Quantifying Natural Selection in Coding Sequences

Alexander G. Lucaci, PhD
Postdoctoral Associate
Institute for Computational Biomedicine @ Weill Cornell Medicine

Email: agl001@med.cornell.edu
Twitter: @aglucaci

Preliminaries

- Please confirm access to **HyPhy**: http://hyphy.org/download/
  - https://youtu.be/fgNrPbOTpxE (How to install HyPhy)
  - You can do a datamonkey.org based tutorial, but if you have Linux or OSX, you can also do a command line tutorial for more features.
  - General user questions and feedback: https://github.com/veg/hyphy/issues
- **Datamonkey** web-app: http://www.datamonkey.org
  - YouTube example videos (channel HyPhy vision)
  - https://www.youtube.com/channel/UCIgRnbJjbOWhshe5ThhaWGw/videos (Tutorials)
- Example datasets at https://github.com/veg/selection-tutorial/
  - Test datasets and practical instructions: www.hyphy.org (search for “Detect Selection”)

Outline

- Brief background and examples of natural selection
- **dN/dS** as a tool to measure the action of natural selection, explained using the first counting method for estimating dN/dS (Nei-Gojobori, 1986) and its extensions.
- Codon substitution models — the basis of modern (1998-) dN/dS estimation approaches
- Confounding processes (synonymous rate variation, recombination, multiple nucleotide substitutions)
- On the suitability of dN/dS for within-species inference

- 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 (SLAC, FEL, FUBAR)
  - Relaxed or intensified selection (RELAX)
  - Detecting **differences** in selective pressure (Contrast-FEL)

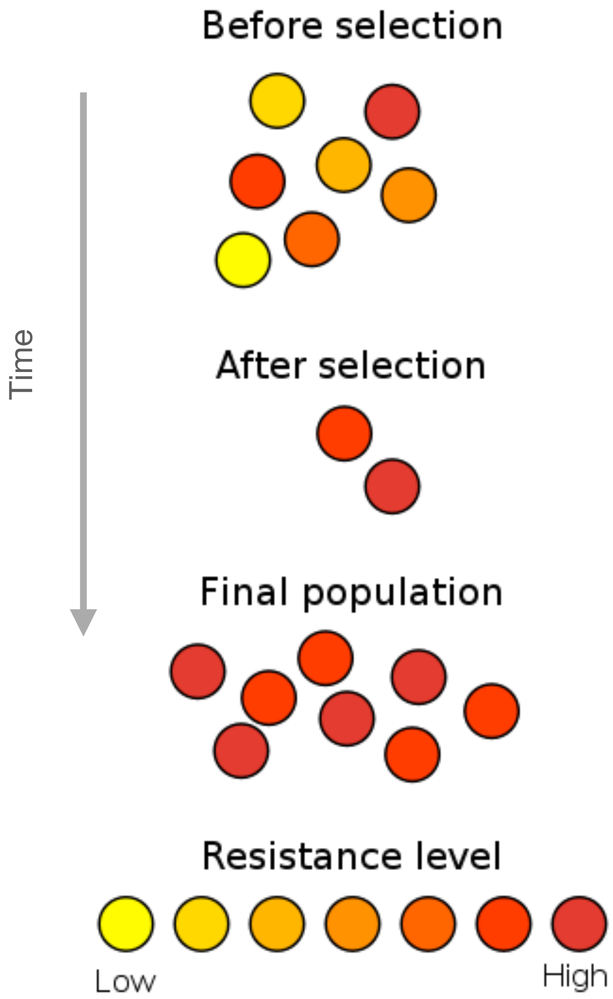
A bit of trivia

- 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
- The same mutation can have different fitness costs in different environments (fitness landscape), and different genetic backgrounds (epistasis)

- 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

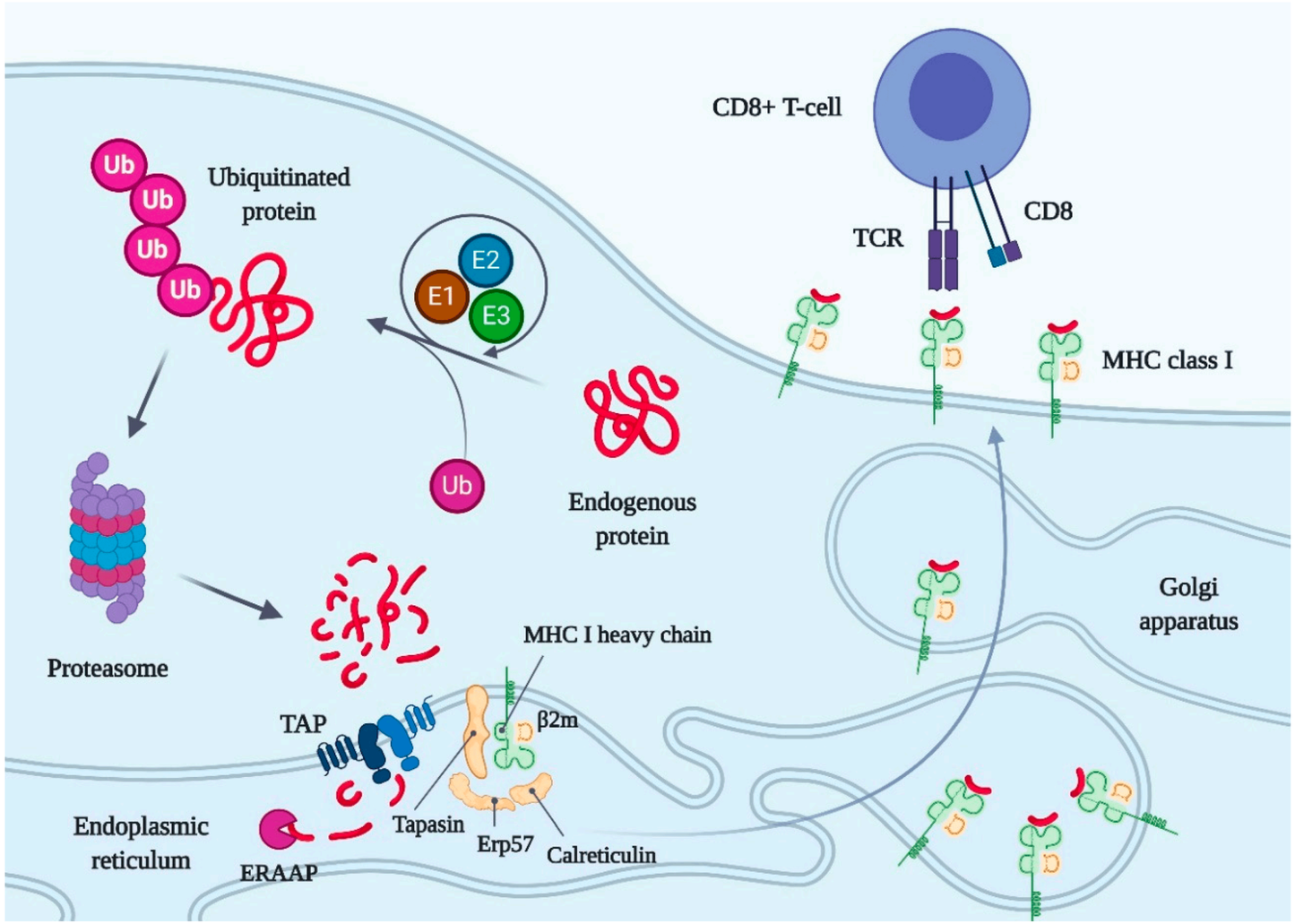


http://en.wikipedia.org/wiki/File:Antibiotic_resistance.svg

https://www.youtube.com/watch?v=plVk4NVIUh8

Example: MHC-restricted CTL killing of infected cells

- CD8+ Cytotoxic T-lymphocytes (CTLs) effect cell-mediated immune response
- Foreign (e.g., viral) proteins are cleaved by the proteasome in infected cells, transported by the Transporter associated with antigen processing (TAP) complex and loaded onto the MHC Class 1 molecule.
- MHC Class 1 presents a restricted polypeptide (**epitope**) on the surface of the cell.
- A CD8+ cells binds to presented foreign peptides via a T cell receptor (TCR) and initiates infected cell apoptosis.



https://www.mdpi.com/1420-3049/25/22/5409

MHC Class 1 Molecules

- Present **linear** foreign peptides which are most commonly 9 or 10 amino acids 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

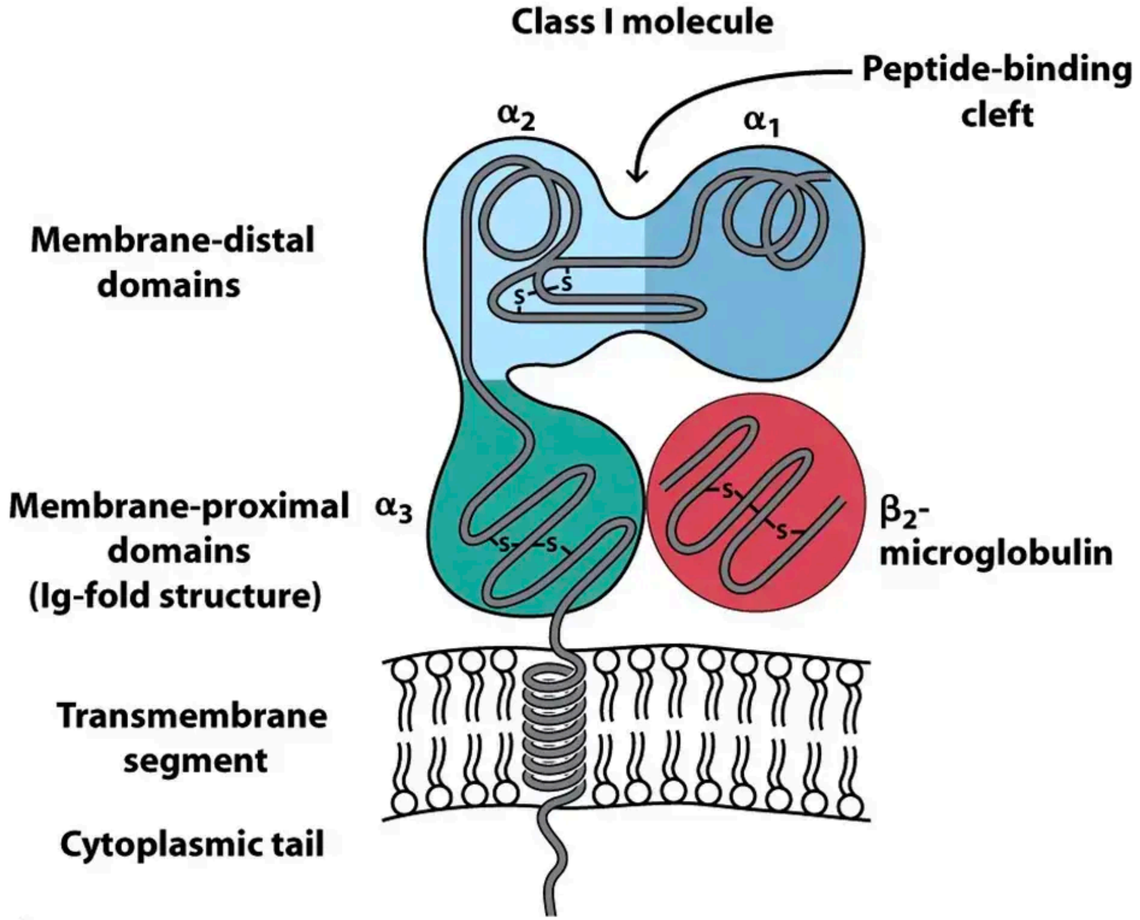
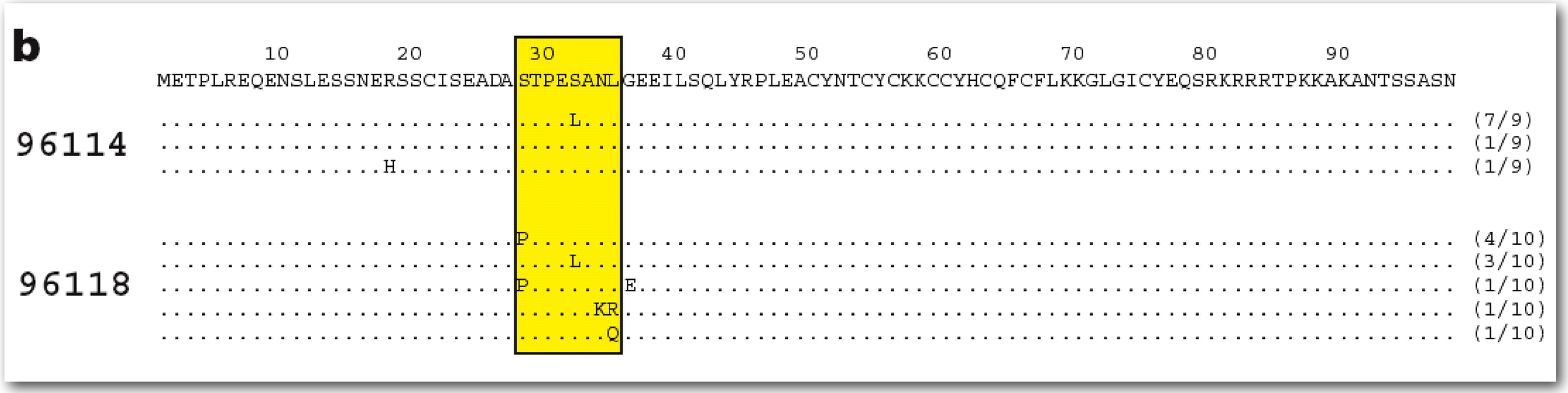


Figure 8-3
Kuby IMMUNOLOGY, Sixth Edition
© 2007 W. H. Freeman and Company

Rapid SIV sequence evolution in macaques in response to CTL-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
  - CTL *escape*

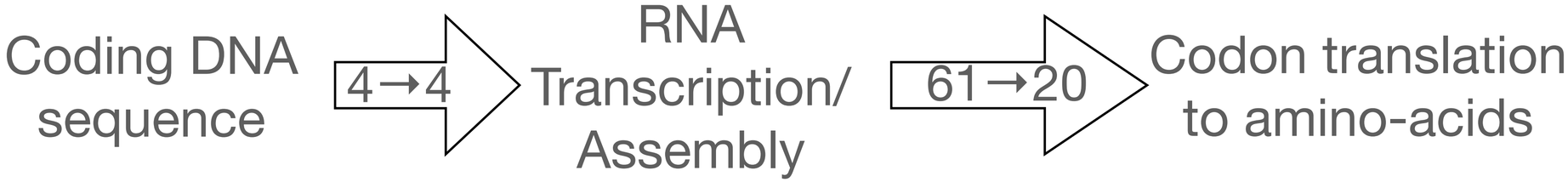


O'Connor et al (2002) Nat Med 8(5):493–499

Key drivers of adaptation in pathogens

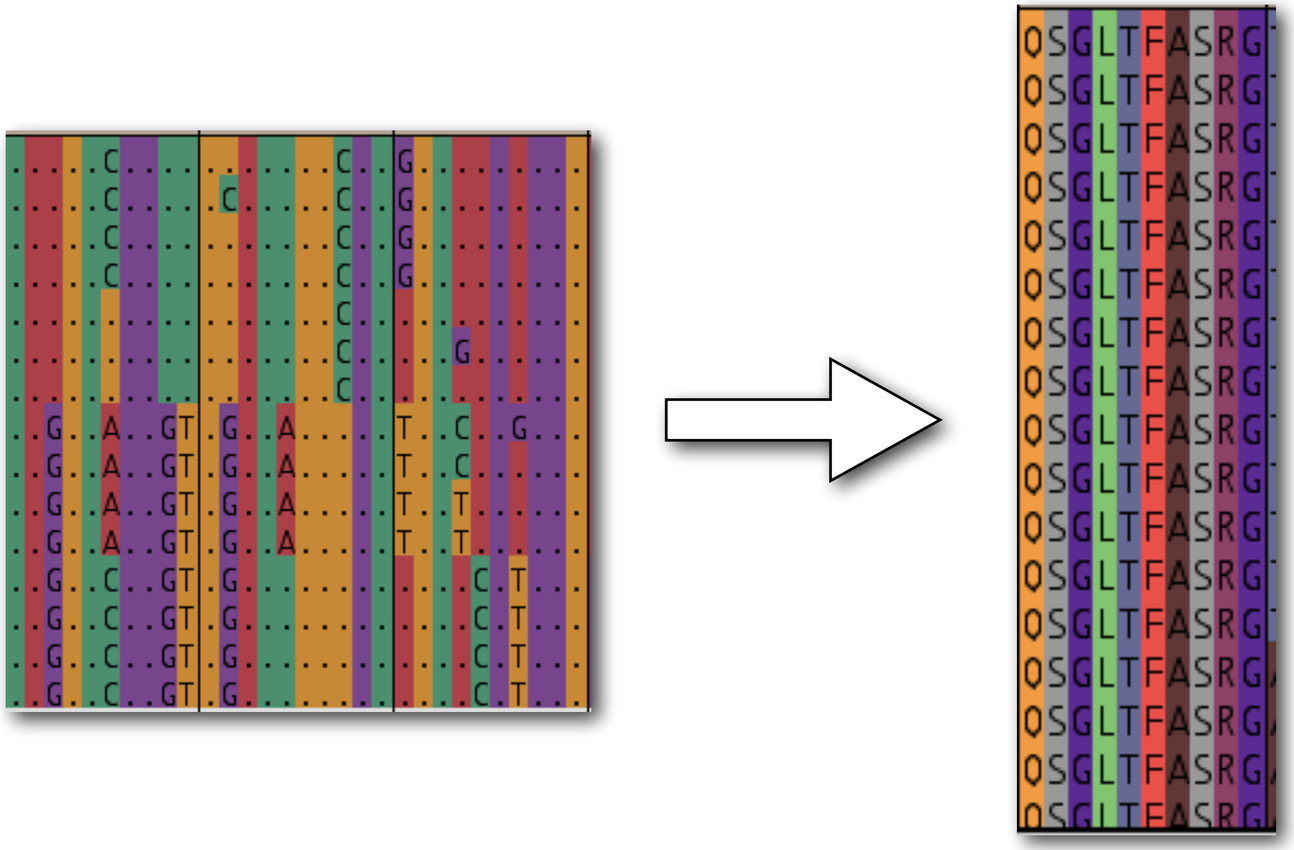
- Zoonoses and transmission to new hosts (both species and individuals)
- Immune selection (CTL, innate, antibody)
- Development of drug resistance
- Virulence/transmissibility
- Host/pathogen arms-races, e.g. host antiviral factors
- **Most** of the time, **most** of the viral genome is **conserved**

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

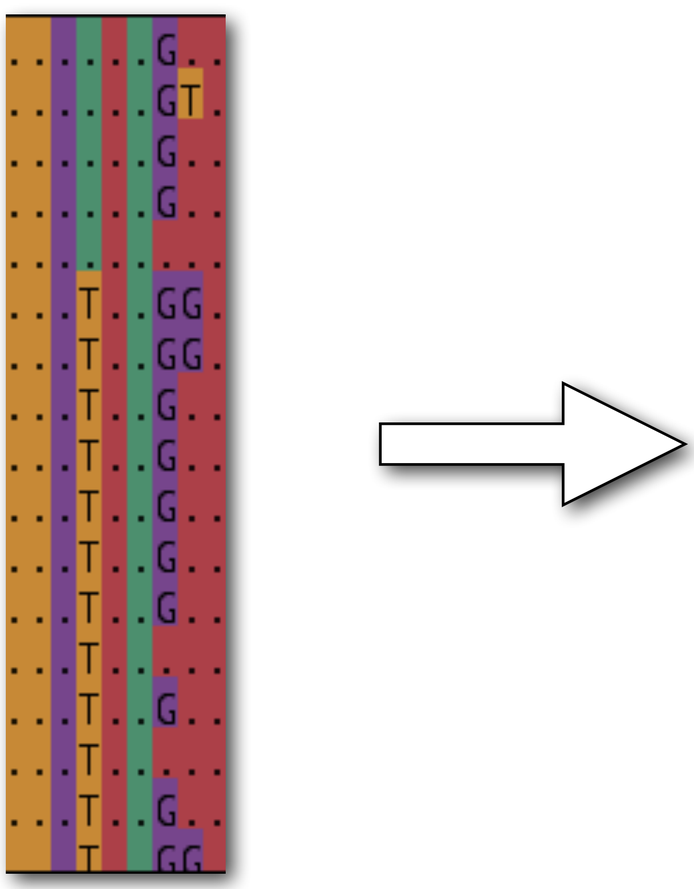
Conservation:
measles, rinderpest, and peste-de-petite ruminant viruses nucleoprotein.



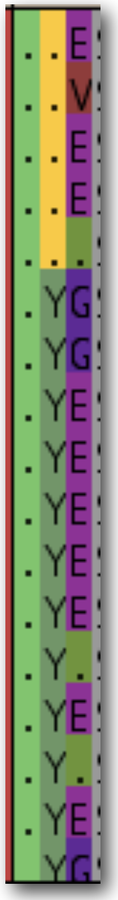
Nucleotides

Amino acids

Diversification:
an antigenic site in H3N2 IAV hemagglutinin



Nucleotides



Amino acids

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**) can serve as the neutral background evolutionary rate
- We can compare the **rate** of accumulation of non-synonymous substitutions (**dN**), which alter the protein sequence, to **dS** and use their ratio to classify the nature of the evolutionary process

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

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

What can the denominator proportions depend on?

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

	Site 1	**Site 2**	**Site 3**	**Site 4**	**Site 5**	**Site 6**
DNA	ACA	ATA	A**T**C	TTT	AA**T**	CAA
AA	T	I	**I**	F	N	Q
DNA	ACA	ATA	A**C**C	TTT	AA**C**	CAA
AA	T	I	**T**	F	N	Q

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

Universal genetic code

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(392 N : 134 S)`** ratio when mutations are generated and **fixed** completely at random
- Non-random distribution over codon positions
  - **All** second position mutations are non-synonymous
  - **Most** (but not all) synonymous mutations are confined to the third position

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)

	1	2	3
Start codon:	**G**	**A**	**A**
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**

8/3 non-synonymous sites (or 7/3 + 1/3 “stop” site)
1/3 synonymous sites

Nei-Gojobori dN/dS estimate (NG86)

- For each codon `C` we define **`ES(C)`** and **`EN(C)`** - the numbers of synonymous and non-synonymous *sites* of a codon
  - e.g., **`ES(GAA) = 1/3, EN(GAA) = 8/3.`**
- May also define them as fractions of substitutions that do not lead to stop codons,
  - e.g., **`ES(GAA) = 1/3, EN(GAA) = 7/3.`**
- 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 (the target of the original method), we average **`ES(C)`** and **`EN(C)`** at each site.
- **`EN/ES`** is thus the ***expected ratio of non-synonymous to synonymous substitutions counts 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)

>5,300 citations

NG86 example

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
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, we **expect** the ratio of non-synonymous to synonymous substitutions of **EN/ES ~ 4.05**

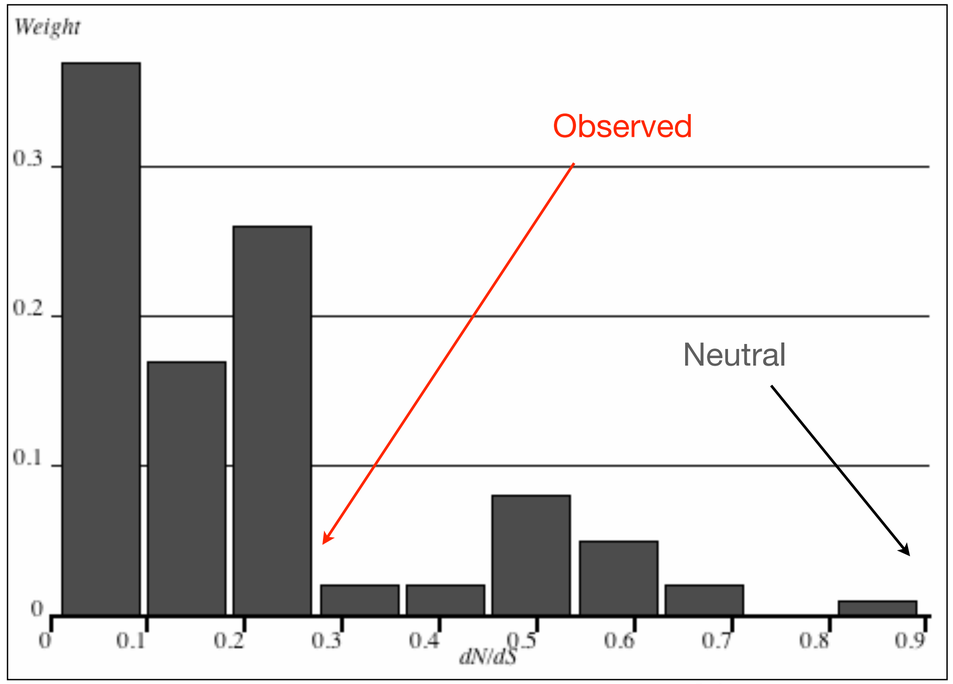
NG86 example

- 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 **`dN/dS = 1`** for neutrally evolving sequences/sites.
- Because there are **fewer** non-synonymous substitutions than expected under neutrality, we conclude that most non-synonymous mutations are **removed by natural selection**, i.e., the sequences are under **negative selection**
- **If there were more** non-synonymous substitutions than expected, we would conclude that many non-synonymous mutations are **fixed due to natural selection**, i.e., the sequences are under **positive selection**

NG86 example

- How reliable is the inference based on only **6** codons?
- Obtain sampling variance via bootstrap (or by limiting approximations)
- In this case, dN/dS is **significantly** less than `1.0 (p ~ 0.01)`

