Evolutionary shortcuts via multi-nucleotide substitutions and their impact on natural selection analyses.

Alexander G Lucaci[1], Jordan D Zehr[1], Sergei L. Kosakovsky Pond[1,*]

[1] Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA

***Corresponding author:** spond@temple.edu

Associate Editor: TBD

Abstract

Inference and interpretation of evolutionary processes - in particular of the types and targets of natural selection affecting coding sequences, are critically influenced by the assumptions built into statistical models for such analyses. If certain aspects of the substitution process (even when they are not of direct interest) are presumed absent or are modeled with too crude of a simplification, estimates of key model parameters can become biased - often systematically, and lead to poor statistical performance. Here, we performed a detailed characterization of how modeling instantaneous multi-nucleotide (or multi-hit, MH) substitutions impacts dN/dS based inference of episodic diversifying selection at the level of the entire alignment. The inclusion of MH reduces the rate (1.37-fold or 26.8%) at which positive selection is called based on the analysis of $N = 9{,}861$ empirical data-sets, while offering significantly better statistical fit to sequence data in 8.37% of cases. Through additional simulation studies, we show that this reduction is not simply due to loss of power because of additional model complexity. After a detailed examination of 21 benchmark alignments and a new high-resolution analysis showing which parts of the alignment provide support for positive selection, we reveal that MH substitutions occurring along shorter branches in the tree are largely responsible for discrepant results in selection detection. Our results add to the growing body of literature which examines decades-old modeling assumptions and finds them to be problematic for biological data analysis. Because multi-nucleotide substitutions have a significant impact on natural selection detection even at the level of an entire gene, we recommend that routine selection analysis of this type consider their inclusion. To facilitate this procedure, we developed a simple model testing selection detection framework able to screen an alignment for positive selection with two biologically important confounding processes: synonymous rate variation, and multi-nucleotide instantaneous substitutions.

Key words: Molecular Evolution, Evolutionary shortcuts, Multi-nucleotide substitutions, codon substitution models

Article

Introduction

Reliable and robust detection of natural selection from coding sequences continues to be of significant interest in comparative genomics and evolutionary biology literature. Estimation of dN/dS (Kosakovsky Pond *et al.*, 2020) analyses using codon-substitution models is a workhorse of selection detection. Its seminal methods have

been published several decades ago (Goldman and Yang, 1994; Muse and Gaut, 1994), and still enjoy wide-spread use, testifying to their utility and longevity. Because computers have gotten faster, and datasets much larger, some of the simplifying assumptions in the seminal dN/dS selection tests have been re-examined, and generally found to be wanting. For example, the assumption that synonymous rates (dS) do not vary across sites in a gene is nearly universally violated, and typically inflates the rates of false positives of selection tests (Pond and Muse, 2005). This led us to recommend the inclusion of variable dS in all selection tests (Wisotsky *et al.*, 2020).

Another important modeling assumption, which is the focus of this study, is that codon substitutions which involve multiple nucleotides (e.g., $ACC \rightarrow AGG$) must be the result of several evolutionary steps, each of which replaces a single nucleotide (e.g., $ACC \rightarrow ACG \rightarrow AGG$). This assumption is encoded in substitution rate matrices of the models as 0 rates for any "multi-hit" (MH) substitutions. The "single-hit" (SH) assumption has been repeatedly investigated in modeling literature and in every instance we found, MH models provided a better fit to the data, and (when examined) the performance of positive selection tests was affected by MH (see a brief review in (Lucaci *et al.*, 2021; Venkat *et al.*, 2018), and also further discussion herein). For a specific type of selection analysis: investigation of genes subject to positive selection on the human branch, standard (SH) branch-site selection tests suffer unacceptably high rates of false positives on neutrally evolving data simulated with MH (Venkat *et al.*, 2018). The intuition is quite simple: a few (or even a single) MH substitutions occurring on a short branch can push dN/dS estimates above 1 with SH models.

So why have MH models not been more widely adopted in selection analyses? Especially since existing literature implies that not modeling MH could invalidate a substantial fraction of selection analyses? All of the following factors appear plausible. First, a general biological mechanism for generating MH substitutions that are sufficiently widespread to matter for selection analyses is not apparent, although candidate processes do exist (e.g., polymerase zeta). Second, it is unclear if a statistical improvement in fit achieved by adding a layer of complexity to already complex models may simply absorb another, unmodelled, evolutionary factor, which may have little to do with MH – an aptly named phenomenological loading (Jones *et al.*, 2018). Third, high throughput analyses on automatically generated alignments, even when carefully curated, may be mistaking upstream alignment or sequencing errors as evidence of non-standard evolutionary features (Di Franco *et al.*, 2019; Rosenberg, 2005). Fourth, popular selection tests which the users are familiar with and are comfortable using do not provide support for MH, and what effect un-modeled MH may have on these

tests has not been systematically explored. Fifth, folding more parameters into a substitution model can lead to loss of statistical power.

In this paper we present a practical approach to handling MH when looking for signatures of episodic diversifying positive selection (EDS) at the level of an entire gene. We modified the BUSTED method for EDS detection (Murrell *et al.*, 2015) to allow multi-nucleotide substitutions (+MH), and investigated its performance compared to the versions of BUSTED without MH support. We based our comparisons on 21 diverse benchmark alignments, i.e., those studied for the purposes of selection detection in literature, a large set of high-quality mammalian gene alignments, and on simulated data. When the inclusion of MH significantly altered the results of selection detection tests, we endeavored to understand the basis of such disagreement, which prompted us to develop a set of exploratory tools for interpreting BUSTED results. Based on the synthesis of empirical and simulated data analyses, we propose a simple model selection framework to decide if the inclusion of MH support in EDS analyses is important. This framework may serve as a recommendation for practical comparative selection analyses, which attempts to balance the impetus to account for increased false positive rates when MH is present but unmodelled, and to mitigate statistical power loss because of increased parametric complexity of the models.

New Approaches

Results

High-level model description

We compared four different BUSTED class models (see Methods for complete details), each of which tests for evidence of a non-zero fraction of branch-site combinations evolving with $\omega > 1$, but makes different assumptions about confounding evolutionary processes. We evaluate four models:

- the baseline model (BUSTED),
- a model which adds site-to-site synonymous rate variation (+S),
- a model with support for instantaneous double- and triple-nucleotide substitutions within a single codon (+MH),
- and a model with support for both (+S+MH),

(cf Table 1 for additional details). These models form a nested hierarchies ($BUSTED \subset +S \subset +S+MH$ and $BUSTED \subset +MH \subset +S+MH$) and can be compared using either information theoretic criteria or pairwise likelihood ratio tests.

Analysis of benchmark alignments

It is informative to begin by examining how the four competing models (Table 2) perform on a collection of empirical sequence alignments. We screened 21 alignments for evidence of EDS. These alignments were chosen because they have each been previously analyzed (many in multiple papers) for evidence of natural selection using a variety of models, and because they represent different alignment sizes, diversity levels, and

Model	Reference	Non-synonymous rates	Synonymous rates	Multi-nucleotide substitutions	Number of parameters
BUSTED	Murrell *et al.* (2015)	Random effects branch-site modeled by a K(=3)-bin discrete distribution	None	None	$B+13+2\times K$
+S	Wisotsky *et al.* (2020)	Random branch-site effects modeled by a K(=3)-bin general discrete distribution	Random site effects modeled by an L(=3)-bin unit mean general discrete distribution	None	$B+11+2\times (K+L)$
+MH	Lucaci *et al.* (2021)	Random branch-site effects modeled by a K(=3)-bin general discrete distribution	None	Alignment-wide double- (δ) and triple-(ψ) nucleotide substitution rates	$B+15+2\times K$
+S+MH	This paper	Random branch-site effects modeled by a K(=3)-bin general discrete distribution	Random site effects modeled by an L(=3)-bin unit mean general discrete distribution	Alignment-wide double- (δ) and triple-(ψ) nucleotide substitution rates	$B+13+2\times (K+L)$

Table 1. Substitution models considered in this paper. B - the number of branches in the phylogenetic tree. K and L are user-tunable parameters, set to 3 each by default.

taxonomic groups, all of which impact selection analyses .

The inclusion of site-to-site synonymous rate variation is strongly supported for all 21 datasets (in agreement with Wisotsky *et al.* (2020)), and further addition of multi-nucleotide substitution (MNS) support is preferred by AIC_c in 12/21 datasets (Table 2). The addition of model complexity reduces the rate at which EDS is detected, with the simplest model (BUSTED) returning significant test results for 14/21 datasets, and the most complex model (+S+MH) – for 9/21. Because our primary analytical endpoint is the detection of EDS, we can categorize the alignments into those where models agree, and those where they disagree. We begin with the **seven** datasets where all four models failed to detect EDS.

`Primate lysozyme.` *(best model: +S)* A version of this dataset was originally used to show lineage specific variation in dN/dS (or ω) in Yang (1998), where tests assuming no site-to-site rate variation (SRV) also identified positive selection (mean $\omega >$ 1) on the hominoid lineage. This evidence is no longer statistically significant if a suitable multiple testing correction is applied to the original results. Overall, this is a low divergence dataset with relatively few substitutions (Table 3).

`Tick-borne flavivirus NS-5 gene.` *(+S+MH)* This dataset was analyzed in Yang *et al.* (2000a) and originally sourced from Kuno *et al.* (1998); no evidence of positive selection was found in the original papers. This is a high-divergence alignment, including 51 events when all three nucleotides are inferred to have changed along a single branch at a particular site (Table 3).

`ADORA3` *(+S)* This alignment of adenosin A3 receptor (placental mammals) was analyzed using Bayesian mutation selection models by Rodrigue *et al.* (2021), who reported weak to no evidence of adaptive evolution.

`COXI` *(+S)* Primate cytochrome oxidase subunit I mitochondrial sequences were previously analyzed in (Seo *et al.*, 2004) using Bayesian methods; they found significant lineage-to-lineage variation in absolute synonymous and non-synonymous rates, but strong conservation ($\omega \ll 1$) overall. We find no evidence of MNS, including 0 point estimates

