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

Alexander G. Lucaci ,[1] Jordan D. Zehr ,[1] David Enard,[2] Joseph W. Thornton ,[3,4] and Sergei L. Kosakovsky Pond *,[1]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA
[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona
[3]Department of Human Genetics, University of Chicago, Chicago, Illinois
[4]Department of Ecology & Evolution, University of Chicago, Chicago, Illinois

***Corresponding author:** E-mail: spond@temple.edu.

Associate editor: Kelley Harris

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 and tests. 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. Previous work established that failing to accommodate multinucleotide (or multihit, MH) substitutions strongly biases dN/dS**-based inference towards false-positive inferences of diversifying episodic selection, as does failing to model variation in the rate of synonymous substitution (SRV) among sites. Here, we develop an integrated analytical framework and software tools to simultaneously incorporate these sources of evolutionary complexity into selection analyses. We found that both MH and SRV are ubiquitous in empirical alignments, and incorporating them has a strong effect on whether or not positive selection is detected (**1.4**-fold reduction) and on the distributions of inferred evolutionary rates. With simulation studies, we show that this effect is not attributable to reduced statistical power caused by using a more complex model. 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 show that MH substitutions occurring along shorter branches in the tree explain a significant fraction of discrepant results in selection detection. Our results add to the growing body of literature which examines decades-old modeling assumptions (including MH) and finds them to be problematic for comparative genomic data analysis. Because multinucleotide substitutions have a significant impact on natural selection detection even at the level of an entire gene, we recommend that selection analyses of this type consider their inclusion as a matter of routine. To facilitate this procedure, we developed, implemented, and benchmarked a simple and well-performing model testing selection detection framework able to screen an alignment for positive selection with two biologically important confounding processes: site-to-site synonymous rate variation, and multinucleotide instantaneous substitutions.**

Key words: **molecular evolution, evolutionary shortcuts, multinucleotide substitutions, codon-substitution models.**

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 using codon-substitution models (e.g., as implemented in Kosakovsky Pond et al. 2020) is a workhorse of selection detection. Its seminal methods were published several decades ago (Goldman and Yang 1994; Muse and Gaut 1994) and still enjoy widespread use. 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 it has been shown to inflate the rates of false-positive inferences in selection tests (Pond and Muse 2005; Davydov et al. 2019; Wisotsky et al. 2020). This finding and its repeated corroborations have led us to recently recommend that synonymous rate variation (SRV) be incorporated into all selection tests as a matter of routine (Wisotsky et al. 2020).

Another important modeling assumption is that codon substitutions involving 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 implemented in classic (and still used in the vast majority of applications) tests of selection by setting the instantaneous rates for all

multinucleotide substitutions to zero. Numerous genetic and molecular studies (Matsuda et al. 2000; Arana et al. 2008; Loeb and Monnat 2008; Hodgkinson and Eyre-Walker 2011; Schrider et al. 2011; Saribasak et al. 2012; Stone et al. 2012; Chen et al. 2014, 2015; Harris and Nielsen 2014; Seplyarskiy et al. 2015; Besenbacher et al. 2016; Assaf et al. 2017; Prendergast et al. 2019) have established that replication errors often occur at adjacent sites, producing multinucleotide (or multihit, MH) mutations that account for ∼1–3% of all mutations through several mechanisms, including polymerase ζ. Multiple studies have shown that realistic levels of MH mutations of just a few percent can produce uncontrolled rates of false-positive inferences of episodic diversifying selection (EDS) and impact selection inference with codon-based models, for example, De Maio et al. (2013). This is because when a standard model "crams" multiple nonsynonymous single-nucleotide substitutions onto a short branch to explain a nonsynonymous MH substitution, the dN/dS ratio estimate will be artificially inflated to accommodate this. Further, in two genome-wide datasets, MH substitutions within the same codon have been found to account for virtually all the support for episodic diversifying selection, and re-analyzing the data with a model that incorporates MHs as single substitution events dramatically reduces the number of positive inferences of diversifying selection (Venkat et al. 2018). A subsequent study confirmed the high rate of false-positive inferences that can be caused by MHs (Dunn et al. 2019); another elaborated the MH model and showed that it provides a significantly improved statistical fit to most genes in a collection of >42,000 empirical alignments (Lucaci et al. 2021).

The combined effect of SRV and MH on selection tests has not been systematically investigated, however. Further, no comprehensive approach has yet been developed that simultaneously incorporates both phenomena for testing hypotheses of positive selection. Here, we present such a framework: we modified the BUSTED method for EDS detection (Murrell et al. 2015) to allow synonymous rate variation among codons (+S) and multinucleotide substitutions (+MH, including both double- and triple-nucleotide substitutions), and developed model-selection and model-averaging approaches that allow either or both forms of heterogeneity to be incorporated according to their fit to the sequence data. We used this method to re-evaluate the evidence for positive selection in 21 diverse benchmark alignments that historically have played an important role in the development of tests for detecting selection. To understand the power and accuracy of this approach, we applied our framework to both simulated data and a large alignment of 9, 181 of proteins in orthologs of 24 mammlian species, previously studied by Enard et al. (2016) in the context selection as a result of interactions with viral proteins. Our practical framework can be used to balance the impetus to prevent false-positive inferences of selection when S and/or MH effects are present but not incorporated, and the desire to prevent the loss of statistical power, which can occur when models of greater parametric complexity are used.

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 nonzero fraction of branch–site combinations evolving with $\omega > 1$, but makes different assumptions about confounding evolutionary processes. We evaluate four models:

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

(cf. table 1 for additional details). These models form a nested hierarchy ($BUSTED \subset +S \subset +S+MH$ and $BUSTED \subset +MH \subset +S+MH$) and can be compared using either information theoretic criteria or pairwise likelihood ratio tests. We test for selection using the entire alignment (all branches, all sites) because that's the most common application of EDS detection, and will

Table 1. Substitution Models Considered in this Paper.

Model	Reference	Nonsynonymous Rates	Synonymous Rates	Multinucleotide 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	$B + 13 + 2 \times (K + L)$

NOTE.—B, the number of branches in the phylogenetic tree. K and L are user-tunable parameters, set to 3 each by default.

Table 2. Selection Analysis on Benchmark Alignments (Arranged by EDS Detection Class First and Further Sorted by Model-Averaged *P*-value).