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
```

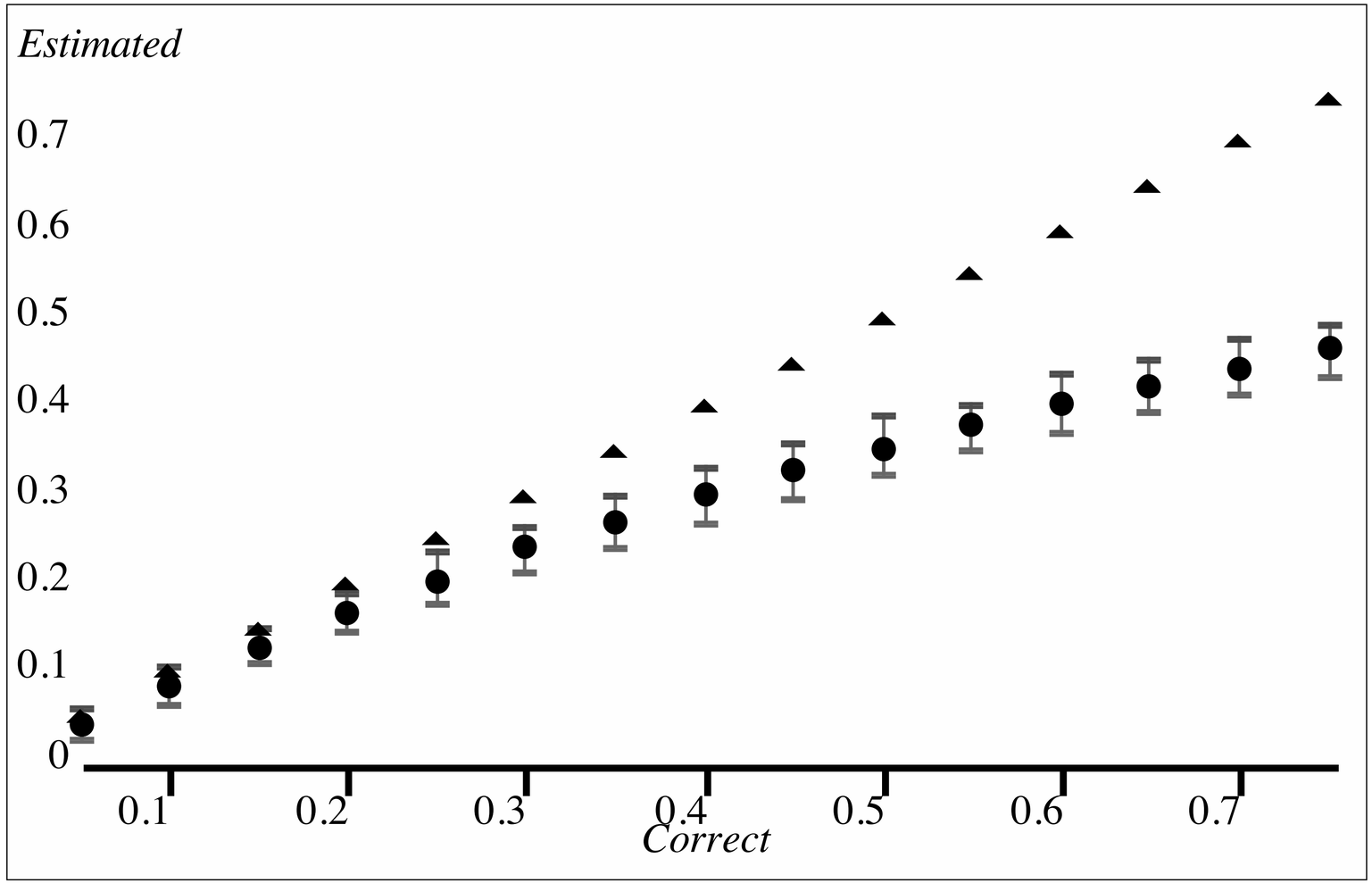
NG86 limitations: multiple substitutions

- How many synonymous and how many non-synonymous substitutions does it take to replace **`CCA`** with **`CAG`**?
- **Assume** the shortest path (minimum of 2 substitutions)
  - **`CCA (Proline)`** ⟹ **`CAA (Histidine)`** ⟹ **`CAG (Glutamine)`**
  - **`CCA (Proline)`** ⟹ **`CCG (Proline)`** ⟹ **`CAG (Glutamine)`**
- Average over the two possible paths: `0.5` synonymous and `1.5` non-synonymous substitutions.
- Intuitively, paths should **not** be equiprobable, e.g., because it should be more expensive to route evolution through (presumably) suboptimal intermediate amino-acids.

NG86 limitations: underestimation of substitution counts for higher divergence levels

- Simulated 100 replicates of 1000 nucleotide long sequences for various divergence levels (substitutions/site)
- 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

Simulated divergence vs that estimated by p-distance



NG86 limitations: ignoring phylogenies

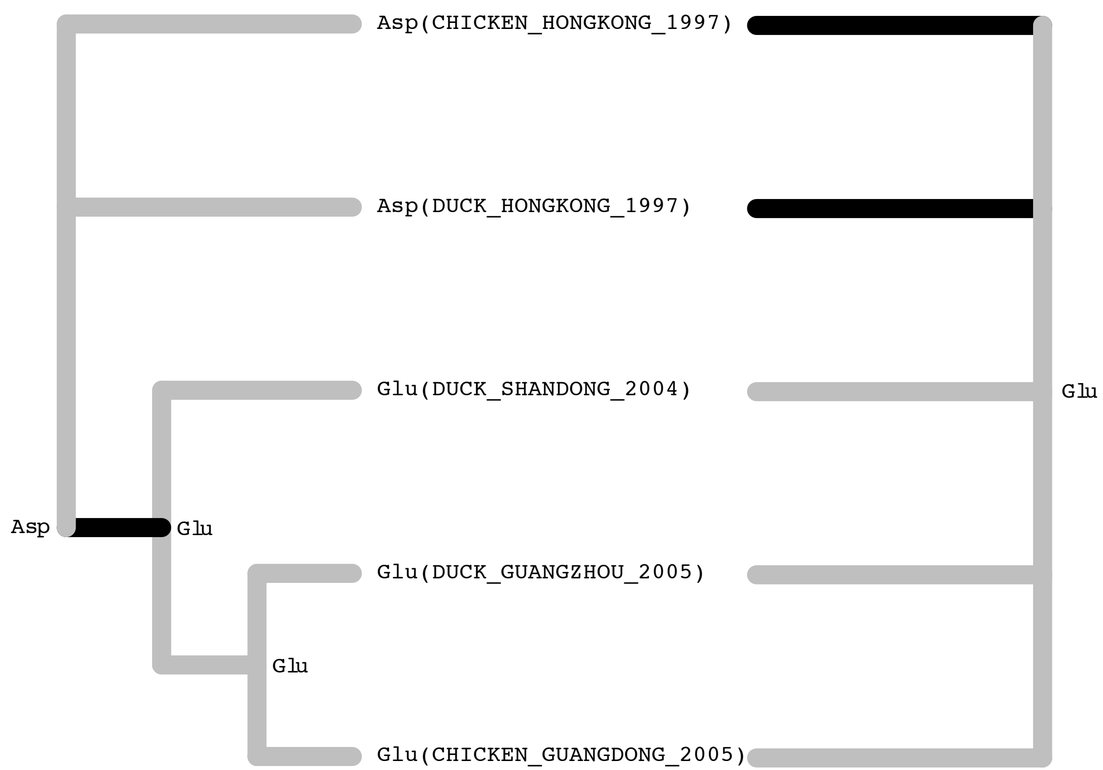


Fig. 1.1. Effect of phylogeny on estimating synonymous and nonsynonymous substitution counts in a dataset of Influenza A/H5N1 haemagglutinin sequences. Using the maximum likelihood tree on the left, the observed variation can be parsimoniously explained with one nonsynonymous substitution along the darker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

http://www.hyphy.org/resources/hyphybook2007.pdf

NG86 limitations: averaging across all sites in 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.
- Negatively selected sites are also of interest, because they point to functional constraint, and could be used to guide drug or vaccine design.
- Must develop methods that are able to disentangle the contributions of individual sites.

Suzuki-Gojobori (SG99): the penultimate extension of NG86

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 $p_e = ES/EN$
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 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)

>500 citations

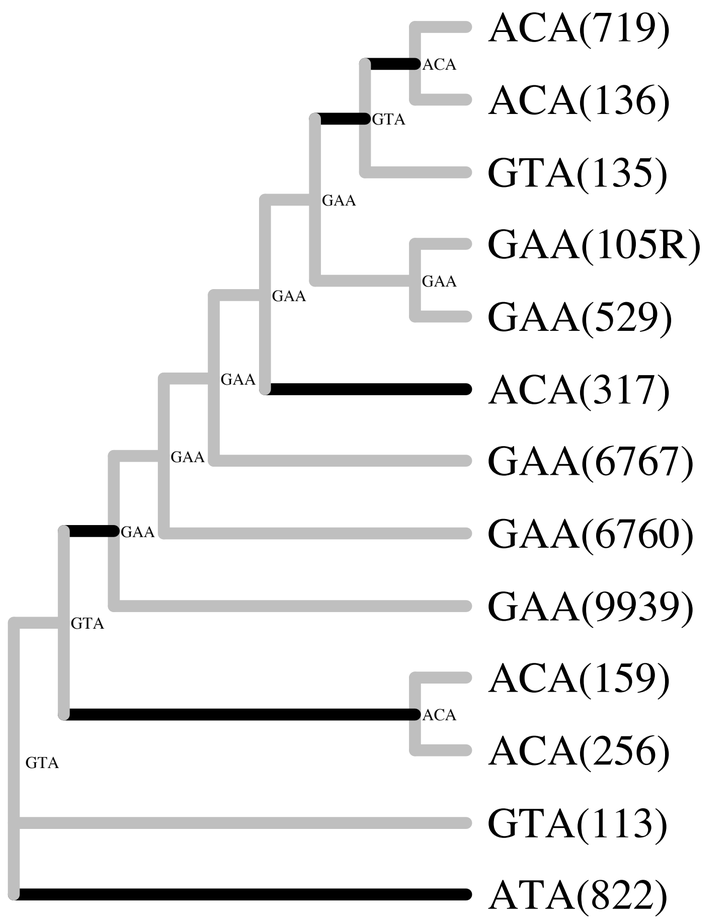


Fig. 1.6. An illustration of SLAC method, applied to a small HIV-1 envelope V3 loop alignment. Sequence names are shown in parentheses. Likelihood state ancestral reconstruction is shown at internal nodes. The parsimonious count yields 0 synonymous and 9 non-synonymous substitutions (highlighted with 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.

Any questions on the previous material?

- **We have covered:**
  - Brief background and examples of natural selection
  - **dN/dS** as a tool to measure the action of natural selection, explained using the first counting method for estimating dN/dS (Nei-Gojobori, 1986) and its extensions
  - **Next section:** Codon substitution models — the basis of modern (1998-) dN/dS estimation approaches

Codon-substitution models

- In 1994, first tractable mechanistic evolutionary models for codon sequences were proposed by **Muse and Gaut** (MG94), and, independently, by **Goldman and Yang** (GY94) [in the same issue of MBE, back to back]
- Markov models of codon substitution provide a powerful framework for **estimating substitution rates** from coding sequence data, as they
  - *encode our mechanistic understanding of the evolutionary process,*
  - *enable one to compute the phylogenetic likelihood,*
  - *permit hypothesis testing or Bayesian inference,*
  - *systematically account for confounding processes (unequal base frequencies, nucleotide substitution biases, etc.),*
  - *afford many opportunities for extension and refinement (still happening today).*

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)

~1000 citations

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

~2250 citations

Rate matrix for an MG-style codon model

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

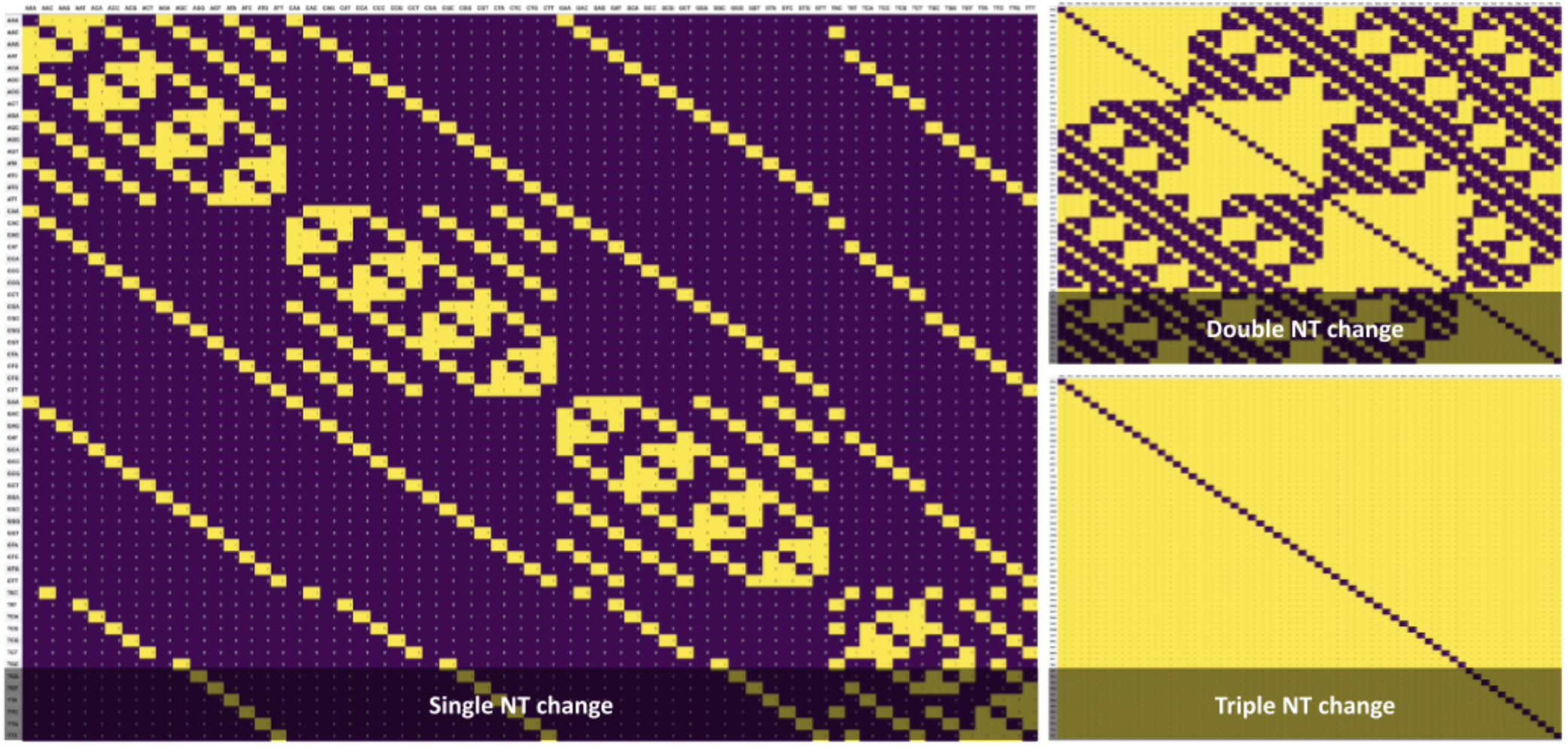
X,Y = AAA...TTT (excluding stop codons),
π_t - frequency of the target nucleotide.

Example substitutions:
AAC→AAT (one step, synonymous - Asparagine)
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

Illuminating the darkness in molecular evolution



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),`** same as with standard nucleotide models, e.g. HKY85 or GTR.

- 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, even though they are relatively straightforward extensions of 4x4 nucleotide models

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`**).
  - This can be further relaxed with models supporting multiple nucleotide changes.
- 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 naive NG86 parsimony approach.

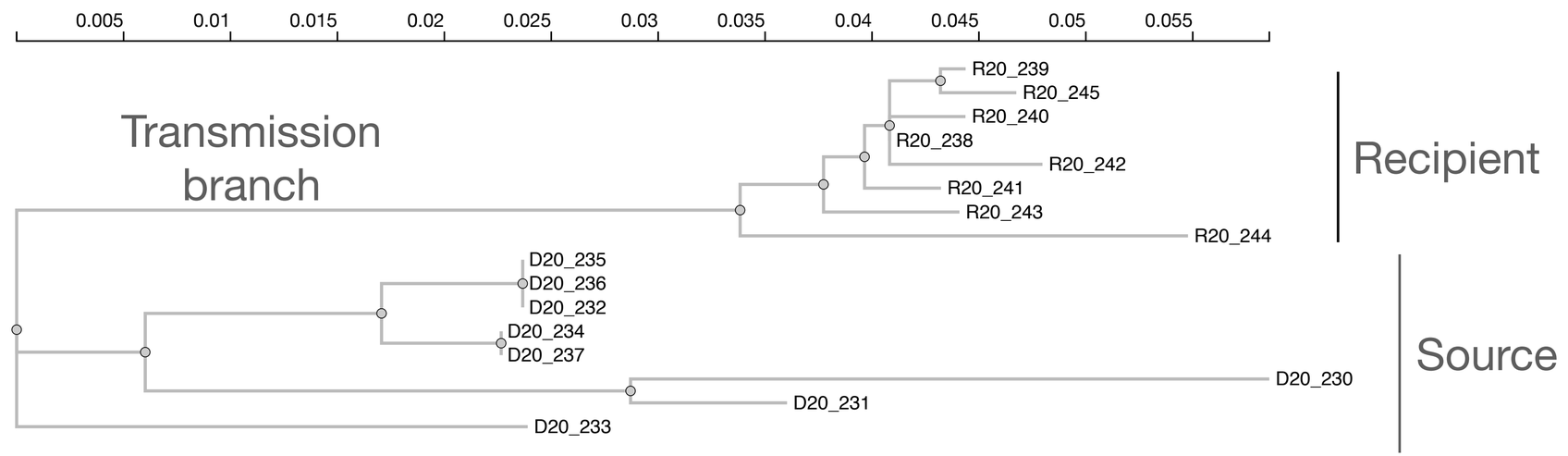
Alignment-wide estimates

- 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)

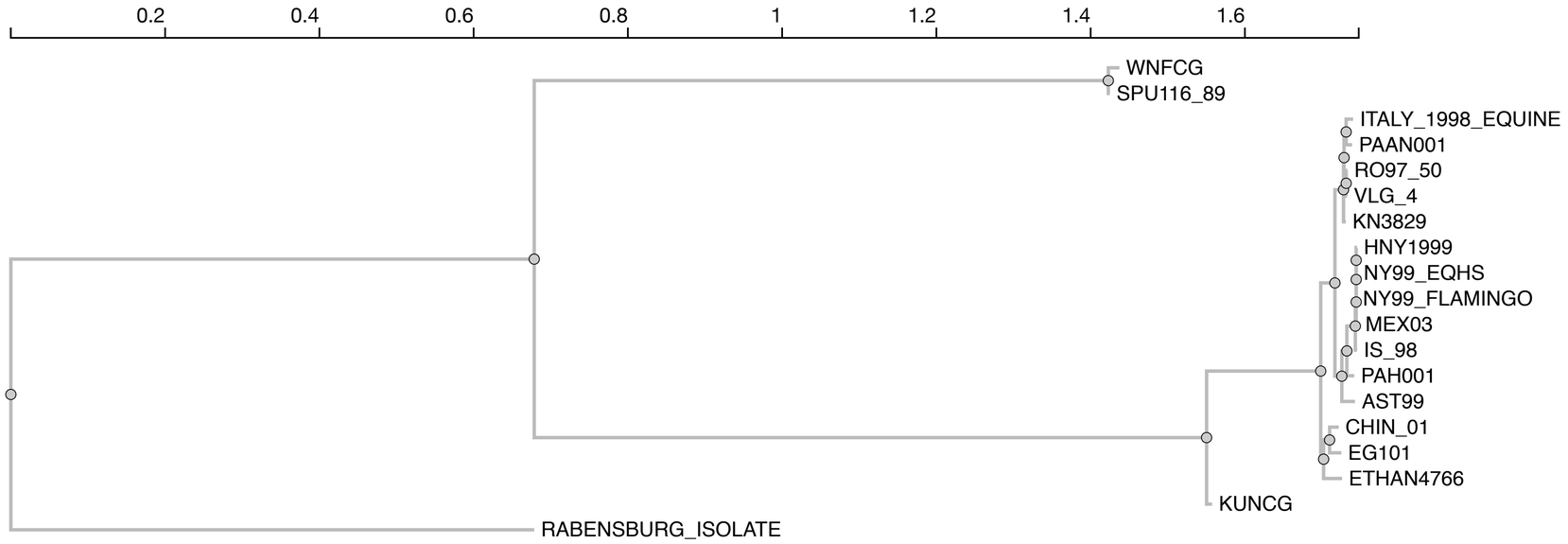
Three example datasets

- **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)
- **SARS-CoV-2 Spike**
  - Full length spike sequences chosen to represent viral diversity (circa mid 2021)
  - Good example for analyzing selection in population samples with many “dead-end” intra-host variants

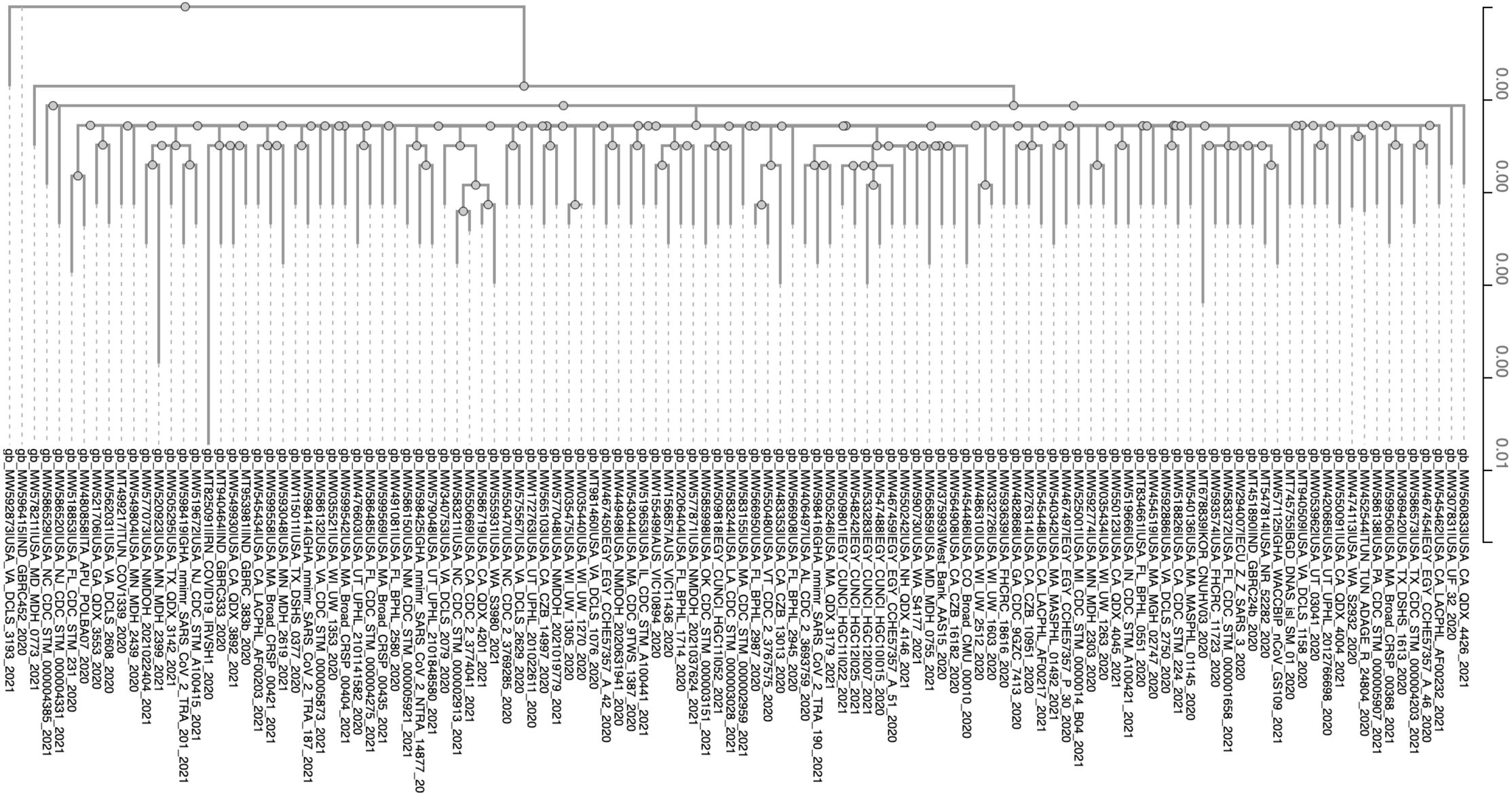
HIV-1 *env*



WN NS3



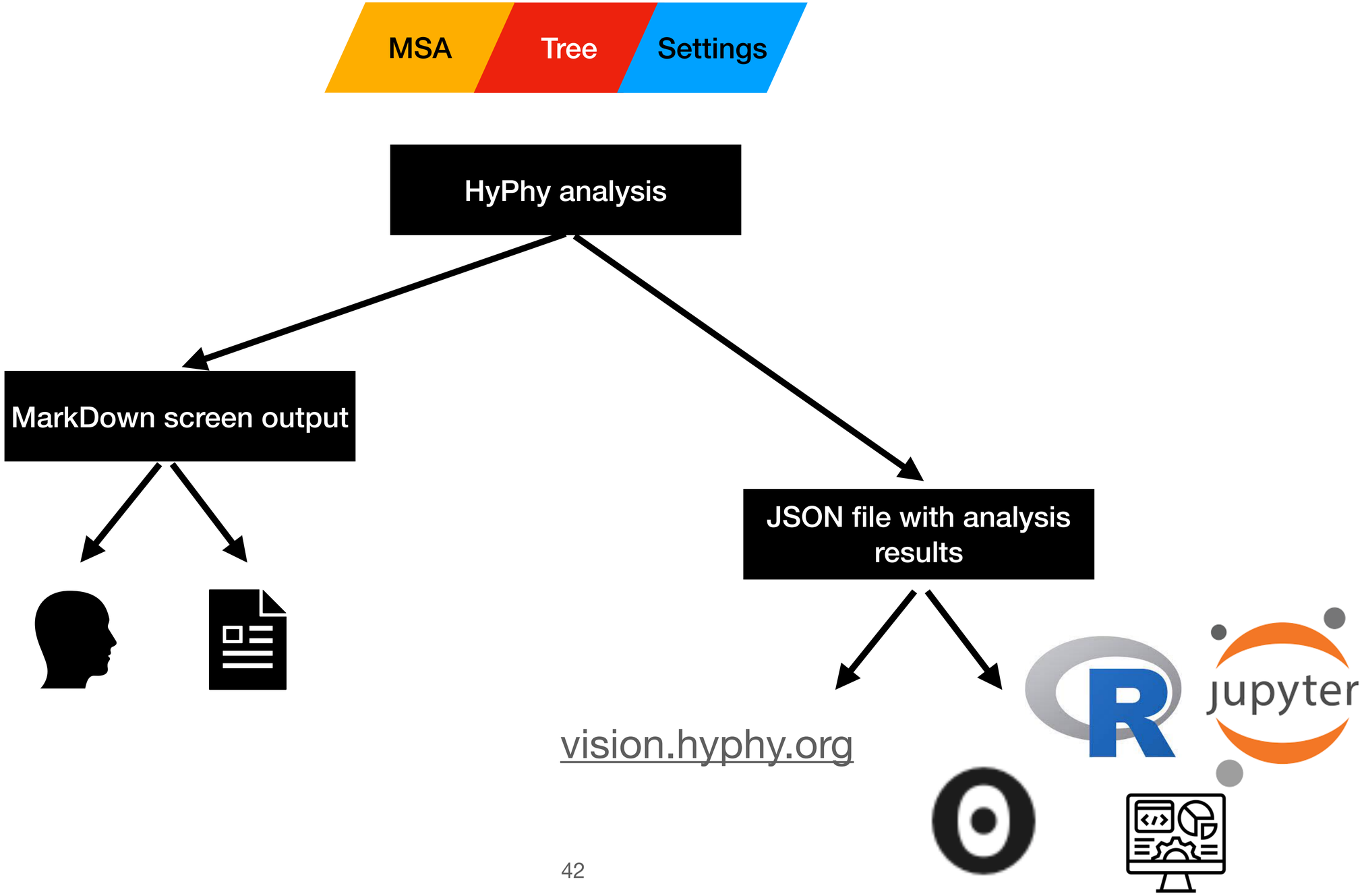
SARS-CoV-2 spike



Information content of the alignments

	WNV NS3	**HIV-1 *env***	**SARS-CoV-2 spike**
Sequences	19	16	118
Codons	619	288	1273
Tree Length *MG94 model, subs/site*	0.67	0.20	0.134

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


MSA
Tree
Settings
HyPhy analysis
MarkDown screen output
JSON file with analysis results
vision.hyphy.org
R
jupyter

$hyphy ~/hyphy-analyses/FitMG94/FitMG94.bf --help

```
Available analysis command line options
---------------------------------------
Use --option VALUE syntax to invoke
If a [required] option is not provided on the command line, the analysis will prompt for its value
[conditionally required] options may or not be required based on the values of other options

rooted
   Accept rooted trees
   default value: No

code
   Which genetic code should be used
   default value: Universal

alignment [required]
   An in-frame codon alignment in one of the formats supported by HyPhy

tree [conditionally required]
   A phylogenetic tree
   applies to: Please select a tree file for the data:

type
   Model type: global (single dN/dS for all branches) or local (separate dN/dS)
   default value: terms.global [computed at run time]
   applies to: Model Type

frequencies
   Equilibrium frequency estimator
   default value: CF3x4

lrt
   Perform LRT to test which for dN/dS == 1 (global model only)
   default value: No

output
   Write the resulting JSON to this file (default is to save to the same path as the alignment file + 'MG94.json')
   default value: fitter.codon_data_info[terms.json.json] [computed at run time]

save-fit
   Save MG94 model fit to this file (default is not to save)
   default value: /dev/null
```


https://github.com/veg/hyphy-analyses/

$hyphy ~/hyphy-analyses/FitMG94/FitMG94.bf --lrt Yes --alignment data/WestNileVirus_NS3.fas

```
Analysis Description
--------------------
Fit an MG94xREV model with several selectable options frequency
estimator and report the fit results including dN/dS ratios, and
synonymous and non-synonymous branch lengths. v0.2 adds LRT test for
dN/dS != 1

- __Requirements__: in-frame codon alignment and a phylogenetic tree

- __Written by__: Sergei L Kosakovsky Pond

- __Contact Information__: spond@temple.edu

- __Analysis Version__: 0.2

rooted: No

>code -> Universal
>Loaded a multiple sequence alignment with **19** sequences, **619** codons, and **1** partitions from `/Users/sergei/Dropbox/Talks/VEME-2021/data/WestNileVirus_NS3.fas`

>type -> global

>frequencies -> CF3x4

>lrt -> Yes


### Obtaining branch lengths and nucleotide substitution biases under the nucleotide GTR model

>kill-zero-lengths -> Yes
```


https://github.com/veg/hyphy-analyses/

```
### Deleted 2 zero-length internal branches: `Node1, Node2`
* Log(L) = -7745.48, AIC-c = 15577.06 (43 estimated parameters)
* 1 partition. Total tree length by partition (subs/site)  0.672

### Fitting Standard MG94
* Log(L) = -6413.46, AIC-c = 12923.31 (48 estimated parameters)
* non-synonymous/synonymous rate ratio =   0.0086 (95% profile CI   0.0068-  0.0106)

### Running the likelihood ratio tests for dN/dS=1

>Testing _non-synonymous/synonymous rate ratio_ == 1

Likelihood ratio test for _non-synonymous/synonymous rate ratio == 1_, **p =   0.0000**.

### **Synonymous tree**
(HNY1999:0.001081256453028512,NY99_EQHS:0.00106649816840513,NY99_FLAMINGO:0,(((((((RABENSBURG_ISOLATE:1.029888927913287,
(WNFCG:0.009422232735390165,SPU116_89:0.006497252142983255)Node11:0.497317793054674)Node9:0.5663849238543818,KUNCG:0.08579616334765416)Node8:0.06828112532356265,
(ETHAN4766:0.02340594686858863,(CHIN_01:0.0118388155978261,EG101:0.01478394721830331)Node17:0.007544440407300843)Node15:0.003379513965204242)Node7:0.01817281158535988
(((ITALY_1998_EQUINE:0.00887303413005913,PAAN001:0.007729652361011649)Node22:0.002644276626382983,
(RO97_50:0.001612279279163659,VLG_4:0.001062750677413841)Node25:0.002742110200748272)Node21:0.0007107517172592428,KN3829:0.003002207398871611)Node20:0.010777674817076
de6:0.009122506962674462,AST99:0.01648683046482152)Node5:0.006360440965550713,PAH001:0.009764197227613279)Node4:0.01061342075646687,IS_98:0.00219787558408969)Node3:0.
4759833708216,MEX03:0.003214366196457822)

### **Non-synonymous tree**
(HNY1999:2.022605013917642e-05,NY99_EQHS:1.994998075348657e-05,NY99_FLAMINGO:0,(((((((RABENSBURG_ISOLATE:0.01926516603476631,
(WNFCG:0.0001762528687761419,SPU116_89:0.0001215380007608477)Node11:0.009302857420415459)Node9:0.01059483144434955,KUNCG:0.001604908342288399)Node8:0.0012772709568442
(ETHAN4766:0.0004378330909313789,
(CHIN_01:0.0002214576173852258,EG101:0.0002765494317789341)Node17:0.0001411267692532066)Node15:6.321739742205773e-05)Node7:0.0003399417383968234,
(((ITALY_1998_EQUINE:0.0001659795256699084,PAAN001:0.0001445913555237784)Node22:4.946400225152157e-05,
(RO97_50:3.01593599470718e-05,VLG_4:1.987987995503712e-05)Node25:5.129408314941099e-05)Node21:1.329536561795865e-05,KN3829:5.615947743729441e-05)Node20:0.00020160785
319)Node6:0.0001706461799189597,AST99:0.0003084036711952811)Node5:0.0001189788024073028,PAH001:0.0001826496777349691)Node4:0.0001985353056318952,IS_98:4.1113596722550
5)Node3:1.916922088107983e-05,MEX03:6.012813303738528e-05)
**Combined tree**
(HNY1999:0.001101482503167688,NY99_EQHS:0.001086448149158617,NY99_FLAMINGO:0,(((((((RABENSBURG_ISOLATE:1.049154093948055,
(WNFCG:0.009598485604166309,SPU116_89:0.006618790143744096)Node11:0.5066206504750898)Node9:0.5769797552987309,KUNCG:0.08740107168994253)Node8:0.06955839628040685,
(ETHAN4766:0.02384377995952001,(CHIN_01:0.01206027321521132,EG101:0.01506049665008225)Node17:0.007685567176554043)Node15:0.003442731362626298)Node7:0.0185127533237567
(((ITALY_1998_EQUINE:0.009039013655729039,PAAN001:0.007874243716535429)Node22:0.002693740628634505,
(RO97_50:0.001642438675158366,VLG_4:0.001082630557368879)Node25:0.002793404283897683)Node21:0.0007240470828772018,KN3829:0.003058366876308903)Node20:0.010979282669332
de6:0.009293153142593342,AST99:0.01679523413601681)Node5:0.006479419767958014,PAH001:0.009946846905348252)Node4:0.01081195606209877,IS_98:0.002238989180812242)Node3:0.
3929054589295,MEX03:0.003274494329495206)

### Writing detailed analysis report to `/Users/sergei/Dropbox/Talks/VEME-current/data/WestNileVirus_NS3.fas.FITTER.json'
```


https://github.com/veg/hyphy-analyses/

results/WestNileVirus_NS3.fna.FITTER.json