Alignment	N	S	L	AIC_c +S+MH	ΔAIC_c vs +S+MH			p-value for EDS					EDS detection
					BUSTED+S	-S	+MH	+S+MH	BUSTED	+S	+MH	Averaged	
Mam. $\beta-$globin	17	144	3.85	7420.9	31.9	-5.2	36.4	0.2219	**0.0000**	**0.0000**	**0.0485**	**0.0154**	Discordant
Primate Lysozyme	19	130	0.25	2149.8	16.0	-4.2	20.3	0.5000	0.3668	0.5000	0.3845	0.5000	All, no
Sperm lysin	25	134	4.46	8765.3	156.3	-4.2	158.8	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	All, yes
HIV *vif*	29	192	0.95	**6911.0**	190.8	2.8	187.2	0.5000	**0.0002**	**0.0239**	**0.0424**	0.4052	Discordant
Drosophila *adh*	23	254	1.76	9357.8	14.1	-3.9	17.2	**0.0255**	**0.0003**	**0.0016**	**0.0197**	**0.0046**	All, yes
Flavivirus NS5	18	342	9.42	**18488.0**	301.9	43.1	280.8	0.5000	0.4218	0.4999	0.5000	0.5000	All, no
Hepatitis D Ag	33	196	2.23	**10416.1**	281.1	8.0	259.9	**0.0314**	**0.0000**	**0.0000**	**0.0019**	**0.0309**	All, yes
ADORA3	67	107	4.61	12612.2	251.0	-3.6	255.2	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
Streptococcus PTS	16	639	11.27	**17344.6**	206.5	31.5	169.1	**0.0000**	**0.0007**	**0.0000**	0.0879	**0.0000**	+S/+S+MH, yes
Primate COXI	21	510	11.25	24292.0	101.2	-3.4	106.3	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
Encephalitis *env*	23	500	0.89	13703.1	42.3	-4.0	44.1	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
Rhodopsin	38	330	5.32	**25902.4**	495.1	21.7	483.6	0.2279	**0.0000**	**0.0000**	**0.0000**	0.2279	Discordant
Camelid VHH	212	96	15.87	**33665.6**	1474.4	28.5	1424.7	**0.0040**	**0.0000**	**0.0000**	**0.0000**	**0.0040**	All, yes
Mammalian RBP3	54	412	4.71	**43083.9**	577.6	1.5	575.8	0.4939	0.4530	0.3321	0.5000	0.4427	All, no
Mammalian VWF	62	392	5.37	**45992.5**	940.3	5.5	931.8	0.5000	0.0767	0.1513	0.4988	0.4786	All, no
Mammalian mtDNA	20	3331	10.09	**179797.7**	1688.2	11.8	1697.3	0.1713	**0.0061**	**0.0346**	**0.0204**	0.1710	Discordant
IAV H3N2 HA	349	329	1.39	**23228.2**	637.2	14.4	630.8	0.5000	**0.0000**	0.1060	0.3637	0.4997	+S/+S+MH, no
HIV rt	476	335	7.19	**52033.6**	1717.5	1.4	1721.4	**0.0006**	**0.0000**	**0.0000**	**0.0000**	**0.0004**	All, yes
rbcL	483	466	11.88	**152988.8**	4341.6	76.5	4315.7	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	All, yes
SARS-CoV-2 S	180	1284	0.13	17817.4	649.3	-3.7	624.7	**0.0002**	**0.0000**	**0.0000**	**0.0002**	**0.0001**	All, yes
IAV H1N1 HA	466	589	2.15	51414.6	912.0	-3.8	913.4	**0.0000**	**0.0000**	**0.0000**	**0.0003**	**0.0000**	All, yes
				12	0	9	0	9	14	13	12	10	

Table 2. Selection analysis on benchmark alignments (sorted by data matrix size, from smallest to largest). N - the number of sequences, S - the number of codons, L - total tree length in expected substitutions/nucleotide, measured under the BUSTED+S+MH model. AIC_c S+MH - small sample AIC score for the BUSTED+S+MH model (shown in boldface if this model is the best fit for the data, i.e. has the lowest AIC_c score) , ΔAIC_c - differences between the AIC_c score for the corresponding model and the BUSTED+S+MH. p-value for ESD: the likelihood ratio test p-value for episodic diversifying selection under the corresponding model (4 digits of precision); shown in boldface if ≤ 0.05. The *A*veraged column shows model averaged p-values (see text). The last column indicates model agreement with respect to detecting ESD at $p \leq 0.05$. The last row shows the number of times each model was preferred by AIC_c, and the number of significant LRT tests for each model and the model averaged approach.

for δ and ψ, despite >100 events with more than one nucleotide being substituted along a single branch at a given site (Table 3). As reported by (Lucaci *et al.*, 2021), standard models are often able to properly account for multiple nucleotide substitution events along long branches.

`Japanese encephalitis env gene.` *(+S)* This alignment was included in Yang *et al.* (2000a), who found it to be subject to strong purifying selection.

`VWF` *(+S+MH)* The von Willbrand factor gene (placental mammals) from Rodrigue *et al.* (2021), who found some evidence of positive selection with mutation-selection Bayesian models (none with standard site-heterogenous codon models), although the authors caution that other unmodeled evolutionary processes (e.g. CpG hypermutability) could confound inference.

`RBP3` *(+S+MH)* Retinol-binding protein 3 (placental mammals) from Rodrigue *et al.* (2021); no evidence of positive selection was found in this gene by the original authors.

Next, we describe the **eight** datasets where all of our models found statistical evidence for EDS (LRT $p \leq 0.05$).

`adh` *(+S)* Drosophila alcohol dehydrogenase ($a$dh) gene (originally from Hudson *et al.* (1987)), studied in numerous selection detection papers, including Yang *et al.* (2000a) and Rodrigue *et al.* (2021). Most analyses failed to detect evidence of diversifying selection, despite long-standing supposition that balancing selection is acting on this gene. Rodrigue *et al.* (2021) reported that mutation-selection models detected numerous sites subject to selection; our methods allocate

Alignment	$\omega_3(p_3)$		$CV(\alpha)$	δ	ψ	Substitutuons (%)			L		
	+MH	+S				1H	2H	3H	1H	2H	3H
$\beta-$globin	2.834 (6.08%)	8.925 (3.70%)	1.263	0.241	-	526 (11.78)	110 (2.46)	24 (0.54)	0.26	0.36	0.52
Lysozyme	1.002 (0.00%)	1.002 (0.00%)	1.242	-	-	81 (2.08)	3 (0.08)	0 (0.00)	0.02	0.04	0.00
Lysin	17.459 (7.57%)	17.020 (7.70%)	0.867	-	-	514 (8.16)	149 (2.37)	33 (0.52)	0.18	0.22	0.30
HIV vif	1.226 (1.00%)	2103.279 (0.05%)	1.049	0.004	0.163	446 (4.30)	20 (0.19)	5 (0.05)	0.03	0.04	0.04
adh	4.056 (2.38%)	4.144 (2.50%)	0.594	0.032	-	693 (6.34)	66 (0.60)	14 (0.13)	0.10	0.15	0.16
Flavivirus NS5	1.105 (0.00%)	1.006 (2.23%)	1.286	0.377	0.986	1956 (17.33)	270 (2.39)	58 (0.51)	0.48	0.58	0.58
Hepatitis D Ag	11.306 (1.71%)	16.249 (1.96%)	0.902	0.143	-	665 (5.39)	122 (0.99)	14 (0.11)	0.08	0.13	0.18
ADORA3	1.013 (0.00%)	1.000 (3.46%)	0.662	0.053	-	1135 (8.35)	75 (0.55)	3 (0.02)	0.06	0.09	0.14
Streptococcus PTS	9.489 (1.56%)	11.871 (1.85%)	1.046	0.310	1.054	1245 (6.72)	293 (1.58)	76 (0.41)	1.79	1.98	1.96
Mammalian COXI	1.021 (1.06%)	1.000 (1.09%)	2.342	-	-	3160 (15.89)	132 (0.66)	6 (0.03)	0.43	0.45	0.59
Encephalitis env	3.166 (0.00%)	1.002 (0.00%)	0.671	0.012	-	1068 (4.97)	26 (0.12)	0 (0.00)	0.04	0.05	0.00
Rhodopsin	5.453 (0.37%)	6.376 (1.31%)	1.403	0.345	0.515	2488 (10.77)	257 (1.11)	45 (0.19)	0.12	0.15	0.16
Camelid VHH	9.193 (2.53%)	24.513 (2.04%)	0.817	0.157	-	2393 (6.91)	528 (1.52)	72 (0.21)	0.09	0.11	0.12
RPB3	1.258 (0.44%)	1.541 (2.03%)	0.593	0.111	0.054	4093 (9.46)	300 (0.69)	18 (0.04)	0.08	0.10	0.11
VWF	1.073 (0.36%)	1.973 (2.35%)	0.643	0.130	-	4608 (9.71)	381 (0.80)	22 (0.05)	0.08	0.11	0.17
Mammalian mtDNA	1.310 (1.04%)	1.434 (1.33%)	1.268	0.227	-	19892 (16.14)	1873 (1.52)	225 (0.18)	0.42	0.57	0.67
IAV H3N2 HA	1.002 (0.00%)	1.550 (29.48%)	1.064	0.062	0.015	1320 (0.74)	29 (0.02)	1 (0.00)	0.00	0.00	0.01
HIV rt	50.714 (0.07%)	48.230 (0.10%)	0.940	0.028	-	4149 (1.35)	129 (0.04)	10 (0.00)	0.02	0.02	0.03
rbcL	37.569 (0.14%)	49.827 (0.19%)	0.831	0.113	0.016	12185 (2.77)	653 (0.15)	72 (0.02)	0.02	0.03	0.02
SARS-CoV-2 S	5.990 (20.12%)	5.746 (29.04%)	3.132	0.012	-	421 (0.13)	13 (0.00)	0 (0.00)	0.00	0.00	0.00
IAV H1N1 HA	1039.054 (0.01%)	862.821 (0.01%)	0.835	-	-	3216 (0.73)	43 (0.01)	2 (0.00)	0.01	0.02	0.00

Table 3. Substitution process characterization on benchmark alignments. $\omega_3(p_3)$ - the maximum likelihood estimate of the ω ratio for the positively selected class, along with its estimated fraction, for +S and +S+MH models. $CV(\alpha)$ - coefficient of variation for the inferred distribution of site-to-site synonymous substitutions rates (+S+MH model). δ – the MLE for the two-hit substitution rate, ψ – the MLE for the three-hit substitution rate; 0 point estimates are shown as - for readability. Substitutions - the counts (and fractions of total branch × sites pairs) where one (1H), two (2H) or three (3H) nucleotides change along the branch under the +S+MH model. L - mean branch lengths for branches experiencing 1H, 2H, and 3H substitutions under the +S+MH model.

2.5% (of branch-site pairs) to the positively selected regime.

`Lysin` *(+S)* This alignment of abalone sperm lysin from Yang *et al.* (2000b) is a canonical example of diversifying positive selection, e.g., due to self-incompatibility constraints. There is no support for MNS in this alignment despite relatively high divergence and numerous multi-nucleotide branch-site substitution events (Table 3).

`Hepatitis D Ag` *(+S+MH)* Anisimova and Yang (2004) analyzed an alignment of Hepatitis Delta virus antigen gene with site-heterogeneous methods, and reported extensive positive selection. Our best fitting model (+S+MH) estimates 1.7% fraction of branch-site pairs to be subject to EDS ($\omega\approx11.3$). The MNS signal is entirely due to double-nucleotide substitutions ($\hat{\delta}=0.143$). While all models have $p\leq0.05$ for EDS, the p-value is highest for the +S+MH model, as we show later, this is a common pattern, when the addition of MNS support reduces (or eliminates) statistical significance levels of tests for EDS.

`Camelid VHH` *(+S+MH)* Su *et al.* (2002) studied this variable regions of immunoglobulin heavy chains in camelids using relatively underpowered counting methods, and found extensive evidence of positive selection. The best fitting model (+S+MH) allocates 2.5% of branch-site pairs to the positively selected class ($\omega\approx9.2$) and the MNS signal is driven by double-nucleotide substitutions ($\hat{\delta}=0.157$).

`HIV-1 rt` *(+S+MH)* A HIV-1 reverse transcriptase alignment comprises pairs of sequences from individuals prior to and following antiretroviral treatment, studied by Seoighe *et al.* (2007) to examine selective pressures due to the development of drug resistance. There is marginal

evidence of MNS based on AIC_c, and very strong ($\omega \sim 50$) positive selection on a small ($\sim 0.1\%$) fraction of branch-site pairs.

`rbcL` *(+S+MH)* Tamuri and Dos Reis (2022) examined this alignment of plant RuBisCO with a penalized likelihood mutation-selection model (no MNS), and identified numerous sites subject to pervasive positive selection. We find strong evidence of MNS based involving both two- and three-nucleotide substitutions, and very strong ($\omega \sim 38$) positive selection on a small ($\sim 0.14\%$) fraction of branch-site pairs.

`SARS-CoV-2 S` *(+S)* A collection of full-length SARS-CoV-2 spike genes from variants of concern and other representative lineages, obtained from GISAID (Shu and McCauley, 2017). Numerous previous studies (e.g. Martin *et al.* (2021, 2022); Viana *et al.* (2022)) detected positive selection on this gene, driven primarily by immune selective pressure and enhanced transmissibility. The best-fitting model (+S) infers that a very large fraction of this gene ($\sim 20\%$) is subject to positive selection ($\omega \approx 5.7$).