Alignment	*N*	*S*	*L*	AIC_c +S +MH	ΔAIC_c vs. +S+MH			*P*-value for EDS					EDS Detection
					BUSTED	+S	+MH	+S+MH	BUSTED	+S	+MH	Averaged	
Mammalian RBP3	54	412	4.71	43,083.9	577.6	1.5	575.8	0.4939	0.4530	0.3321	0.5000	0.4427	All, no
Mammalian VWF	62	392	5.37	45,992.5	940.3	5.5	931.8	0.5000	0.0767	0.1513	0.4988	0.4786	All, no
Flavivirus NS5	18	342	9.42	18,488.0	301.9	43.1	280.8	0.5000	0.4218	0.4999	0.5000	0.5000	All, no
Primate Lysozyme	19	130	0.25	2,149.8	16.0	−4.2	20.3	0.5000	0.3668	0.5000	0.3845	0.5000	All, no
Primate COXI	21	510	11.25	24,292.0	101.2	−3.4	106.3	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
Encephalitis *env*	23	500	0.89	13,703.1	42.3	−4.0	44.1	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
ADORA3	67	107	4.61	12,612.2	251.0	−3.6	255.2	0.5000	0.5000	0.5000	0.5000	0.5000	All, no
Sperm lysin	25	134	4.46	8,765.3	156.3	−4.2	158.8	0.0000	0.0000	0.0000	0.0000	0.0000	All, yes
IAV H1N1 HA	466	589	2.15	51,414.6	912.0	−3.8	913.4	0.0000	0.0000	0.0000	0.0003	0.0000	All, yes
rbcL	483	466	11.88	152,988.8	4,341.6	76.5	4,315.7	0.0000	0.0000	0.0000	0.0000	0.0000	All, yes
SARS-CoV-2 S	180	1,284	0.13	17,817.4	649.3	−3.7	624.7	0.0002	0.0000	0.0000	0.0002	0.0001	All, yes
HIV rt	476	335	7.19	52,033.6	1,717.5	1.4	1,721.4	0.0006	0.0000	0.0000	0.0000	0.0004	All, yes
Camelid VHH	212	96	15.87	33,665.6	1,474.4	28.5	1,424.7	0.0040	0.0000	0.0000	0.0000	0.0040	All, yes
Drosophila *adh*	23	254	1.76	9,357.8	14.1	−3.9	17.2	0.0255	0.0003	0.0016	0.0197	0.0046	All, yes
Hepatitis D Ag	33	196	2.23	10,416.1	281.1	8.0	259.9	0.0314	0.0000	0.0000	0.0019	0.0309	All, yes
IAV H3N2 HA	349	329	1.39	23,228.2	637.2	14.4	630.8	0.5000	0.0000	0.1060	0.3637	0.4997	+S/+S+MH, no
Streptococcus PTS	16	639	11.27	17,344.6	206.5	31.5	169.1	0.0000	0.0007	0.0000	0.0879	0.0000	+S/+S+MH, yes
Mam. β-globin	17	144	3.85	7,420.9	31.9	−5.2	36.4	0.2219	0.0000	0.0000	0.0485	0.0154	Discordant
Mammalian mtDNA	20	3,331	10.09	179,797.7	1,688.2	11.8	1,697.3	0.1713	0.0061	0.0346	0.0204	0.1710	Discordant
Rhodopsin	38	330	5.32	25,902.4	495.1	21.7	483.6	0.2279	0.0000	0.0000	0.0000	0.2279	Discordant
HIV *vif*	29	192	0.95	6,911.0	190.8	2.8	187.2	0.5000	0.0002	0.0239	0.0424	0.4052	Discordant
				12	0	9	0	9	14	13	12	10	

NOTE.—*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, that is, 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 EDS under the corresponding model (4 digits of precision); shown in boldface if ≤0.05. The Averaged 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.

enable direct comparisons with published results. Our implementation does allows for a priori selection of branches to include for testing, treating the rest as a nuisance background.

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 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 multinucleotide 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 arrived at a negative result. Some of these examples confirm previous negative findings, and some contradict previous positive (potentially weak) findings.

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 (see tables 2 and 3).

Tick-borne flavivirus NS-5 gene (+S+MH). This dataset was analyzed in Yang, Nielsen, et al. (2000) 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

Table 3. Substitution Process Characterization on Benchmark Alignments.

Alignment	$\omega_3(p_3)$		$CV(\alpha)$	δ (% exp.)	ψ (% exp.)	Substitutions (%)			L		
	+MH	+S				1H	2H	3H	1H	2H	3H
Mammalian RPB3	1.258 (0.44%)	1.541 (2.03%)	0.593	0.111 (1.5%)	0.054 (0.07%)	4093 (9.46)	300 (0.69)	18 (0.04)	0.08	0.10	0.11
Mammalian VWF	1.073 (0.36%)	1.973 (2.35%)	0.643	0.130 (2.2%)	—	4,608 (9.71)	381 (0.80)	22 (0.05)	0.08	0.11	0.17
Flavivirus NS5	1.105 (0.00%)	1.006 (2.23%)	1.286	0.377 (2.5%)	0.986 (0.6%)	1,956 (17.33)	270 (2.39)	58 (0.51)	0.48	0.58	0.58
Primate 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
Primate COXI	1.021 (1.06%)	1.000 (1.09%)	2.342	—	—	3,160 (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 (0.1%)	—	1,068 (4.97)	26 (0.12)	0 (0.00)	0.04	0.05	0.00
ADORA3	1.013 (0.00%)	1.000 (3.46%)	0.662	0.053 (0.8%)	—	1,135 (8.35)	75 (0.55)	3 (0.02)	0.06	0.09	0.14
Sperm 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
IAV H1N1 HA	1,039.05 (0.01%)	862.82 (0.01%)	0.835	—	—	3,216 (0.73)	43 (0.01)	2 (0.00)	0.01	0.02	0.00
rbcL	37.569 (0.14%)	49.827 (0.19%)	0.831	0.113 (2.6%)	0.016 (0.06%)	12,185 (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 (0.6%)	—	421 (0.13)	13 (0.00)	0 (0.00)	0.00	0.00	0.00
HIV rt	50.714 (0.07%)	48.230 (0.10%)	0.940	0.028 (0.4%)	—	4,149 (1.35)	129 (0.04)	10 (0.00)	0.02	0.02	0.03
Camelid VHH	9.193 (2.53%)	24.513 (2.04%)	0.817	0.157 (6.4%)	—	2,393 (6.91)	528 (1.52)	72 (0.21)	0.09	0.11	0.12
Drosophila adh	4.056 (2.38%)	4.144 (2.50%)	0.594	0.032 (0.8%)	—	693 (6.34)	66 (0.60)	14 (0.13)	0.10	0.15	0.16
Hepatitis D Ag	11.306 (1.71%)	16.249 (1.96%)	0.902	0.143 (4.9%)	—	665 (5.39)	122 (0.99)	14 (0.11)	0.08	0.13	0.18
IAV H3N2 HA	1.002 (0.00%)	1.550 (29.48%)	1.064	0.062 (1.5%)	0.015 (0.04%)	1,320 (0.74)	29 (0.02)	1 (0.00)	0.00	0.00	0.01
Streptococcus PTS	9.489 (1.56%)	11.871 (1.85%)	1.046	0.310 (6.7%)	1.054 (3.9%)	1245 (6.72)	293 (1.58)	76 (0.41)	1.79	1.98	1.96
Mam. β-globin	2.834 (6.08%)	8.925 (3.70%)	1.263	0.24 (7.0%)	—	526 (11.78)	110 (2.46)	24 (0.54)	0.26	0.36	0.52
Mammalian mtDNA	1.310 (1.04%)	1.434 (1.33%)	1.268	0.227 (1.7%)	—	19,892 (16.14)	1,873 (1.52)	225 (0.18)	0.42	0.57	0.67
Rhodopsin	5.453 (0.37%)	6.376 (1.31%)	1.403	0.345 (4.4%)	0.515 (0.9%)	2,488 (10.77)	257 (1.11)	45 (0.19)	0.12	0.15	0.16
HIV *vif*	1.226 (1.00%)	2,103.28 (0.05%)	1.049	0.004 (0.1%)	0.163 (0.7%)	446 (4.30)	20 (0.19)	5 (0.05)	0.03	0.04	0.04