```
    "Standard MG94":{
      "AIC-c":12923.31190785199,
      "Confidence Intervals":{
        "non-synonymous/synonymous rate ratio":{
          "LB":0.006844594453196991,
          "UB":0.01055858911944712
         }
       },
      "Log Likelihood":-6413.455134253863,
      "Rate Distributions":{
        "Substitution rate from nucleotide A to nucleotide C":0.2433728512340663,
        "Substitution rate from nucleotide A to nucleotide G":1,
        "Substitution rate from nucleotide A to nucleotide T":0.3060223183303938,
        "Substitution rate from nucleotide C to nucleotide G":0.02087667707983453,
        "Substitution rate from nucleotide C to nucleotide T":1.979133076079397,
        "Substitution rate from nucleotide G to nucleotide T":0.2305847598841325,
        "non-synonymous/synonymous rate ratio":0.008575064843083209
       },
      "display order":1,
      "estimated parameters":48
     }
   },
  "input":{
    "file name":"/Users/sergei/Dropbox/Talks/VEME-current/data/WestNileVirus_NS3.fas",
    "number of sequences":19,
    "number of sites":619,
    "partition count":1,
    "trees":{
      "0":"(HNY1999,NY99_EQHS,NY99_FLAMINGO,(((((((RABENSBURG_ISOLATE,(WNFCG,SPU116_89)Node11)Node9,KUNCG)Node8,(ETHAN4766,
          (CHIN_01,EG101)Node17)Node15)Node7,(((ITALY_1998_EQUINE,PAAN001)Node22,
          (RO97_50,VLG_4)Node25)Node21,KN3829)Node20)Node6,AST99)Node5,PAH001)Node4,IS_98)Node3,MEX03)"
     }
   },
  "test results":{
    "non-synonymous/synonymous rate ratio":{
      "LRT":2512.58476730381,
      "p-value":0
     }
   },
```


https://github.com/veg/hyphy-analyses/

WNV NS3

Model	Log L	# p	dN/dS	LRT	p-value
Null	-7745.48	49	1		
Alternative	-6413.5	50	0.009 [0.007-0.011]	2512.6	~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	1.122 [0.94-1.33]	0.33	~0.6

Not significantly different from neutral

SARS-CoV-2 spike

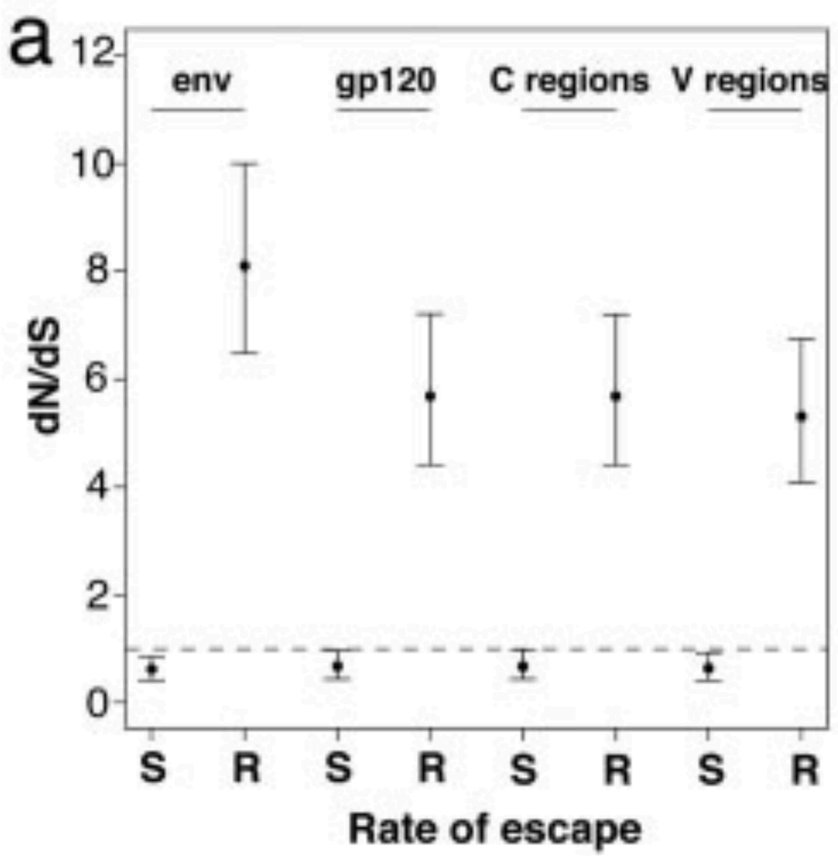
Model	Log L	# p	dN/dS	LRT	p-value
Null	-9311.0	176	1		
Alternative	-9292.0	177	0.54 [0.48-0.61]	37.94	~0

Very strongly conserved

Mean gene-wide dN/dS estimates

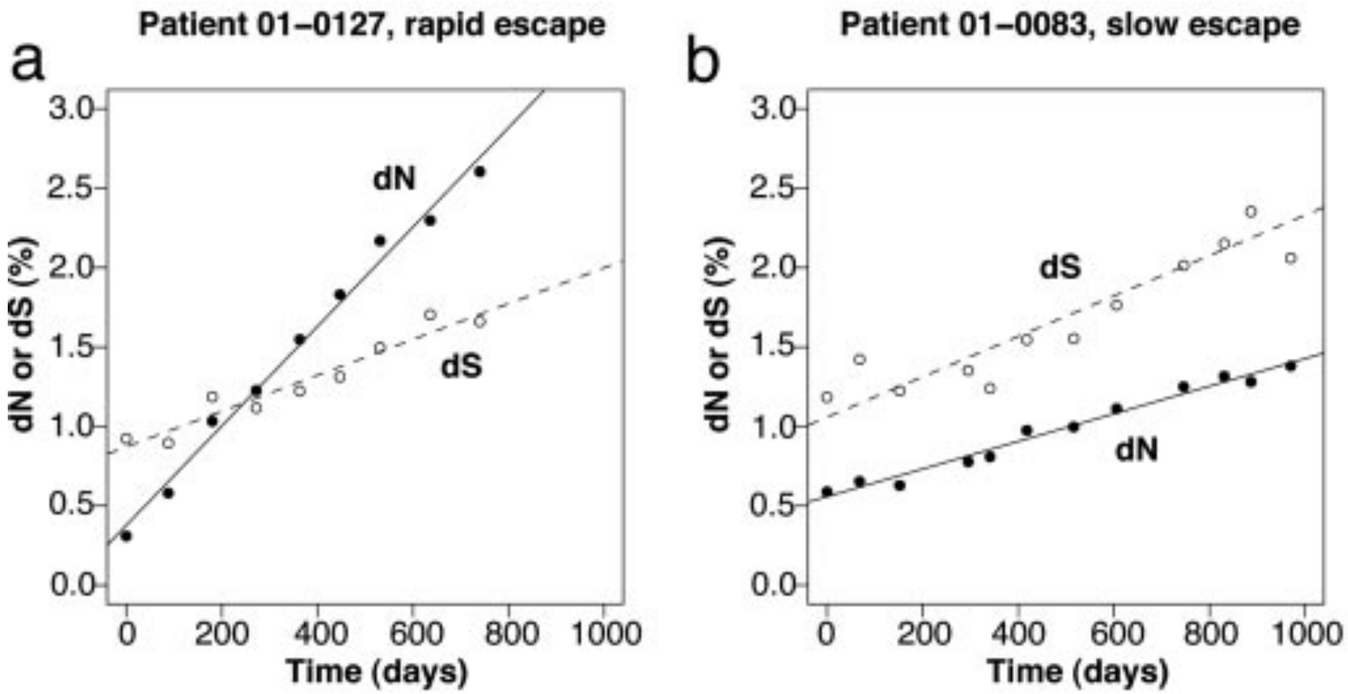
- Are not the way to go, **except** when you have very small (2-3 sequence) datasets
- **For example:**
  - The humoral arm of the immune system mounts a potent defense against viral infections
  - Existing successful vaccines are based on raising a neutralizing antibody (nAb) response to the pathogen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

An example of mean dN/dS utility



Slow (S) and Rapid (R)

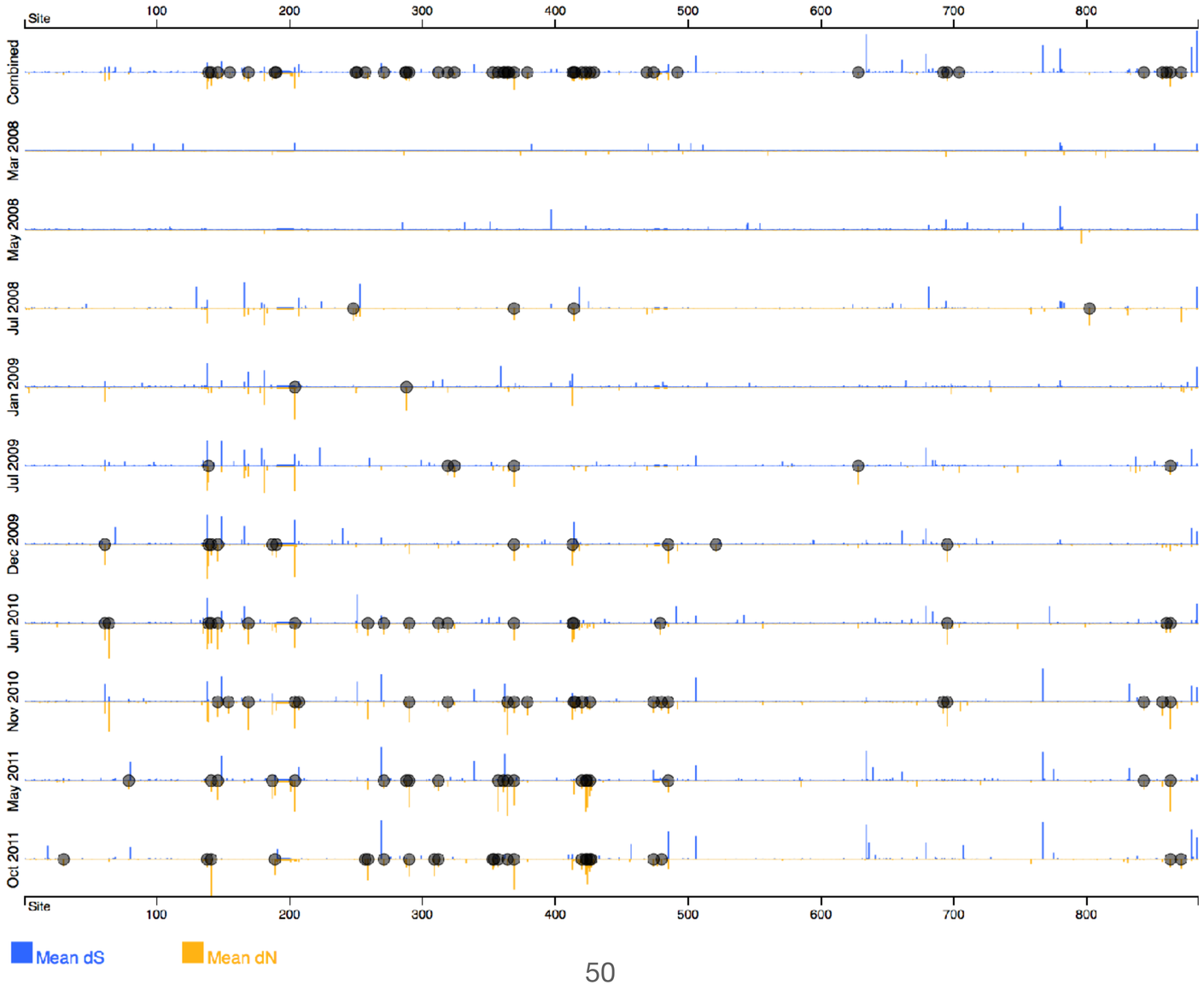
The extent of immune selection pressure drives intra-host evolution in HIV-1



Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection

Simon D. W. Frost*†, Terri Wrin‡, Davey M. Smith*§, Sergei L. Kosakovsky Pond*, Yang Liu‡, Ellen Paxinos‡, Colombe Chappey‡, Justin Galovich‡, Jeff Beauchaine‡, Christos J. Petropoulos‡, Susan J. Little*, and Douglas D. Richman*§

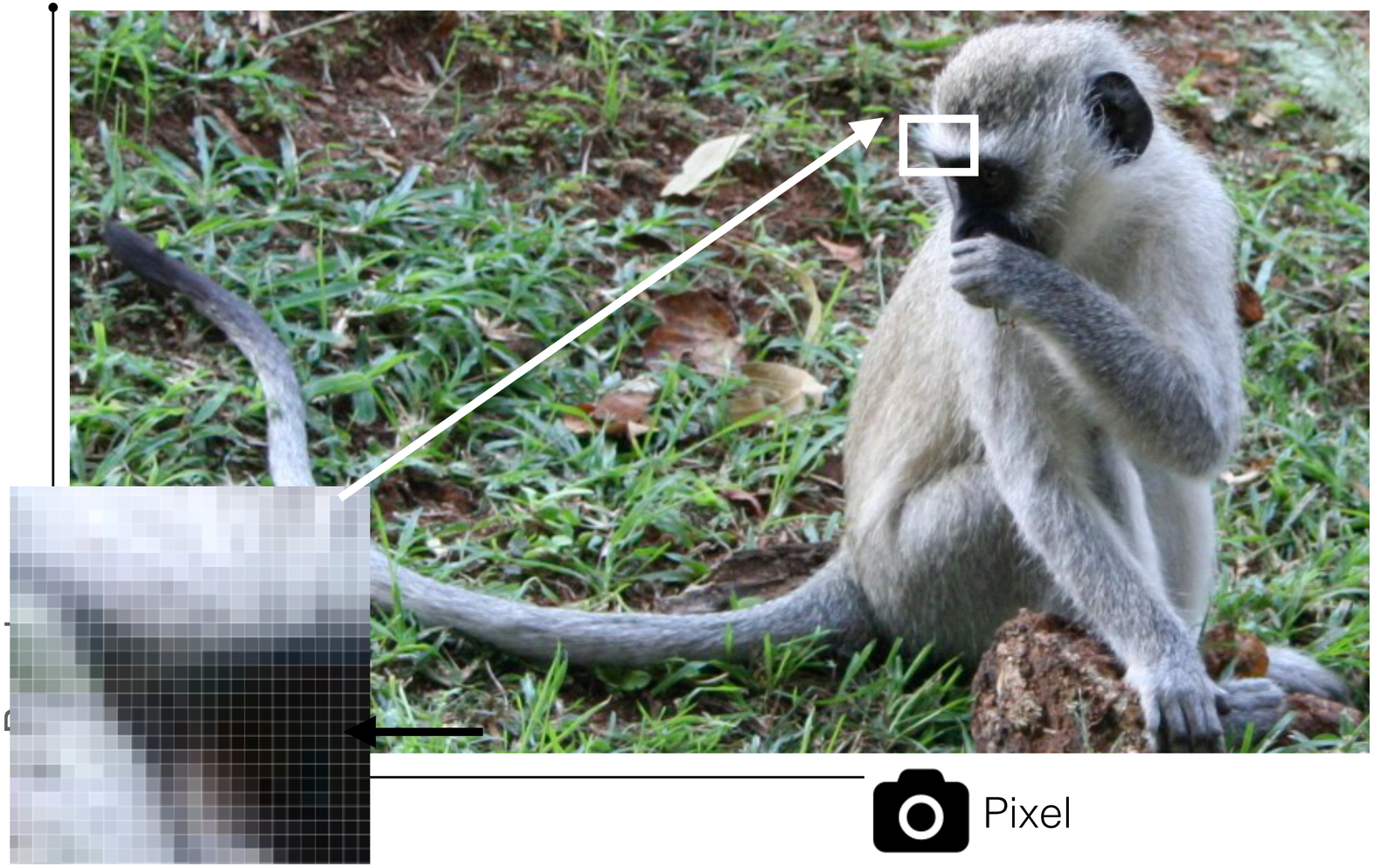
But upon closer look, this pattern is highly variable both across a gene and through time.



PLoS Pathog 12(1): e1005369. Patient 064

Break

Selection inference as image processing



Pixel

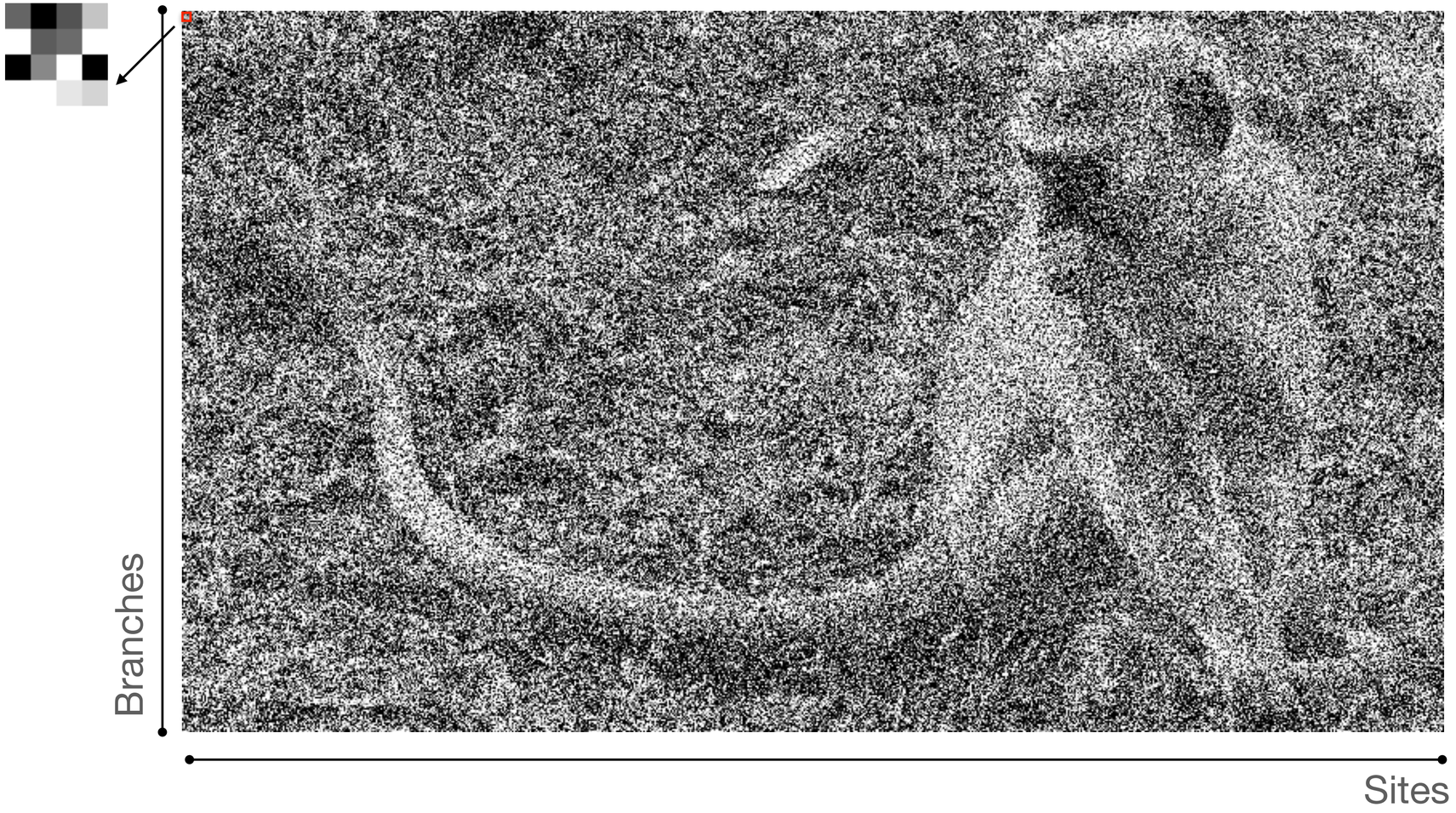
Evolutionary process along a single branch at a single site

Forget about the color

Intensity/brightness Color

Evolutionary rate (dN/dS) Type of evolutionary/function/property change

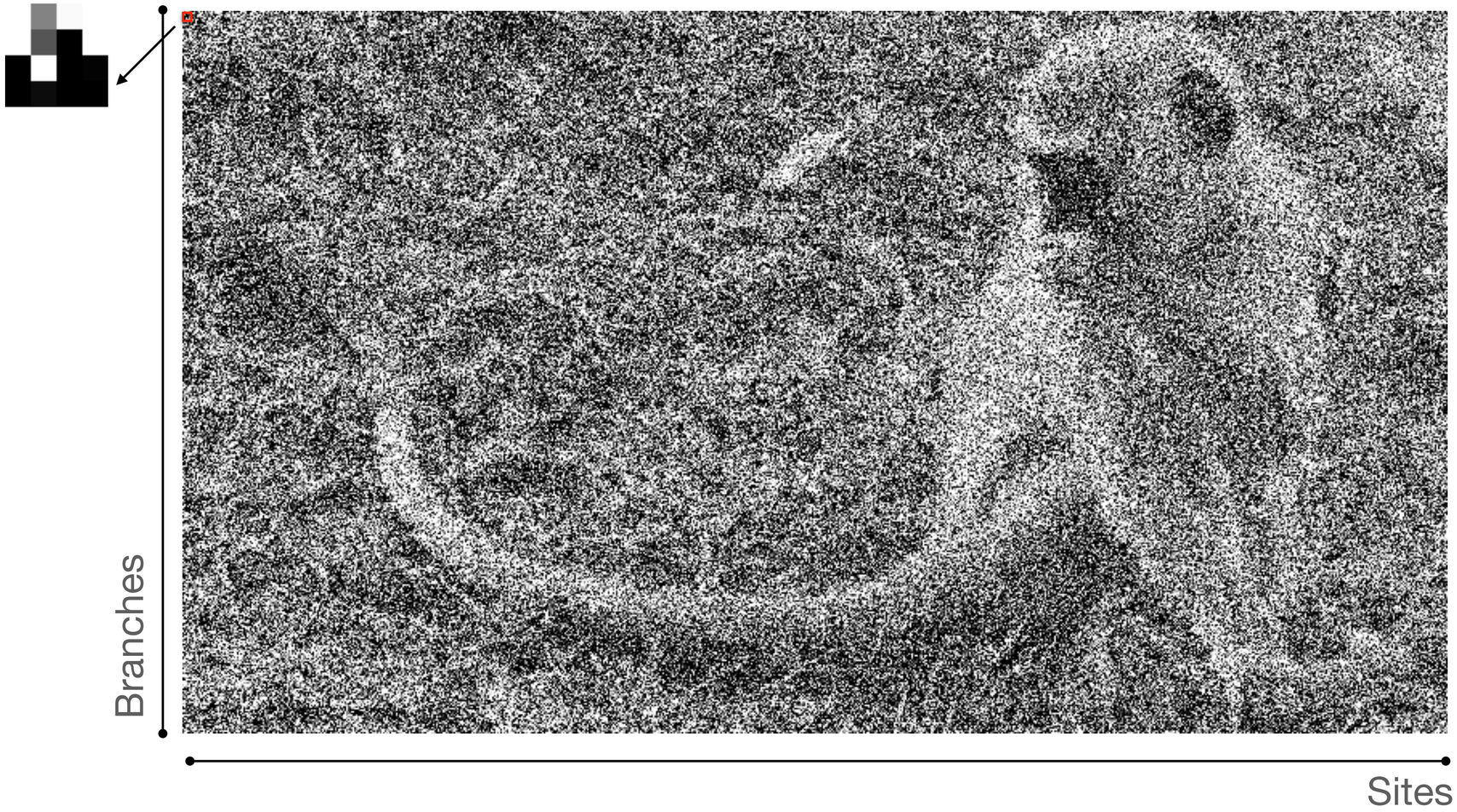
Evolution is largely unobserved and noisy



Visual noise

Saturation, missing data, model misspecification, sampling variation

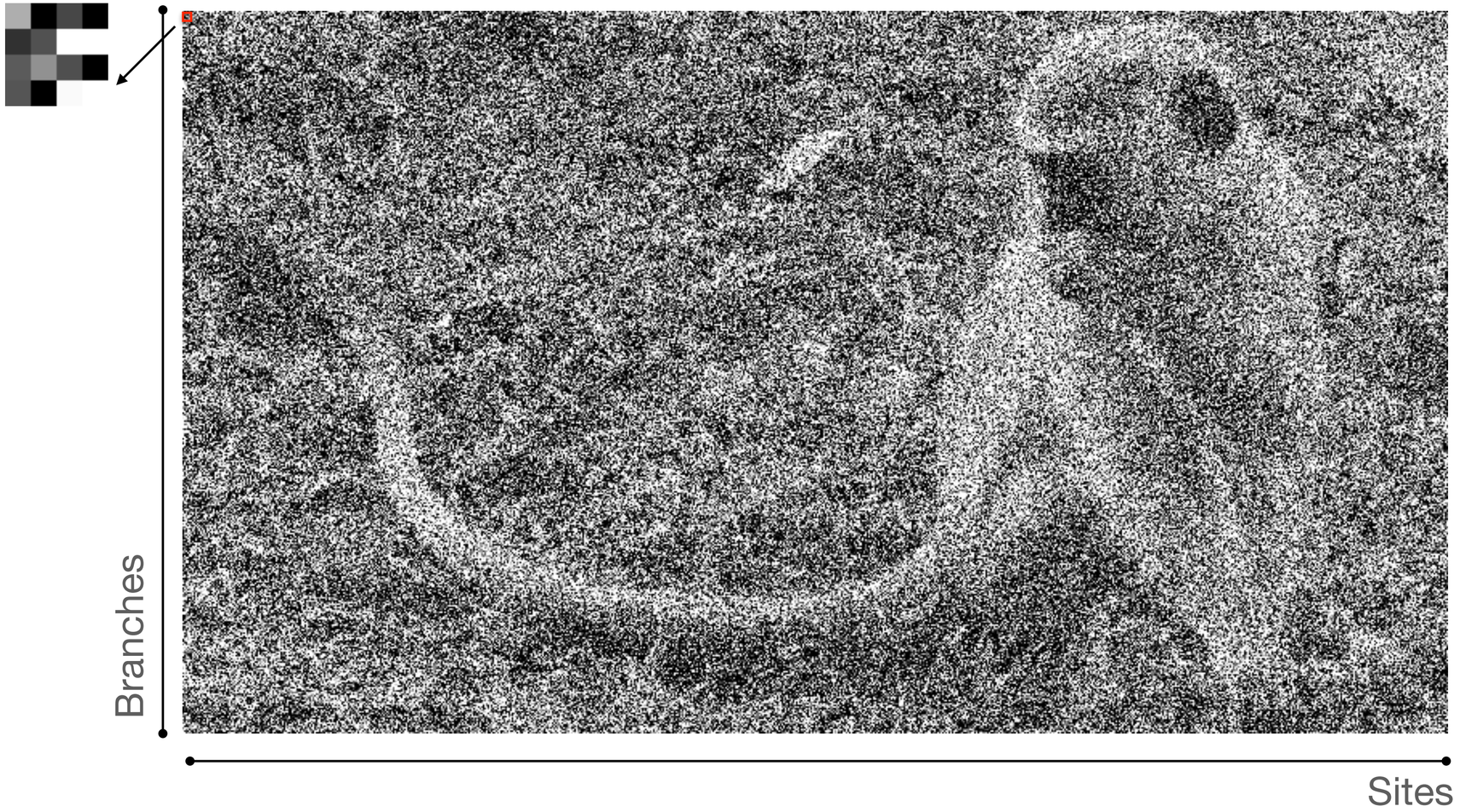
Evolution is largely unobserved and noisy (another replicate)



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

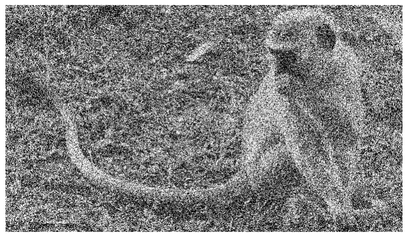
🌲 Statistical inference draws power from sample (and effect) size, need to aggregate data to gain power

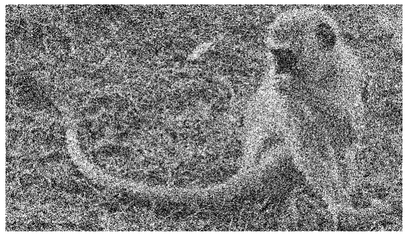
Gene-wide selection (mean dN/dS)

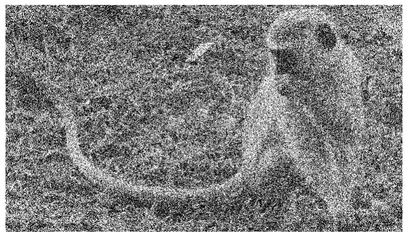
Is the average color sufficiently “bright”

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

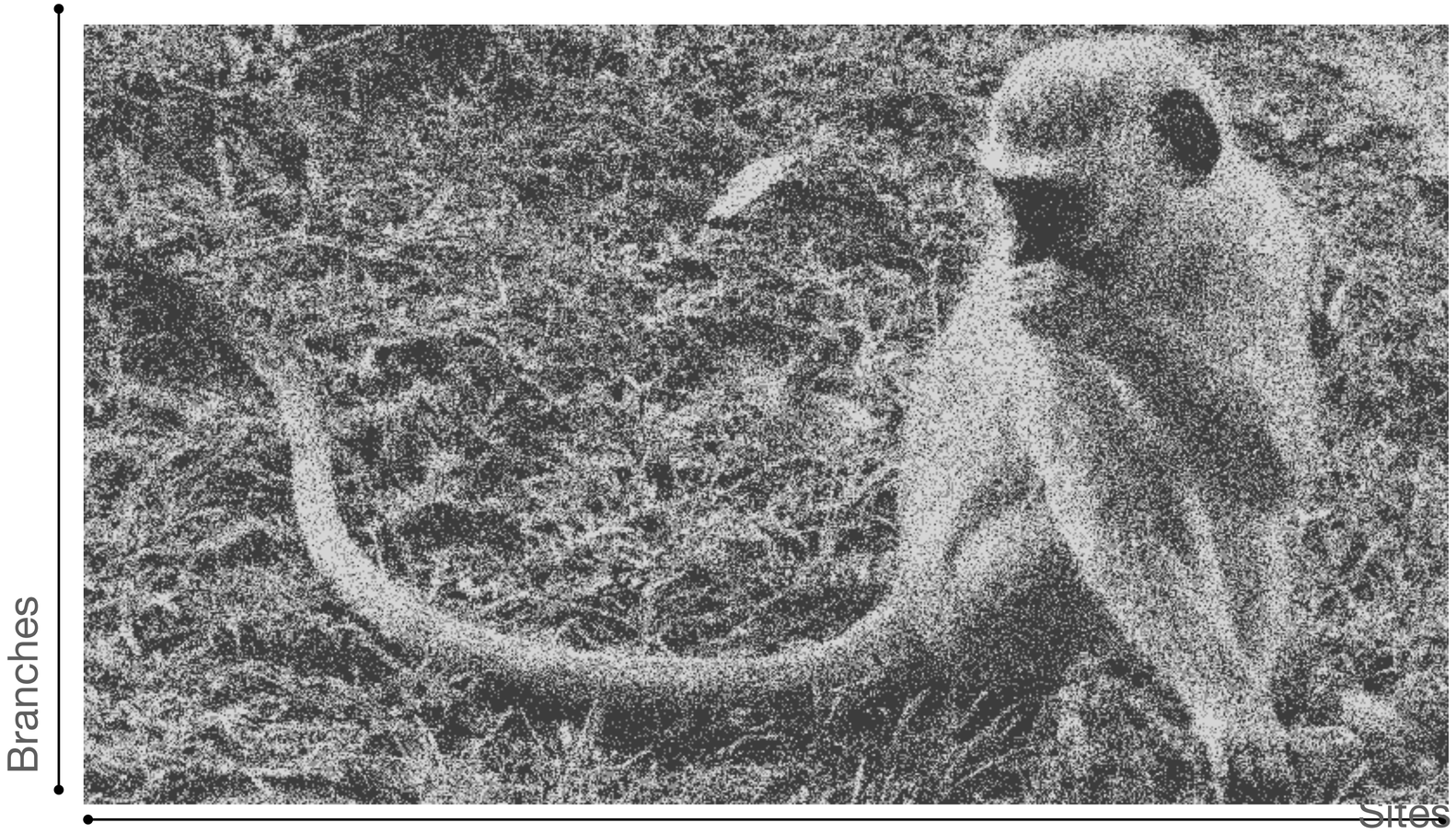
- Simple
  - single rate parameter
  - relatively compute-light
- Very robust to local variation
- Sample size ~ sites x branches
- Very low power
  - most genes are **on average** conserved
- No resolution
  - if selection occurred, how much of the gene was involved, and when did it happen
- Rate variation 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 K (=3) colors, chosen adaptively, e.g. to minimize perceptual differences

[BUSTED]: each branch-site combination is a drawn from a K-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*

```
hyphy busted --srv No --alignment
data/HIV-sets.nex --starting-
points 5
```

Produces *HIV-sets.nex.BUSTED.json* file
View in http://vision.hyphy.org/BUSTED or
https://observablehq.com/@spond/busted

Based on the likelihood ratio test, there **is** evidence of *episodic diversifying selection* in this dataset (p=0.000).

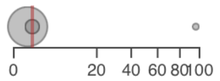
BUSTED analysis (v4.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/HIV-sets.fas using HyPhy v2.5.40. This analysis **did not include** site-to-site synonymous rate variation.

Suggested citation: *Gene-wide identification of episodic selection*, Mol Biol Evol. 32(5):1365-71, *Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril*, Mol Biol Evol. 37(8):2430-2439

Evidence ratio threshold 10 Update

16 sequences in the alignment	**288** codon sites in the alignment	**1** partitions
26 median branches/partition used for testing	**3 classes** non-synonymous rate variation	**None** synonymous rate variation
0.0 p-value for episodic diversityfing selection	**13** Sites with ER≥10 for positive selection	**N/A:N/A** Multiple hit rates (2H:3H)

Alignment-wide results

Model	Log (L)	AIC-c	Params.	Rate distribution	Rate plot
Unconstrained model	−2039.96	4170.83	45	**Tested ω** 0.5596 (86.941%) 0.9885 (10.960%) 96.09 (2.0981%) Mean = **2.611**, CoV = **5.242**	
Constrained model	−2078.31	4245.48	44	**Tested ω** 1.000 (14.819%) 1.000 (20.229%) 1.000 (64.952%) Mean = **1.000**, CoV = **NaN**	

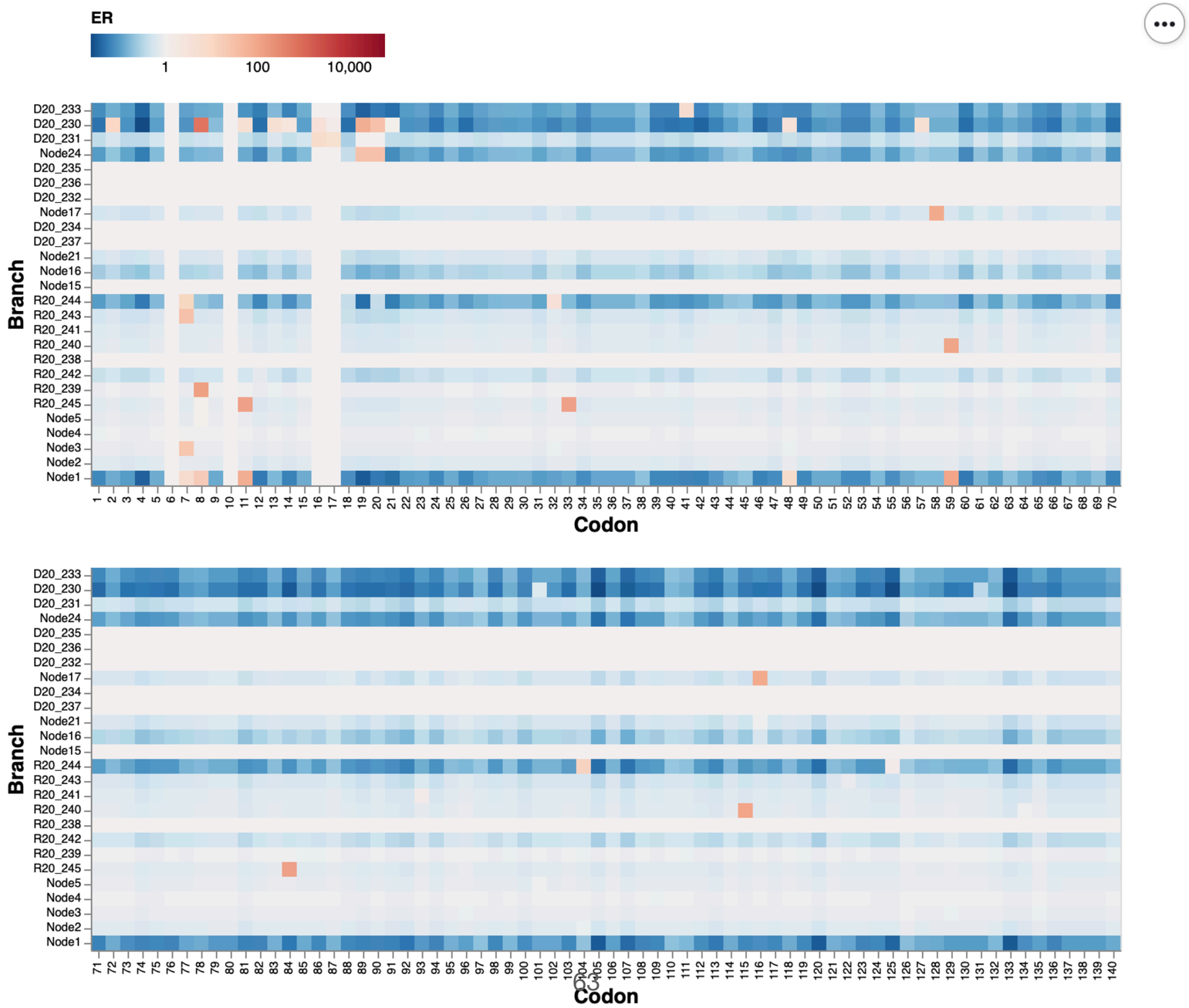
Murrell et al | Mol. Biol. Evol | 32(5) | 1365–1371

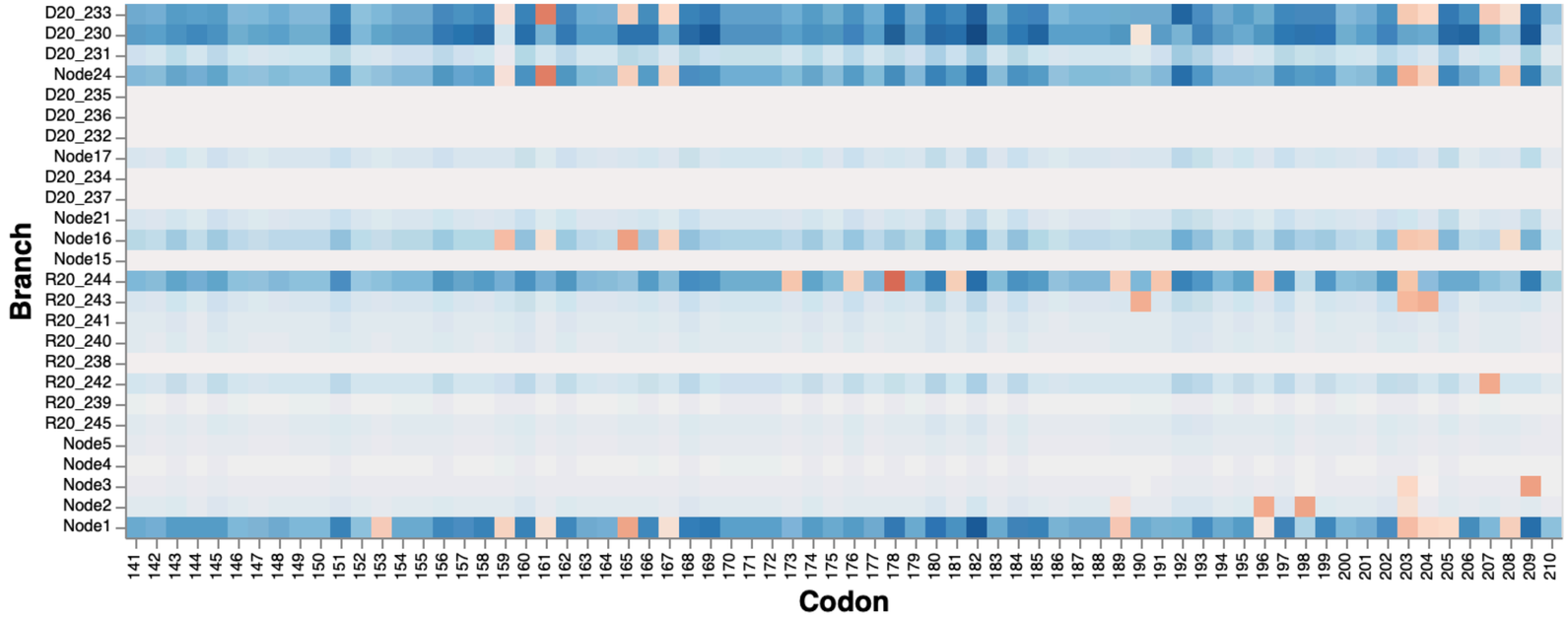
BUSTED inference

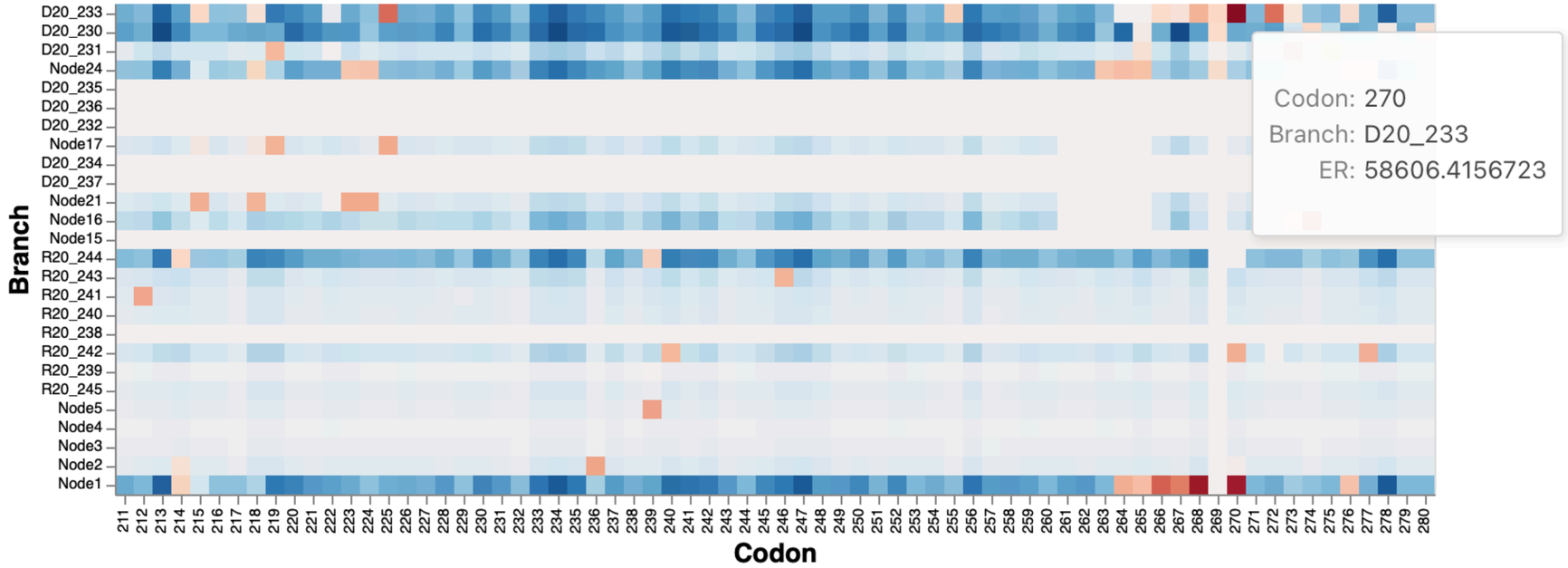
- Because BUSTED is a random-effects method, it **pools** information across multiple sites and branches to gain power
- The cost to this pooling is lack of site-level **resolution**, i.e., it is not immediately obvious which sites and/or branches drive the signal
- Standard ways to extract individual 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
- Can also compute exploratory evidence for selection support along individual branches at specific sites

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

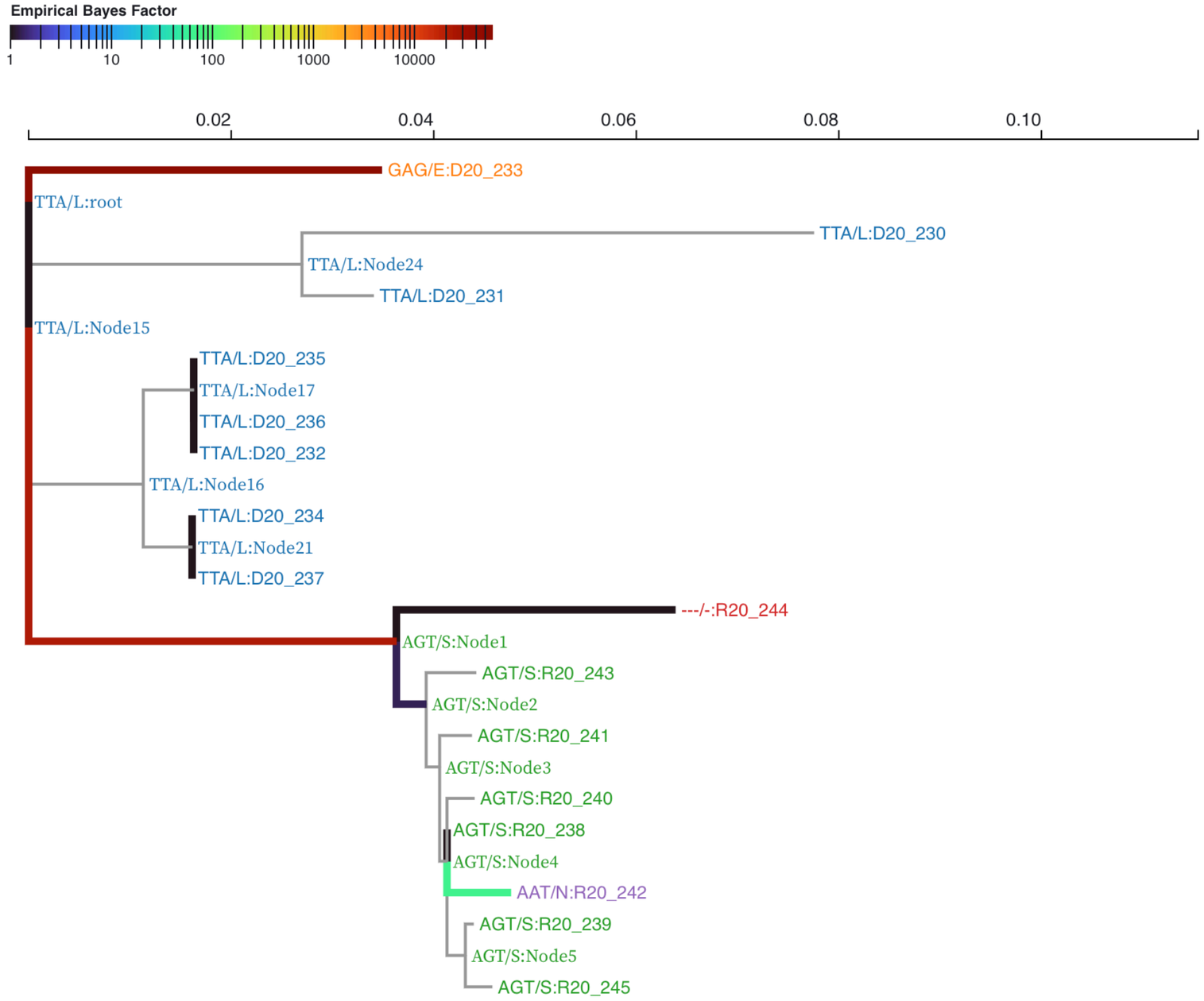
Figure 1. Empirical Bayes Factors for ω>1 at a particular branch and site (only tested branches are included).




Branch
D20_233
D20_230
D20_231
Node24
D20_235
D20_236
D20_232
Node17
D20_234
D20_237
Node21
Node16
Node15
R20_244
R20_243
R20_241
R20_240
R20_238
R20_242
R20_239
R20_245
Node5
Node4
Node3
Node2
Node1
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
Codon


Branch
D20_233
D20_230
D20_231
Node24
D20_235
D20_236
D20_232
Node17
D20_234
D20_237
Node21
Node16
Node15
R20_244
R20_243
R20_241
R20_240
R20_238
R20_242
R20_239
R20_245
Node5
Node4
Node3
Node2
Node1
211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280
Codon
Codon: 270
Branch: D20_233
ER: 58606.4156723

Codon 270


Empirical Bayes Factor
1
10
100
1000
10000
0.02
0.04
0.06
0.08
0.10
GAG/E:D20_233
TTA/L:root
TTA/L:D20_230
TTA/L:Node24
TTA/L:D20_231
TTA/L:Node15
TTA/L:D20_235
TTA/L:Node17
TTA/L:D20_236
TTA/L:D20_232
TTA/L:Node16
TTA/L:D20_234
TTA/L:Node21
TTA/L:D20_237
---/-:R20_244
AGT/S:Node1
AGT/S:R20_243
AGT/S:Node2
AGT/S:R20_241
AGT/S:Node3
AGT/S:R20_240
AGT/S:R20_238
AGT/S:Node4
AAT/N:R20_242
AGT/S:R20_239
AGT/S:Node5
AGT/S:R20_245

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