`IAV H1NA1` *(+S)* Tamuri and Dos Reis (2022) performed a detailed analysis of this H1N1 Influenza A virus (human hosts) hemagluttinin dataset and found 14-18 (depending on model parameters) sites under selection. Our analysis detects very strong ($\omega \sim 1000$) positive selection on a very small ($\sim 0.01\%$) fraction of branch-site pairs, and no evidence of MNS.

Lastly, we discuss the **six** remaining datasets, where EDS detection depends on the model. These are the most important to address, because they represent the cases where selection inference is, in some sense, not robust to modeling assumptions.

$\beta-$`globin` *(+S)* Mammalian $\beta-$globin is one of the datasets from Yang *et al.* (2000a) where positive selection has been inferred, and confirmed using many other studies and methods (e.g., Rodrigue *et al.* (2021)). All of our models, except for +S+MH, including the best fitting model (+S), also infer positive selection. However, the addition of MNS (+S+MH) model results in the elimination of statistical significance; this appears to be the case of overfitting, because +S+MH is supported neither by AIC_c, nor by direct nested LRT between the two models ($p \sim 0.5$).

`HIV-1 vif` *(+S+MH)* HIV-1 viral infectivity factor (*vif*) was inferred to be under positive selection in Yang *et al.* (2000a), but not according to our best-fitting model (+S+MH). The second best fitting model (+S), whose AIC_c is only slightly higher, returns a significant p-value for EDS. To better understand which features of the dataset leads to discordant conclusions, we applied fit profiling techniques (see Methods), and found that a single codon in the alignment (codon 6) contributes the majority of the cumulative likelihood ratio test signal (Figure S1) for the +S model. Furthermore, a single triple-nucleotide substitution along a terminal tree branch at that

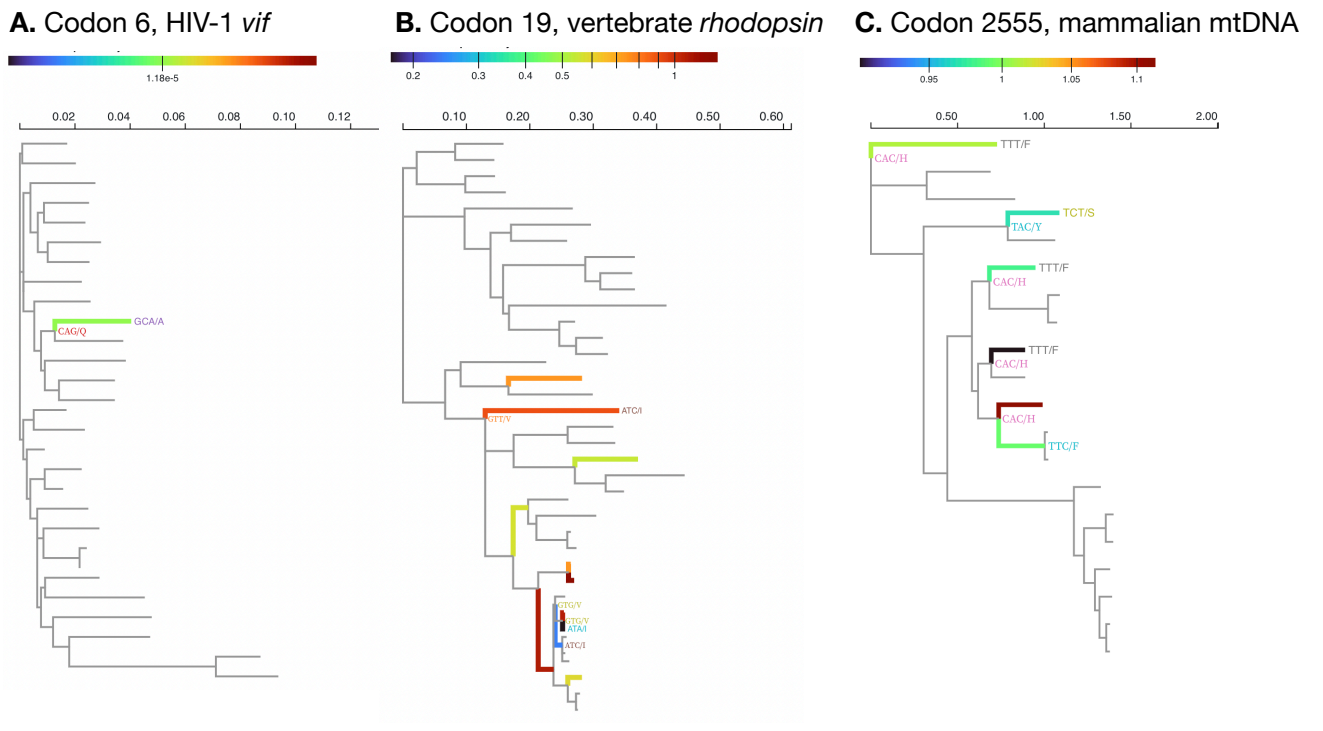

FIG. 1. Example sites from benchmark alignments with discordant selection signal. Only substitutions involving multiple nucleotides are labeled (codon/amino-acid). Coloring of the branches represents the ratio of empirical Bayes factors for $\omega > 1$ at this branch/site (see Methods) between the +S+MH and +S models. Values < 1 imply that the +S+MH model has less support for $\omega > 1$ than the +S model. The scales are different for each of the examples because they have dramatically different ranges.

site, `CAG (Q)` → `GCA (A)`, contributes the bulk of statistical support for EDS in the +S model and the addition of MHS the model completely eliminates this support (Figure 1.A). Multi-nucleotide substitutions along short branches have been shown to return false positive selection detection results in simulations (Lucaci *et al.*, 2021; Venkat *et al.*, 2018). Masking a single codon (GCA) with gaps in the alignment and rerunning BUSTED+S yields a non-significant p-value for EDS. The fact that a single codon can be responsible for the detection of gene-wide positive selection does not inspire confidence in the positive result with the +S model.

`Streptococcus` *(+S+MH).* Dunn *et al.* (2019) analyzed this trehalose-specific PTS sugar transporter system alignment (gene 2 in their study) using parameter rich models including MNS and found evidence of positive selection ($\omega_+ = 4.9, p_+ = 0.028$). Our best fitting model (+S+MH) infers a 1.6% fraction of branch-site pairs to be subject to EDS ($\omega \approx 9.5$), and so does the second best fitting model (+S). The contrarian model (+MH) is a much poorer fit to the data, and can be discounted.

`Vertebrate Rhodopsin` *(+S+MH).* This dim-light vision protein was exhaustively analyzed by Yokoyama *et al.* (2008) with comparative methods and via experimental assays. They found that amino-acid substitutions at 12 sites altered a key phenotype (absorption wavelength, $\lambda_{\max}$

of some sequences, but that traditional site-level methods for diversifying selection detection found fewer sites without significant phenotypic impact. Our best (+S+MH) and second best (+S) fitting models return strongly discordant results for EDS ($p=0.23$ and $p<0.0001$, respectively). Both double- (257 events) and triple-nucleotide (45 events) substitution rates have non-zero MLE (Table 3). Compared to the +S model, the +S+MH model infers a smaller fraction of branch site combinations (0.37% vs 1.31%) with lower ω (5.5 vs 6.4). We noticed a similar trend with simpler rate variation models in Lucaci *et al.* (2021) – the inclusion of MNS reduces ω estimates. Most of the sites which contribute signal to EDS detection with the +S model, contribute less (or no) signal under the +S+MH model (Figure S1), with strong reduction occurring at short branches which harbor multi-nucleotide substitutions (Figure 1.B). One obvious explanation for model discordance is loss of power for the more parameter rich +S+MH model, but it seems unlikely. When we simulate under the +S model (using parameter fits from the data, which includes EDS), the power to detect selection is comparable between the models (0.99 for +S+MH vs 1.00 for +S, please see the Simulated Data section for more details).

`Mammalian mtDNA` *(+S+MH).* This concatenated alignment of mammalian mitochondrial genomes ships as a test dataset with the PAML package and has been recently re-analyzed by Jones *et al.* (2018) using several models including those supporting MNS, which were preferred. Our analyses also indicate support for MNS (both double- and triple-nucleotide), but the +S+MH model (best-fitting) and +S model (second best fitting) disagree on the presence of EDS. The +S model (Table 3) allocates 1.3% of branch-site combinations to a weakly selected component ($\omega=1.4$), but the source of this support is diffuse across many sites, with relatively little signal contributed by individual sites (Figure S1). Similarly, the reduction is EDS support under +S+MH is also diffuse and less pronounced for individual sites. Because of longer branches, even sites with extensive MNS have a minor decrease in inferred local support for EDS when comparing +S+MH and +S models (Figure 1.C). Analysis of 100 replicates generated under the +S model shows that the lack of detection under S+MH is probably not due to because of significant power loss (0.50 for +S+MH vs 0.65 for +S, please see the Simulated Data section for more details)

`IAV H3N2` *(+S+MH).* Yang (2000) examined an alignment of human isolates of H3N2 Influenza A virus hemagluttinin sequences, originally studied by Bush *et al.* (1999), for evidence of EDS using site-level methods and found support for it. With the exception of the poorly-fitting BUSTED model, our analyses fail to find evidence of EDS, potentially because of extensive synonymous rate

variation (Wisotsky *et al.*, 2020), although the addition of MNS support without SRV (+MH model), also removes the selection signal.

In summary, there is a good degree of agreement between models in detecting episodic diversifying selection on 21 benchmark datasets, with 15/21 agreements among all models and 17/21 for the best fitting models (+S and +S+MH), Cohen's $\kappa = 0.63$ (substantial agreement). In all four substantively discordant cases, +S+MH did not find evidence for selection, while +S – did find such evidence. This greater "conservatism" on the part of +S+MH is unlikely to be due to significant loss of power relative to +S (see Simulations), and manual examination of discordant datasets points towards events which involve multi-nucleotide changes along shorter tree branches as a main driver of the differences. In nearly all of the datasets, +S+MH model infers a smaller proportion of sites subject to weaker (smaller ω) selection, implying that the +S model, at least for the datasets where +S+MH is preferred by AIC_c may be absorbing some of the un-modeled multi-nucleotide substitutions into the ω distribution (Jones *et al.*, 2018; Lucaci *et al.*, 2021).

Model averaged p-values

As a simple and interpretable approach to synthesize the results different models fitted to the same dataset, and account for different goodness-of-fit, we propose a model averaged p-value. It is defined as $p_{MA} = \sum_{m=1}^{M} p_m w_m$, where the sum is taken over all models considered, p_m is the p-value returned by model m and w_m is the Akaike weight for model m (Wagenmakers and Farrell, 2004). $w_m = \exp([AIC_c^{\text{best}} - AIC_c^m])/2)$, where AIC_c^{best} is the score of the best-fitting model normalized to sum to 1 over all M models. The Akaike weight, w_m, can be interpreted as $\sim P(\text{model} = \text{m}|\text{data})$, when M models are being compared. Consequently, if model m returns the likelihood ratio test of LRT_m, then $p_{MA} \sim \sum_{m=1}^{M} P(LRT \geq LRT_m|\text{null model } m) P(\text{model} = \text{m}|\text{data})$.

The model averaged approach detects the same 9 datasets as the +S+MH model, and also the $\beta-$globin dataset, where the +S model (with EDS signal) has a sufficiently significant edge in goodness of fit to "outvote" the +S+MH model (Table 2). As our simulation results show (next section), the model averaged approach is a simple and automated way to control false positives, while maintaining very good power.

Analysis of simulated alignments

Four taxon tree null simulations.

We generated synthetic alignments of 4 sequences with 800 codons each, using the tree shown in Figure 2, subject to negative selection or neutral evolution ($\omega_1 = 0.1(50\%), \omega_2 = 0.5(25\%), \omega_3 = 1.0(25\%)$), under the +S or +S+MH models. We varied the 2H rate (δ), and the 3H rate (ψ) as well as lengths of two of the 5 branches in the tree, generating 100 replicates for each parameter combination

considered. We then fitted +S and +S+MH models to all of the replicates, and tabulated false positive rates (FPR).