Note.—$\omega_3(p_3)$, 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. Numbers in parentheses show fractions (as %) of expected substitutions attributable to 2H and 3H substitutions, respectively. 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. Row ordering is that same as in table 2.

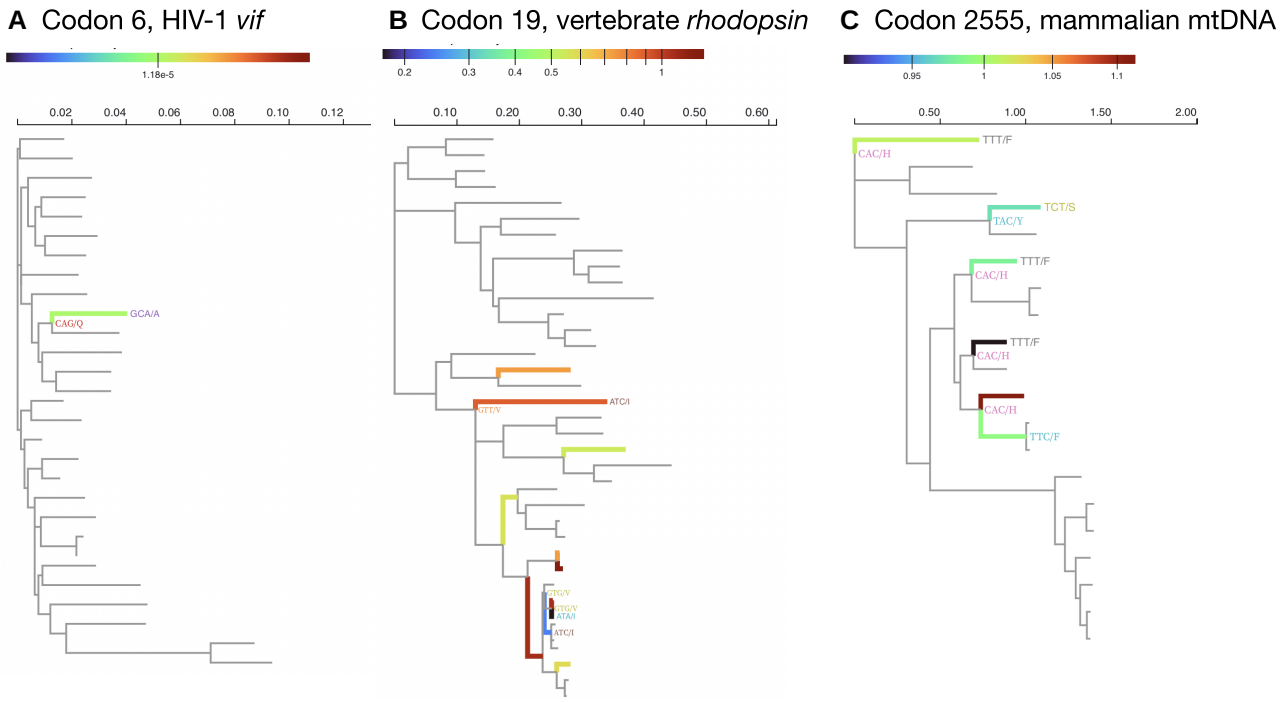

Fig. 1. Example sites from benchmark alignments with discordant selection signal. Only substitutions involving multiple nucleotides are labeled (codon/amino acid). Coloring of the branches represents the ratio of empirical Bayes factors for $\omega > 1$ at this branch/site (see Methods) between the +S+MH 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. Multiple nucleotide substitutions can have impacts on a broad set of biological protein-coding datasets including examples such as (*A*) virus, (*B*) vertebrate, and (*C*) mammalian.

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 nonsynonymous rates, but strong conservation ($\omega \ll 1$) overall. We find no evidence of MNS, including 0 point estimates 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, Nielsen, et al. (2000), 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, Nielsen, et al. (2000) and Rodrigue et al. (2021). Most analyses failed to detect evidence of diversifying selection, despite a 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 2.5% (of branch–site pairs) to the positively selected regime.

Lysin (+S). This alignment of abalone sperm lysin from Yang, Swanson, et al. (2000) is a canonical example of diversifying positive selection, for example, due to self-incompatibility constraints. There is no support for MNS in this alignment despite relatively high divergence and numerous multinucleotide 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 ($\omega \approx 11.3$). The MNS signal is entirely due to double-nucleotide substitutions ($\hat{\delta} = 0.143$). While all models have $P \leq 0.05$ 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 ($\omega \approx 9.2$) 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 (∼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 (∼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 (∼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 ($\omega \sim 1,000$) positive selection on a very small (∼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.