```
hyphy busted --srv No --alignment
data/WestNileVirus_NS3.fna --
starting-points 5
```

Produces *WestNileVirus_NS3.fna.BUSTED.json* file

View in http://vision.hyphy.org/BUSTED

Based on the likelihood ratio test, there **is no** evidence of *episodic diversifying selection* in this dataset (p=0.2691).

BUSTED analysis (v4.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/WestNileVirus_NS3.fas using HyPhy v2.5.40. This analysis **did not include** site-to-site synonymous rate variation.

Suggested citation: *Gene-wide identification of episodic selection*, Mol Biol Evol. 32(5):1365-71, *Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril*, Mol Biol Evol. 37(8):2430-2439

Evidence ratio threshold 10 Update

19 sequences in the alignment	**619** codon sites in the alignment	**1** partitions
33 median branches/partition used for testing	**3 classes** non-synonymous rate variation	**None** synonymous rate variation
0.27 p-value for episodic diversityfing selection	**0** Sites with ER≥10 for positive selection	**N/A:N/A** Multiple hit rates (2H:3H)

Alignment-wide results

Model	Log (L)	AIC-c	Params.	Rate distribution	Rate plot
Unconstrained model	−6396.17	12896.8	52	**Tested ω** 0.003895 (98.862%) 0.004316 (0.40762%) 1.859 (0.73075%) Mean = **0.01745**, CoV = **9.054**	0.0 0.5 1.0
Constrained model	−6396.79	12896.0	51	**Tested ω** 0.003674 (0.10704%) 0.003690 (98.824%) 1.000 (1.0691%) Mean = **0.01434**, CoV = **7.145**	0.0 0.2 0.4 0.

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

Gene-wide selection analysis using a branch-site method (BUSTED), SARS-CoV-2 spike

```
hyphy busted --srv No --alignment
data/spike.fas --tree data/
spike.tree --starting-points 5
```

Produces *spike.fas.BUSTED.json* file
View in http://vision.hyphy.org/BUSTED

Based on the likelihood ratio test, there **is** evidence of *episodic diversifying selection* in this dataset (p=0.01901).

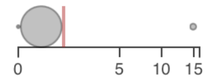
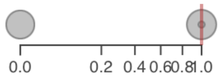
BUSTED analysis (v4.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/spike.fas using HyPhy v2.5.40. This analysis **did not include** site-to-site synonymous rate variation.

Suggested citation: *Gene-wide identification of episodic selection*, Mol Biol Evol. 32(5):1365-71, *Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril*, Mol Biol Evol. 37(8):2430-2439

Evidence ratio threshold 10 Update

118 sequences in the alignment	**1273** codon sites in the alignment	**1** partitions
162 median branches/partition used for testing	**3 classes** non-synonymous rate variation	**None** synonymous rate variation
0.019 p-value for episodic diversityfing selection	**11** Sites with ER≥10 for positive selection	**N/A:N/A** Multiple hit rates (2H:3H)

Alignment-wide results

Model	Log (L)	AIC-c	Params.	Rate distribution	Rate plot
Unconstrained model	-9287.35	18937.1	181	**Tested ω** 0.000 (0.45262%) 0.2561 (97.532%) 14.88 (2.0156%) Mean = **0.5497**, CoV = **3.740**	
Constrained model	-9290.62	18941.7	180	**Tested ω** 0.000 (46.919%) 1.000 (2.2974%) 1.000 (50.784%) Mean = **0.5308**, CoV = **0.9402**	

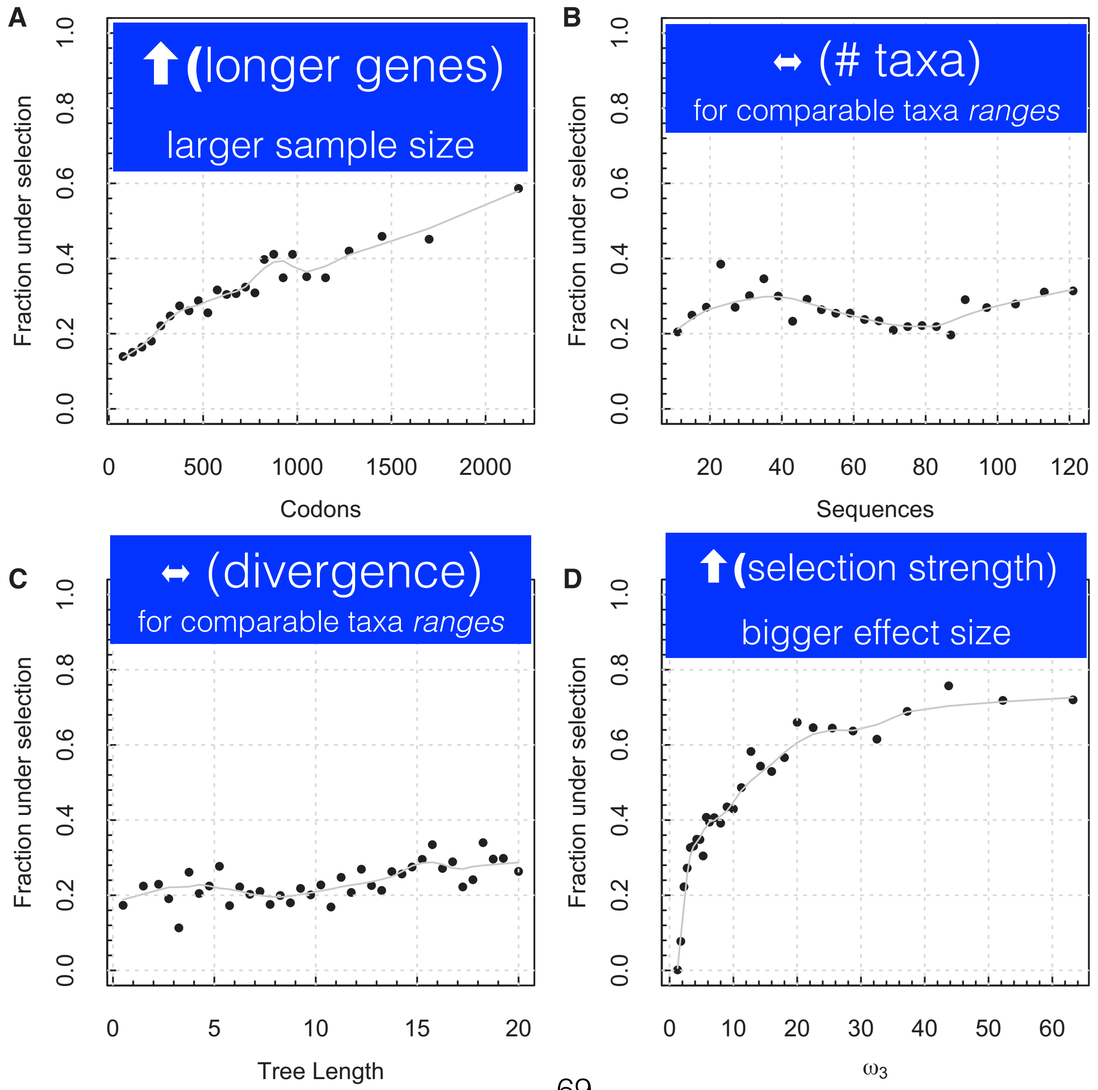
BUSTED analysis

- **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`)
- **SARS-CoV-2 spike**
  - Evidence of episodic diversification (`dN/dS ~ 15`) on a small proportion of sites (`~2%`)
  - Most of the rest of the gene evolves with purifying selection (`dN/dS = 0.2`)

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) | 136

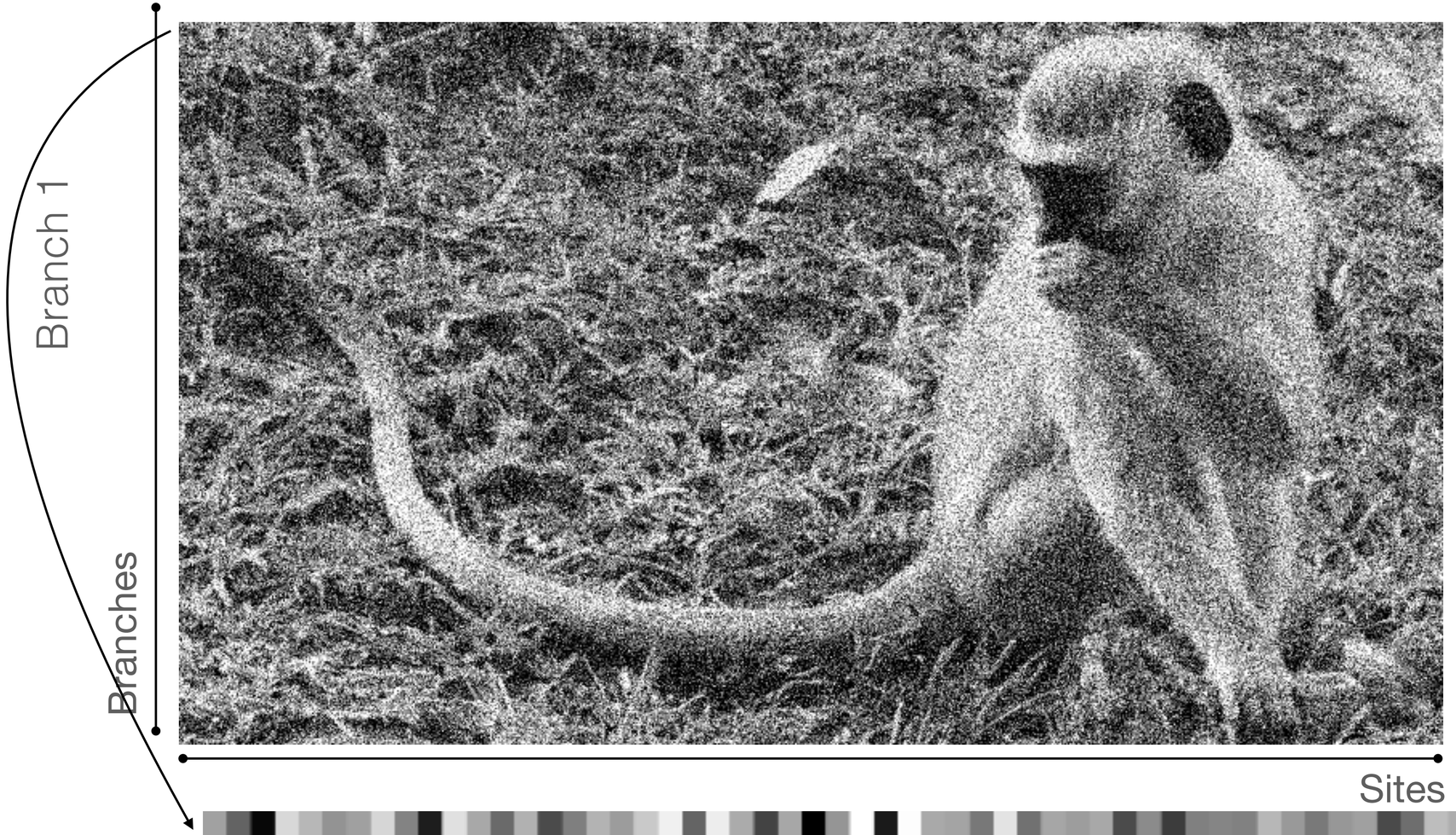
Any questions on the previous material?

We covered the following methods in HyPhy: FitMG94, BUSTED

Which estimates a mean gene-wide dN/dS (FitMG94)

Or estimates dN/dS through a branch-site method (BUSTED)

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

- Fix the tree; estimate and fix some of the nuisance model parameters that are shared by all branches
(GTR biases, frequency counts)
- Fit a simple baseline model (one ω per branch); use this model to get initial guesses for all other parameters
- Perform a greedy step-up procedure (like forward variable selection in regression models, but not as statistically bad)
- For each branch (longest first) try two ω rate classes, then three ω rate classes etc, until no more goodness-of-fit improvement (AIC-c)
- Fix the number of rates and move on to the next longest branch
- Perform selection testing on the overall model (different number of ω classes on branches), using the likelihood ratio test
- Each branch specified a priori (could be all branches)
- Appropriate multiple testing correction

HIV-1 *env*

adaptive Branch Site REL
results summary

INPUT DATA | **HIV-sets.fas** | **16** sequences | **288** sites

Export

aBSREL **found evidence** of episodic diversifying selection on **3** out of **26** branches in your phylogeny.

A total of **26** branches were formally tested for diversifying selection. Significance was assessed using the Likelihood Ratio Test at a threshold of $p \leq 0.05$, after correcting for multiple testing. Significance and number of rate categories inferred at each branch are provided in the detailed results table.

See here for more information about this method.

Please cite PMID 25697341 if you use this result in a publication, presentation, or other scientific work.

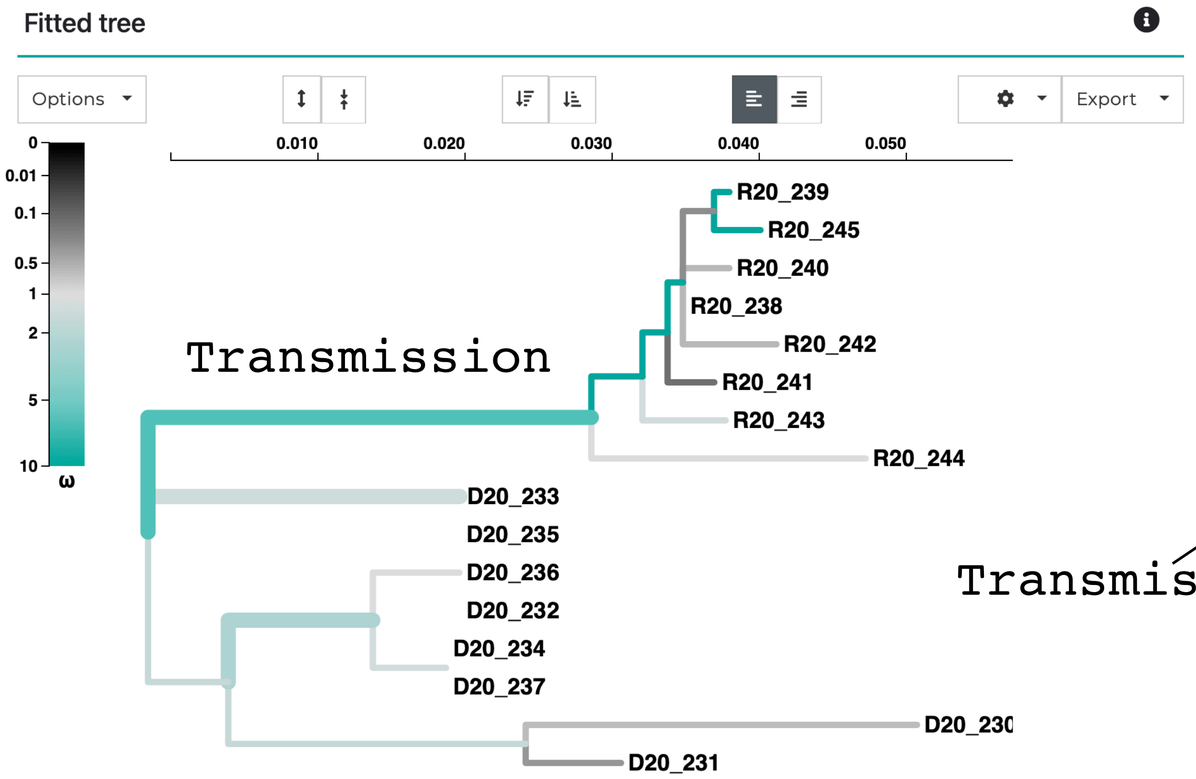
Tree summary

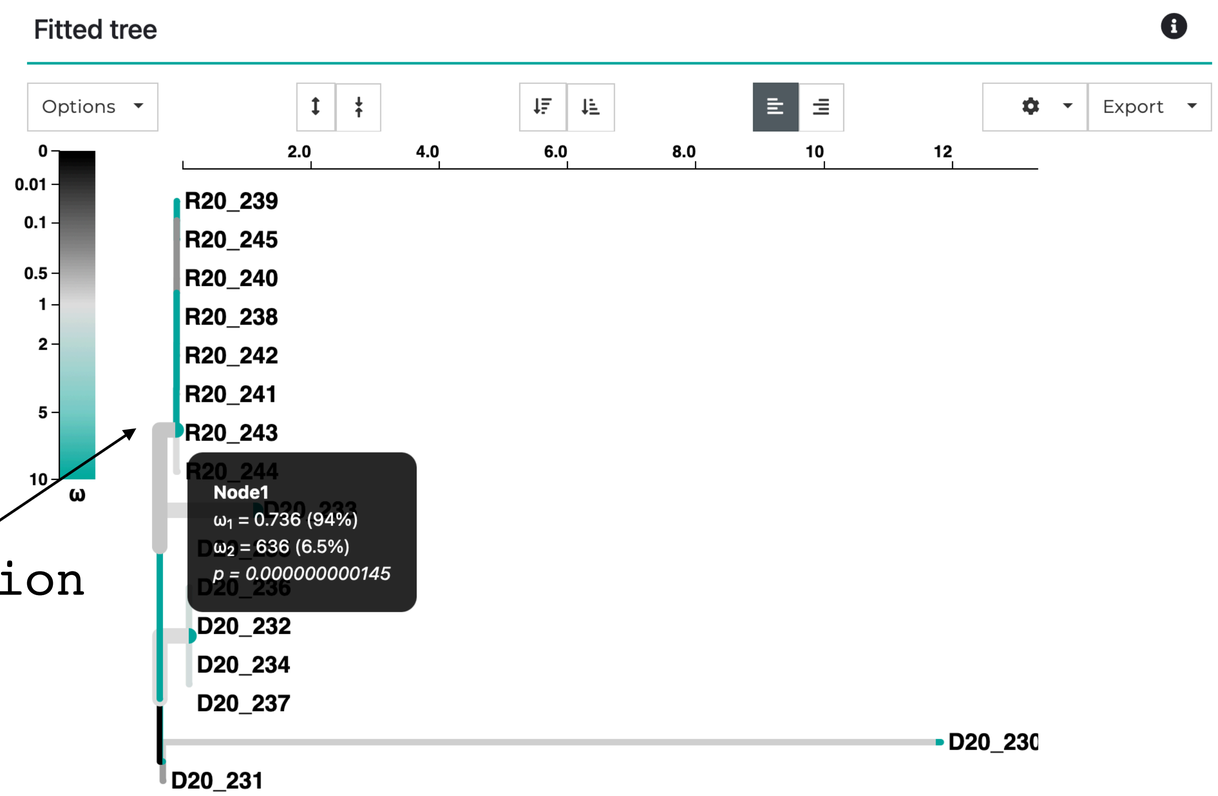
ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	21	81%	0.49%	0
2	5	19%	100%	3

This table contains a summary of the inferred aBSREL model complexity. Each row provides information about the branches that were best described by the given number of ω rate categories.

```
hyphy absrel --alignment data/HIV-sets.nex
```

HIV-1 *env*





Transmission

One dN/dS per branch

Adaptive dN/dS per branch

WNV NS3

adaptive Branch Site REL

results summary

INPUT DATA | **WestNileVirus_NS3.fas** | **19** sequences | **619** sites

Export

aBSREL **found no evidence** of episodic diversifying selection in your phylogeny.

A total of **33** branches were formally tested for diversifying selection. Significance was assessed using the Likelihood Ratio Test at a threshold of $p \leq 0.05$, after correcting for multiple testing. Significance and number of rate categories inferred at each branch are provided in the detailed results table.

See here for more information about this method.

Please cite PMID 25697341 if you use this result in a publication, presentation, or other scientific work.

Tree summary

ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	30	91%	37%	0
2	3	9.1%	63%	0

This table contains a summary of the inferred aBSREL model complexity. Each row provides information about the branches that were best described by the given number of ω rate categories.