False positive rates. As the 2H rate (δ) increases (Figure 2), the +S model shows progressively higher FPR (reaching 100%), coupled with increasingly biased estimates of ω_3 – the positive selection model component. On the other hand, the +S+MH model shows nominal or conservative FPR, and generally consistent estimates of ω_3. Because the +S+MH model has increasingly better fit to the data as δ becomes larger, the model averaged p-value (which gives progressively more weight to +S+MH), also has controlled FPR, with the exception of slightly elevated rates for $0.15 \leq \delta \leq 0.25$. Therefore, the +S model appears to "absorb" unmodeled multiple-hit substitutions into biased ω estimates, which leads to catastrophically high rates of false positives. An identical pattern is observed for a fixed δ and increasing rates of 3H substitutions (ψ), seen in Figure 2. Finally, FPR of the +S model also depends on branch lengths of the tree. In these simple simulations branch lengths $\sim$0.05 show an elevation in +S FPR rates. The intuition is simple: very short branches do not accumulate many substitutions (no signal), sufficiently long branches do not benefit as much from access to instantaneous 2H substitutions, because over longer branches it is nearly as easy to obtain a 2H substitution via two (or more) consecutive 1H substitutions allowed in the standard models. Short branches with multi-nucleotide substitutions force the +S model to absorb these unmodeled changes into the ω rate, and have the largest effect on FPR rates.

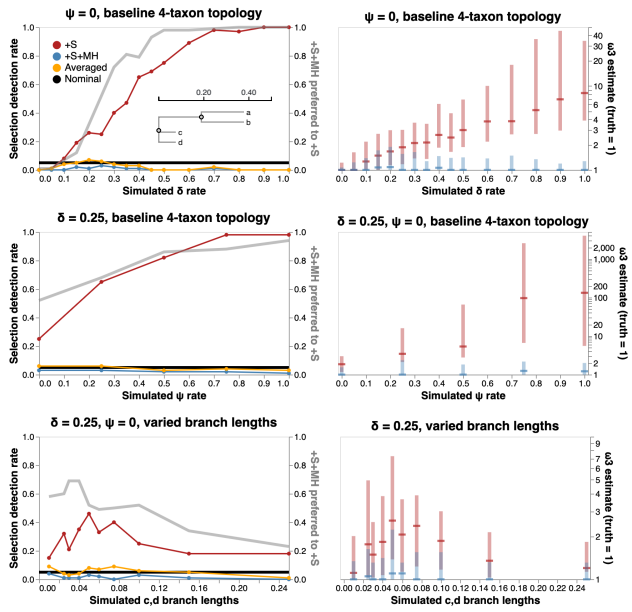

FIG. 2. Model performance on null simulated data. Left column : false positive detection rate for EDS (at $p \leq 0.05$) as a function of rate parameters and branch lengths, and the rate at which +S+MH is preferred to +S by a nested LRT test. Right column: ω_3 estimates (median, IQR) for various simulation scenarios. 100 replicates were generated using the four-taxon tree shown as inset in the top left plot for each parameter combination. For the bottom row, we varies the lengths of branches leading to `c` and `d` in the tree

Power. On the same 4-taxon tree, we next simulated alignments with a non-zero fraction of the alignment subject to EDS, with the distribution of rates ($\omega_1 = 0.1(50\%), \omega_2 = 0.5(40\%), \omega_3 > 1(10\%)$. We iterated ω_3 over the set $\{1.25, 1.5, 2, 4, 8, 16\}$, set $\psi = 0$, and iterated δ over the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, for a total of 36 simulation scenarios. The three methods for EDS detection (+S, +S+MH, and model averaged), all gained power as the effect size (ω_3) increased, reaching 100% (Figure 3). When no multiple-hits are allowed ($\delta = 0$), +S+MH shows a small

loss of power compared to the +S model, but because +S has better fit in nearly every dataset, model averaging rescues most of the power. Both +S and +S+MH return consistent estimates of ω_3. For $\delta > 0$ and for $\omega_3 < 8$ the +S model has progressively higher power, but that power comes at the cost of progressively more and more biased estimates of ω_3. This behavior mirrors what we saw for null data, except, for data simulated with positive selection (low or moderate effect sizes), the bias results in a desirable outcome (higher power). Model averaging becomes less effective as δ grows, because the +S model loses goodness-of-fit compared to the +S+MH model. Increasing the fraction of alignments subject to selection to 25% ($\omega_1 = 0.1(50\%), \omega_2 = 0.5(25\%), \omega_3 > 1(25\%)$) shows the same qualitative behavior, except that all methods have higher power for a given value of δ and ω_3 (Figure S2).

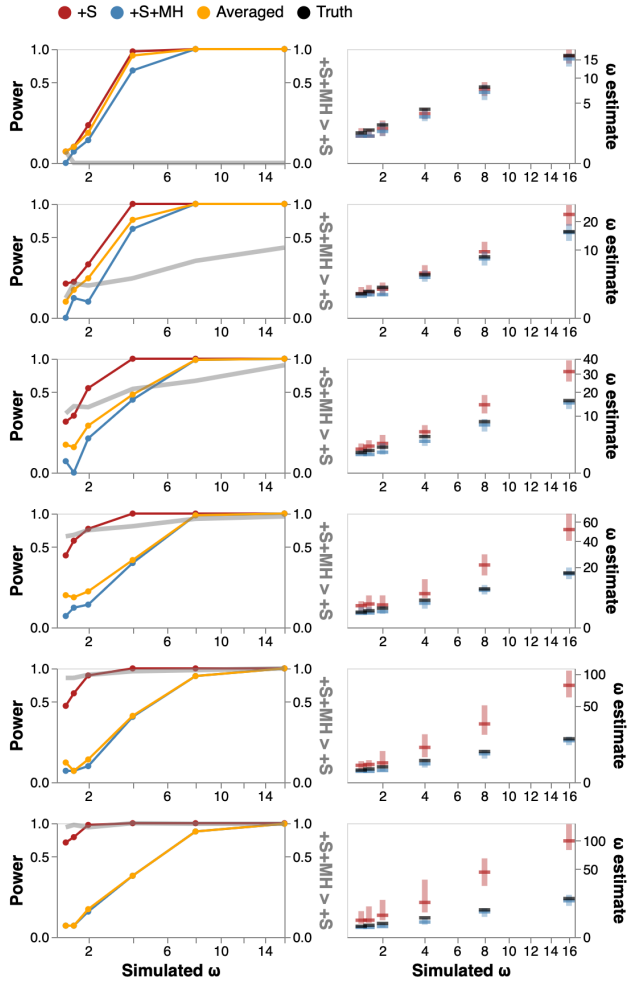

FIG. 3. Model performance on data simulated with EDS. Left column : detection rate for EDS (at $p \leq 0.05$) as a function of rate ω_3 (effect size) and δ (confounding parameter), and the rate at which +S+MH is preferred to +S by a nested LRT test. Right column: ω_3 estimates (median, IQR) for various simulation scenarios.

Benchmark datasets

We generated an additional 9 null and 18 power simulations (100 replicates each) based on empirical data sets (details shown in Table S1). These scenarios are more representative of biological data because they use alignment sizes, tree topologies, branch lengths, nucleotide substitution biases, and other model parameters based on biological data. We fixed all model parameters except ω_3, δ, ψ, and p_3. These data recapitulate the patterns found in the simple 4-taxon tree simulations.

1. For null data, +S loses control of FPR as δ and/or ψ are increased. +S+MH and model averaging maintain FPR control regardless of the values of 2H and 3H rates.

2. For data with EDS but without 2H or 3H, +S has a slight power edge over +S+MH, but model averaging rescues the power because +S has a better goodness of fit.

3. For data with EDS and with 2H and/or 3H, +S has a power edge over +S+MH,

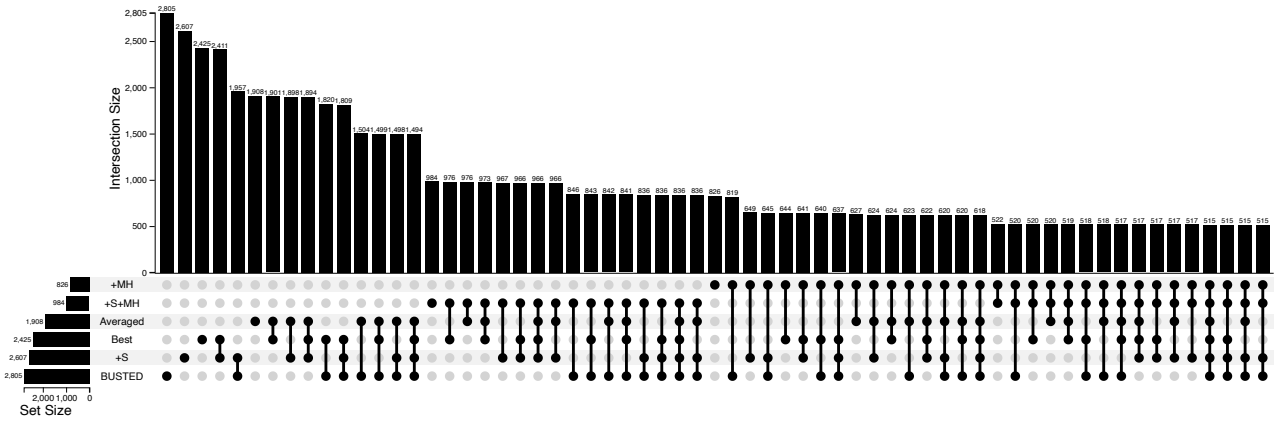

FIG. 4. Alignments with evidence of episodic diversifying selection in the Enard *et al* dataset. The number of alignments (9,861 total) which returned LRT $p \leq 0.05$ under each of the following scenarios: individual model (BUSTED, +S, +MH, +S+MH), best model selected by AIC_c (Best), model averaged "p-value" (Averaged). The sizes of each of the non-empty intersections of these six combinations are also shown. For example, 515 datasets are found to be subject to EDS under all of the six considered criteria.

and model averaging is only partially able to rescue the power because +S+MH has a much better goodness of fit. This gain in power for +S comes at a cost of significant (often dramatic) upward biases in ω_3 estimates.

Because in real biological data, the presence of selection is the object of inference and it is not expected to be prevalent (e.g., a typical gene is more likely to not be subject to EDS), therefore controlling FP rates should be the prevailing concern. As our simulations demonstrate, unmodeled multi-nucleotide substitutions dramatically inflate the estimates of ω rates, and result in significant and often catastrophic FPR.

A Large-Scale Empirical Screen

We compared the inferences made by the four BUSTED class models on a large-scale empirical dataset (Enard *et al.*, 2016) with 9,861 alignments and phylogenetic trees of mammalian species (cf Methods). This collection was originally prepared to assess the influence of viruses on mammalian protein evolution and includes sequences from 24 species.

As with the benchmark datasets, only two out of four model (+S and +S+MH) had the best goodness-of-fit (AIC_c) measures for most of the alignments (Table 4). BUSTED and +MH were the top model for 97($<$1%) of the alignments which were either very short alignments ($<$150 codons, e.g., SF3B6 in Table 5) or with minimal divergence (tree length $<$0.01 substitutions/site). Alignment length and total tree length were not significantly associated (Mann-Whitney U test) with the predilection towards the +S or +S+MH model.