β-globin (+S). Mammalian β-globin is one of the datasets from Yang, Nielsen, et al. (2000) 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, Nielsen, et al. (2000), 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 supplementary figure S2, Supplementary Material online for the +S model. Furthermore, a single triple-nucleotide substitution along a terminal tree branch at that 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 (fig. 1A). Multinucleotide substitutions along short branches have been shown to return false-positive selection detection results in simulations (Venkat et al. 2018; Lucaci et al. 2021). Masking a single codon (GCA) with gaps in the alignment and rerunning BUSTED+S yields a nonsignificant *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. Using this dataset as a motivating example, we computed two simple heuristic statistics: how many (branch, site) pairs produce empirical Bayes factors >100 for selection, and how many sites contribute the majority (>80%) of the alignment-wide likelihood ratio test statistic. For this dataset (see supplementary table S2, Supplementary Material online) the entirety of EBF support for the +S model is distributed across only 11 branch–site pairs and a single site. We hypothesize that datasets where only a few sites contribute to a significant EBF signal under the +S model, and there is no evidence of EBD under the +S +MH model could benefit from subsequent manual inspection for possible local alignment errors.

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

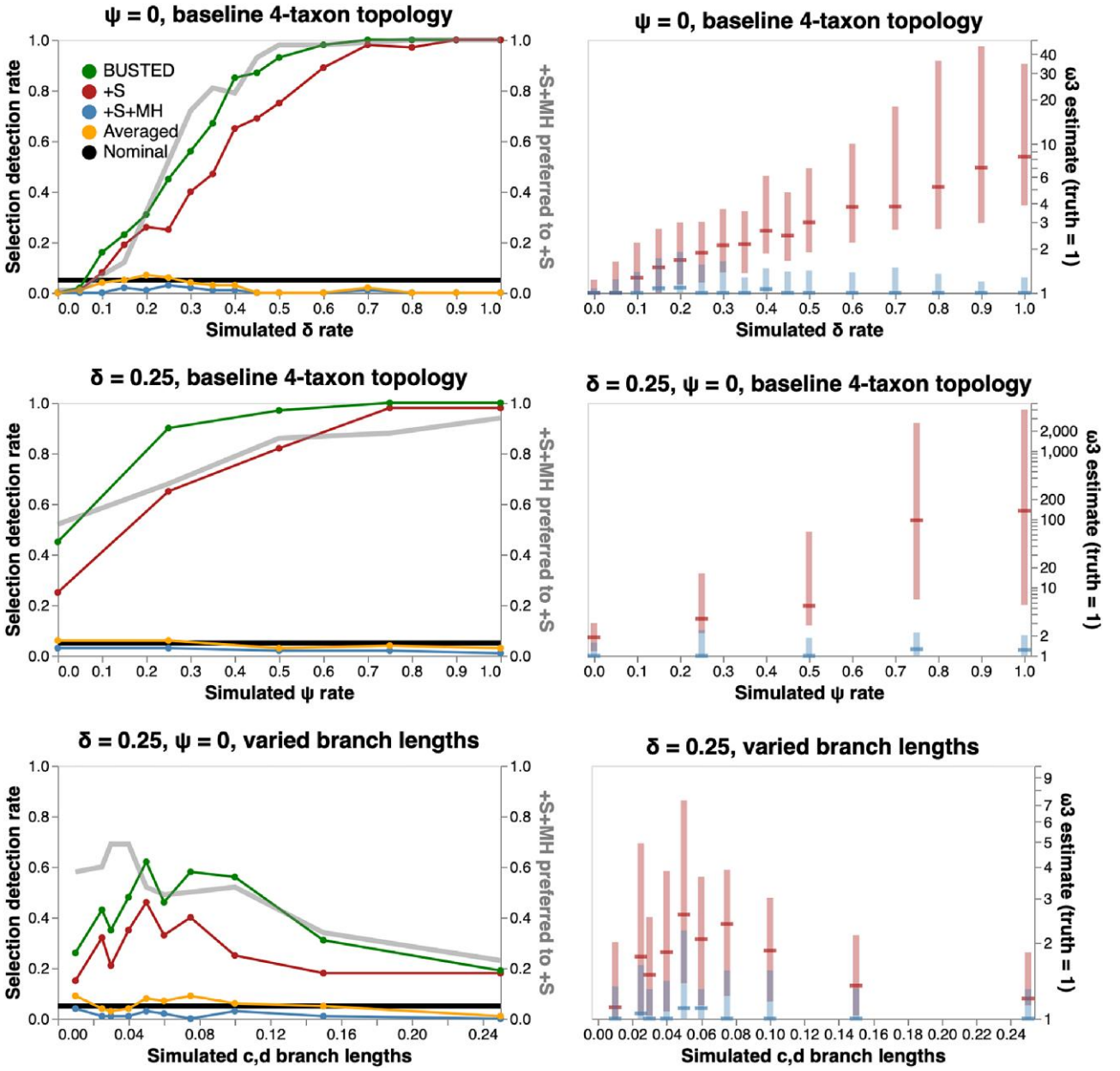

FIG. 2. Model performance on null simulated data. Left column: false-positive detection rate for EDS (at $P \leq 0.05$) as a function of rate parameters and branch lengths, and the rate at which +S+MH is preferred to +S by a nested LRT test. Right column: ω_3 estimates (median, IQR) for various simulation scenarios. One hundred 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.