```
hyphy absrel --alignment data/WestNileVirus_NS3.fna
```

SARS-CoV-2 spike

adaptive Branch Site REL
results summary

INPUT DATA | **spike.fas** | **118** sequences | **1273** sites

Export

aBSREL **found no evidence** of episodic diversifying selection in your phylogeny.

A total of **44** branches were formally tested for diversifying selection. Significance was assessed using the Likelihood Ratio Test at a threshold of $p \leq 0.05$, after correcting for multiple testing. Significance and number of rate categories inferred at each branch are provided in the detailed results table.

See here for more information about this method.

Please cite PMID 25697341 if you use this result in a publication, presentation, or other scientific work.

Tree summary

ω rate classes	# of branches	% of branches	% of tree length	# under selection
1	161	99%	61%	0
2	1	0.62%	39%	0

This table contains a summary of the inferred aBSREL model complexity. Each row provides information about the branches that were best described by the given number of ω rate categories.

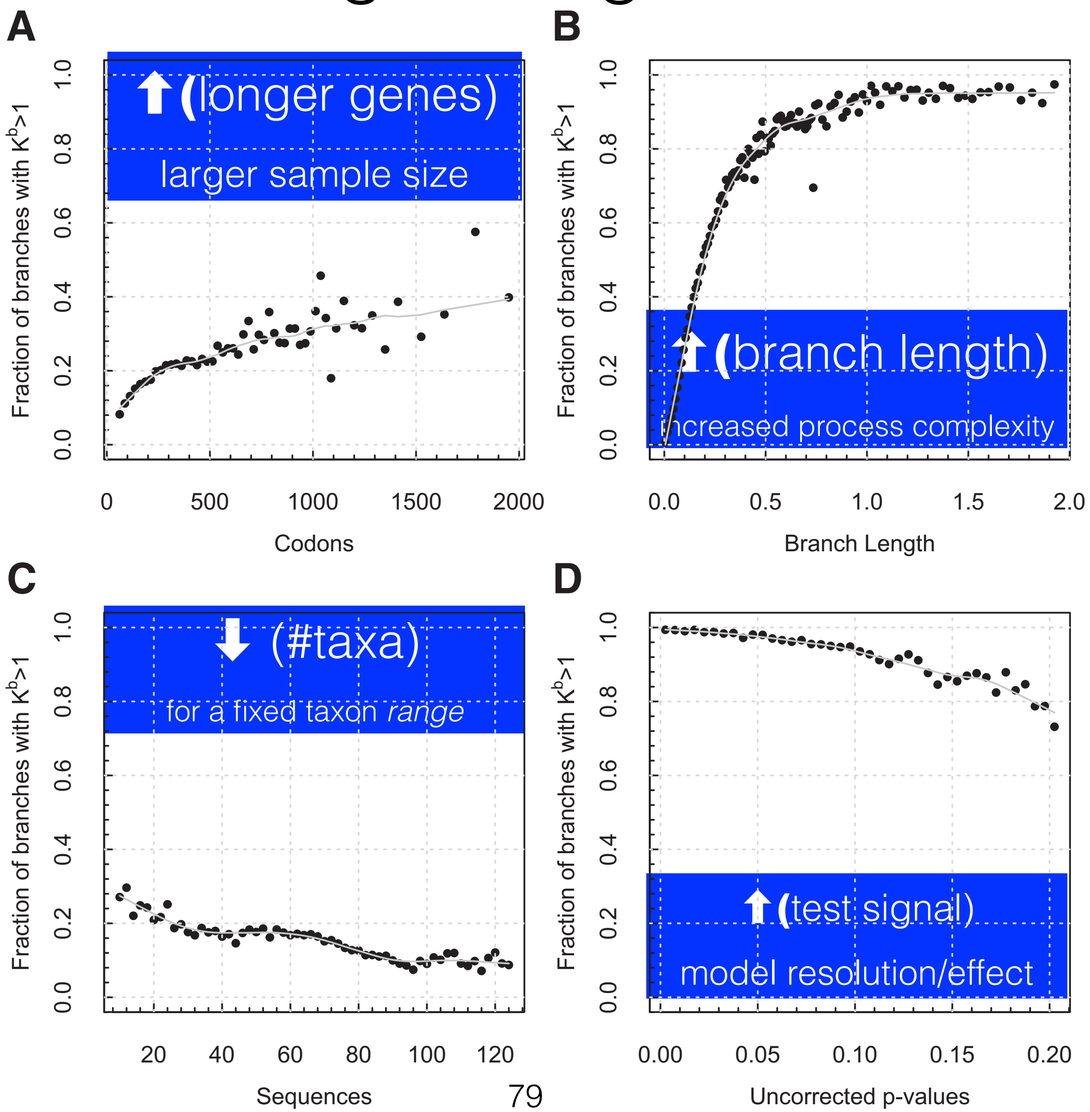
```
hyphy absrel --alignment data/spike.fas --tree data/spike.tree --branches Internal
```

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**
  - `76%` branches can be explained with simple (single `dN/dS`) models
  - `5` branches (`24%, ~100%` 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
- **SARS-CoV-2 spike**
  - All but **one** branch can be explained with simple (single `dN/dS`) models
  - `1` long terminal branch (`~34%` of tree length) has evidence of multiple `dN/dS` rate classes over sites
  - No evidence of branch level selection on internal branches.

Correlates of evolutionary complexity

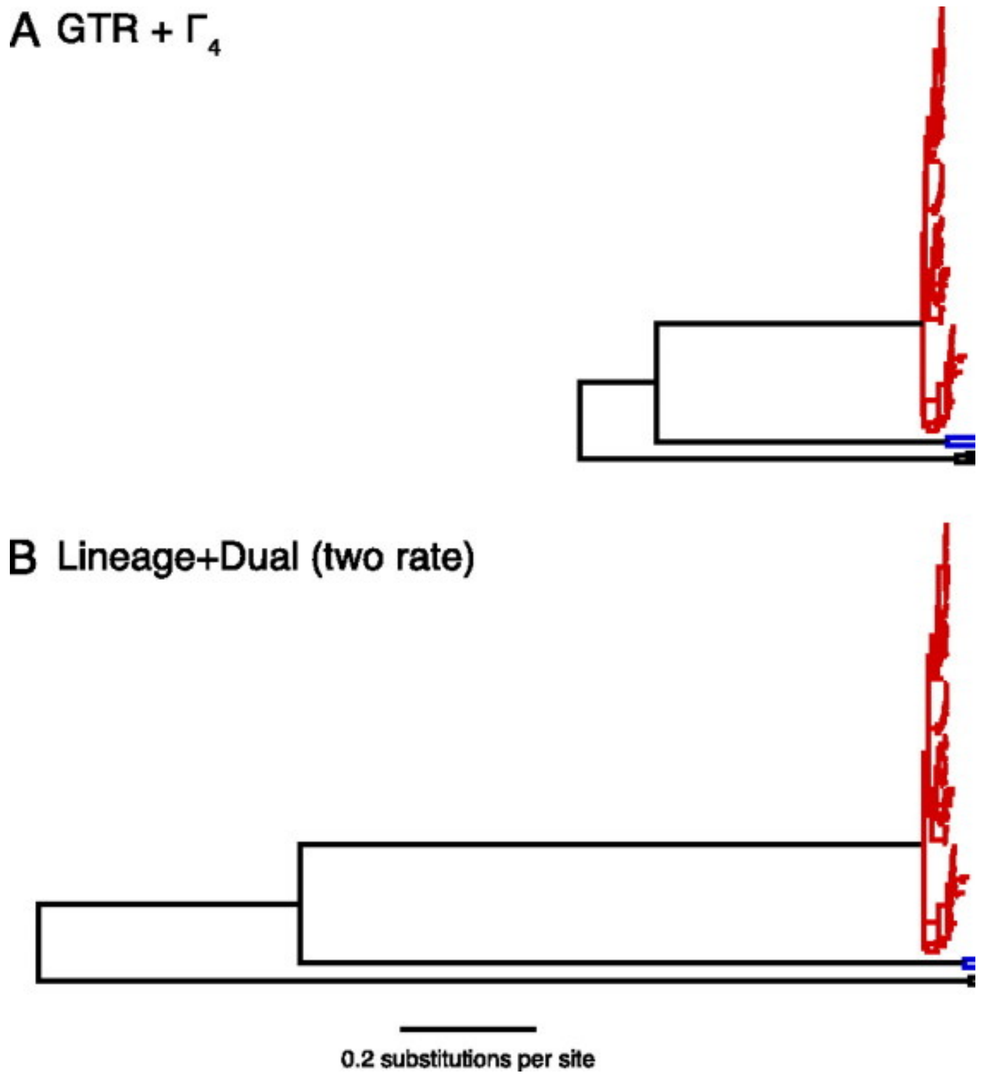
An analysis of ~9,000 curated gene alignments from *selectome.unil.ch*



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



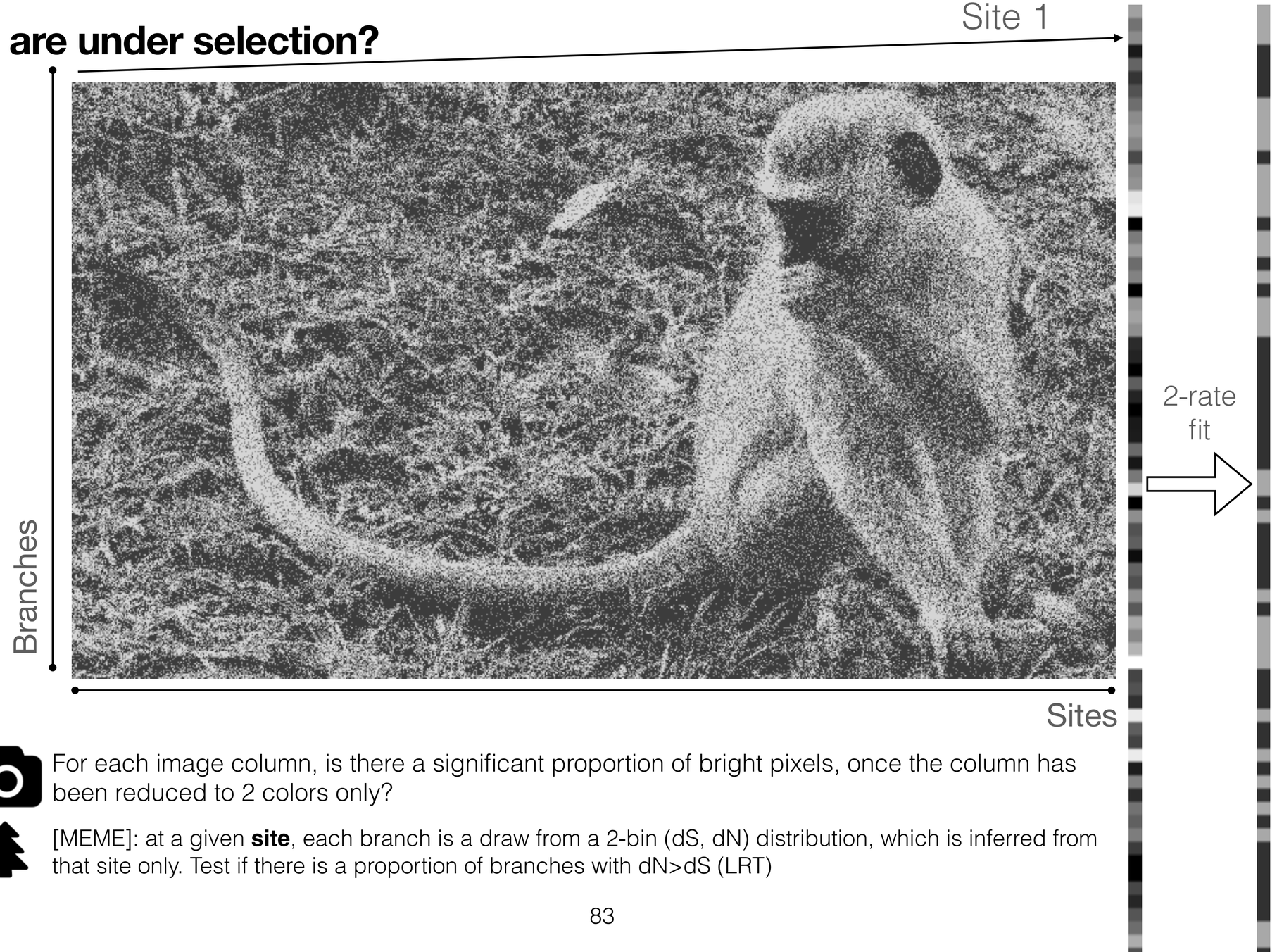
Wertheim and Pond (2011) Mol Biol Evol. 28(12):3355-65

Questions about the previous material?

We covered the aBSREL method, which detects episodes of selection across branches in your phylogeny.

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 with any precision
- Does not scale especially well with the number of sequences

https://observablehq.com/@spond/meme

HIV-1 env

Based on the likelihood ratio test, *episodic diversifying selection* has acted on **8** sites in this dataset (p≤0.1).

MEME analysis (v3.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/HIV-sets.fas using HyPhy v2.5.40.

p-value threshold 0.1 Update

Suggested citation: Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet* 8(7): e1002764.

16 sequences in the alignment

288 codon sites in the alignment

1 partitions

26 median branches/partition used for testing

N/A bootstrap replicates

8 sites subject to episodic diversifying selection

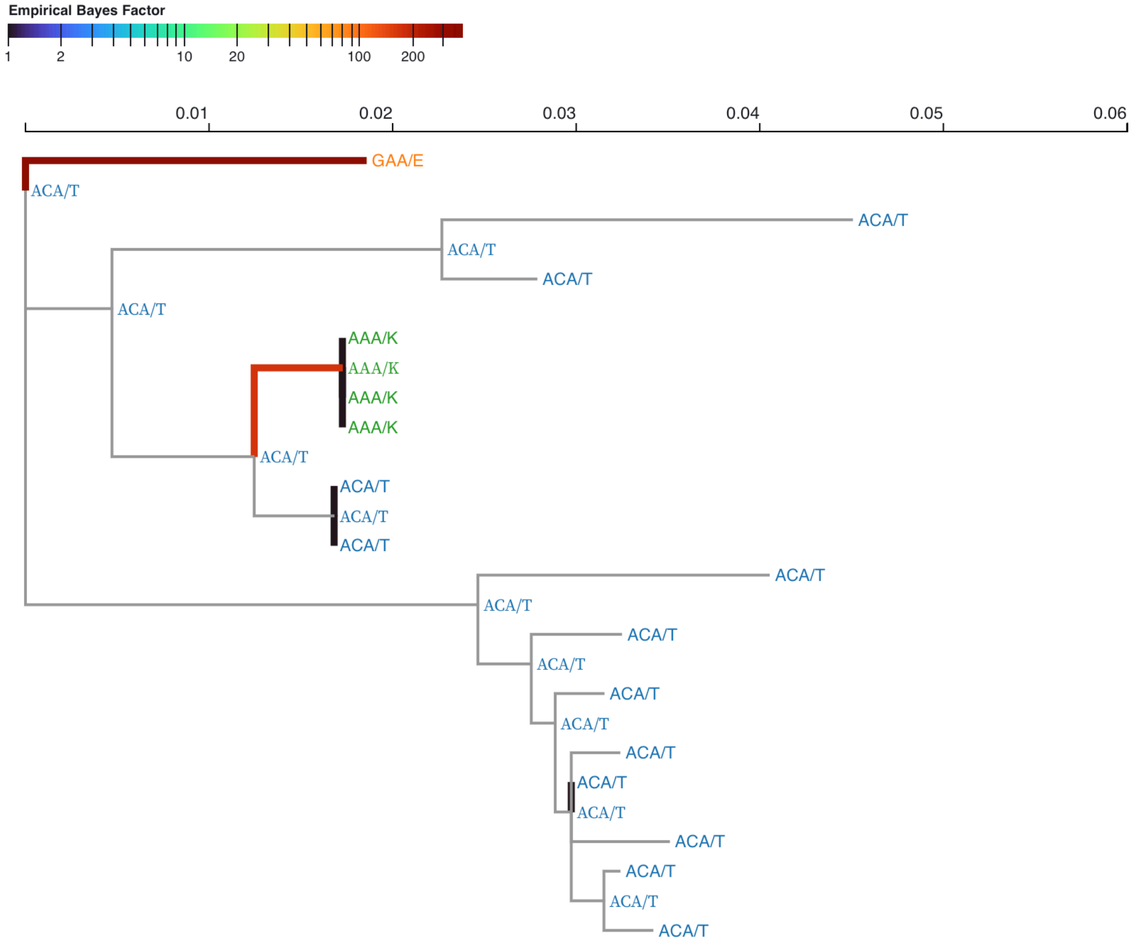
0.50 median branches with support for selection/selected site

3 sites with variable ω across branches

Table 1. Detailed site-by-site results from the MEME analysis

Part.	Codon	α	β^-	p^-	β^+	p^+	LRT	p-value	# branches under selection	MEME LogL	FEL LogL	Variation p
1	19	0	0	0.945	2,749.03	0.055	6.602	0.017	0	-14.847	-11.568	0.038
1	161	0	0	0.82	114.386	0.18	7.579	0.01	0	-16.568	-14.144	0.089
1	165	0	0	0.774	52.349	0.226	4.247	0.056	0	-15.506	-14.441	0.345
1	225	0	0	0.747	47.804	0.253	3.708	0.074	2	-13.869	-13.061	0.445
1	264	0	0	0.894	168.831	0.106	3.613	0.077	0	-11.753	-10.172	0.206
1	272	0	0	0.853	40.182	0.147	3.325	0.09	1	-10.449	-9.374	0.341
1	274	2.785	2.785	0.947	10,000	0.053	4.981	0.038	1	-20.161	-17.673	0.083
1	282	0	0	0	8.192	1	3.652	0.076	0	-19.326	-19.324	0.999

hyphy meme --alignment data/HIV-sets.nex



- Where in the tree is there evidence for selection?
- Can use exploratory Empirical Bayes Factor analysis to find “hotspots”
- Not a “strict” statistical test!

WNV NS3

Based on the likelihood ratio test, *episodic diversifying selection* has acted on **4** sites in this dataset (p≤0.1).

MEME analysis (v3.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/WestNileVirus_NS3.fas using HyPhy v2.5.40.

p-value threshold 0.1 Update

Suggested citation: Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet* 8(7): e1002764.

19 sequences in the alignment	**619** codon sites in the alignment	**1** partitions
33 median branches/partition used for testing	**N/A** bootstrap replicates	**4** sites subject to episodic diversifying selection
0.75 median branches with support for selection/selected site	**6** sites with variable ω across branches	

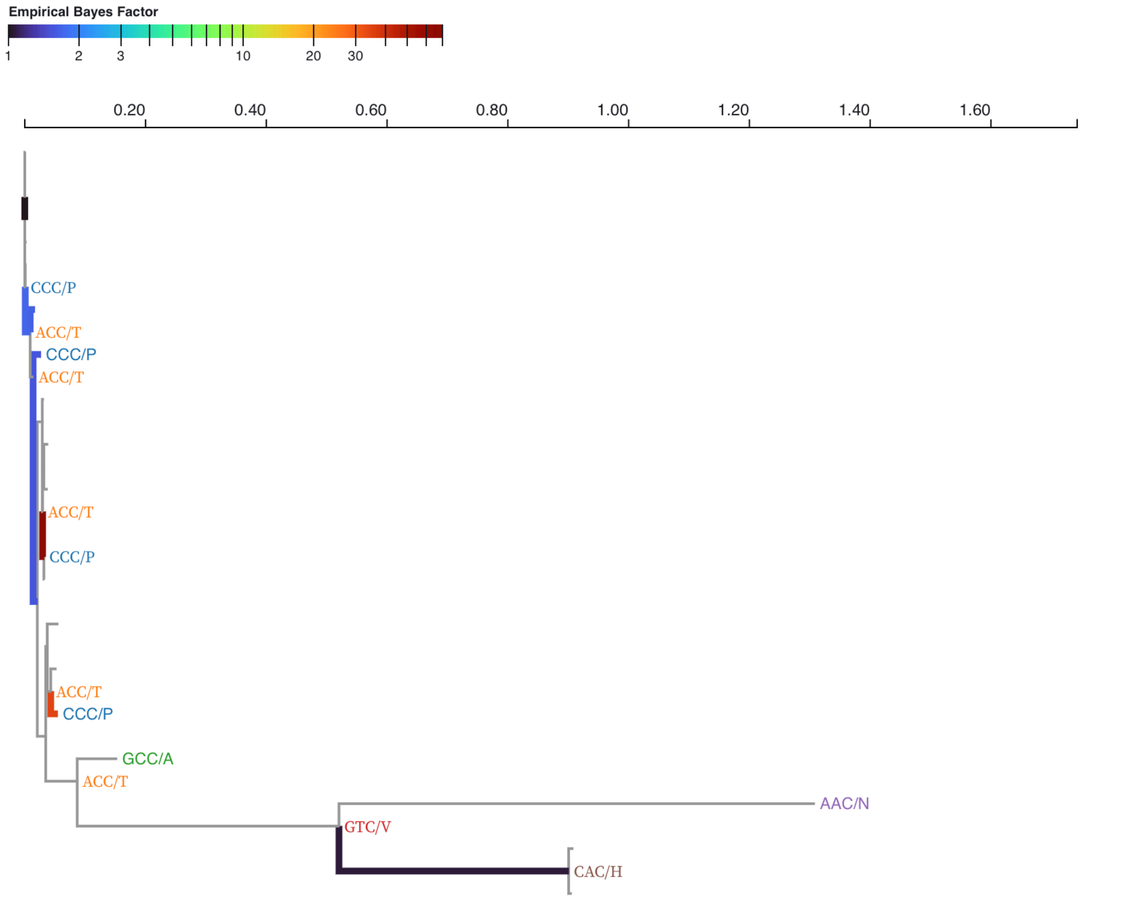


Table 1. Detailed site-by-site results from the MEME analysis

Part.	Codon	α	β⁻	p⁻	β⁺	p⁺	LRT	▲p-value	# branches under selection	MEME LogL	FEL LogL	Variation p
1	249	0	0	0	2.708	1	7.883	0.009	0	-34.231	-34.232	1
1	557	0.234	0	0.965	140.484	0.035	5.517	0.029	1	-17.698	-14.167	0.029
1	521	0.922	0	0.961	103.466	0.039	3.6	0.078	1	-17.268	-14.308	0.052
1	87	1.972	0	0.948	29.804	0.052	3.455	0.084	1	-23.521	-16.735	0.001

```
hyphy meme --alignment data/WestNileVirus_NS3.fna
```

SARS CoV-2 Spike

Based on the likelihood ratio test, *episodic diversifying selection* has acted on **6** sites in this dataset (p≤0.1).

MEME analysis (v3.0) was performed on the alignment from /Users/sergei/Dropbox/Talks/VEME-current/data/spike.fas using HyPhy v2.5.40.

p-value threshold 0.1 Update

Suggested citation: Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet* 8(7): e1002764.

Codon	α	β⁻	p⁻	β⁺	p⁺	LRT	▴p-value	# branches under selection	MEME LogL	FEL LogL
1	1.213	0.543	0.994	2,732.12	0.006	9.76	0.003	1	-19.143	-14.262
1,243	0	0	0.983	552.73	0.017	8.846	0.005	2	-23.278	-19.431
452	0	0	0.01	14.48	0.99	8.771	0.005	5	-36.485	-36.486
470	3.168	0.727	0.994	10,000	0.006	8.012	0.008	1	-24.317	-19.832
501	0.004	0.002	0.925	343.752	0.075	5.017	0.037	5	-37.067	-36.202
157	0	0	0.01	7.313	0.99	3.708	0.074	4	-29.546	-29.547

118 sequences in the alignment

1273 codon sites in the alignment

1 partitions

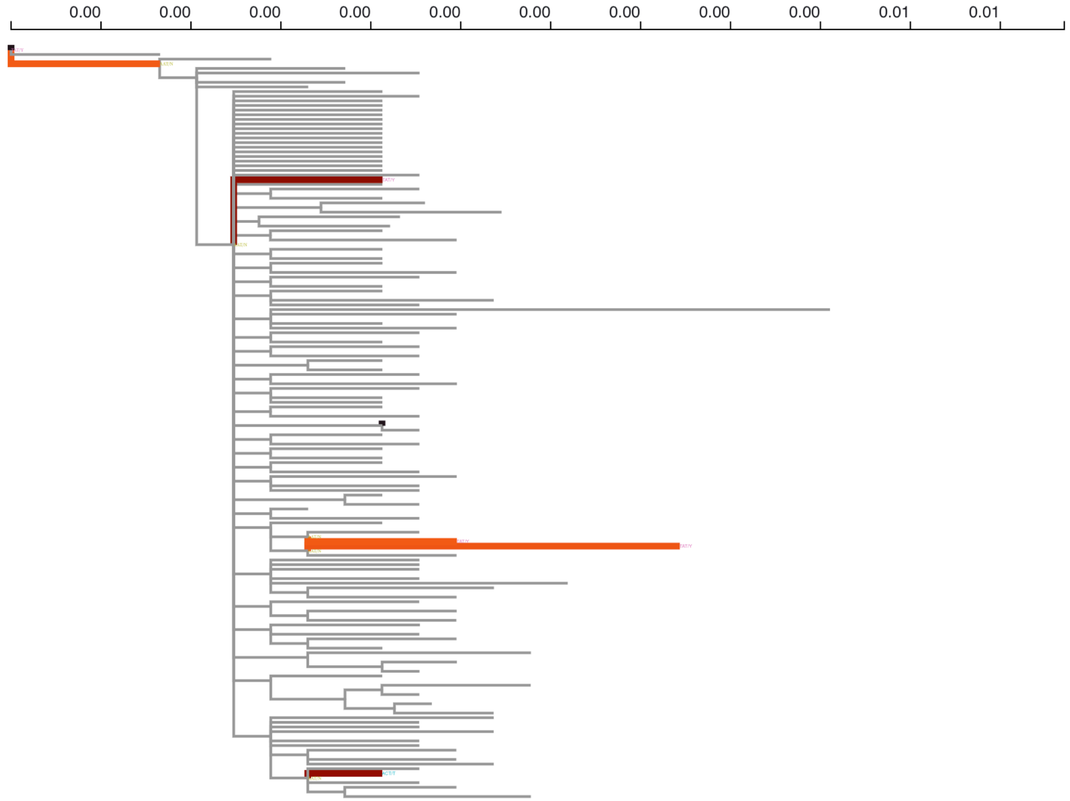
162 median branches/partition used for testing

N/A bootstrap replicates

6 sites subject to episodic diversifying selection

3.00 median branches with support for selection/selected site

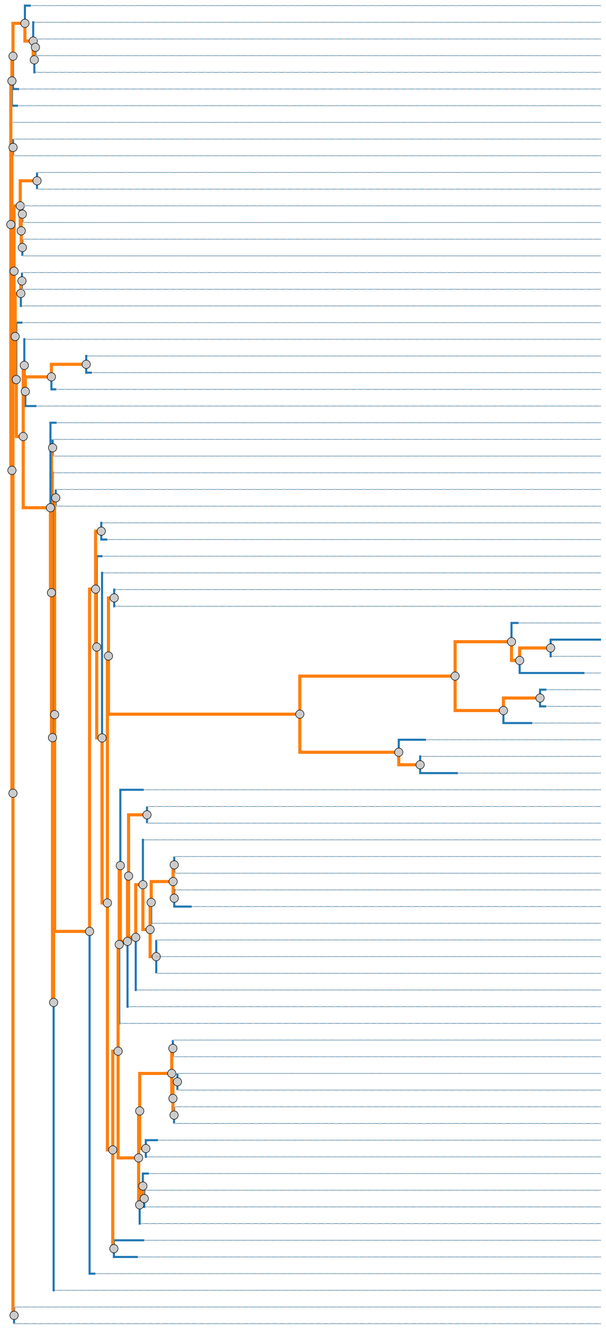
3 sites with variable ω across branches



hyphy meme --alignment data/spike.fas --tree data/spike.tree

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

SARS CoV-2 Spike
Internal Branches Only

Codon	Partition	alpha	beta+	p+	LRT	Episodic selection detected?	# branches	Most common codon substitutions at this site
367	1	0.000	98.114	0.459	9.062	Yes, p = 0.0047	2	[2]Gtc>Ttc
439	1	0.000	35.271	1.000	4.990	Yes, p = 0.0379	1	[1]aaC>aaA
452	1	0.000	30.877	1.000	5.520	Yes, p = 0.0288	1	[4]cTg>cGg|[1]Ctg>Atg
477	1	0.000	51.490	0.460	4.334	Yes, p = 0.0532	1	[1]aGc>aAc
501	1	0.000	271.405	0.145	3.460	Yes, p = 0.0839	1	[3]Aat>Tat|[1]aAt>aCt,Tat>Aat
614	1	0.000	49.307	0.464	3.187	Yes, p = 0.0968	1	[1]Gat>Aat,gGt>gAt

hyphy meme --alignment data/spike.fas --tree data/spike.tree --branches Internal

MEME results

- **West Nile Virus NS3 protein**
  - **`Four`** sites (incl. **`249, previously reported`**) with significant evidence of **episodic** (or pervasive) diversifying selection.
- **HIV-1 transmission pair**
  - **`Nine`** sites with significant evidence of **episodic** (or pervasive) diversifying selection. HIV-1 transmission pair
- **SARS-CoV-2 spike (all)**
  - **`Six`** sites with significant evidence of **episodic** (or pervasive) diversifying selection.
- **SARS-CoV-2 spike (internal)**
  - **`Six`** sites with significant evidence of **episodic** (or pervasive) diversifying selection.

Questions about the previous material?

We covered the MEME method:

which detects episodes of selection and is used as a more sensitive method for estimating diversifying selection.

More on site-level selection

- Three more methods in HyPhy
- Fixed Effects Likelihood **(FEL)**
  - A simpler alternative to MEME (looks for pervasive selection)
  - May be more suited for smaller datasets or datasets of low divergence
- Single Likelihood Ancestor Counting (**SLAC**)
  - A counting-based approach
  - Good for data exploration and visualization
- Fast Unrestricted Bayesian AppRoximation (**FUBAR**)
    - A novel statistical approach for detecting pervasive adaptive evolution on large datasets (scales to 10000s of sequences)

FEL on internal branches of Spike finds most selected sites, including many known to be of functional significance

Codon	Partition	alpha	beta	LRT	Selection detected?
5	1	0.000	19.047	2.891	Pos. p = 0.0891
12	1	0.000	20.331	2.990	Pos. p = 0.0838
18	1	0.000	19.108	2.886	Pos. p = 0.0893
138	1	0.000	26.726	2.738	Pos. p = 0.0980
367	1	0.000	44.298	9.049	Pos. p = 0.0026
439	1	0.000	34.504	4.989	Pos. p = 0.0255
452	1	0.000	30.455	5.519	Pos. p = 0.0188
477	1	0.000	23.695	4.327	Pos. p = 0.0375
501	1	0.000	38.294	3.319	Pos. p = 0.0685
570	1	0.000	21.078	3.049	Pos. p = 0.0808
614	1	0.000	22.071	3.101	Pos. p = 0.0783
681	1	0.000	18.297	2.820	Pos. p = 0.0931
1176	1	0.000	21.975	3.040	Pos. p = 0.0812

```
hyphy fel --alignment data/spike.fas --tree data/spike.tree --branches Internal
```

More accurate testing via parametric bootstrap

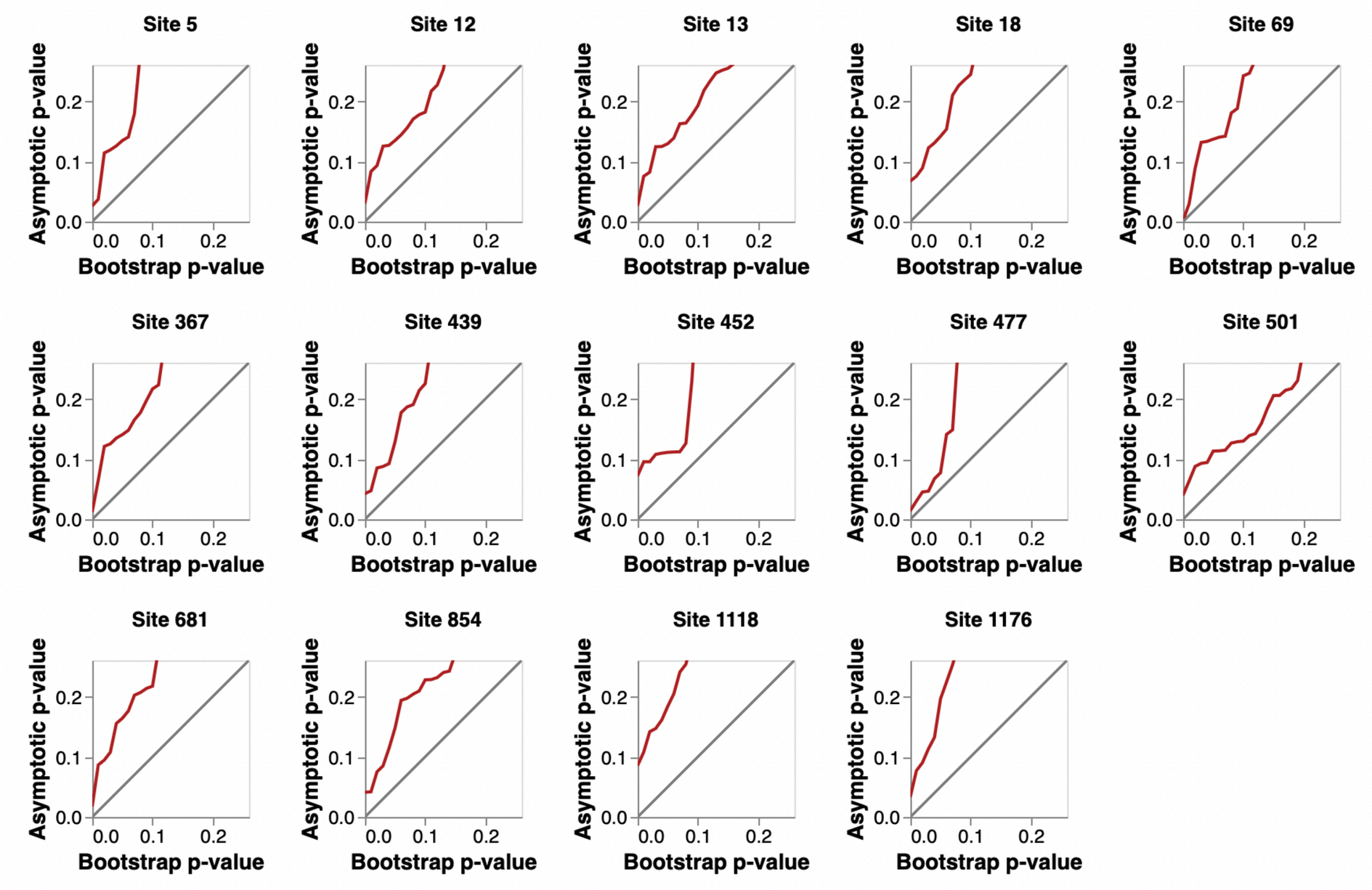
- P-values for MEME/FEL etc are derived from asymptotic approximations (large N)
- Not clear how well these hold for smaller and low-divergence datasets
- Can use a **much slower** simulation based method to derive more accurate p-values at each site
- Can result both in improved power and lower rates of false positives

FEL on internal branches of Spike finds most selected sites

Codon	Partition	alpha	beta	LRT	Selection detected?
1	1	10000.000	0.075	1.004	Neg. p = 0.0099
5	1	0.000	19.044	2.890	Pos. p = 0.0297
12	1	0.000	20.323	2.989	Pos. p = 0.0297
13	1	0.000	23.877	2.589	Pos. p = 0.0396
18	1	0.000	19.110	2.885	Pos. p = 0.0396
49	1	0.000	22.768	2.519	Pos. p = 0.0594
69	1	0.000	22.573	2.464	Pos. p = 0.0396
138	1	0.000	26.821	2.736	Pos. p = 0.0693
177	1	10000.000	0.000	1.187	Neg. p = 0.0099
245	1	0.000	22.499	2.501	Pos. p = 0.0792
367	1	0.000	44.535	9.044	Pos. p = 0.0099
439	1	0.000	34.481	4.986	Pos. p = 0.0099
452	1	0.000	30.321	5.518	Pos. p = 0.0099
477	1	0.000	23.678	4.325	Pos. p = 0.0297
501	1	0.000	38.373	3.317	Pos. p = 0.0297
570	1	0.000	21.039	3.047	Pos. p = 0.0594
614	1	0.000	22.215	3.099	Pos. p = 0.0594
681	1	0.000	18.344	2.819	Pos. p = 0.0297
701	1	0.000	20.245	1.998	Pos. p = 0.0792
716	1	0.000	25.450	2.349	Pos. p = 0.0792
769	1	0.000	28.107	2.470	Pos. p = 0.0693
854	1	0.000	25.908	2.506	Pos. p = 0.0495
941	1	0.000	25.251	2.344	Pos. p = 0.0891
1118	1	3.081	27.290	2.023	Pos. p = 0.0495
1176	1	0.000	22.141	3.039	Pos. p = 0.0297
1237	1	10000.000	0.151	0.876	Neg. p = 0.0099
1248	1	0.000	20.075	2.336	Pos. p = 0.0891

CAUTION: A VERY TIME CONSUMING ANALYSIS (SEVERAL HOURS)

```
hyphy fel --alignment data/spike.fas --tree data/spike.tree --branches Internal --
output Spike-pbs.FEL.json --resample 100
```



Asymptotic p-value is too conservative in this case for most sites

Obtaining site-level dN/dS estimates with FEL

- dN/dS estimates at individual sites are not **precise**
- They are estimated from relatively small samples
- Precision improves with the number of sequences and divergence levels
- One approach to correct for this is to compute approximate site-level confidence intervals.

Codon	Partition	alpha	beta	LRT	Selection detected?	dN/dS with confidence intervals
2	1	1.843	0.000	7.521	Neg. p = 0.0061	0.000(0.00- 0.09)
3	1	0.786	0.000	3.161	Neg. p = 0.0754	0.000(0.00- 0.16)
4	1	2.174	0.000	10.742	Neg. p = 0.0010	0.000(0.00- 0.07)
7	1	1.105	0.000	7.537	Neg. p = 0.0060	0.000(0.00- 0.11)
8	1	0.422	0.000	3.173	Neg. p = 0.0749	0.000(0.00- 0.28)
9	1	1.353	0.000	8.638	Neg. p = 0.0033	0.000(0.00- 0.08)
10	1	1.353	0.000	8.369	Neg. p = 0.0038	0.000(0.00- 0.09)
…						
247	1	1.353	0.000	8.088	Neg. p = 0.0045	0.000(0.00- 0.10)
248	1	0.451	0.000	3.496	Neg. p = 0.0615	0.000(0.00- 0.28)
249	1	0.000	2.700	7.881	Pos. p = 0.0050	10000.000(7599.84-10000.00)
250	1	0.220	0.000	2.797	Neg. p = 0.0945	0.000(0.00- 0.61)
388	1	0.220	0.000	2.797	Neg. p = 0.0945	0.000(0.00- 0.61)

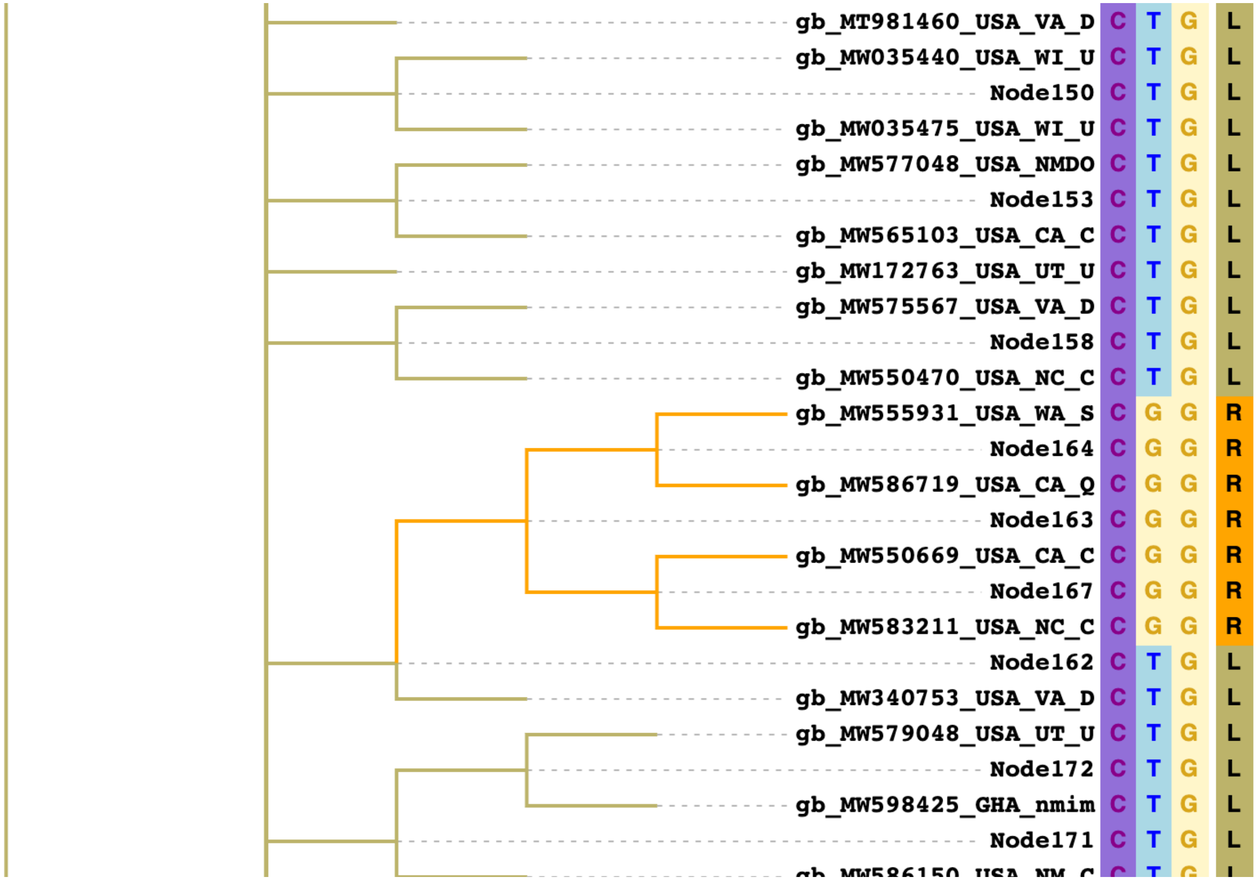
hyphy fel --alignment data/WestNileVirus_NS3.fna –ci Yes

Mapping substitutions with SLAC

- SLAC capable of detecting selection, is fast, but generally lacks power
- It provides a number of intuitive metrics for interpreting selection results
- SLAC recovers ancestral states and allows one to “map” evolutionary history onto a tree.

