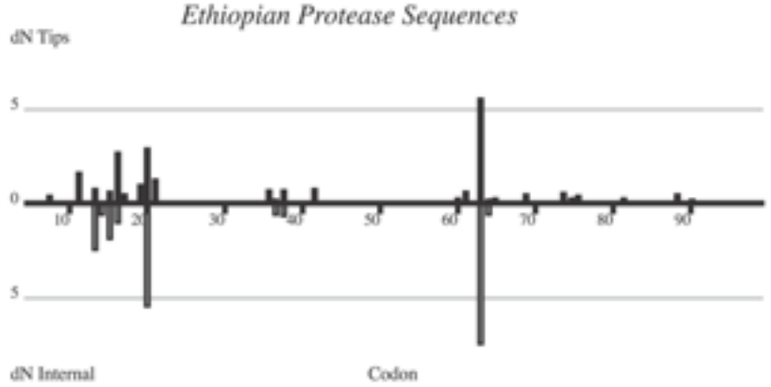
Interpreting dN/dS for intra-host and intra-species pathogen

- **dN/dS** can be estimated for all sorts of sequence data (e.g., it has been done for cancer SNP data)
- Traditional interpretation of dN/dS is based on the assumption that **substitution ~ fixation**
- Not the same for intra-species / intra-host pathogens
  - Much of variation is due to polymorphism, or even dead-end mutations
  - This is because selection has not had a chance to “filter” mutations (except for patently deleterious ones)
  - This often manifests as differences in selective “regimes” between tips and internal branches



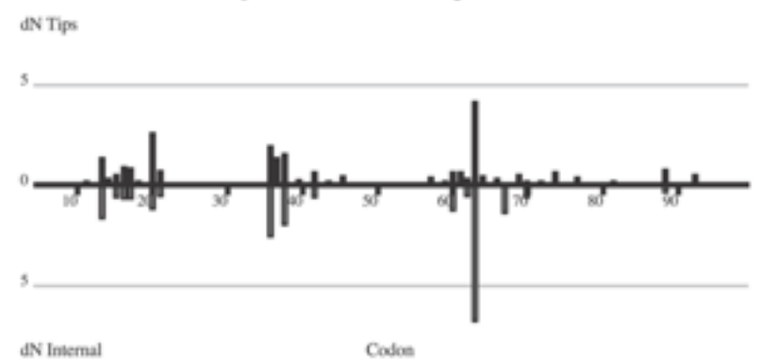
- Partition a pathogen tree into terminal and internal branches
- Terminal branches potentially include “dead-end” lineages, i.e. those which are maladaptive
- Internal branches include at least one “*transmission*” (intra-species) or “*replication*” (intra-host) events: stronger action of selection
- Focusing on a subset of branches can allow one to interpret dN/dS more precisely

Ethiopian Protease Sequences



Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0
Only dN Tips >0	17	0.49	LRT = 5.55
dN Internal >0	9	1.42	$p_A = 0.018$, **$p_B = 0.02$**

South African Protease Sequences



Site Class	Codon Count	dN/dS along Terminal Branches	Is dN/dS elevated when dN internal > 0
Only dN Tips >0	18	0.42	LRT = 10.46
dN Internal >0	15	1.19	$p_A = 0.001$, **$p_B = <0.01$**

- *"… at least half of the amino acid sites selected within individuals are not selected at a population level"*

- *"… Based on the elevated rate of adaptation within individuals detected at codons subject to population-level selection, relative to the codons where only recent substitutions have been inferred, we conclude that recent substitutions are, on average, maladaptive at the level of the human population"*

PLoS Comput Biol 2(6): e62. doi:10.1371/journal.pcbi.0020062