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, λ_{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 nonzero 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 [fig. 1]), with strong reduction

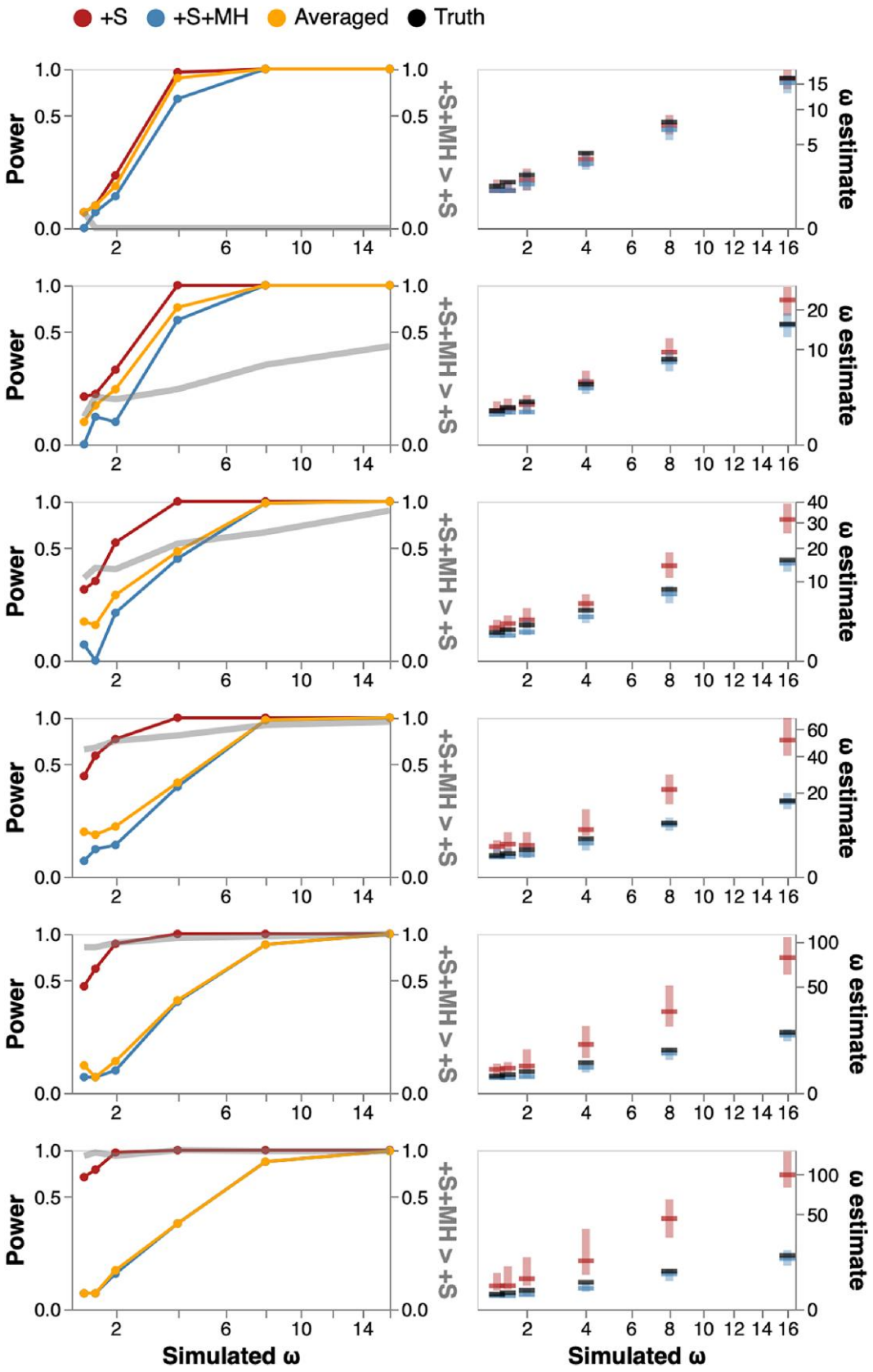

FIG. 3. Model performance on data simulated with EDS. Left column: detection rate for EDS (at $P \leq 0.05$) as a function of rate ω_3 (effect size) and δ (confounding parameter), and the rate at which +S+MH is preferred to +S by a nested LRT test. Right column: ω_3 estimates (median, IQR) for various simulation scenarios. The 2H rate parameter (δ) is varied in increments of 0.1, from 0 (top row) to 0.5 (bottom row).

occurring at short branches which harbor multinucleotide substitutions (fig. 1B). 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 Synthetic 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 (supplementary fig. S2, Supplementary Material online). 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 (fig. 1C). Analysis of 100 replicates generated under the +S model shows that the lack of detection under S+MH is probably not 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 multinucleotide 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 unmodeled multinucleotide 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). $w_m = \exp([AIC_c^{best} - AIC_c^m])/2)$, 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 $p_{MA} \sim \sum_{m=1}^{M} P(LRT \geq LRT_m|\text{null} = m)P(\text{model} = \text{m}|\text{data})$.

The model-averaged approach detects the same 9 datasets as the +S+MH model, and also the β-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.

Decomposing the Contribution of Double and Triple Hits

Because +S+MH model defines separate parameters for 2H and 3H events, we can test which of the two processes contributes to improved data fit. For example, by considering an intermediate model (+S+2H), where δ is estimated, but $\psi := 0$, we can test (1) the hypothesis that $\delta > 0$, by comparing +S+2H to +S and (2) the hypothesis that $\psi > 0$ by comparing +S+2H to +S+MH. For example, among the seven datasets where the maximum likelihood estimate $\hat{\psi} > 0$ in table 1.

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 2 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 (fig. 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.

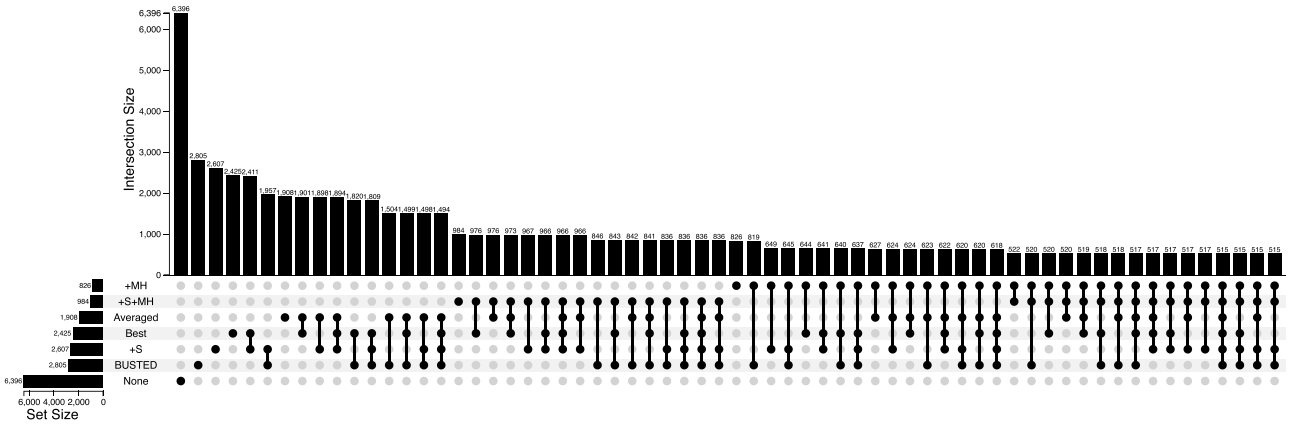

FIG. 4. Alignment classification with respect to diversifying selection for the Enard *et al.* dataset. The number of alignments (9,861 total) which returned LRT $P \leq 0.05$ under each of the following scenarios: individual model (BUSTED, +S, +MH, +S+MH), best model selected by AIC_c (Best), model-averaged "*P*-value" (Averaged). The sizes of each of the nonempty 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. The category "None" shows all those datasets on which all six approaches failed to achieve significance.