```
hyphy slac --alignment data/spike.fas --tree data/spike.tree --branches Internal
```

Partition	Site	ES	EN	S	N	P[S]	dS	dN	dN-dS	P [dN/dS > 1]	P [dN/dS < 1]	Total branch length
1	452	1.75	1.25	0.00	1.00	0.584	0.00	0.801	49.6	0.416	1.00	0.0162
1	367	0.996	2.00	0.00	2.00	0.332	0.00	0.998	61.8	0.446	1.00	0.0162



Site 452

Analysis summary

	WNV NS3	**HIV-1 *env***	**SARS-CoV-2 spike**
Gene-wide episodic selection (BUSTED)	No	Yes	Yes
Branch-level selection (aBSREL)	No	Yes, three branches, including transmission	No
Site-level episodic selection (MEME)	Yes, **1** site	Yes, **8** sites	Yes, sites found depend on which branches are included

Questions about the previous material?

We covered the FEL and SLAC selection analyses, site-level methods

- Fixed Effects Likelihood **(FEL)**
  - A simpler alternative to MEME (looks for pervasive selection)
  - May be more suited for smaller datasets or datasets of low divergence
- Single Likelihood Ancestor Counting (**SLAC**)
  - A counting-based approach
  - Good for data exploration and visualization

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 k-bin (k=3) 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

Purpose-built models

- It is tempting to “hack” existing tools to answer questions that they are not designed to answer
- A recent example we tackled is a rigorous test for relaxation of selection (or more generally a difference in selective regimes) in a part of the tree, relative to the rest of the tree
- Typical approaches have been to estimate dN/dS rations from two sets of branches, and interpret an *elevation* in dN/dS as evidence of selective constraint relaxation

- Two problems with this approach:
  - An increase in mean dN/dS could also be caused by an **intensification** of selective forces.
  - *Post-hoc* analyses (e.g., estimate branch-level dN/dS and then compare [t-test, etc] them as if they were observed quantities) discard a lot of information (e.g., variance of individual estimates), and make obviously wrong assumptions (e.g., estimates are uncorrelated).

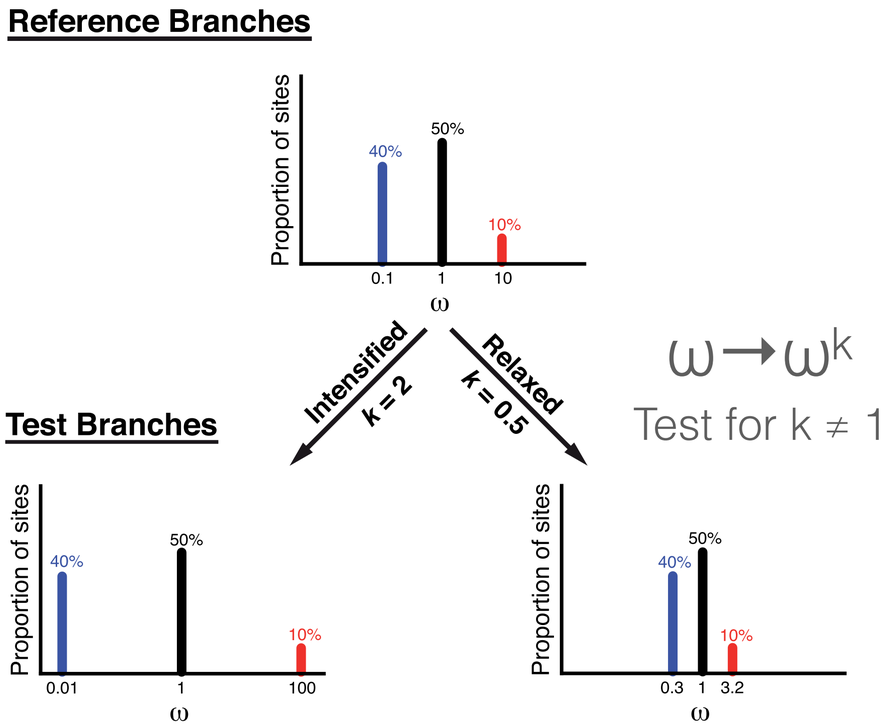


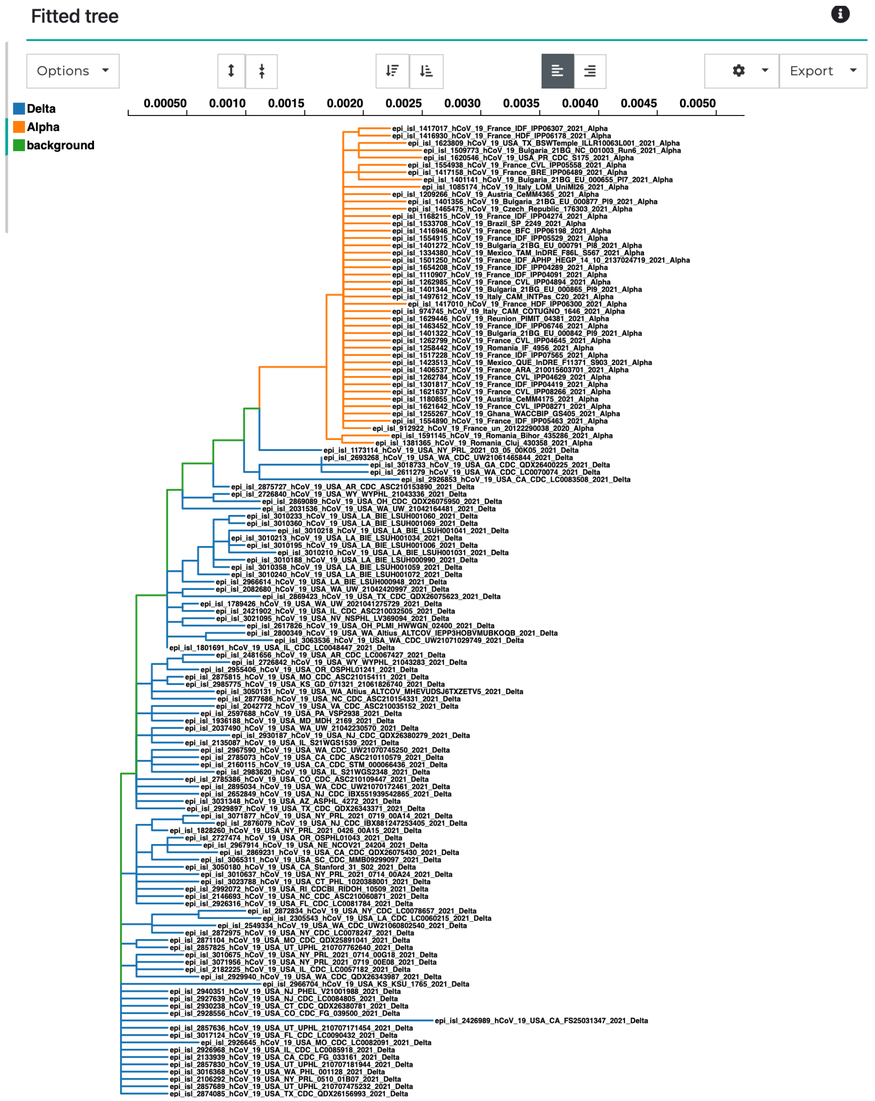
Table 1. Test for Relaxed Selection Using RELAX in Various Taxonomic Groups.

Taxa	Gene/Genes	Test Branches	Reference Branches	k[a]	*P*-Value
γ-proteobacteria	Single-copy orthologs	Primary/secondary endosymbionts	Free-living γ-proteobacteria	0.30	< 0.0001
		Primary endosymbionts	Free-living γ-proteobacteria	0.28	< 0.0001
		Secondary endosymbionts	Free-living γ-proteobacteria	0.61	< 0.0001
		Primary endosymbionts	Secondary endosymbionts	0.56	< 0.0001
Bats	*SWS1*	HDC echolocating and cave roosting (pseudogenes)	LDC echolocating and tree roosting (functional genes)	0.16	< 0.0001
		LDC echolocating	Tree roosting	1.07	0.577
	M/LWS1	HDC echolocating and cave roosting	LDC echolocating and tree roosting	0.70	0.495
		Echolocating species	Tree- and cave-roosting species	0.21	0.0005
		HDC echolocating	LDC echolocating	0.84	0.427
Bornavirus	Nucleoprotein	Endogenous viral elements	Exogenous virus	0.02	< 0.0001
Daphnia pulex	Mitochondrial protein-coding genes	Asexual	Sexual	0.63	< 0.0001

[a]Estimated selection intensity.

Comparing alpha vs delta clades in SARS-CoV-2

- Are selective pressures on the Delta SARS-CoV-2 clade relaxed or intensified compared to the Alpha clade?
- Partition the tree into corresponding clades.
- See http://www.hyphy.org/tutorials/CL-prompt-tutorial/#preparing-labeled-phylogenies for how to label phylogenies



RELAX(ed selection test)

results summary

INPUT DATA | **AlphaDeltaSpike.fas** | **133** sequences | **1273** sites

Export

Test for selection **intensification** (K = **1.31**) was **not significant** (p = **0.558**, LR = **0.34**).

See here for more information about this method.

Please cite PMID 123456789 if you use this result in a publication, presentation, or other scientific work.

Model fits

Model	*log* L	#. params	AIC_c	Branch set	ω_1	ω_2	ω_3
General descriptive	-8790.1	367	18315.9	Shared	0.00 (11.36%)	0.86 (88.62%)	1288.13 (0.02%)
RELAX alternative	-8876.3	199	18151.1	Reference	1.00 (97.76%)	1.00 (2.24%)	1450.96 (0.00%)
				Test	1.00 (97.76%)	1.00 (2.24%)	13744.44 (0.00%)
RELAX null	-8876.5	198	18149.4	Reference	1.00 (98.05%)	1.00 (1.95%)	11625.16 (0.00%)
				Test	1.00 (98.05%)	1.00 (1.95%)	11625.16 (0.00%)

```
hyphy relax --alignment data/AlphaDeltaSpike.fas --tree data/AlphaDeltaSpike.nwk --test Delta --reference Alpha  --starting-points 5
```

Which sites are evolving differentially?

- We have established that in the HIV example, donor, recipient, and transmission branches evolve differently.
- Can we identify specific sites where this may be occurring?
  - Why is this of interest?
- More generally, given a tree with N sets of branches, we fish to find sites where evolution is different between these N sets, with a degree of statistical significance.

- Solution: use a fixed effects method (Contrast-FEL)
  - For each branch set **i**, estimate a dN/dS ration (**N** total ratios)
  - Test whether or not any of the ratios are different (group test)
  - For each pair of ratios, test if they are different [up to N (N-1) / 2 tests]
- Can identify subtle differences among selective pressures.

Molecular Biology and Evolution, Volume 38, Issue 3, March 2021, Pages 1184–1198,

Contrast-FEL

results summary

INPUT DATA | AlphaDeltaSpike.fas | 133 sequences | 1273 sites

Export

Contrast-FEL **found evidence** of

Found 0 sites with different dN/dS

with q-value threshold of 0.2 .

See here for more information about this method.

Please cite PMID 15703242 if you use this result in a publication, presentation, or other scientific work.

ContrastFEL Table

Showing entries 1 through 20 out of 1273.

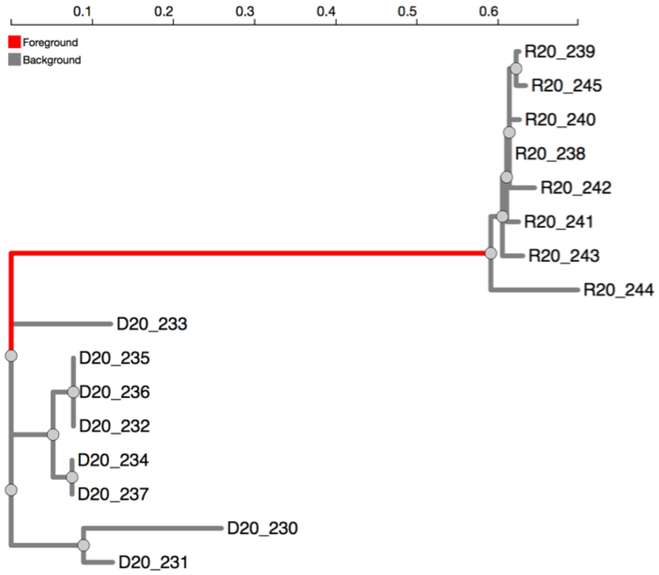
Export Table to CSV

Site	Partition	alpha	beta (background)	beta (Delta)	beta (Alpha)	subs (Delta)	subs (Alpha)	P-value (overall)	Q-value (overall)	Permutation p-value	b
1118	1	0.000	3.764	0.553	281.807	0.000	1.000	0.005	1.000	1.000	
1191	1	0.091	8.275	0.280	275.724	0.000	1.000	0.071	1.000	-1.000	
70	1	4.841	23.558	0.000	869.606	0.000	1.000	0.081	1.000	-1.000	