We next considered whether an alignment was detected subject to EDS, using LRT $p \leq 0.05$, under different selection criteria: model fixed

Model	Rank 1	Rank 2	Rank 3	Rank 4
BUSTED	93	60	7436	2272
+S	8943	853	61	4
+MH	4	62	2307	7488
+S+MH	821	8866	57	97

Table 4. Model goodness-of-fit ranking for the Enard dataset. How each of the four models ranked for each of the 9,861 alignments.

a priori, best fitting-model selected by AIC_c, and the model averaged "p-value" (Fig 4). The simplest model (BUSTED) has the highest raw detection rate of 2805/9861 alignments (28.4%), while the models with MH support have much lower detection rates: 984/9861 (10%) for +S+MH and 826/9861 for +MH (8.4%). Detection of EDS is quite sensitive to which model/approach is being used: there are only 515 alignments, where EDS is detected by all of the models (including the best fitting model), and by model averaging. Requiring a complete model consensus does not strike us as a sensible approach – why, for example, should we give an equal vote to models that do not describe the data well? Indeed, even for the 515 unanimous datasets, the median difference in AIC_c (ΔAIC_c) scores between the best and the worst fitting model was >200 points, implying that the worst models had much worse fits to the data than the best, and should be discounted. One solution, which has found common use in comparative analysis is to simply pick the best fitting model (such as ModelTest (Posada and Buckley, 2004), and call EDS based on it. Here, the best-fitting model detects EDS on 2,425 datasets. One danger with simply picking the best-fitting model, is that in cases when it detects EDS, but the second-best model does not and the second-best model is not dramatically worse fitting, we are discounting a discordant signal from a credible alternative model. The model averaging approach is a simple way to account for this: if two models have similar goodness-of-fit, and one has a low EDS p-value, but the other has a high EDS p-value, averaging the two will result in a non-significant (conservative) call. The "averaging" approach detects EDS on 1,908 datasets. On 524 datasets when the best model detects EDS, but model averaging does not, the median Akaike weight difference AIC_c for the second best fitting model (defined as $w = e^{-\Delta AIC_c/2}$, normalized to sum to 1 over all four models) was 0.18, hence a high p-value from the second-best fitting model is sufficient to push the averaged p-value above 0.05.

In Table 5, we show examples of patterns for comparative model fit and EDS inference, and discuss them (below) in terms of the selective patterns:

No selection detected by any method, pattern (000000). Nearly two thirds (6396, or 64.9%) of the datasets have no evidence of episodic diversifying selection under any of the six possible detection criteria : (+S+MH, BUSTED, +S, +MH, Averaged, Best). These alignments (e.g., RCOR1), tended to be shorter compared to the alignments with some selection signal (median 427 codons, vs median 599 codons, $p < 10^{-16}$, Wilcoxson test), have lower overall divergence

(median tree length, 1.04 vs 1.27, $p<10^{-16}$), and have a smaller fraction of datasets where a model with support for 2H or 3H (odds ratio 0.6, $p<10^{-12}$, Fisher exact test).

Selection detected by every method (111111). A total of 515 datasets supported EDS with every detection approach (e.g., PDIK1L), For 489 of those, +S was the best fitting model, and for the remaining 26 – +S+MH was the best fitting model, with longer and more divergent alignments falling into the second bin (+S+MH model), with Wilcoxson p-values of <0.02. Compared to datasets where only some of the methods detected EDS (not consensus), the consensus collection had a larger estimated EDS effect size, approximated by the $\sqrt{\omega_3}p_3$ (scaled weight assigned to the positive selection regime), median 0.0133 vs 0.0024 (Wilcoxson $p=0.001$).

Selection detected by all but one model. These datasets are "near-consensus" in that all but one of the individual models (e.g., +MH for the (111011) pattern), including the best fitting model and the model-averaged p-value, support EDS (e.g., TIMM50). There are 428 alignments in this bucket, including 401 for which +S is the best fitting model, 25 – +S+MH, and 2 – BUSTED. The most common "outlier" model was +MH (321), followed +S+MH (103), and 2 each for BUSTED and +S.

The best model drives EDS selection detection. CDC123 is a prototypical example, where +S is the best model, is the only model that shows evidence of EDS, yet is sufficient for both the Best model and the Averaged model criteria to also indicate EDS. Of the 268 alignments in this group, EDS detection was driven by the +S model for all but two datasets, where the +S+MH model drove detection (DRC7, for example). For all 266 datasets with +S as the best-fitting model, +S+MH was the second-best model, receiving a median Akaike weight of only 0.046, i.e. making it irrelevant for model averaged p-value calculations.

+S and +S+MH models both detect EDS. For genes like ADAMTS1and ELP2, +S and +S+MH are the two credible models, which both detect EDS, together with the Best and Averaged approaches. There are 125 of such datasets for which +S is the best-fitting model, and 3 with +S+MH as the best-fitting model.

The best model drives EDS selection detection, but model averaging disagrees. The first class of datasets where important disagreement occurs, are those where EDS is detected with the best fitting model but not detected with the model averaged approach. ODF1 is an example: the best fitting model (+S) supports EDS with $p=0.0187$, but the second-best model (+S+MH), finds no evidence EDS ($p=0.5$). Model-averaging takes both of those indications into account and arrives at a non-significant p-value of 0.06. There are 524 datasets in this bucket, and for all but 9 of those, +S is the best model, and +S+MH is second best

Gene	S	L	AIC_c +S+MH	ΔAIC_c vs +S+MH			p-value for EDS					EDS		
				BUSTED	+S	+MH	+S+MH	BUSTED	+S	+MH	Averaged	Pattern	Best	Count
RCOR1	429	1.00	10966.0	116.3	**-4.0**	120.1	0.5000	0.5000	0.5000	0.5000	0.5000	(000**0**0)	*	6396
PDIK1L	341	0.47	6374.4	23.0	**-4.1**	27.1	**0.0080**	**0.0038**	**0.0035**	**0.0118**	**0.0041**	(111**1**1)	+S	489
TIMM50	439	1.38	12513.2	102.1	**-3.6**	124.0	**0.0008**	**0.0000**	**0.0004**	0.5000	**0.0005**	(111**0**1)	+S	297
CDC123	336	0.93	9464.4	59.0	**-3.9**	63.2	0.1573	0.0575	**0.0244**	0.5000	**0.0409**	(001**0**1)	+S	268
ODF1	250	1.56	7735.3	108.7	**-4.7**	111.1	0.5000	0.5000	**0.0187**	0.5000	0.0609	(001**0**1)	+S	202
ADAMTS1	967	1.44	33306.6	301.0	**-4.0**	305.0	**0.0443**	0.0798	**0.0464**	0.1532	**0.0462**	(101**0**1)	+S	125
SDC1	310	1.94	**12722.5**	166.9	0.8	162.5	0.0697	**0.0000**	**0.0010**	0.1243	**0.0419**	(****10)	*	7
SF3B6	125	0.64	2646.7	**-6.6**	-4.9	-3.7	0.3904	**0.0070**	**0.0060**	0.1240	**0.0308**	(011**0**1)	BUSTED	6
DRC7	868	2.66	**30617.4**	298.5	3.2	300.8	**0.0169**	0.2034	0.1472	0.5000	**0.0390**	(100**0**1)	+S+MH	2
ELP2	886	1.18	**27640.2**	204.8	1.8	207.2	**0.0240**	0.1074	**0.0184**	0.2524	**0.0224**	(101**0**1)	+S+MH	3

Table 5. Examples of patterns of agreement and disagreement of different approaches to detecting EDS on the Enard et al dataset, sorted from most to least common. There are 24 sequences in each alignment. **Gene** – gene name, S - the number of codons, L - total tree length in expected substitutions/nucleotide, measured under the BUSTED+S+MH model. AIC_c S+MH - small sample AIC score for the BUSTED+S+MH model (shown in boldface if this model is the best fit for the data, i.e. has the lowest AIC_c score) , ΔAIC_c - differences between the AIC_c score for the corresponding model and the BUSTED+S+MH. p-value for ESD: the likelihood ratio test p-value for episodic diversifying selection under the corresponding model (4 digits of precision); shown in boldface if ≤ 0.05. **Averaged** - the model averaged p-value for ESD (bolded if f ≤ 0.05. **Pattern** - a bit vector of whether or not the EDS was detected at $p \leq 0.05$ with each of the six models: (+S+MH, BUSTED, +S, +MH, Averaged, Best); * denotes a wildcard (0 or 1). **Best**, best fitting model (AIC_c). **Count** - the number of datasets matching this detection pattern.

and plays the role of spoiler. Median MLEs for MH rates were higher in the datasets than where +S and +S+MH disagreed (only +S supports EDS), compared to where they agreed (both models support EDS): 0.03 vs 0.0, $p < 10^{-10}$ for δ; 0.03 vs 0.05, $p < 10^{-10}$ for ψ). The models also had significantly different ($p < 10^{-10}$) estimates for ω_3, with median differences $\omega_3^{+S} - \omega_3^{+S+MH}$ of 24.7 (EDS only for +S) vs 0.01 (EDS in both). These patterns are consistent with false positive EDS detection by the +S model as seen on simulated data.

Model averaging finds EDS, but the best-fitting model disagrees. There are only 7 datasets (e.g., SDC1), in this counter-intuitive class of an important disagreement. For these types of datasets, the best fitting model has a borderline significant p-value, the second best fitting model has a highly significant p-value (and is a very close in terms of AIC_c), and the model averaging approach arrives at a significant p-value.

Discussion

Evolutionary substitution models that are practically useful and computable must make many simplifying assumptions about the biological processes. Many, if not most, of these assumptions are not justifiable on biological grounds. However, certain classes of inference problems appear to be quite robust to even severe model mis-specifications. Examples include phylogenetic inference (Abadi *et al.*, 2019), and relative evolutionary rate estimates for individual sites (Spielman and Kosakovsky Pond, 2018). Other inference problems, including selection detection, seem to be highly sensitive to modeling assumptions (Kosakovsky Pond *et al.*, 2011; Venkat *et al.*, 2018). Such sensitivity is not surprising for methods that are tuned to extract statistical signal from a small subset of branches and sites in a sequence alignment. In extreme cases, a single substitution event is sufficient to power selection detection, as seen in the HIV-1

*v*if example in this paper and for human lineage selection detection in Venkat *et al.* (2018).

Statistical tests which compare the ω ratio of non-synonymous and synonymous substitution rates to 1 and interpret significant differences as evidence of non-neutral evolution are susceptible to confounding processes which bias ω estimates. We have previously demonstrated that ω estimates are strongly biased (and resulting in high Type 1 and 2 error rates) when the distribution used to model ω variation across branches and sites is too restrictive (Kosakovsky Pond *et al.*, 2011), and when synonymous substitution rates are assumed to be constant across sites in the alignment (Wisotsky *et al.*, 2020). Furthermore, these confounding processes are not rare, but instead are very likely present in biological data. Because "the scientist must be alert to what is importantly wrong" (Box, 1976), and these models are clearly wrong in important ways, as they misinterpret widespread confounding evolutionary processes as evidence of selection, continued use of such models is unsound.

In this study, we demonstrate that not accounting for instantaneous multi-nucleotide substitutions or "hits" (MH) when looking for evidence of positive selection can be similarly fraught with statistical error (Venkat *et al.*, 2018; Wisotsky *et al.*, 2020). Estimates of ω become inflated with standard codon substitution models when they are used to analyze data with MH, and progressively more so as the degree of MH is increased. This bias, in turn, produces uncontrolled rates of false positives for positive selection on simulated data for MH parameter values that appear realistic. However, the inclusion of MH in models can lead to some loss of statistical power in cases as compared to standard models. A large-scale empirical analysis of mammalian genes (Enard *et al.*, 2016) suggests that 10% of alignments are best fit with models supporting MH, and that roughly 80% of positively selected genes are robustly detected even when accounting for MH using a model-averaging procedure. Consequently, confounding due to MH can be viewed as a "second-order" effect, compared, for example, to the inclusion of synonymous site-to-site rate variation. We argue that the effect is sufficiently important to be considered in routine analyses of selection. Our practical recommendation, supported by simulated data and empirical analyses, is to fit multiple flavors of selection models followed up by model-averaged selection detection to obtain a good tradeoff between power and false positive rate control. We also developed a series of visual tools to assist researchers in interpreting selection analysis results, exploring which branches and sites in the alignment provide support for various evolutionary processes (selection and/or MH), and understanding how much a positive selection result is influenced by information from a small number of sites.