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 ~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 multinucleotide substitutions force the +S model to absorb these unmodeled changes into the ω rate, and have the largest effect on FPR rates.

Power. On the same 4-taxon tree, we next simulated alignments with a nonzero 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% (fig. 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 (fig. 3). When the 3H rate is varied and the 2H rate is zero, the same qualitative behavior is observed: power gains for the +S model, but severe biases in ω parameter estimates (supplementary fig. S6, Supplementary Material online). When both 2H and 3H are positive, the 2H component (δ) produces the dominant effect with increasing 3H rates only having a relatively minor additional influence, at least for those parameter values that we considered here.

Biological Interpretability of 2H and 3H Parameter Ranges

For these simulations, the values of δ used (0, 0.1, 0.2, . . . , 1.0) correspond to the range in the expected fraction of substitutions involving 2H changes of 0% through ~30%, when $\psi = 0$ (see supplementary fig. S3, Supplementary Material online). For example, when $\delta = 0.25$ and $\omega_3 = 1.0$, the expected fraction of 2H substitutions is 8.5% which is at the high end of empirically derived values, cf. (table 2). For the scenarios with nonzero 3H rates, and $\delta = 0.25$, the expected fraction of 3H ranges from 0.5% to 5.2%, which is comparable with the range of values inferred from the empirical alignments. The purpose of these simplified simulations is to establish the qualitative behavior of the tests, and some of the larger 2H and 3H (e.g., for $\delta \geq 0.5$ are unlikely to be seen in biological data. However, false-positive rates and parameter estimation biases become significant for biologically common ranges of 2H and 3H rates (e.g., $\delta \leq 0.3$).

Benchmark Datasets

We generated 9 null and 18 power simulations (100 replicates each) based on empirical datasets (details shown in supplementary table S1, Supplementary Material online). 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, 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 multinucleotide substitutions dramatically inflate the estimates of ω rates, and result in significant and often catastrophic FPR.

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.

Point estimates for 2H and 3H rate parameters were positive (at least one of the two) for 5794 alignments, and the means of fractions of substitutions attributable to 2H and 3H were 0.3% and 0.07%, respectively.

Table 4. Model Goodness-of-Fit Ranking for the Enard Dataset.

Model	Rank 1	Rank 2	Rank 3	Rank 4
BUSTED	93	60	7,436	2,272
+S	8,943	853	61	4
+MH	4	62	2,307	7,488
+S+MH	821	8,866	57	97

NOTE.—How each of the four models ranked for each of the 9,861 alignments.

Among the alignments where *+S + MH* or *+MH* were the best-fitting models, the means of these fractions were 6.2% (2H) and 5.6% (3H). Among the alignments where model-averaging detected the presence of EDS (see below), the means of these fractions were 0.3% (2H) and 0.09% (3H). Estimates on the order of a few percent of all substitutions coming from 2H or 3H changes are consistent with those derived from other empirical and mechanistic models (Kosiol et al. 2007; De Maio et al. 2013; Dunn et al. 2019) (see supplementary fig. S7, Supplementary Material online for the distribution of estimated 2H and 3H substitution fractions).

We next considered whether an alignment was detected subject to EDS, using LRT $P \leq 0.05$, under different selection criteria: model fixed 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 2,805/9,861 alignments (28.4%), while the models with MH support have much lower detection rates: 984/9,861 (10%) for +S+MH and 826/9,861 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 nonsignificant (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