```
hyphy contrast-fel --alignment data/AlphaDeltaSpike.fas --tree data/AlphaDeltaSpike.tree --branch-set Alpha --branch-set Delta
```

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 (8% vs 1%), 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

Task	Test	Site strategy	Branch strategy	Complexity	Effective sample size	Parallelization	Practical # sequences limit
Gene-wide selection	**BUSTED**	Random Effects	Random Effects	Fixed	~sites x taxa	SMP	~2,000
Site-level selection / episodic	**MEME**	Fixed Effects	Random Effects	Fixed	~ taxa	SMP/MPI	~25000 (cluster)
Site-level selection / pervasive	**FEL**	Fixed Effects	Fixed Effects	Fixed	~ taxa	SMP/MPI	~25000 (cluster)
Branch-level selection	**aBSREL**	Random Effects	Fixed Effects	Adaptive	~ sites	SMP/MPI	~ 1,000
Compare selective regimes between sets of branches	**RELAX**	Random Effects	Mixed Effects	Fixed	~sites x (branch set size)	SMP	~ 1,000
Compare selective pressure between sets of branches on individual sites	**Contrast-FEL**	Fixed Effects	Fixed Effects	Fixed	~ branch set	SMP/MPI	~25000 (cluster)

Current suggested best practices.

There are lots of methods you could use to study positive selection, including >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
Intra-species viral analyses for sites under selection	MEME/FEL internal branches	• 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.

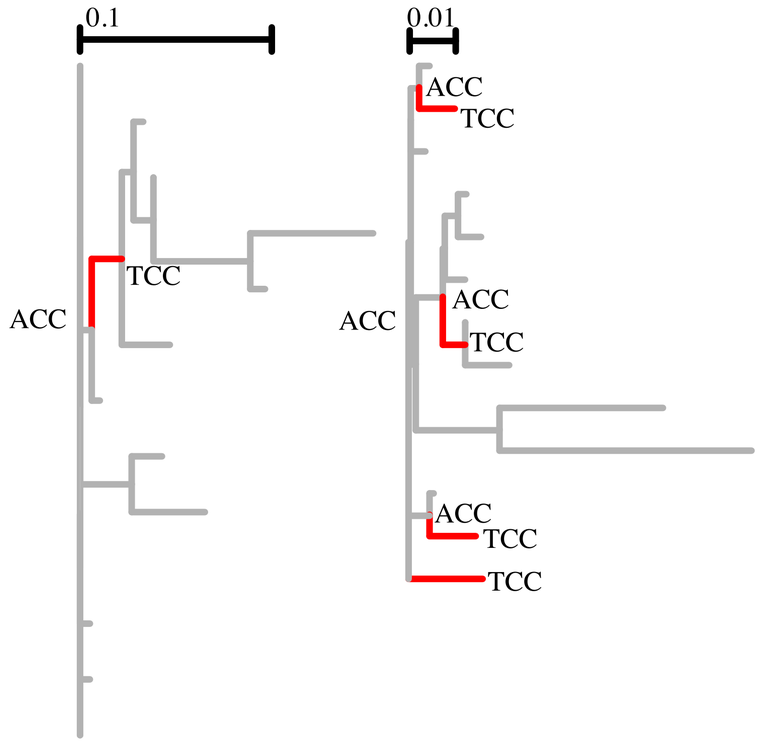


Figure 4.2: The effect of recombination on inferring diversifying selection. Reconstructed evolutionary history of codon 516 of the Cache Valley Fever virus glycoprotein alignment is shown according to GARD inferred segment phylogeny (left) or a single phylogeny inferred from the entire alignment (right). 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).

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

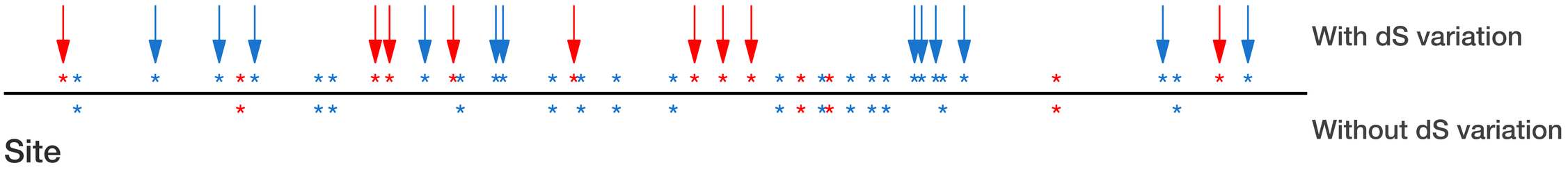
Table 4. Effect of correcting for recombination when using fixed effects likelihood to detect positively selected sites.

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.

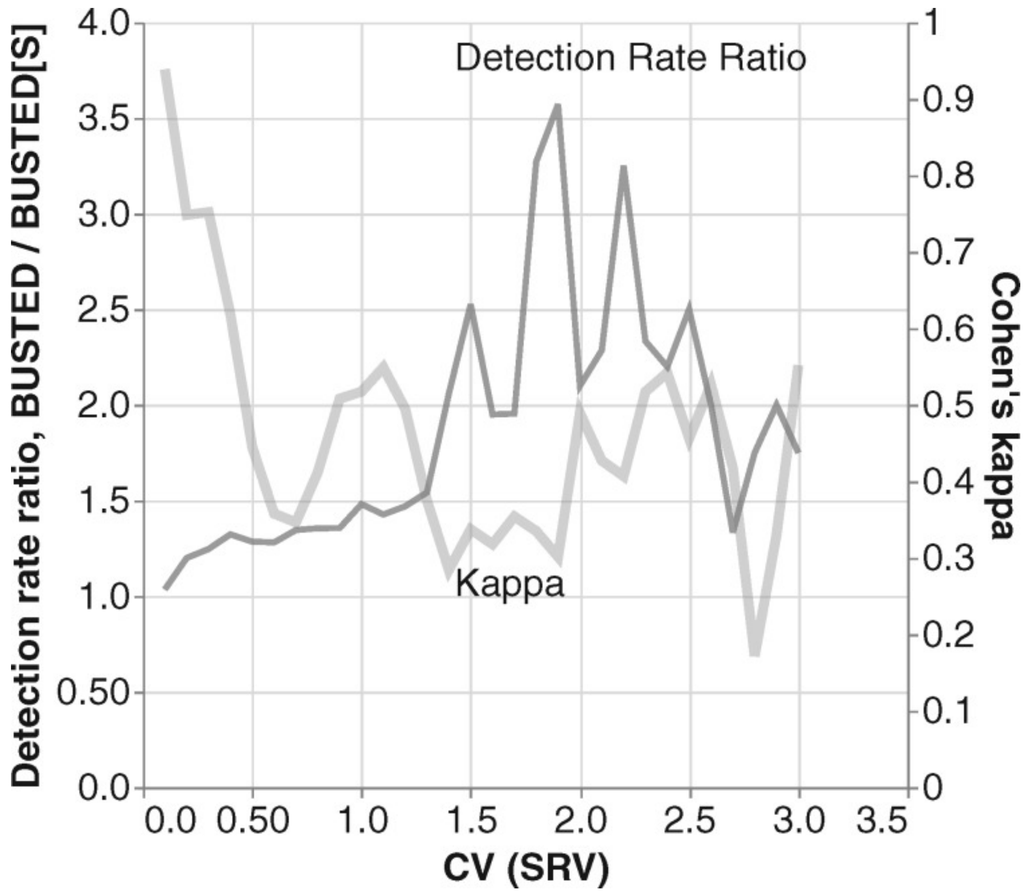
Mol. Biol. Evol. | 23(10):1891-901 | 2006

Sites detected by FEL with and without dS variation



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 and hight false positive rates
- Most of HyPhy methods provide support for including dS



Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril

Sadie R Wisotsky [1 2], Sergei L Kosakovsky Pond [2], Stephen D Shank [2], Spencer V Muse [1 3]

Affiliations + expand

PMID: 32068869 PMCID: PMC7403620 DOI: 10.1093/molbev/msaa037

Free PMC article

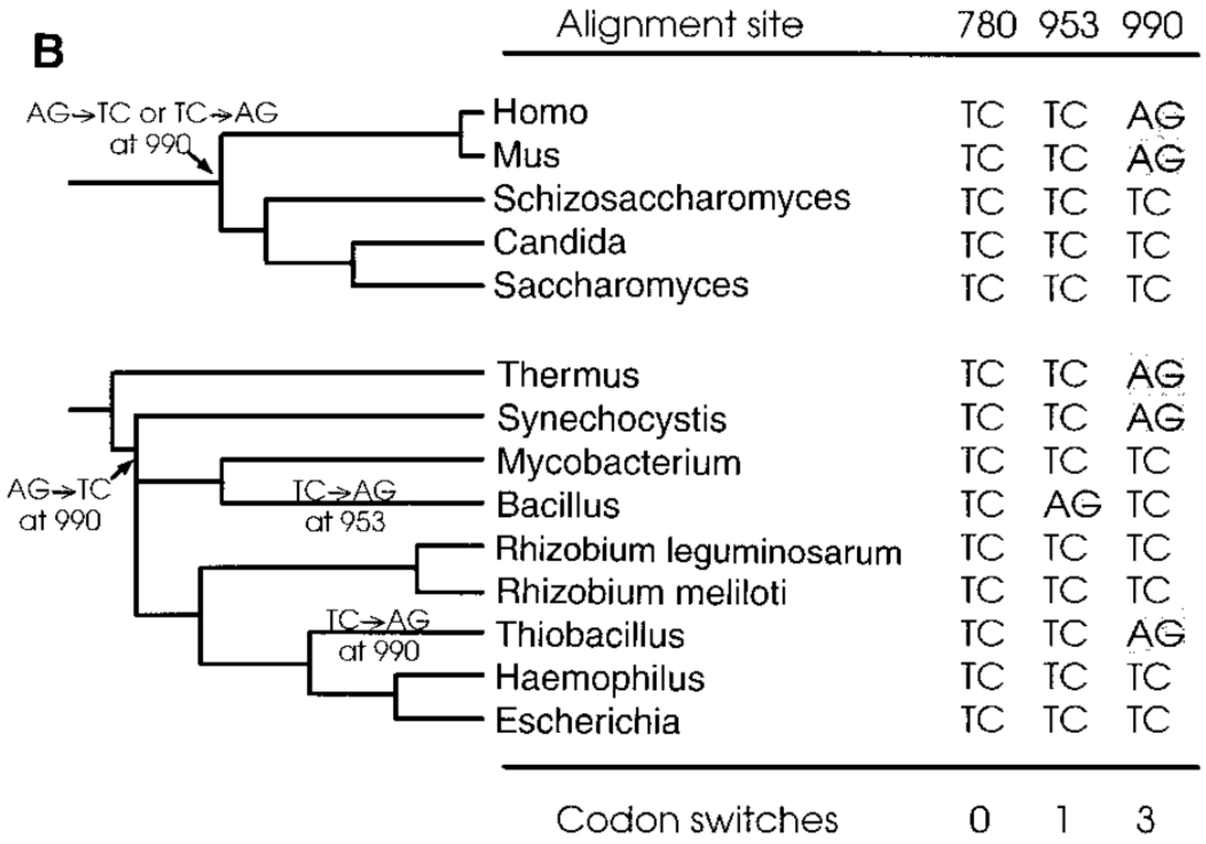
Mol. Biol. Evol. 22(12):2375–2385. 2005

Biological rationale for MH

Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions

Michalis Averof,[1]* Antonis Rokas,[2] Kenneth H. Wolfe,[3] Paul M. Sharp[4]*

Point mutations are generally assumed to involve changes of single nucleotides. Nevertheless, the nature and known mechanisms of mutation do not exclude the possibility that several adjacent nucleotides may change simultaneously in a single mutational event. Two independent approaches are used here to estimate the frequency of simultaneous double-nucleotide substitutions. The first examines switches between TCN and AGY (where N is any nucleotide and Y is a pyrimidine) codons encoding absolutely conserved serine residues in a number of proteins from diverse organisms. The second reveals double-nucleotide substitutions in primate noncoding sequences. These two complementary approaches provide similar high estimates for the rate of doublet substitutions, on the order of 0.1 per site per billion years.



Science, Feb. 2000

Allowing multi-nucleotide substitutions

- Some of the methods (e.g. BUSTED, aBSREL, RELAX) can extend substitution models to allow instantaneous double- and triple-“hits” (e.g. ACC to AGG)
- Sometimes multi-nucleotide changes along short branches at a single site can drive selection signal (possible false positives?)
- HyPhy includes a simple standard analysis for estimating alignment-wide multiple-hit rates.

Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes

Alexander G. Lucaci, Sadie R. Wisotsky, Stephen D. Shank, Steven Weaver, Sergei L. Kosakovsky Pond

Published: March 12, 2021 • https://doi.org/10.1371/journal.pone.0248337

See the preprint

JOURNAL ARTICLE

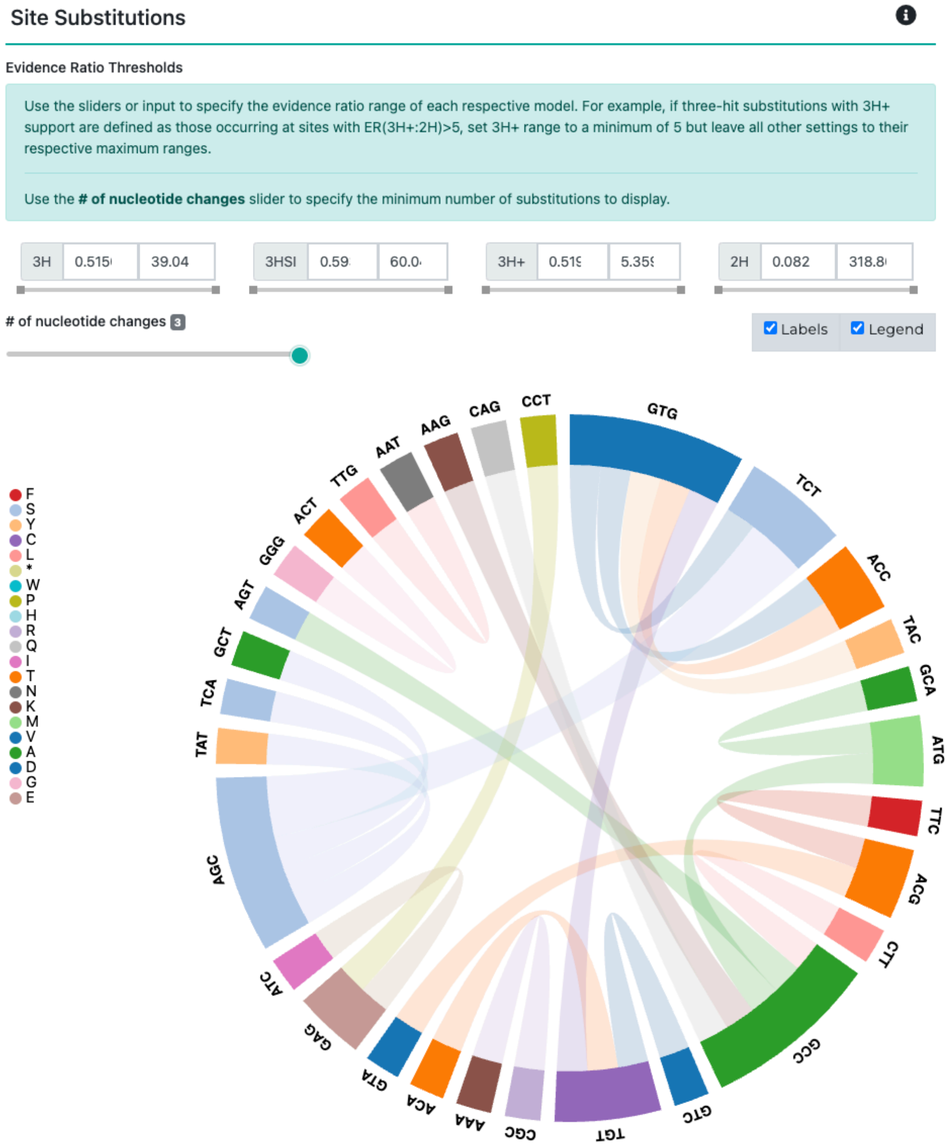
Evolutionary Shortcuts via Multinucleotide Substitutions and Their Impact on Natural Selection Analyses

Alexander G Lucaci, Jordan D Zehr, David Enard, Joseph W Thornton, Sergei L Kosakovsky Pond

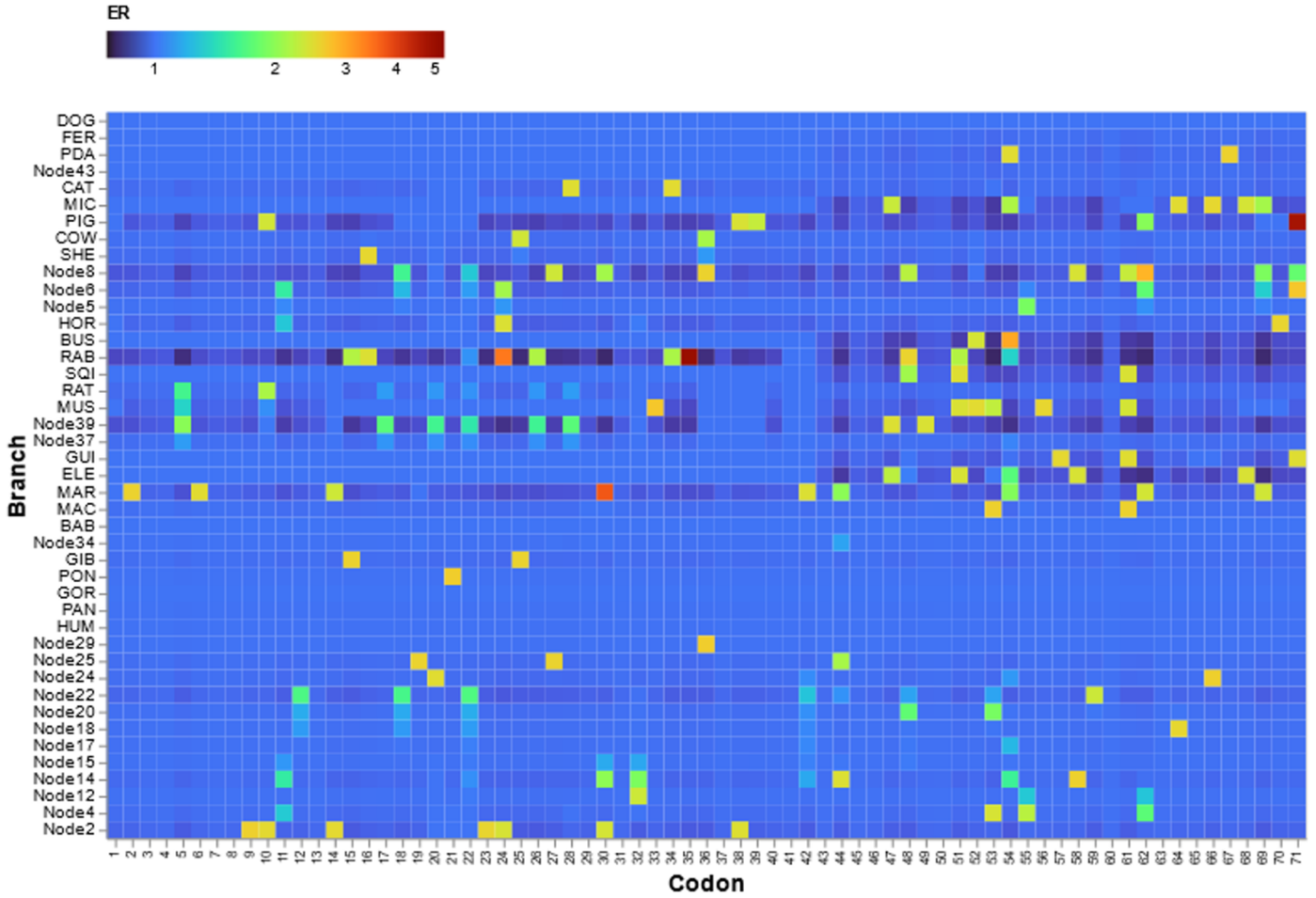
Molecular Biology and Evolution, Volume 40, Issue 7, July 2023, msad150, https://doi.org/10.1093/molbev/msad150

Published: 03 July 2023

Examining multinucleotide substitutions

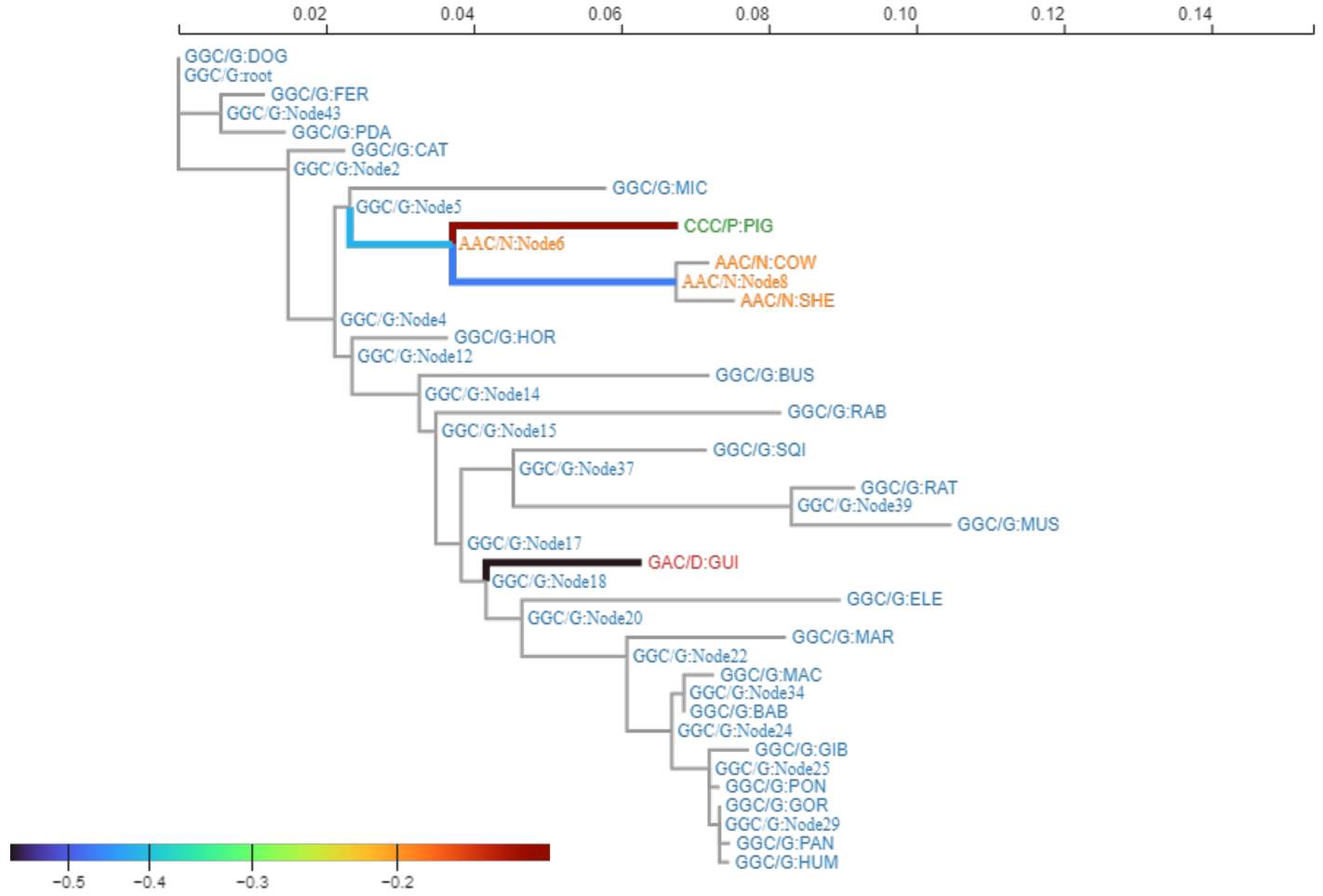


Examining support for selection in RNA Binding Motif Protein 3 (RBM3)

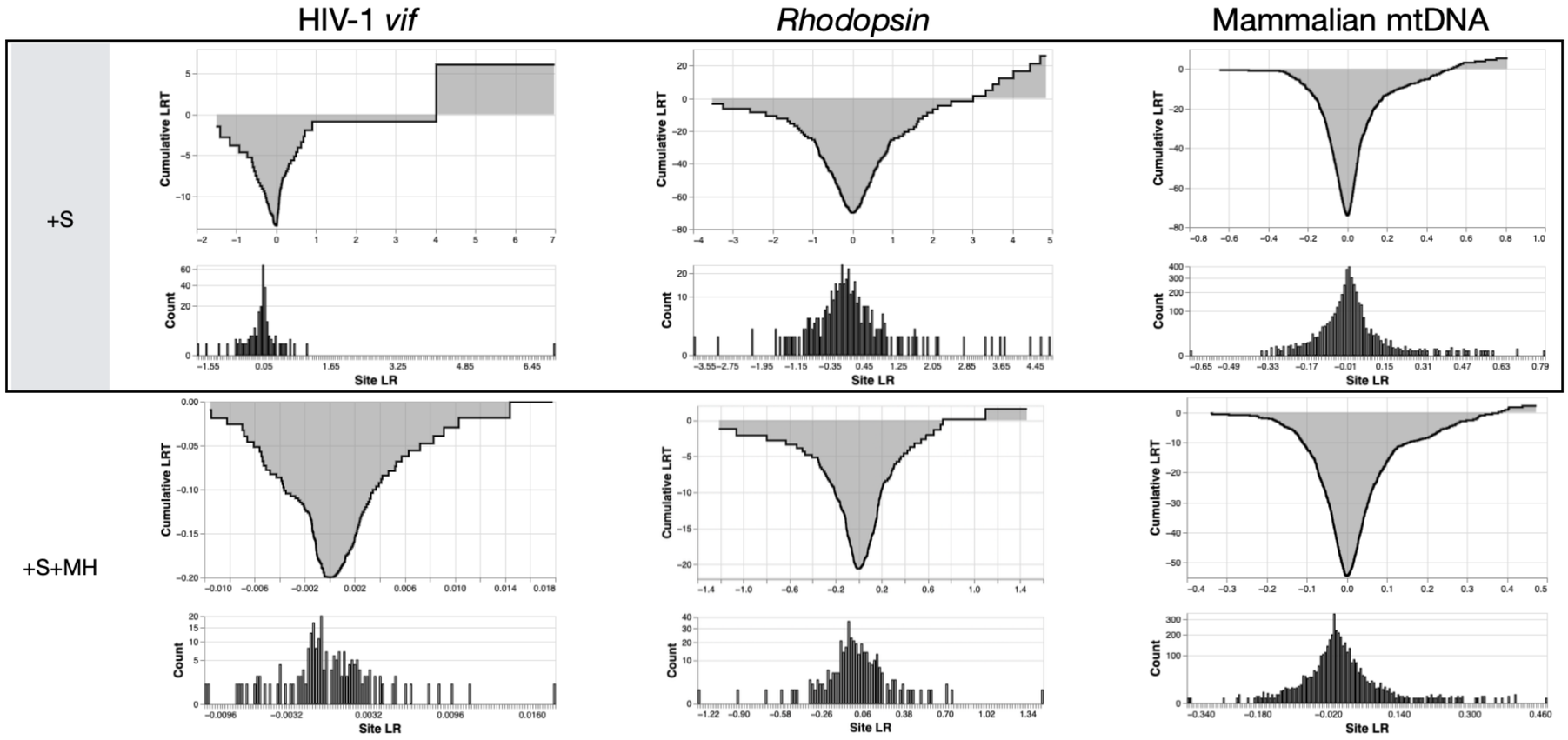


...

RBM3, Codon 71, A difference in selection



Site-level support for Episodic Diversifying Selection in three benchmark alignments.



Break

Using sensitive methods for positive selection detection to detect *(and filter out)* coding sequence alignment errors

HyPhy
Hypothesis Testing using Phylogenies

HyPhy

- home
- news and releases
- about
- download
- installation
- getting started
- methods
- tutorials
- batch language
- resources

Hypothesis Testing using Phylogenies

An open-source software package

for comparative sequence analysis
using stochastic evolutionary models

Download

- Evolutionary analyses of multiple sequence aligmments
- Specific focus on coding sequence evolution

Datamonkey

A Collection of State of the Art Statistical Models and Bioinformatics Tools

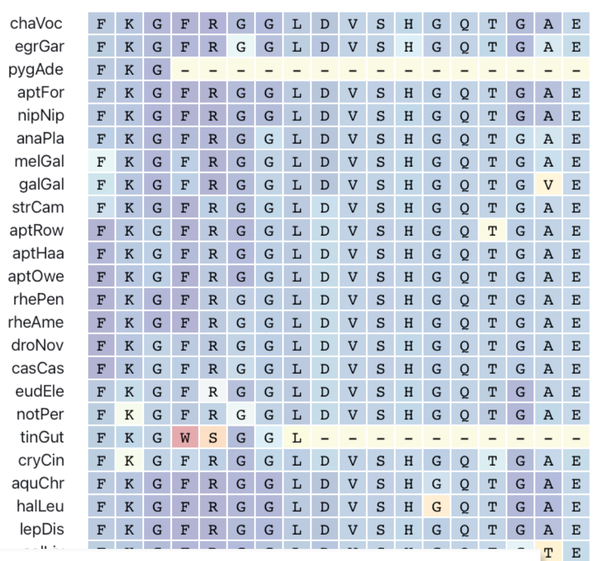
What evolutionary process would you like to detect?

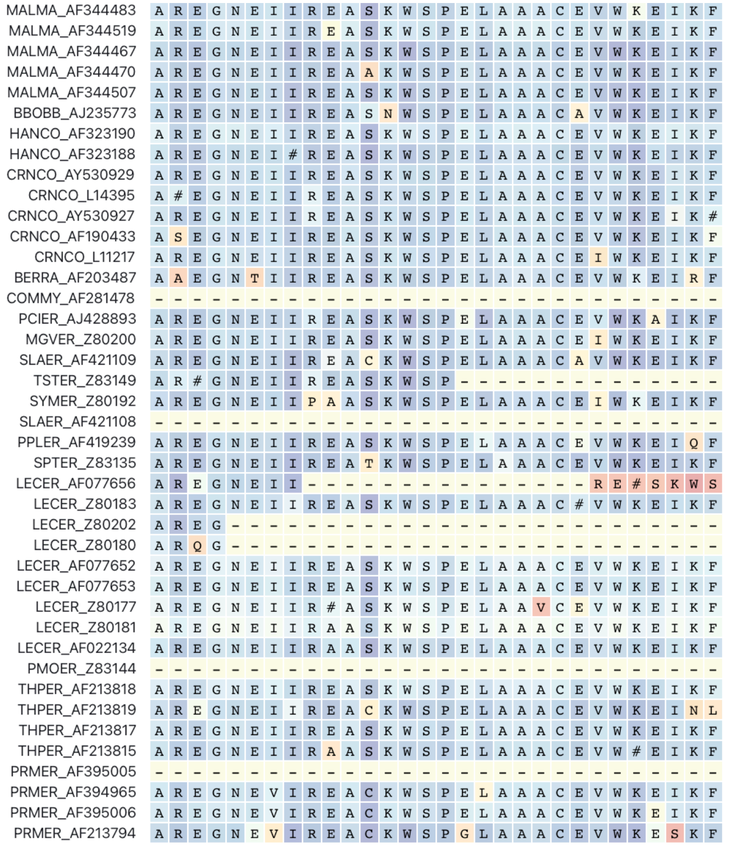
Selection	Recombination

Datamonkey is funded jointly by **MIDAS** and **NIH award R01 GM093939**

- dN/dS analyses on coding sequence MSAs are ubiquitous (1000s of papers each year): very useful biological insights.
- A very large fraction of analyzed MSAs contain “local” alignment errors.
- Sensitive statistical methods often interpret these errors as evolutionary signal.
- The problem is only exacerbated in larger MSAs with more divergent species
- Can we turn the sensitivity of these methods to our advantage to find and “filter” the errors.
- Based on very simple intuition and extensions of “tried-and-true” methods.

Sample alignments from recently published papers





These 45,367 hierarchical orthologous groups, or HOGs, were filtered to retain 16,151 HOGs with sequences for at least four species. Protein sequences were aligned with MAFFT v. 7.245 (Katoh and Standley, 2013), and filtered in three steps. First, entire columns were excluded if missing in more than 30% of species, had sequence in fewer than 10 taxa, or was missing in two of the three of the main taxonomic groups (paleognaths, neognaths, or non-avian outgroups). Second, poorly aligned regions were masked according to Jarvis et al. (2014) using a sliding-window similarity approach. Third, columns were removed using the same criteria as the first round. Next, entire sequences were removed from each alignment if they were over 50% shorter than their pre-filtered length or contained excess gaps. Finally, entire HOGs were removed if they contained more than three sequences for any species, did not have more than 1.5x sequences for the given number of species present in the alignment, or were less than 100 base pairs long. Nucleotide sequences for all remaining HOGs were aligned with the codon model in Prank v. 150803 (Löytynoja and Goldman, 2008). In total, 11,247 HOGs remained after all alignment and filtering steps.

2021

After the application of “due diligence” alignment masking and filtering techniques

WDR72

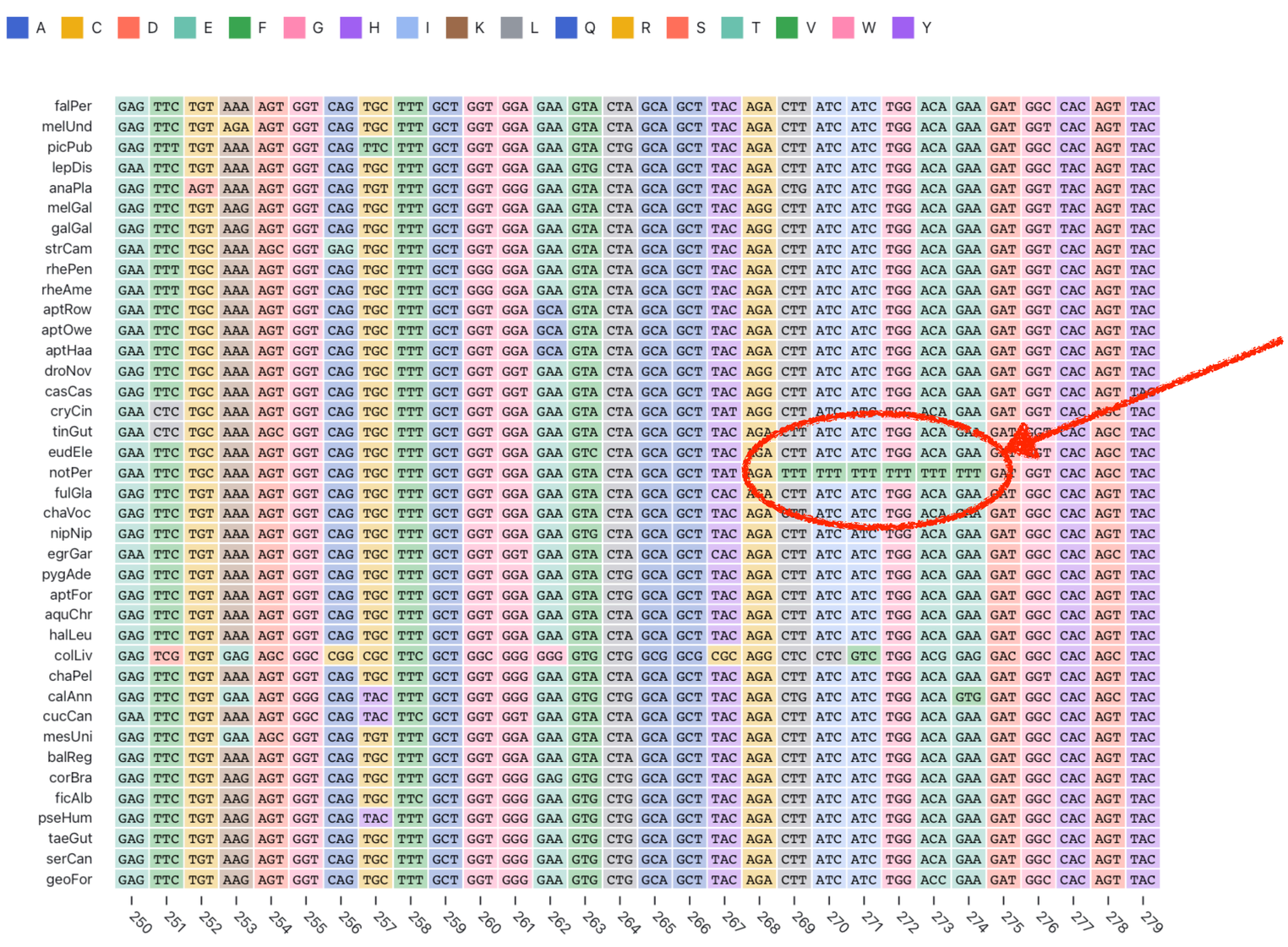
Research Article
Evolutionary Biology, Genetics and Genomics

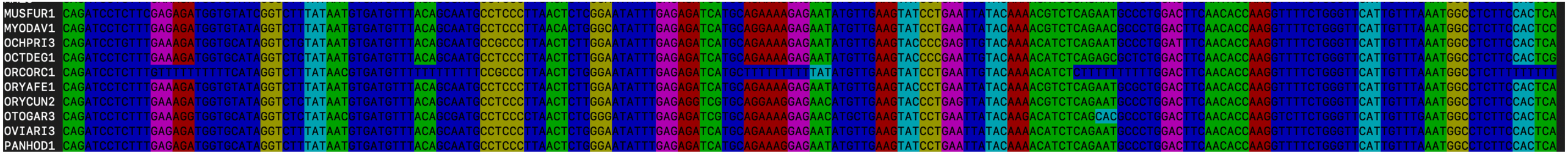
Immune genes are hotspots of shared positive selection across birds and mammals

Allison J Shultz, Timothy B Sackton

Harvard University, United States

Jan 8, 2019 · https://doi.org/10.7554/eLife.41815

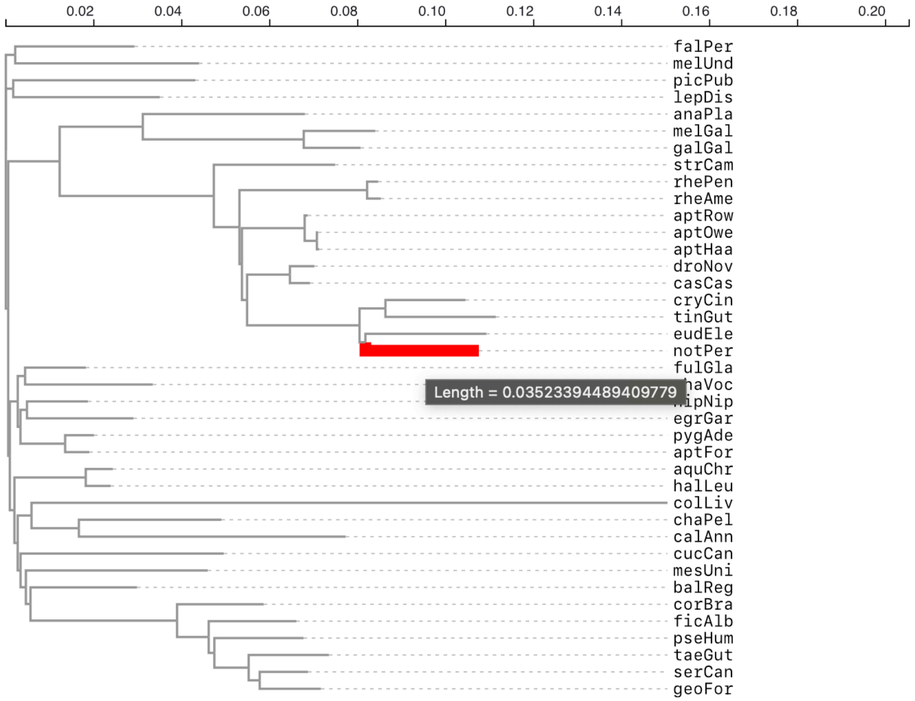




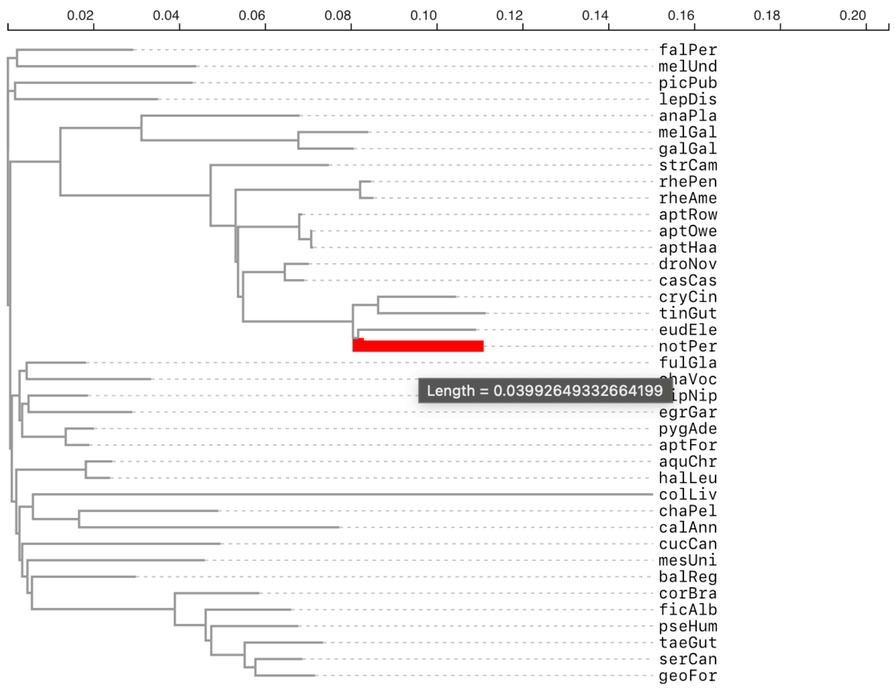
Can you identify regions of potential misalignment or sequencing-error?

BRANCH LENGTH

+original (0.035)



+error (0.040)



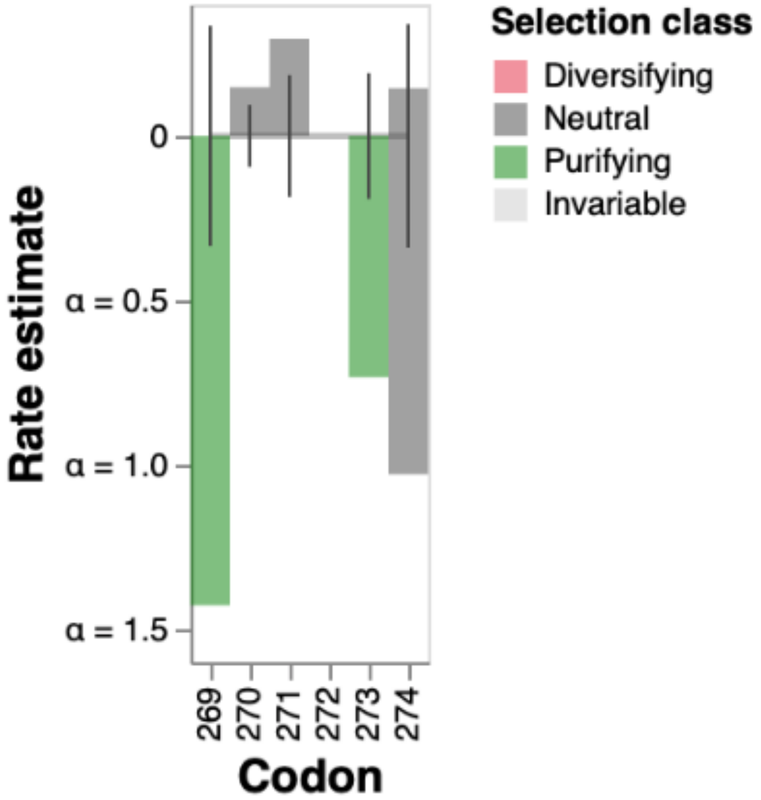
MEAN dN/dS

0.3245

0.3256

SITE-LEVEL RATES (FEL METHOD)

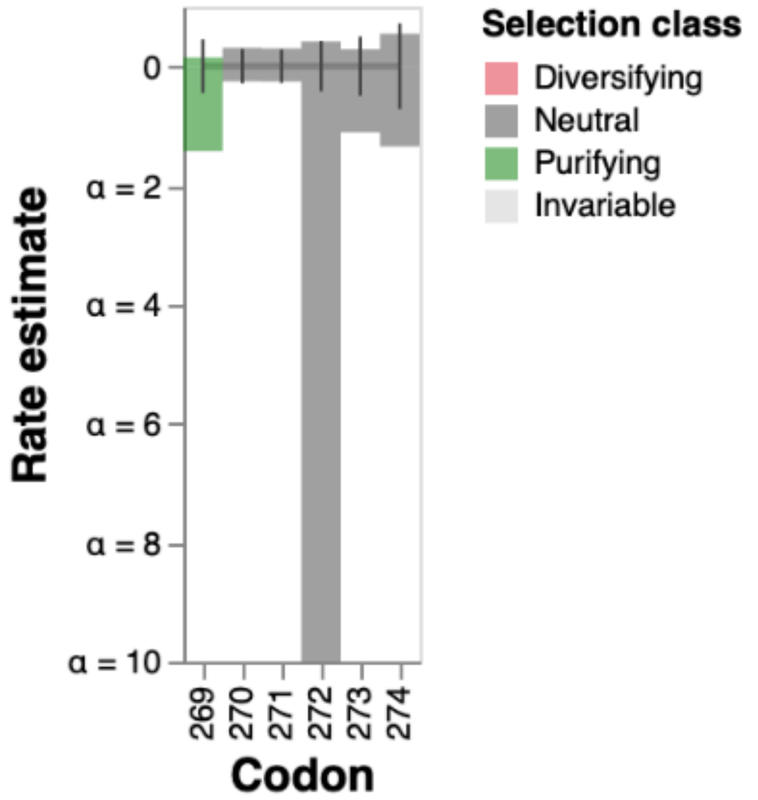
Original



Synonymous rate

Non-synonymous rate

+error



BRANCH-SITE RANDOM EFFECTS dN/dS METHOD

Original (no selection, p = 0.50)

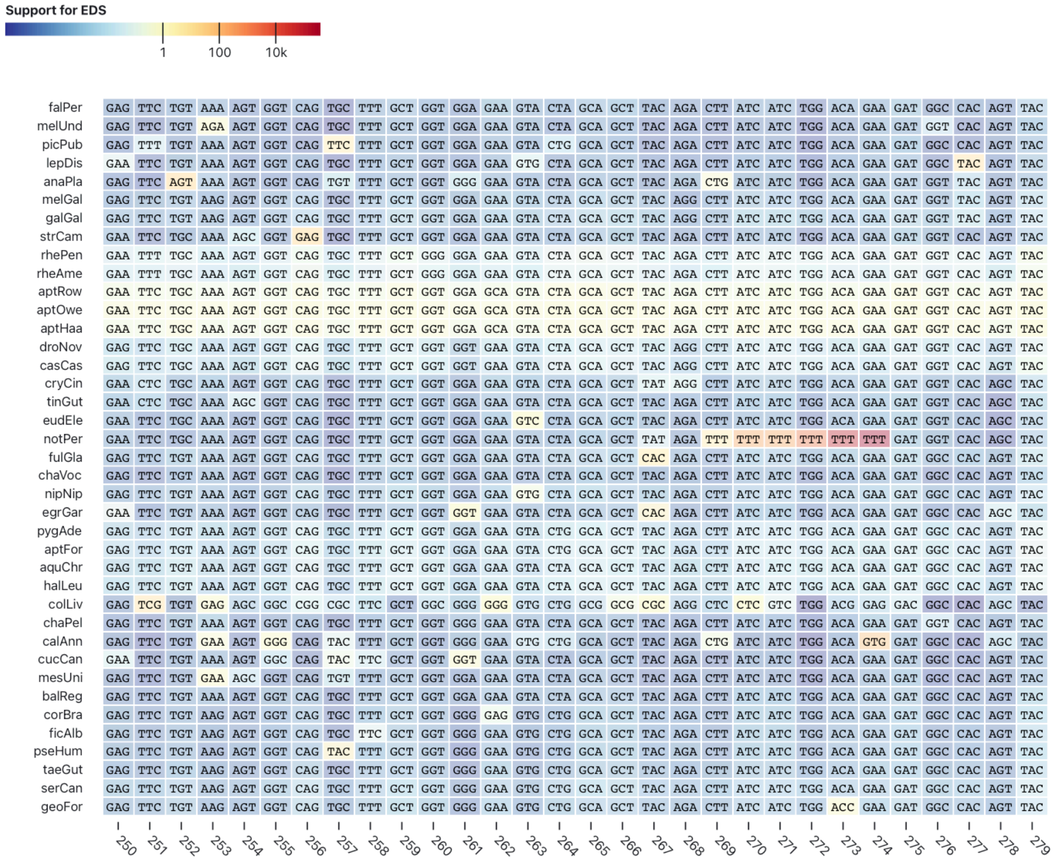
ω distribution

0.1333 (44.644%)
0.3708 (48.792%)
1.118 (6.5581%)

+error (selection, p = 0.0026)

ω distribution

0.02043 (28.451%)
0.4251 (71.522%)
127.8 (0.026700%)



JOURNAL ARTICLE

Gene-Wide Identification of Episodic Selection

Ben Murrell, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P. Martin, Davey M. Smith, Konrad Scheffler, Sergei L. Kosakovsky Pond
Author Notes

Molecular Biology and Evolution, Volume 32, Issue 5, May 2015, Pages 1365–1371,
https://doi.org/10.1093/molbev/msv035
Published: 19 February 2015

- Sensitive dN/dS (ω) models can “zoom-in” on small fractions of alignments
- Those could (should*) be real biological features indicative of accelerated evolutionary rates
- But they can (and often appare to be) artifacts due to MSA/sequence quality issues
- These apparent issues remain even after applying current state-of-the-art MSA/filtering tools (e.g. PRANK, PREQUAL) in this space
- Need a method to separate the signal from the noise
- What are some obvious **<u>evolutionary</u>** features of the noise?
- We developed this method out of necessity, because the signal was being overwhelmed by the noise in high-throughput screens for selection.

- Let’s examine another example (from Zoonomia), where we downsample the genes down to 16 species (using **treemer**)

- Two genes that show evidence of episodic positive diversifying selection