It remains unclear just how pervasive MH is in different types of biological sequence data, although our current and previous results (Lucaci *et al.*, 2021), and other studies (Cohen *et al.*, 2021; Freitas and Nery, 2022; Hensley *et al.*, 2021; MacLean *et al.*, 2021; Steward *et al.*, 2022) suggest that MH occurs broadly over diverse taxonomic groups. We expect that future research with interdisciplinary design combining computational and experimentally-informed approaches may shed light on the application of our method(s) and the patterns and processes underlying the contribution of MH to gene evolution. Creative investigation may help discover additional mechanisms and interpretations of the biological underpinnings of the mutational spectrum as it applies to rare mutations in natural populations. Additionally, we see a strong tailwind in this field as technological improvements for functional studies designed with the precise manipulation of DNA (Wang *et al.*, 2020) including CRISPR-Cas9, and detection of MH polymorphisms (J. Huang *et al.*, 2014) continue to emerge, draw interest, and be fine-tuned. Downstream innovations and technological design are critical in the detection of natural selection, where models such as ours can be of particular interest to researchers interested in gene-drug target design for particular fitness effects. Additionally, our work supports an emerging body of information on the underlying trends, biological mechanisms, and genetic signaling pathways under selective pressure. These results can feed directly into a number of *post hoc* analyses to qualify or quantify an exploratory genetic profile and evolutionary history across lineages.

Methods

Statistical Methodology

We adapted two existing models: the BUSTED model, a test of episodic diversifying selection, by (Murrell *et al.*, 2015), and the +S model by (Wisotsky *et al.*, 2020), which was created as a modification of the BUSTED model, to account for the presence of synonymous rate variation (SRV). The +S+MH model is a straightforward extension of +S which allows it to account for instantaneous multiple nucleotide changes occurring within a codon (MH) and SRV, while the BUSTED+MH model is an extension of the BUSTED model where SRV is not modeled (Table 1). In this framework, the nucleotide substitution process is described using the standard discrete-state continuous-time Markov process approach of (Muse and Gaut, 1994), with entries in the instantaneous rate matrix (Q) corresponding to substitutions between sense codons i and j and defined as follows:

Type	**Expression for Q_{ij}**
1 step synonymous change	$\alpha^s \theta_{ij} \pi_j^p$
1 step nonsynonymous change	$\alpha^s \omega^{bs} \theta_{ij} \pi_j^p$
2 step synonymous change	$\delta \alpha^s \prod_{n=1}^{2} \theta_{ij}^n \pi_j^n$
2 step nonsynonymous change	$\delta \alpha^s \omega^{bs} \prod_{n=1}^{2} \theta_{ij}^n \pi_j^n$
3 step synonymous change	$\psi \alpha^s \prod_{n=1}^{3} \theta_{ij}^n \pi_j^n$
3 step nonsynonymous change	$\psi \alpha^s \omega^{bs} \prod_{n=1}^{3} \theta_{ij}^n \pi_j^n$

Here, $\theta_{ij}(=\theta_{ji})$ denote nucleotide substitution bias parameters. For example, $\theta_{ACT,AGT}=\theta_{CG}$ and because we incorporate the standard nucleotide general time-reversible (GTR) (Tavare, 1986) model there are five identifiable θ_{ij} parameters: θ_{AC}, θ_{AT}, θ_{CG}, θ_{CT}, and θ_{GT} with $\theta_{AG}=1$. The position-specific equilibrium frequency of the target nucleotide of a substitution is π_j^p; for example, it is π_G^2 for the second-position change associated with $q_{ACT,AGT}$. The π_j^p and the stationary frequencies of codons under this model are estimated using the CF3×4 procedure (Pond *et al.*, 2010), adding nine parameters to the model. The ratio of nonsynonymous to synonymous substitution rates for site s along branch b is ω^{bs}, and this ratio is modeled using a 3-bin general discrete distribution (GDD) with five estimated hyperparameters: $0\leq\omega_1\leq\omega_2\leq1\leq\omega_3$, $p_1=P(\omega^{bs}=\omega_1)$, and $p_2=P(\omega^{bs}=\omega_2)$. The procedure for efficient computation of the phylogenetic likelihood function for these models was described in Kosakovsky Pond *et al.* (2011).

The quantity α^s is a site-specific synonymous substitution rate (no branch-to-branch variation is modeled) drawn from a separate 3-bin GDD. The mean of this distribution is constrained to one to maintain statistical identifiability, resulting in four estimated hyperparameters: $0\leq c\alpha_1\leq\alpha_2=c\leq c\alpha_3$, $f_1=P(\alpha^s=\alpha_1)$, and $f_2=P(\alpha^s=\alpha_2)$, with c chosen to ensure that $\mathrm{E}[\alpha_s]=1$.

The key parameters are global relative rates of multiple hit substitutions: δ is the rate for two substitutions relative to the one substitution synonymous rate (baseline), ψ is the relative rate for non-synonymous three substitutions. All parameters, except π, including branch lengths are fitted using a directly optimized phylogenetic likelihood in HyPhy.

Typical implementations, including ours, allow the number of α and ω rate categories to be separately adjusted by the user, for example, to minimize cAIC or to optimize some other measure of model fit. The default setting of three categories generally provides a good balance between fit and performance when using this GDD approach for modeling. Our implementation of +S+MH, and BUSTED+MH will warn the user if there is evidence of model overfitting, such as the appearance of rate categories with very similar estimated rate values or very low frequencies.

The +S+MH procedure for identifying positive selection is the likelihood ratio test comparing the full model described above to the constrained model formed when ω_3 is set equal to 1 (i.e., no positively selected sites). Critical values of the test are derived from a 50:50 mixture distribution of χ_0^2 and χ_2^2 (Murrell *et al.*, 2015; Wisotsky *et al.*, 2020). Both +S and +S+MH analyses in the current work use the same 50:50 mixture test statistic. +S+MH reduces to +S by setting the MH rates to 0. The method is implemented as a part of HyPhy (version 2.5.42 or later) (Kosakovsky Pond *et al.*, 2020) .

Empirical Data and Alignments

The (Enard *et al.*, 2016) data collection includes 9,861 orthologous coding sequence alignments of 24 mammalian species and is available at `https://datadryad.org/stash/dataset/doi:10.5061/dryad.fs756`. Phylogenetic trees were inferred for each alignment using RAxML (Kozlov *et al.*, 2019).

Synthetic data

Simulated data sets can be downloaded from https://data.hyphy.org/web/busteds-mh/.

Additional information is present in the `README.md` file, including details of how to generate alignments under the +S and +S+MH models.

Implementation

All analyses were performed in HyPhy version 2.5.41 or later. The BUSTED+MH and +S+MH models are implemented as part of the standard HyPhy library. You can run this option using the "–multiple-hits" option from the command line with either "Double" to consider DH substitutions or "Double+Triple" to consider DH and TH substitutions. The HyPhy Batch Language (HBL) implementation is located in a dedicated GitHub repository at `https://github.com/veg/hyphy`

Site-level support

In order to identify which individual sites show preference for MH models, we use evidence ratios (ER), defined as the ratio of site likelihoods under two models being compared. We previously showed that ERs are useful for identifying the sites driving support for one model over another, and they incur trivial additional overhead to compute once model fits have been performed.

Empirical Bayes support

We can estimate statistical support for selection ($\omega_3>1$) or multiple hit substitutions ($\delta>0$ or $\psi>0$) at a particular site (s) and branch (b), using a straightforward empirical Bayes calculation. For example, $P\left(\omega_3^{bs}>1|D_s\right)=P(D_s|\omega_3^{bs}>1)\times P(\omega_3^{bs}>1)/P(D_s)$, where $P(D_s)$ is the standard phylogenetic likelihood of D_s (summed over all ω combinations), $P(D_s|\omega_3^{bs}>1)$ is the phylogenetic likelihood at site s, computed by setting the distribution of ω at branch b to assign all weight to $\omega_3>1$, and $P(\omega_3^{bs}>1)$ is the mixture weight estimated from the entire alignment (MLE for the corresponding hyperparameter). The corresponding empirical Bayes factor (EBF) is $\frac{P\left(\omega_3^{bs}>1|D_s\right)/(1-P\left(\omega_3^{bs}>1|D_s\right))}{P(\omega_3^{bs}>1)/(1-P(\omega_3^{bs}>1))}$. As discussed in Murrell *et al.* (2012), these empirical estimates are quite noisy and should only be used for exploratory purposes, e.g., to look for "hot-spots" in a tree (cf Figure 1).

Hypothesis testing

Nested models are compared using likelihood ratio tests with asymptotic distribution used to assess significance. A conservative χ^2_2 asymptotic distribution is used to compare the fit of +S and +S+MH (null hypothesis : $\delta=\psi=0$).

Computational complexity

Treating BUSTED as a baseline, we expect the +S model to require about a relative $\times L$ (L = number of synonymous rate classes) more time per likelihood calculation and longer convergence time due to an extra random effects distribution. Because +MH models have dense rate matrices, there is a computational cost incurred for computing transition matrices since optimizations available for standard (sparse) matrices no longer apply. +S+MH models are expected to be the slowest, but have the same order of complexity as $+S$. On 24-sequence alignments from Enard *et al.* (2016), we observed the following performance for each of the four models.

Model	Median, sec	Mean, sec	Relative
BUSTED	50	60	-
+S	174	219.1	3.3
+MH	279	379.1	6.1
+S+MH	1405	1788.5	28.6

Table 6. Model run-times for the Enard dataset. Run times on 4 cores on an AMD EPYC 7702 CPU compute node are shown. `Relative` : median of the relative run times on the same dataset compared to BUSTED.

BUSTED ModelTesting

We recommend our BUSTED model testing and averaging procedure (see main text) in order to select the best fitting model, to interpret the results of natural selection acting on your gene of interest. Our goal is to understand which underlying model and its parameters are able to detect the areas of the dataset which drive the greatest degree of evolutionary signals. We screen the dataset for episodic diversifying selection acting on the whole gene while accounting for SRV across the alignment, and MH substitutions.

Analysis is conducted as a series of experiments in the BUSTED framework of selection analysis with our methods under analysis in the hierarchical structure described in Table 1 and includes BUSTED, +S, BUSTED+MH, and +S+MH.

We implement a Snakemake (Mölder *et al.*, 2021) version of our model testing procedure, available at `https://github.com/veg/BUSTED_ModelTest`. This application takes the same input as a normal BUSTED analysis, a multiple sequence alignment and inferred phylogenetic, and returns JavaScript Object Notation (JSON) files (one for each model described above).

We recommend performing model averaging to determine whether or not an alignment is subject to episodic diversifying selection. Alternative approaches could include selecting the best fitting model, or model consensus, however, as shown by our simulations, these approaches are less statistically efficient (lower power and/or higher rate of false positives).

Acknowledgments

We thank members of the HyPhy and Datamonkey teams for their contribution to this work. This research was supported in part by grants GM144468 (NIH/NIGMS), AI140970 (NIH/NIAID), and AI134384 (NIH/NIAID).

References

Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun*, 10(1): 934.

Number: 1 Publisher: Nature Publishing Group.

Anisimova, M. and Yang, Z. 2004. Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? *J Mol Evol*, 59(6): 815–26.

Box, G. E. P. 1976. Science and Statistics. *Journal of the American Statistical Association*, 71(356): 791–799. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. 1999. Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Mol Biol Evol*, 16(11): 1457–65.