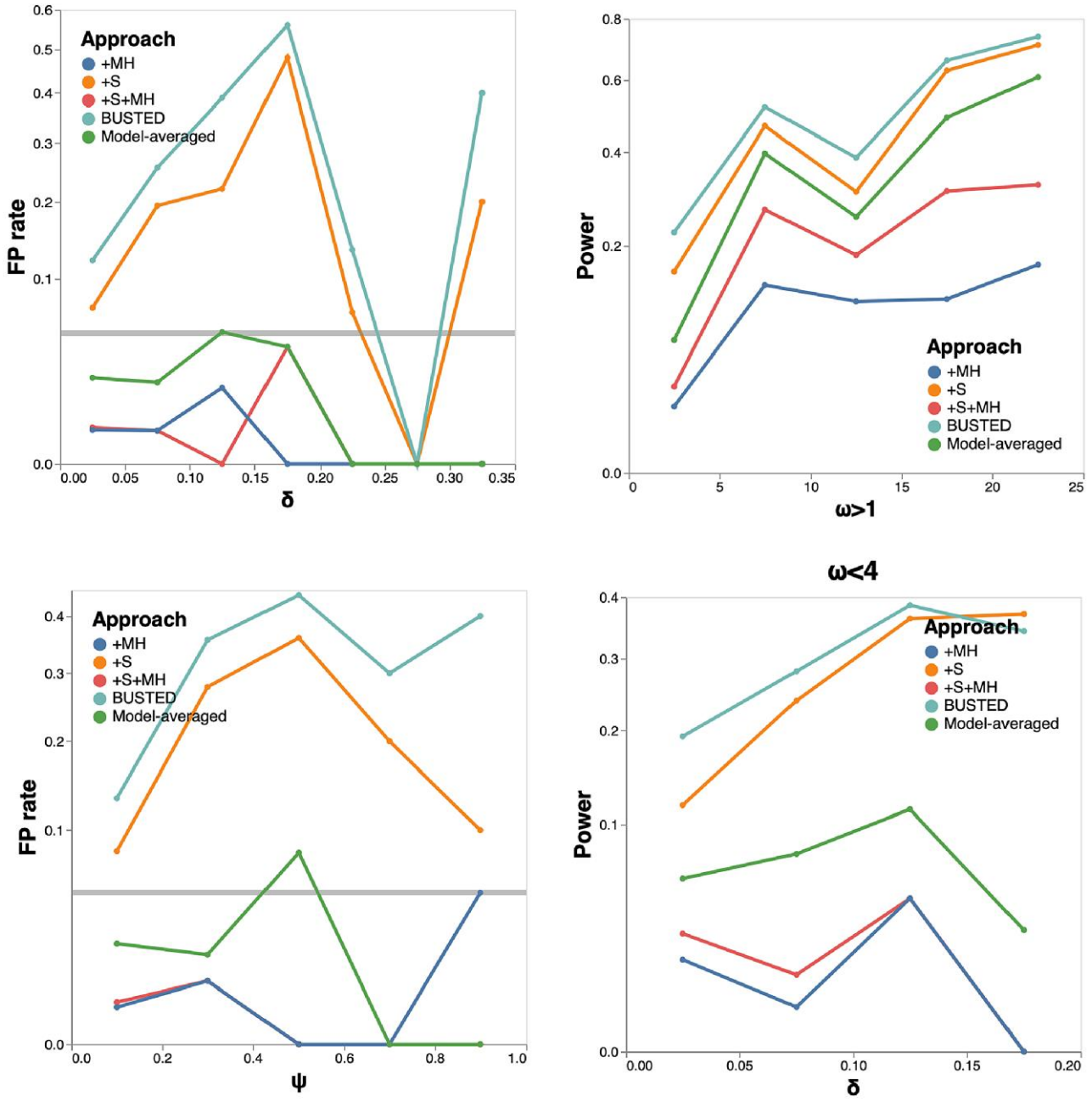

FIG. 5. Simulations based on the Enard *et al.* dataset. False-positive rate of EDS selection for the four individual models and the model-averaged approach, as a function of the 2H rate (δ) and the 3H rate (ψ) used to generate the sequences (left top and bottom). Power for EDS detection as a function of the strength of selection effect ω (top right), and the 2H rate (δ) for the subset of alignments with weak selection effects ($\omega < 4$).

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 (6,396, 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

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.

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	10,966.0	116.3	−4.0	120.1	0.5000	0.5000	0.5000	0.5000	0.5000	(000000)	*	6396
PDIK1L	341	0.47	6,374.4	23.0	−4.1	27.1	0.0080	0.0038	0.0035	0.0118	0.0041	(111111)	+S	489
TIMM50	439	1.38	12,513.2	102.1	−3.6	124.0	0.0008	0.0000	0.0004	0.5000	0.0005	(111011)	+S	297
CDC123	336	0.93	9,464.4	59.0	−3.9	63.2	0.1573	0.0575	0.0244	0.5000	0.0409	(001011)	+S	268
ODF1	250	1.56	7,735.3	108.7	−4.7	111.1	0.5000	0.5000	0.0187	0.5000	0.0609	(001001)	+S	202
ADAMTS1	967	1.44	33,306.6	301.0	−4.0	305.0	0.0443	0.0798	0.0464	0.1532	0.0462	(101011)	+S	125
SDC1	310	1.94	12,722.5	166.9	0.8	162.5	0.0697	0.0000	0.0010	0.1243	0.0419	(****10)	*	7
SF3B6	125	0.64	2,646.7	−6.6	−4.9	−3.7	0.3904	0.0070	0.0060	0.1240	0.0308	(011011)	BUSTED	6
DRC7	868	2.66	30,617.4	298.5	3.2	300.8	0.0169	0.2034	0.1472	0.5000	0.0390	(100011)	+S+MH	2
ELP2	886	1.18	27,640.2	204.8	1.8	207.2	0.0240	0.1074	0.0184	0.2524	0.0224	(101011)	+S+MH	3

NOTE.—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 EDS under the corresponding model (4 digits of precision); shown in boldface if ≤0.05. *Averaged* - the model-averaged *P*-value for ESD (bolded if $f \leq 0.05$. *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), with * used to denote "any model." *Count* - the number of datasets matching this detection pattern, and the specified best-fitting model.

(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 versus 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, that is, 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, that 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 nonsignificant *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 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) versus 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 seven 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 very close in terms of AIC_c), and the model-averaging approach arrives at a significant *P*-value.

The Effect of EDS Detection Strategies Under Muptiple Testing Correction. Depending on the specific type of analysis, large-scale screens of genes for a specific feature (e.g., EDS) often require some form of multiple-testing

correction. Such correction can be directly applied to *P*-values from of the EDS detection methods we described above. For example, table 6 shows the rates of EDS detection for all 6 methods considered here using raw *P*-values ($P \leq 0.05$) and also false-discovery rate corrected *q*-values (Benjamini-Hochberg FDR, $q \leq 0.2$). For models without MH (BUSTED, +S), the effect of FDR correction is effectively nil, at least for these thresholds, whereas for models with MH (+MH, +MH+S), performing FDR dramatically reduces already low detection rates by about 3-fold. Encouragingly, the model-averaged approach is much less affected by the FDR correction, with the detection rate dropping from 19.3% to 15.5%.

Heuristic Measures of the Extent of EDS Support

Regardless of the model or levels of model agreement, a median of only ~3 sites per alignment contributed the majority of statistical signal for EDS, and ~20–30 individual (branch, site) pairs showed empirical Bayes evidence of EDS support (supplementary table S3, Supplementary Material online). Mean estimates for mixture model proportions for the category $\omega > 1$ directly estimated by ML are in the 0.5–2% range (results not shown). This observation brings into focus that EDS detection is frequently based on substitution patterns in a small subset of sites and branches, and could be influenced by alignment errors, genome assembly, or other abiological processes, especially in high throughput screens of automatically generated alignments. For ~5–10% of EDS detections, most of the statistical support comes from a single site. Recalling the HIV-1 *vif* example, where masking a single codon can completely ablate statistical support for EDS, more careful inspection of alignments for individual substitutions whose presence/absence determines EDS detection is advisable. One straightforward, but computationally expensive option is to perform alignment-level nonparametric bootstrap or jack-knife resampling to investigate the stability of ω_3, p_3 and other rate estimates. Another is to simply look for outlier datasets, like those mimicking HIV-1 *vif*; we identified 172 datasets where +S detected EDS with a single site as the primary source of support, but +S+MH did not, and neither did model-averaging.