  - One likely “real”

  - One likely “not real”

  - How can we tell?

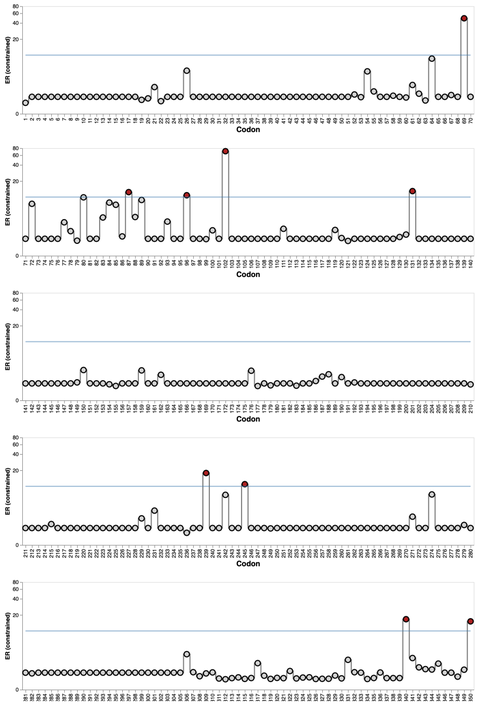
Standard BUSTED method results, p ≤ 0.001 for both

REAL

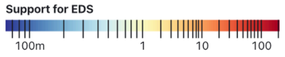
IQCF1

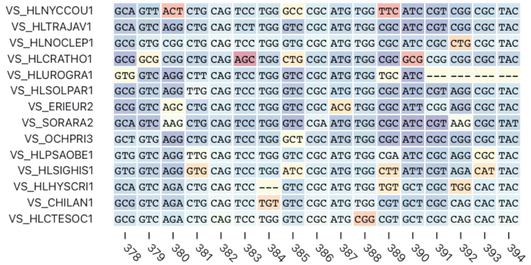
ω_1=0.4085 (79.597%)
ω_2=0.4078 (15.784%)
ω_3=11.16 (4.6197%)

Reasonable ω value with broad support



Nicely dispersed sites with evidence of selection





No clear issues with an area of the alignment where there's support for selection

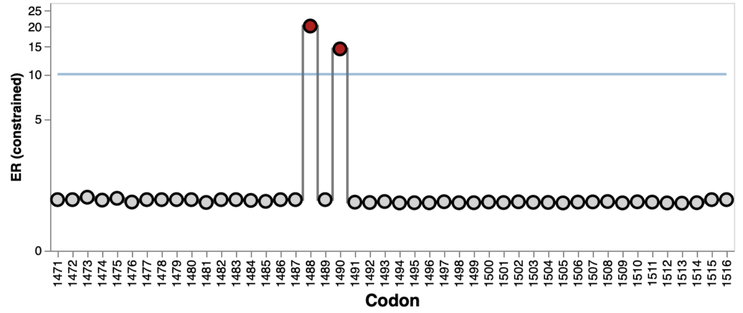
NOT REAL

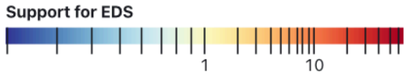
KRT8

ω_1=0.02107 (80.763%)
ω_2=0.6961 (18.533%)
ω_3=136.3 (0.70395%)

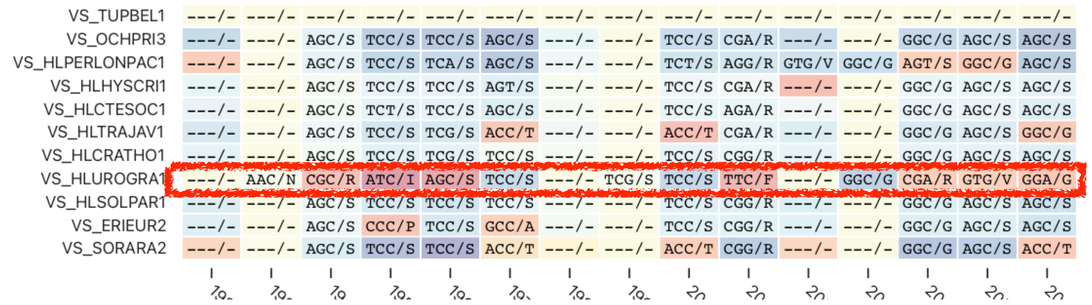
A large ω value with narrow support

Evidence of selection is a localized clump





Obvious alignment/ homology issues, here in one sequence



- The simple fix here is to include an **explicit error component** in the model
- We simply allow a small fraction of the aligment (e.g ≤ 1%) to evolve with abiologically high rates (e.g. dN/dS ≥ 100)
- This is in addition to the standard model which allows negative, neutral, and positive selection regimes.
- The primary goal of this analysis is to classify the selective regime on a gene (in the presence of specific types of errors)
- So error detection (and filtering) is a byproduct of an already useful analysis

Calculation of dN/dS

Multiple-sequence alignments from each gene family were back-translated into codon alignments to reconstruct phylogenetic trees using FastTree2 with default parameters. The entire workflow was executed using ETE3 (ref. [64]) with options ete3 build --nt-switch-threshold 0.0 --noimg --clearall --nochecks -w clustalo_default-none-none-none --no-seq-rename. For calculation of selective pressure per family we ran HyPhy using the BUSTED model[65] with default parameters, codon-based nucleotide alignment and the phylogenetic tree generated previously, retrieving the dN/dS ratio under the full codon model. We discarded gene families with dN/dS values higher than 0.5.

Article | Open access | Published: 18 December 2023

Functional and evolutionary significance of unknown genes from uncultivated taxa

Álvaro Rodríguez del Río, Joaquín Giner-Lamia, Carlos P. Cantalapiedra, Jorge Botas, Ziqi Deng, Ana Hernández-Plaza, Martí Munar-Palmer, Saray Santamaría-Hernando, José J. Rodríguez-Herva, Hans-Joachim Ruscheweyh, Lucas Paoli, Thomas S. B. Schmidt, Shinichi Sunagawa, Peer Bork, Emilia López-Solanilla, Luis Pedro Coelho & Jaime Huerta-Cepas

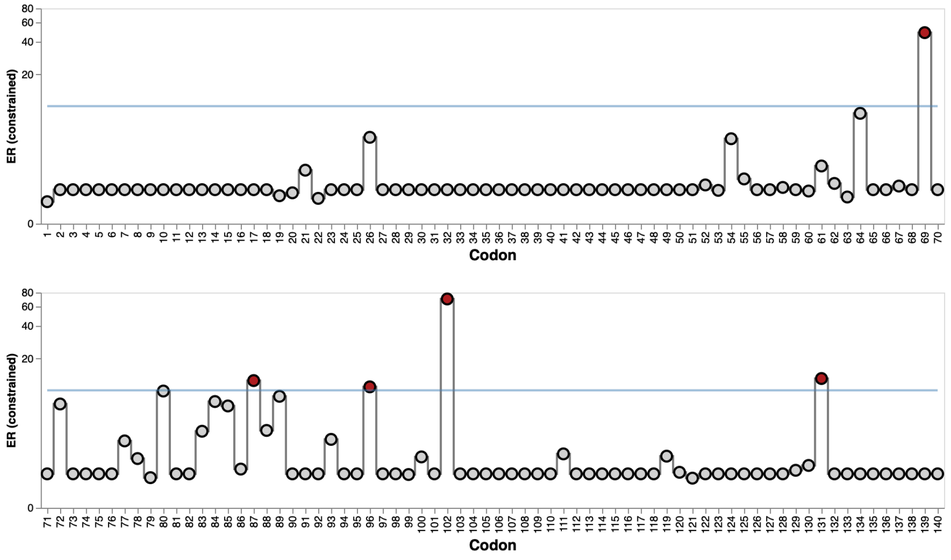
BUSTED-E (error corrected) results

REAL

IQCF1

ω_1=0.4063 (1.5589%)
ω_2=0.4076 (93.816%)
ω_3=11.09 (4.6249%)+
0.0% error

14 sequences in the alignment	**475** codon sites in the alignment	**1** partitions
25 median branches/partition used for testing	**3 classes** non-synonymous rate variation	**3 classes** synonymous rate variation
0.0000043 p-value for episodic diversifying selection	**10** Sites with ER≥10 for positive selection	**N/A:N/A** Multiple hit rates (2H:3H)
15 (branch, site) pairs with EBF ≥ 100	**10** Sites contributing most signal to EDS detection	**N/A:N/A** Expected fractions of MH subs (2H:3H)



NOT REAL

KRT8

ω_1 = 0.00001848 (11.039%)
ω_2 = 0.04232 (75.696%)
ω_3 = 1.000 (12.675%)+
0.58% error (ω_e >10^6)

11 sequences in the alignment	**1516** codon sites in the alignment	**1** partitions
18 median branches/partition used for testing	**3 classes** non-synonymous rate variation	**3 classes** synonymous rate variation
0.50 p-value for episodic diversifying selection	**0** Sites with ER≥10 for positive selection	**N/A:N/A** Multiple hit rates (2H:3H)
N/A (branch, site) pairs with EBF ≥ 100	**N/A** Sites contributing most signal to EDS detection	**N/A:N/A** Expected fractions of MH subs (2H:3H)

- What does this do overall for (tree-wide) positive selection detection rates?
    - It’s reduced quite a bit!

Research Article
Evolutionary Biology, Genetics and Genomics

Immune genes are hotspots of shared positive selection across birds and mammals

Allison J Shultz, Timothy B Sackton
Harvard University, United States

Jan 8, 2019 · https://doi.org/10.7554/eLife.41815

- What does this do overall for (tree-wide) positive selection detection rates?
- **Without** the error component, **40.4%** of all genes have signal for selection
- **With** the error component, **4.5%** of all genes have signal for selection
  - Nearly a 10x reduction in yield
  - This is actually more in line with what biologists are comfortable with.

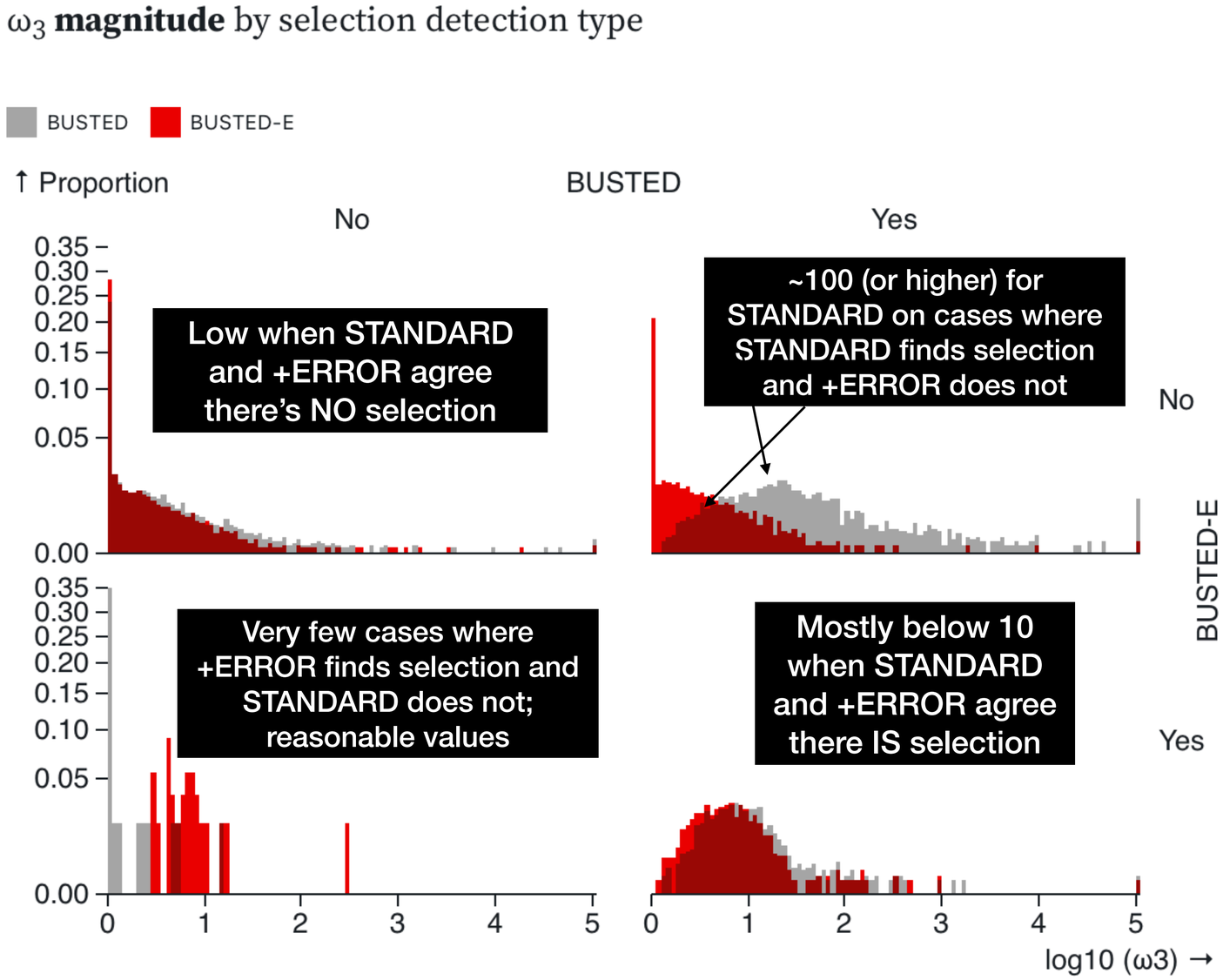
Some of the ZOONOMIA data

N of taxa	Alignments analyzed	Fraction (%) where selection was detected P ≤ 0.05 (FDR q ≤ 0.10)		
		BUSTED	BUSTED-E	FDR discovery reduction X
8	7062	15.9 (5.8)	4.7 (0.1)	58.0
16	7165	24.9 (17.7)	7.4 (0.7)	25.3
32	7229	36.3 (33.4)	10.5 (2.3)	14.5
64	3950	57.8 (58.7)	17.7 (9.0)	6.5

- What is a key feature of genes that may be falsely flagged as positively selected?
    - (Very) low proportions of (very) high omega values!

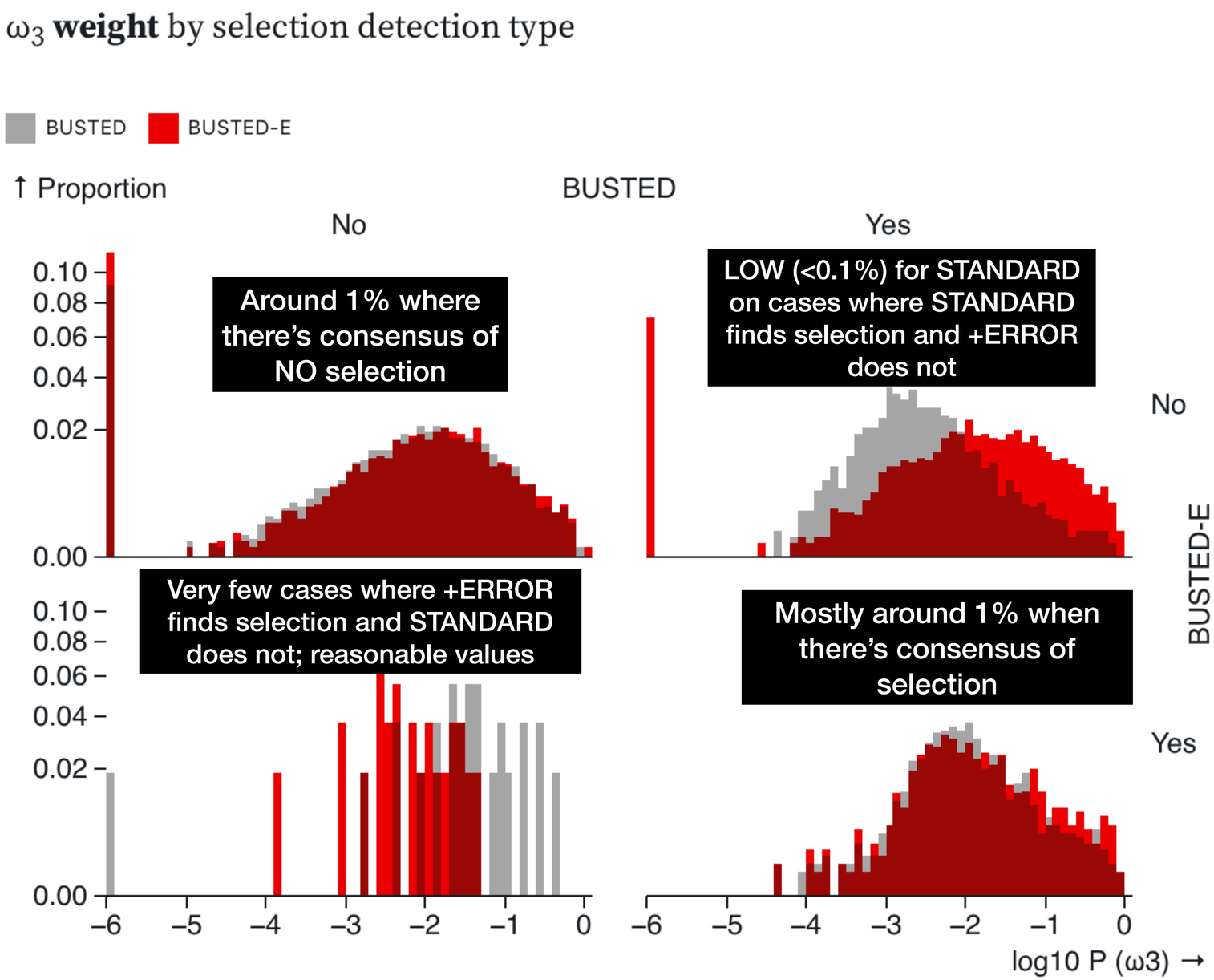
32-taxon Zoonomia tree

The magnitude of ω in the positive selection component



32-taxon Zoonomia tree

The weight of ω in the positive selection component

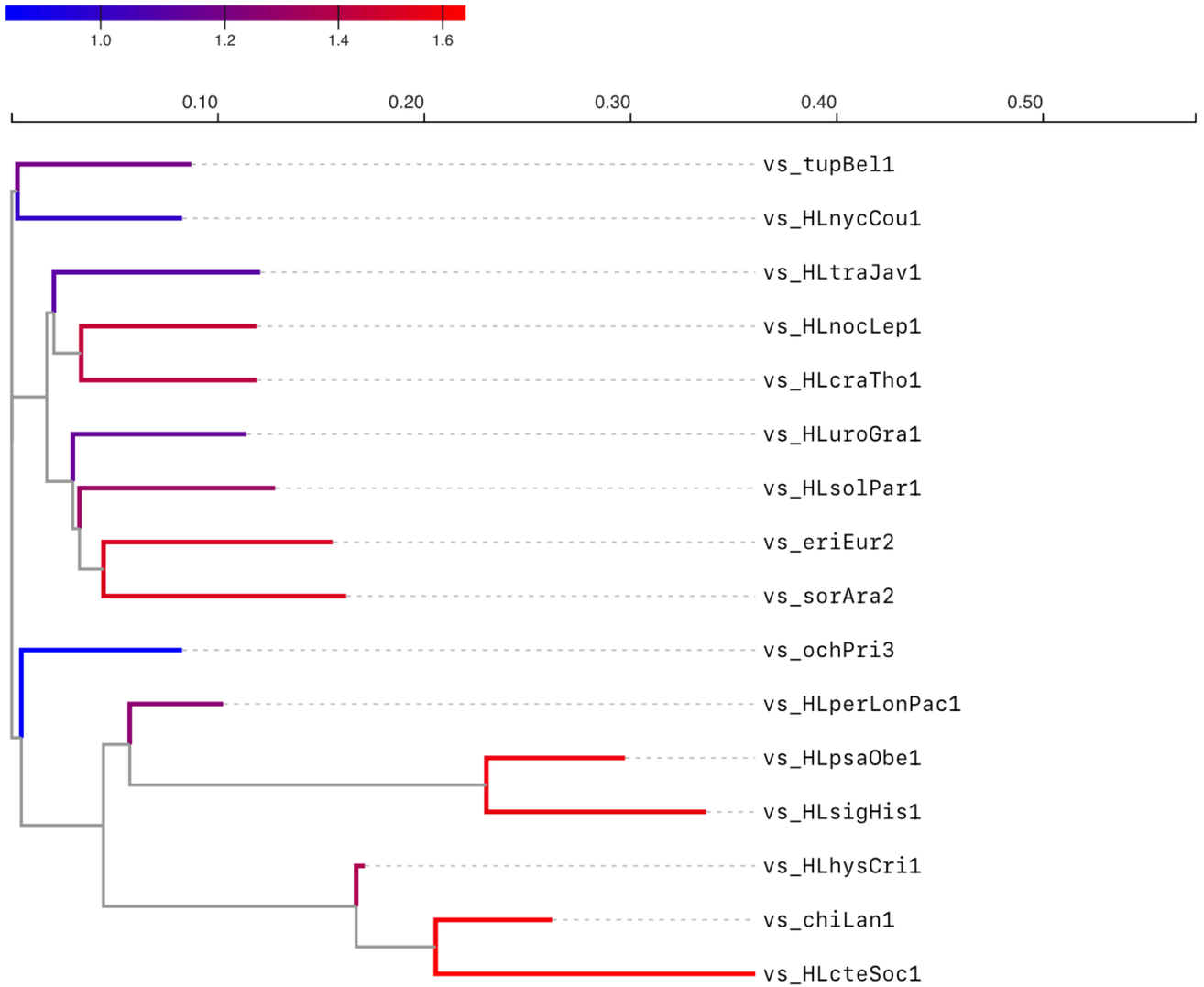


- BUSTED-E can annotate each input alignment with what it identifies as potentially erroneous codon positions (specific sequence, specific column)

- These can be further filtered (e.g. replaced with ‘- - -‘) for downstream analyses with other tools

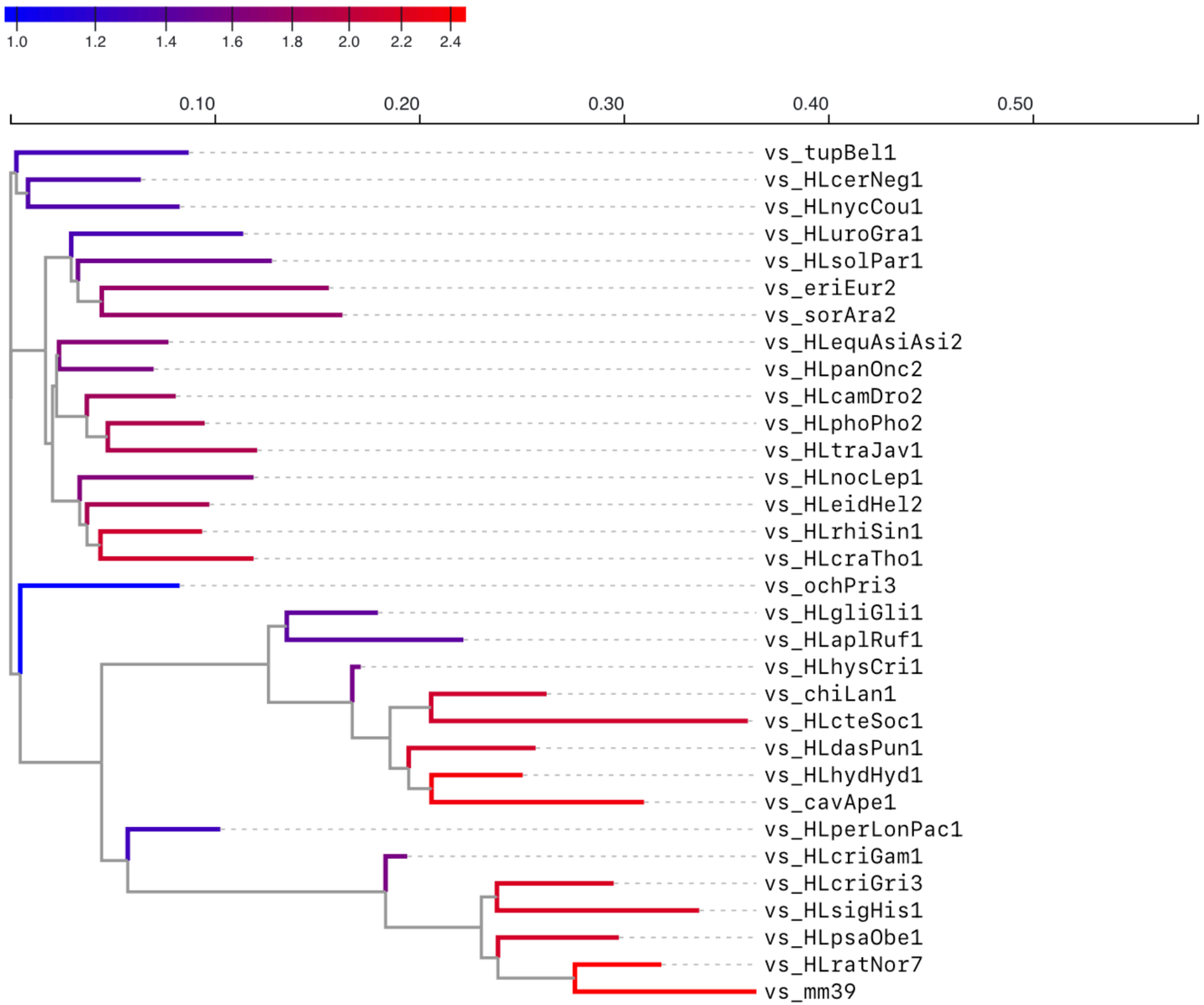
- The stringency of filtering is tunable

- Which species/genomes tend to have the most “putative” errors?



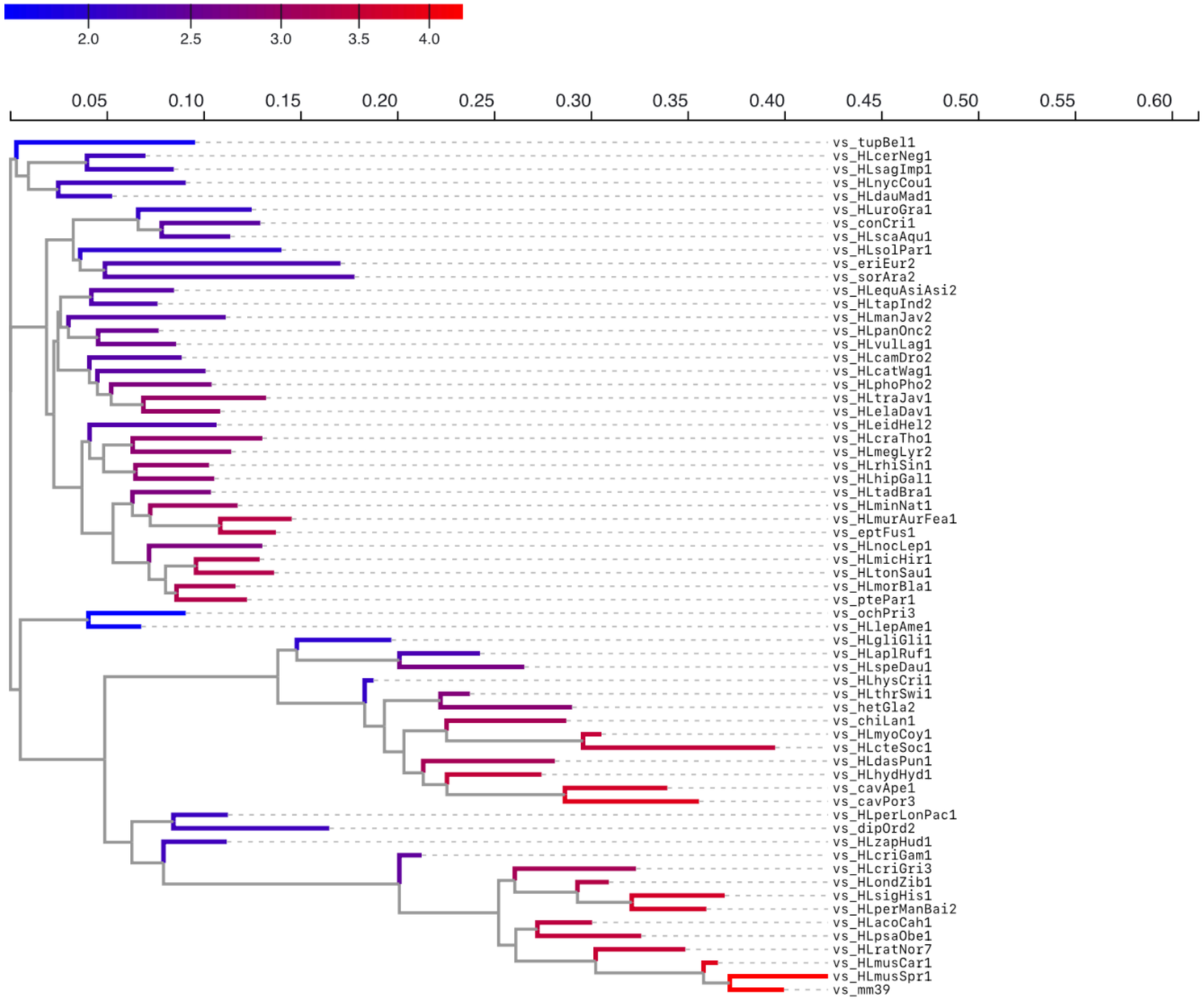
16-taxon Zoonomia tree
Codons filtered / 1000 codons

- Which species/genomes tend to have the most “putative” errors?



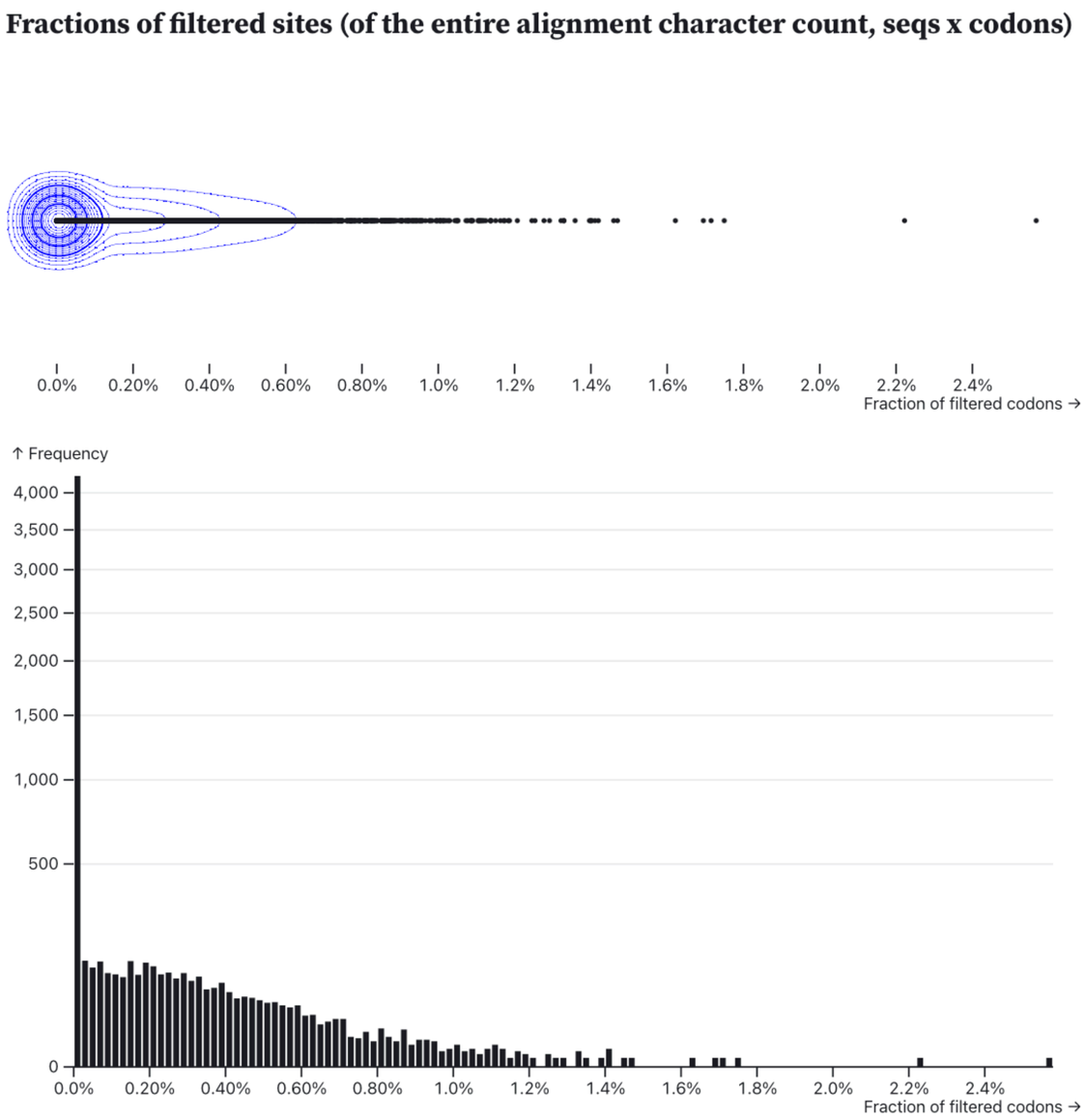
32-taxon Zoonomia tree
Codons filtered / 1000 codons

- Which species/genomes tend to have the most “putative” errors?



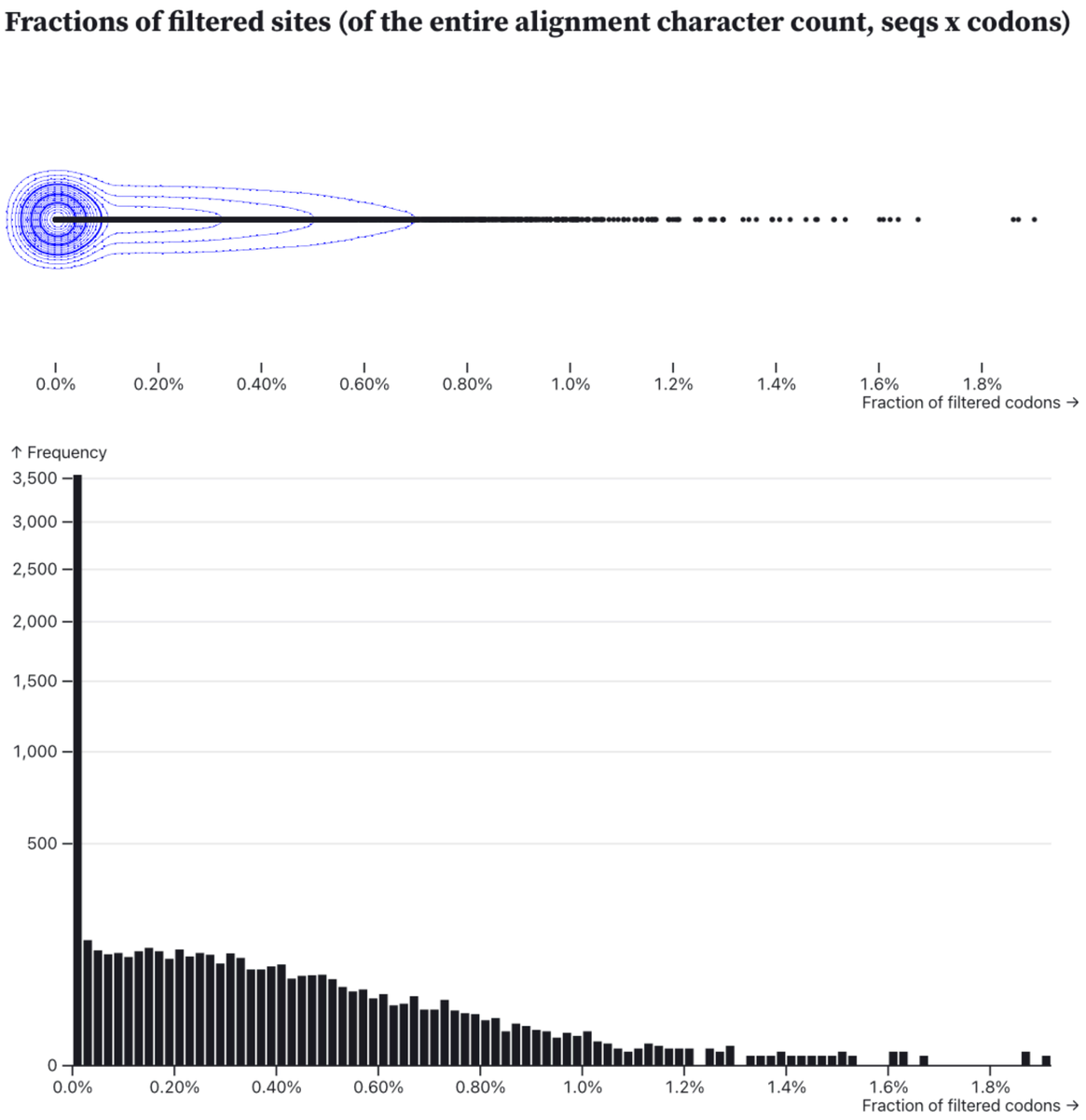
64-taxon Zoonomia tree
Codons filtered / 1000 codons

- The distribution of per MSA filtered codon counts



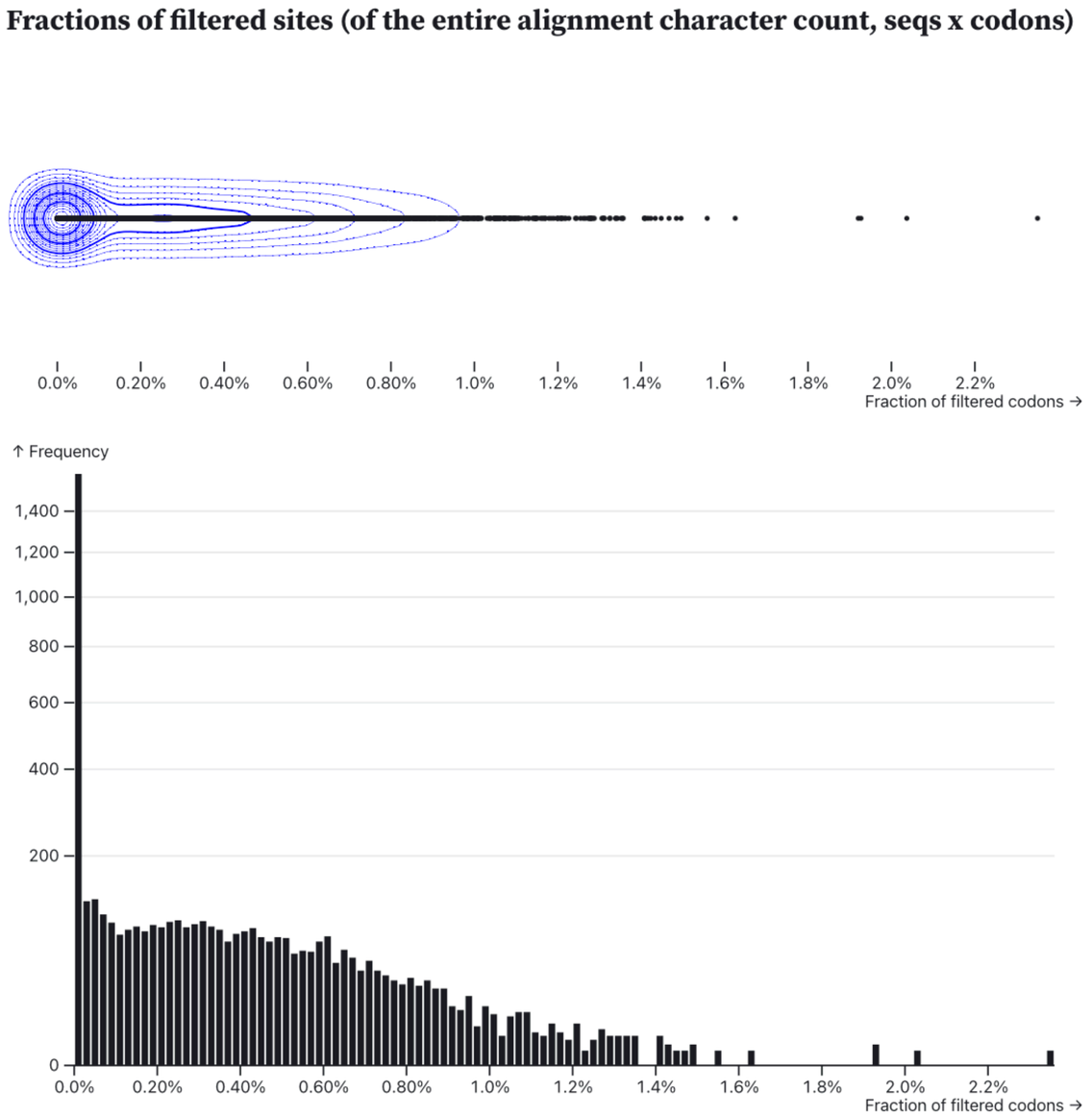
16-taxon Zoonomia tree
Codons filtered / 1000 codons

- The distribution of per MSA filtered codon counts



32-taxon Zoonomia tree
Codons filtered / 1000 codons

- The distribution of per MSA filtered codon counts



64-taxon Zoonomia tree
Codons filtered / 1000 codons

- dN/dS methods can be used to both handle some residual MSA error during selection screens and also find it in alignments
- The structure of the underlying substitution model can be adjusted to reflect more biological realism (e.g. MNM)
- Can identify specific genes and/or species which have higher putative errors

- Implemented in the HyPhy package
- Not computationally cheap, but delivers a useful result while screening for errors
  - The main cost growth dimension is the # of genomes
  - For 32 genomes, per MSA run time is on the order of 1-10 minutes per core

Break

Hands-on workshop component

- Getting familiar with evolutionary analyses
- Run data through the command-line
- Run data through Datamonkey
- View results on Hyphy Vision
- If there is time, we will write custom scripts to parse our results and plot.

Hands-on workshop component

- Choose a dataset (one of your own, or one of the examples we provide)
  - Use examples and code provided throughout the slides
- Run several different selection analyses on this file
- Use HyPhy Vision to compile results and visualizations
- Think about your results, do they make sense?

Assignment: Detecting selection pressures in viral genomes using HyPhy

Objective:

To understand how to use Hyphy for detecting selection pressures on viral genomes and to interpret the results in the context of viral evolution and pathogenesis.

Preparation

Ensure you have access to a computer with Hyphy installed.

Review the slides presented today on on the basics of molecular evolution and selection analyses and decide which analysis best fits your hypothesis

Each student or group will choose a specific viral gene dataset (e.g., from HIV, Influenza, SARS-CoV-2) available from databases like GenBank or specific viral genome repositories. I recommend - https://www.bv-brc.org/

Data retrieval

Download the nucleotide sequences of their assigned viral gene.

Align the sequences using tools like MACSE or codon-msa

It is important to generate codon-aware alignments, otherwise the selection analyses will fail.

Hyphy Analysis

- Students will go through the following steps in Hyphy
  - **Input Preparation:** Import the aligned sequences into Hyphy.
  - **Model Selection:** Select an appropriate substitution model.
  - **Selection Analysis:** Run selection analysis using methods such as SLAC (Single Likelihood Ancestor Counting), FEL (Fixed Effects Likelihood), and MEME (Mixed Effects Model of Evolution).

Results Interpretation

Students will interpret the results of their analysis, focusing on:

- Identifying sites under positive or negative selection.
- Understanding the biological significance of these sites in the context of viral evolution.
- Comparing the results from different methods (SLAC, FEL, MEME) and discussing any discrepancies.

Report

- Each student or group should prepare a report that includes
  - A brief introduction to the viral gene and its significance
  - A methodology section outlining the steps taken in Hyphy.
  - Results with tables and/or graphs showing selected sites.
  - A discussion on the implications of the findings, including any potential functional or clinical relevance

Presentation

Students present their findings to the class, highlighting key points from their analysis and discussing any challenges they faced during the assignment.

Evaluation Criteria

- Accuracy and completeness of the data retrieval and alignment.
- Correct application and execution of Hyphy selection analyses.
- Depth of interpretation and biological insights drawn from the results.
- Clarity and coherence of the written report and oral presentation.