Cohen, Z. P., Brevik, K., Chen, Y. H., Hawthorne, D. J., Weibel, B. D., and Schoville, S. D. 2021. Elevated rates of positive selection drive the evolution of pestiferousness in the Colorado potato beetle (Leptinotarsa decemlineata, Say). *Molecular Ecology*, 30(1): 237–254. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15703.

Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, 19(1): 21.

Dunn, K. A., Kenney, T., Gu, H., and Bielawski, J. P. 2019. Improved inference of site-specific positive selection under a generalized parametric codon model when there are multinucleotide mutations and multiple nonsynonymous rates. *BMC Evol Biol*, 19(1): 22.

Enard, D., Cai, L., Gwennap, C., and Petrov, D. A. 2016. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5.

Freitas, L. and Nery, M. F. 2022. Positive selection in multiple salivary gland proteins of Anophelinae reveals potential targets for vector control. *Infection, Genetics and Evolution*, 100: 105271.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, 11(5): 725–736. ISBN: 0737-4038 (Print)\textbackslashn0737-4038 (Linking).

Hensley, N. M., Ellis, E. A., Leung, N. Y., Coupart, J., Mikhailovsky, A., Taketa, D. A., Tessler, M., Gruber, D. F., De Tomaso, A. W., Mitani, Y., Rivers, T. J., Gerrish, G. A., Torres, E., and Oakley, T. H. 2021. Selection, drift, and constraint in cypridinid luciferases and the diversification of bioluminescent signals in sea fireflies. *Molecular Ecology*, 30(8): 1864–1879. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15673.

Hudson, R. R., Kreitman, M., and Aguadé, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116(1): 153–9.

J. Huang, C., F. Fang, W., S. Ke, M., E. Chou, H. Y., and T. Yang, J. 2014. A biocompatible open-surface droplet manipulation platform for detection of multi-nucleotide polymorphism. *Lab on a Chip*, 14(12): 2057–2062. Publisher: Royal Society of Chemistry.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol Biol Evol*, 35(6): 1473–1488.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D. W., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*, 28(11): 3033–43.

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S. J., Frost, S. D. W., and Muse, S. V. 2020. Hyphy 2.5-a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol*, 37(1): 295–299.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics (Oxford, England)*, 35(21): 4453–4455.

Kuno, G., Chang, G. J., Tsuchiya, K. R., Karabatsos, N., and Cropp, C. B. 1998. Phylogeny of the genus

flavivirus. *J Virol*, 72(1): 73–83.

Lucaci, A. G., Wisotsky, S. R., Shank, S. D., Weaver, S., and Kosakovsky Pond, S. L. 2021. Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes. *PLoS One*, 16(3): e0248337.

MacLean, O. A., Lytras, S., Weaver, S., Singer, J. B., Boni, M. F., Lemey, P., Kosakovsky Pond, S. L., and Robertson, D. L. 2021. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS biology*, 19(3): e3001115.

Martin, D. P., Weaver, S., Tegally, H., San, J. E., Shank, S. D., Wilkinson, E., Lucaci, A. G., Giandhari, J., Naidoo, S., Pillay, Y., Singh, L., Lessells, R. J., NGS-SA, COVID-19 Genomics UK (COG-UK), Gupta, R. K., Wertheim, J. O., Nekturenko, A., Murrell, B., Harkins, G. W., Lemey, P., MacLean, O. A., Robertson, D. L., de Oliveira, T., and Kosakovsky Pond, S. L. 2021. The emergence and ongoing convergent evolution of the sars-cov-2 n501y lineages. *Cell*, 184(20): 5189–5200.e7.

Martin, D. P., Lytras, S., Lucaci, A. G., Maier, W., Grüning, B., Shank, S. D., Weaver, S., MacLean, O. A., Orton, R. J., Lemey, P., Boni, M. F., Tegally, H., Harkins, G. W., Scheepers, C., Bhiman, J. N., Everatt, J., Amoako, D. G., San, J. E., Giandhari, J., Sigal, A., NGS-SA, Williamson, C., Hsiao, N.-Y., von Gottberg, A., De Klerk, A., Shafer, R. W., Robertson, D. L., Wilkinson, R. J., Sewell, B. T., Lessells, R., Nekrutenko, A., Greaney, A. J., Starr, T. N., Bloom, J. D., Murrell, B., Wilkinson, E., Gupta, R. K., de Oliveira, T., and Kosakovsky Pond, S. L. 2022. Selection analysis identifies clusters of unusual mutational changes in omicron lineage ba.1 that likely impact spike function. *Mol Biol Evol*, 39(4).

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Pond, S. L. K. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genetics*, 8(7): e1002764. Publisher: Public Library of Science.

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., and Kosakovsky Pond, S. L. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*, 32(5): 1365–71.

Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5): 715–724.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Köster, J. 2021. Sustainable data analysis with Snakemake. *F1000Res*, 10: 33.

Pond, S. K. and Muse, S. V. 2005. Site-to-Site Variation of Synonymous Substitution Rates. *Molecular Biology and Evolution*, 22(12): 2375–2385.

Pond, S. K., Delport, W., Muse, S. V., and Scheffler, K. 2010. Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. *PLOS ONE*, 5(7): e11230. Publisher: Public Library of Science.

Posada, D. and Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, 53(5): 793–808.

Rodrigue, N., Latrille, T., and Lartillot, N. 2021. A bayesian mutation-selection framework for detecting site-specific adaptive evolution in protein-coding genes. *Mol Biol Evol*, 38(3): 1199–1208.

Rosenberg, M. S. 2005. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics*, 6: 278.

Seo, T.-K., Kishino, H., and Thorne, J. L. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Mol Biol Evol*,

21(7): 1201–13.

Seoighe, C., Ketwaroo, F., Pillay, V., Scheffler, K., Wood, N., Duffet, R., Zvelebil, M., Martinson, N., McIntyre, J., Morris, L., and Hide, W. 2007. A model of directional selection applied to the evolution of drug resistance in hiv-1. *Mol Biol Evol*, 24(4): 1025–31.

Shu, Y. and McCauley, J. 2017. Gisaid: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*, 22(13).

Spielman, S. J. and Kosakovsky Pond, S. L. 2018. Relative Evolutionary Rates in Proteins Are Largely Insensitive to the Substitution Model. *Molecular Biology and Evolution*, 35(9): 2307–2317.

Steward, R. A., de Jong, M. A., Oostra, V., and Wheat, C. W. 2022. Alternative splicing in seasonal plasticity and the potential for adaptation to environmental change. *Nature Communications*, 13(1): 755. Number: 1 Publisher: Nature Publishing Group.

Su, C., Nguyen, V. K., and Nei, M. 2002. Adaptive evolution of variable region genes encoding an unusual type of immunoglobulin in camelids. *Mol Biol Evol*, 19(3): 205–15.

Tamuri, A. U. and Dos Reis, M. 2022. A mutation-selection model of protein evolution under persistent positive selection. *Mol Biol Evol*, 39(1).

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Some mathematical questions in biology / DNA sequence analysis edited by Robert M. Miura*. Publisher: Providence, R.I. American Mathematical Society, c1986.

Venkat, A., Hahn, M. W., and Thornton, J. W. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol*, 2(8): 1280–1288.

Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Althaus, C. L., Anyaneji, U. J., Bester, P. A., Boni, M. F., Chand, M., Choga, W. T., Colquhoun, R., Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., Giovanetti, M., Hardie, D., Hill, V., Hsiao, N.-Y., Iranzadeh, A., Ismail, A., Joseph, C., Joseph, R., Koopile, L., Kosakovsky Pond, S. L., Kraemer, M. U. G., Kuate-Lere, L., Laguda-Akingba, O., Lesetedi-Mafoko, O., Lessells, R. J., Lockman, S., Lucaci, A. G., Maharaj, A., Mahlangu, B., Maponga, T., Mahlakwane, K., Makatini, Z., Marais, G., Maruapula, D., Masupu, K., Matshaba, M., Mayaphi, S., Mbhele, N., Mbulawa, M. B., Mendes, A., Mlisana, K., Mnguni, A., Mohale, T., Moir, M., Moruisi, K., Mosepele, M., Motsatsi, G., Motswaledi, M. S., Mphoyakgosi, T., Msomi, N., Mwangi, P. N., Naidoo, Y., Ntuli, N., Nyaga, M., Olubayo, L., Pillay, S., Radibe, B., Ramphal, Y., Ramphal, U., San, J. E., Scott, L., Shapiro, R., Singh, L., Smith-Lawrence, P., Stevens, W., Strydom, A., Subramoney, K., Tebeila, N., Tshiabuila, D., Tsui, J., van Wyk, S., Weaver, S., Wibmer, C. K., Wilkinson, E., Wolter, N., Zarebski, A. E., Zuze, B., Goedhals, D., Preiser, W., Treurnicht, F., Venter, M., Williamson, C., Pybus, O. G., Bhiman, J., Glass, A., Martin, D. P., Rambaut, A., Gaseitsiwe, S., von Gottberg, A., and de Oliveira, T. 2022. Rapid epidemic expansion of the sars-cov-2 omicron variant in southern africa. *Nature*, 603(7902): 679–686.

Wagenmakers, E.-J. and Farrell, S. 2004. AIC model selection using Akaike weights. *Psychon Bull Rev*, 11(1): 192–196.

Wang, S., Zong, Y., Lin, Q., Zhang, H., Chai, Z., Zhang, D., Chen, K., Qiu, J.-L., and Gao, C. 2020. Precise, predictable multi-nucleotide deletions in rice and wheat using APOBEC–Cas9. *Nature Biotechnology*, 38(12): 1460–1465. Number: 12 Publisher: Nature Publishing Group.

Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D., and Muse, S. V. 2020. 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.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.

Mol Biol Evol, 15(5): 568–73.

Yang, Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *J Mol Evol*, 51(5): 423–32.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1): 431–49.

Yang, Z., Swanson, W. J., and Vacquier, V. D. 2000b. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol*, 17(10): 1446–55.

Yokoyama, S., Tada, T., Zhang, H., and Britt, L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci U S A*, 105(36): 13480–5.

TITLE · doi:10.1093/molbev/mst012 MBE

Supplementary Material

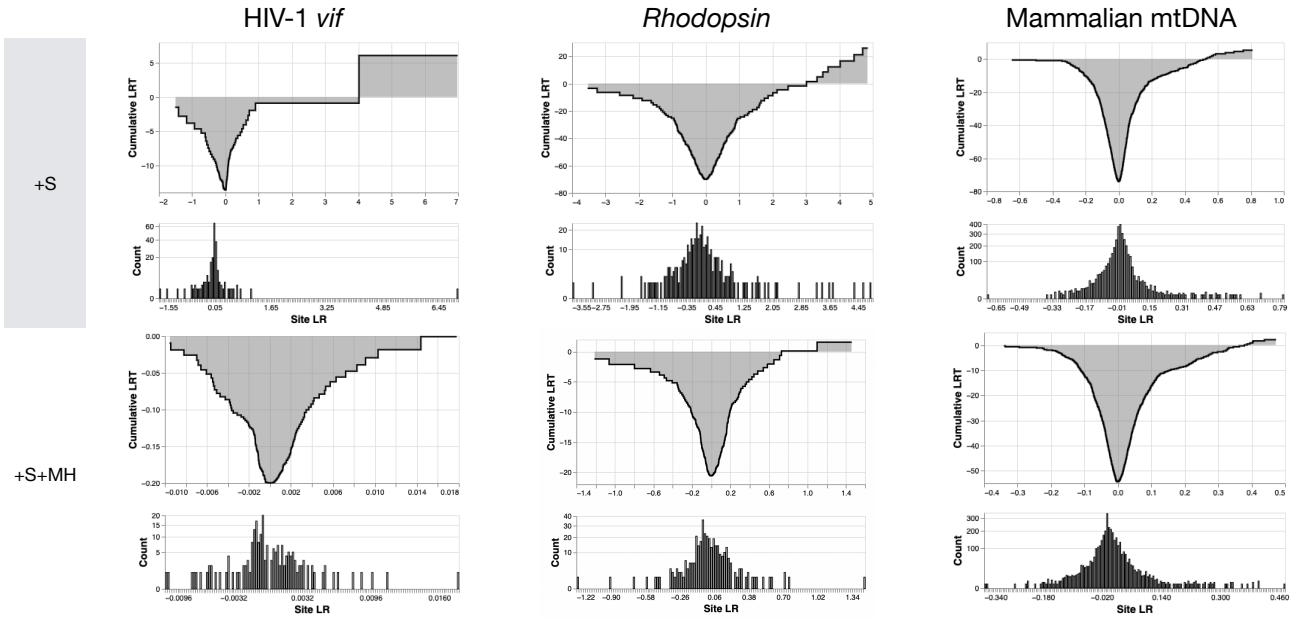

FIG. S1. Site-level support for Episodic Diversifying Selection in three benchmark alignments. Each dataset / model panel includes two views of the same data: the top plot is the cumulative value of the likelihood ratio test statistic (LRT) for the EDS test over sites, where site-level LRT are sorted from smallest to largest; the bottom plot is the histogram of site-level LRTs.

Scenario	$\omega_3(p_3)$			δ		ψ		Detection (# if AIC_c is better)			+S+MH pref
	Truth	+S	+S+MH	Truth	+S+MH	Truth	+S+MH	+S	+S+MH	Averaged	by LRT
Null simulations (no positive selection, MH present)											
adh/N1	**1.0**(2.38%)	1.00 – 2.42 (1.6%)	1.00 – 2.63 (1%)	0.003	0.00 – 0.04	0.0	0.00 – 0.00	0.03 (97)	0.01 (3)	0.01	0
adh/N2	**1.0**(2.38%)	1.00 – 3.60 (1.7%)	1.00 – 1.93 (0.98%)	0.1	0.00 – 0.12	0.0	0.00 – 0.07	0.1 (80)	0 (20)	0.04	0.14
adh/N3	**1.0**(2.38%)	1.30 – 5.21 (2.3%)	1.00 – 2.26 (1.2%)	0.25	0.10 – 0.27	0.0	0.00 – 0.08	0.26 (47)	0.03 (53)	0.05	0.39
adh/N4	**1.0**(2.38%)	2.07 – 7.98 (2.3%)	1.00 – 2.69 (1.3%)	0.5	0.33 – 0.56	0.0	0.00 – 0.06	0.61 (6)	0.02 (94)	0.02	0.9
adh/N5	**1.0**(2.38%)	3.05 – 19.29 (2.1%)	1.00 – 2.32 (1.1%)	0.75	0.57 – 0.79	0.0	0.00 – 0.12	0.87 (0)	0.03 (100)	0.03	1
Hepatitis D Ag/N1	**1.0**(1.71%)	1.19 – 10.93 (13%)	1.00 – 1.04 (6.9%)	0.14	0.08 – 0.15	0.0	0.00 – 0.00	0.34 (32)	0.01 (68)	0.06	0.54
HIV *vif*/N1	**1.0**(1.00%)	1.07 – 50.21 (14%)	1.00 – 1.90 (7.2%)	0.004	0.00 – 0.00	0.16	0.08 – 0.22	0.32 (32)	0.01 (68)	0.01	0.63
Rhodopsin/N1	**1.0**(0.37%)	3.37 – 14.11 (1.1%)	1.00 – 1.69 (0.6%)	0.35	0.27 – 0.37	0.52	0.27 – 0.64	0.9 (1)	0.02 (100)	0.03	0.99
Strep. PTS/N1	**1.0**(1.56%)	1.48 – 12.41 (1.6%)	1.00 – 5.59 (0.93%)	0.31	0.15 – 0.40	1.1	0.70 – 1.46	0.31 (1)	0.03 (99)	0.03	0.99
Power simulations (positive selection, MH absent)											
adh/P1	4.14(2.50%)	3.21 – 4.79 (2.9%)	3.30 – 4.91 (2.8%)	0.0	0.00 – 0.00	0.0	0.00 – 0.00	0.91 (99)	0.85 (1)	0.92	0.01
β-globin/P1	8.925(3.70%)	5.53 – 10.96 (5%)	5.73 – 9.67 (4.8%)	0.0	0.00 – 0.00	0.0	0.00 – 0.03	1 (91)	0.98 (9)	1	0.05
HIV *vif*/P1	2103(0.05%)	1.09 – 3142.84 (9.7%)	1.00 – 2.54 (6.2%)	0.0	0.00 – 0.02	0.0	0.00 – 0.11	0.56 (80)	0.1 (20)	0.53	0.14
Mam. mtDNA/P1	1.434(1.33%)	1.29 – 1.63 (1.3%)	1.21 – 1.49 (1.4%)	0.0	0.00 – 0.03	0.0	0.00 – 0.05	0.65 (90)	0.5 (10)	0.64	0.06
Rhodopsin/P1	6.376(1.31%)	5.13 – 7.38 (1.4%)	5.16 – 7.50 (1.4%)	0.0	0.00 – 0.00	0.0	0.00 – 0.10	1 (99)	0.98 (2)	0.99	0.01
Power simulations (positive selection, MH present)											
adh/P2	4.05(2.38%)	3.23 – 5.40 (2.6%)	3.04 – 5.05 (2.5%)	0.003	0.00 – 0.03	0.0	0.00 – 0.06	0.9 (94)	0.75 (6)	0.86	0.02
β-globin/P2	2.834(6.08%)	2.71 – 6.40 (7%)	1.88 – 4.54 (7.4%)	0.24	0.01 – 0.28	0.19	0.00 – 0.14	0.88 (56)	0.4 (44)	0.55	0.33
Hepatitis D Ag/P1	11.3(1.71%)	8.93 – 19.65 (3.7%)	2.40 – 9.86 (9.4%)	0.14	0.07 – 0.17	0.0	0.00 – 0.07	1 (32)	0.58 (68)	0.65	0.57
HIV *vif*/P2	1.226(1.00%)	1.64 – 319.54 (9.4%)	1.00 – 3.12 (6.2%)	0.004	0.00 – 0.00	0.16	0.08 – 0.23	0.46 (35)	0.01 (65)	0.04	0.52
Rhodopsin/P2	5.453(0.37%)	5.61 – 15.94 (0.96%)	1.00 – 6.27 (0.73%)	0.35	0.26 – 0.35	0.52	0.25 – 0.81	1 (1)	0.24 (100)	0.25	0.99
Rhodopsin/P3	5.453(0.37%)	4.14 – 9.24 (1.1%)	1.00 – 5.47 (0.61%)	0.35	0.26 – 0.37	**0.0**	0.00 – 0.18	0.96 (0)	0.22 (100)	0.22	0.97
Rhodopsin/P4	5.453(0.37%)	3.59 – 8.65 (0.72%)	2.23 – 7.11 (0.71%)	**0.10**	0.03 – 0.13	**0.0**	0.00 – 0.08	0.87 (70)	0.32 (30)	0.55	0.18
Rhodopsin/P5	5.453(0.37%)	3.15 – 8.12 (1.1%)	1.37 – 6.64 (0.83%)	**0.20**	0.14 – 0.24	**0.0**	0.00 – 0.09	0.85 (21)	0.27 (79)	0.37	0.7
Rhodopsin/P6	5.453(0.37%)	4.47 – 20.37 (0.6%)	1.62 – 6.06 (0.71%)	**0.00**	0.00 – 0.01	0.52	0.35 – 0.60	0.96 (26)	0.28 (74)	0.39	0.62
Rhodopsin/P7	5.453(**2.1**%)	7.21 – 11.00 (2.2%)	4.65 – 6.46 (2.2%)	0.35	0.28 – 0.38	**0.0**	0.00 – 0.12	1 (3)	0.98 (97)	0.99	0.97
Rhodopsin/P8	5.453(**4.2**%)	7.51 – 9.64 (4.1%)	4.78 – 6.15 (4.4%)	0.35	0.26 – 0.36	**0.0**	0.00 – 0.06	1 (3)	0.99 (97)	1	0.97
Strep. PTS/P1	9.489(1.56%)	17.10 – 71.54 (1.4%)	7.55 – 12.88 (1.6%)	0.31	0.15 – 0.43	1.1	0.65 – 1.45	0.99 (0)	0.97 (100)	0.97	1
SARS-CoV-2 S/P1	5.990(20.12%)	4.64 – 7.63 (27%)	4.86 – 8.94 (22%)	0.012	0.00 – 0.00	0.0	0.00 – 0.00	0.99 (100)	0.97 (0)	0.99	0

Table S1. BUSTED test performance on synthetic data, under model fits from benchmark datasets to parametrize various simulation scenarios simulations (100 replicates each). `Truth` – values used for data generation; parameters changed from their MLE values from the corresponding empirical dataset are shown in **boldface**. For model rate estimates, interquartile range is shown. For proportion estimates, mean value is shown. `Detection` columns shows the fraction of replicates where the LRT for episodic diversifying selection yields $p \leq 0.05$; and the value in parentheses – the number of replicates where the corresponding model was preferred by AIC_c. `Detection / Averaged` – the fraction of replicates where model-averaged LRT p-value was ≤ 0.05. The last column shows the fraction of replicates for which the +S+MH model was preferred to the +S model, using the χ^2_2 based LRT $p \leq 0.05$.

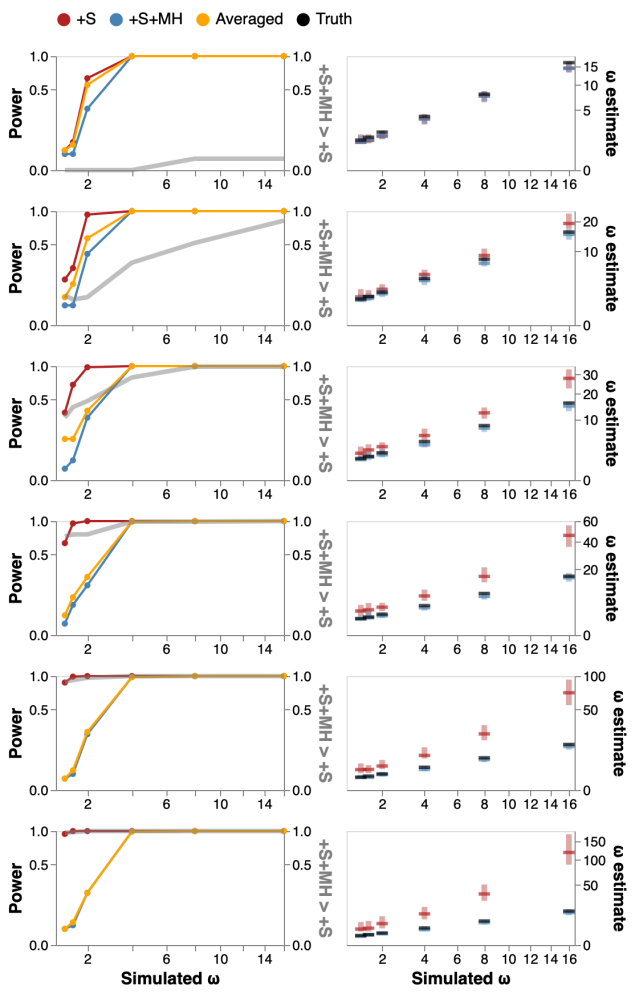

FIG. S2. Model performance on data simulated with EDS (25% selected fraction). Left column : detection rate for EDS (at $p \leq 0.05$) as a function of rate ω_3 (effect size) and δ (confounding parameter), and the rate at which +S+MH is preferred to +S by a nested LRT test. Right column: ω_3 estimates (median, IQR) for various simulation scenarios.