Additional Empirical-driven Simulations

We randomly selected two sets of 500 alignments from the Enard et al. (2016) collection (Null and Power), and used model parameter estimates obtained under the +S+MH model to generate 5 synthetic replicates per alignment, yielding a total of 5,000 simulated datasets mimicking real data in terms of base compositions, tree shapes and branch lengths, and rate estimates. For the Null collection, we set $\omega_3 = 1$, thereby enforcing neutral evolution, while for the Power collection, we retained the maximum likelihood $\omega_3 \geq 1$ estimate. We then fitted the four models and performed model averaging, focusing on false-positive rates (Null), and power to detect EDS (Power). The qualitative behavior of the models mirrored what we saw with 4-taxon simulations.

1) Unmodeled multiple hit rates yielded higher than expected false-positive rates for BUSTED and +S, but not +MH, +S+MH, or model-averaged approaches, with higher values of δ and/or ψ corresponding to increased rates of false positives, exceeding 50% for some parameter combinations (fig. 5).
2) Power to detect selection increases with the effect size (e.g., as measured by the magnitude of ω_3, fig. 5). BUSTED and +S models have higher apparent power for lower ω, but this is confounded with δ and ψ rates (fig. 5). This is especially apparent for datasets with weaker selection ($\omega_3 < 4$), where the power of BUSTED and +S models grows significantly as δ is increased. While +S+MH exhibits reduced power compared to +S, using model-averaging rescues most of this power. For example, for $\omega \geq 4$, +S has 52% power, +S+MH—25% power, and model-averaging—42% power.

Table 6. EDS Detection Rates With and Without Multiple-testing Correction.

Method	$P \leq 0.05$	$q \leq 0.20$
BUSTED	**0.284**	**0.2985**
+S	**0.264**	**0.272**
+MH	**0.0838**	**0.0271**
+S+MH	**0.0998**	**0.0348**
Best model	**0.246**	**0.244**
Model averaged	**0.193**	**0.155**

Discussion

Evolutionary substitution models that are practically useful and computable must make many simplifying assumptions about 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 misspecifications. 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 *vif* example in this paper and for human lineage selection detection in Venkat et al. (2018).

Statistical tests which compare the ω ratio of nonsynonymous and synonymous substitution rates to 1 and interpret significant differences as evidence of nonneutral 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 Type 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.

Using simulations and empirical data, our study corroborates the conclusions of multiple previous reports (e.g., Whelan and Goldman 2004; Venkat et al. 2018; Wisotsky et al. 2020) on the impact that MH has on evolutionary process inference and characterization. Not accounting for instantaneous multinucleotide substitutions or "hits" (MH) when estimating evolutionary process parameters, especially related to natural selection, leads to serious statistical misbehavior of widely used tests, and biases rate estimates. 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. These findings add to the already significant body of literature in this space. This is the first study, however, to demonstrate how to model MH and synonymous site-to-site rate variation (SRV) jointly, and that MH remains an important factor even if SRV is accounted for.

In a large-scale empirical analysis of mammalian genes (Enard et al. 2016) we found 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 (SRV), which impacts the vast majority of genes. However, we argue that even second-order effects are 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.

There are other forms of potential model violation that we did not address here, which could also bias tests of EDS. For example, MNMs are known to be enriched for transversions, which tend to be nonsynonymous, and failure to incorporate this bias has been shown previously to inflate false positive rates of EDS tests Venkat et al. (2018). Other potentially important forms of complexity are heterogeneity in preference among amino acids and heterogeneity among pairs, for example, De Maio et al. (2013), Dunn et al. (2019), as well as among-site differences in these factors (e.g., Bloom 2014). These approaches broadly fall into the category of models where the ω ratio depends of the aminoacids being exchanged (Kosiol et al. 2007; Delport, Scheffler, Botha, et al. 2010; De Maio et al. 2013). The key issue for implementing such models is selecting a biologically or empirically justified parameterization, which remains an open problem because it is relatively trivial to improve the goodness-of-fit over the baseline model (Delport, Scheffler, Gravenor, et al. 2010), and such improvement should not be viewed as *prima facie* evidence of biological significance. We also do not directly model multiple hits which span codon boundaries (Whelan and Goldman 2004), thereby breaking up the assumption of codon-level independence of sites and increasing model and computational complexity dramatically. Further research is necessary to incorporate these and other kinds of complexity into tractable Markov models and/or evaluate their effects on EDS analyses.

Another important consideration that we do not address here is data and alignment quality, both of which are known to affect the results of comparative selection analyses (e.g., Spielman et al. 2014). Sensitive methods, such as BUSTED, can be influenced by data from just a few codons, as illustrated by the HIV-1 *vif* example. Post hoc analysis exploration tools we develop here (heatmaps of EBFs, site-level likelihood ratio contributions) give users some tools with which to investigate the robustness of inference and the influence of local data features on their results. Ultimately, however, the issues of alignment generation and data quality assurance are challenging and recalcitrant issues, outside the scope of our study.

A plethora of studies (Whelan and Goldman 2004; Kosiol et al. 2007; De Maio et al. 2013; Venkat et al. 2018; Cohen et al. 2021; Hensley et al. 2021; Lucaci et al. 2021; MacLean et al. 2021; Freitas and Nery 2022; Steward et al. 2022) suggest that MH occurs broadly over diverse taxonomic groups. We expect that future research with an 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 (Huang 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 EDS, 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

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	$\psi \alpha^s \omega^{bs} \prod_{n=1}^{3} \theta_{ij}^n \pi_j^n$

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:

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) (Tavaré and Miura 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: $0 \leq \omega_1 \leq \omega_2 \leq 1 \leq \omega_3$, $p_1 = P(\omega^{bs} = \omega_1)$, 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 $E[\alpha_s] = 1$.

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 nonsynonymous three substitutions. All parameters, except π, including branch lengths, are fitted using a directly optimized phylogenetic likelihood in HyPhy.

Parameter values of δ, ψ are not directly biologically interpretable, because the corresponding q_{ij} rate matrix terms are additionally multiplied by nucleotide rate parameters and frequency terms. However, relative contributions of two- and three-hit components to the expected number of substitutions per site per unit time can be measured using the relative branch length quantity, discussed in De Maio et al. (2013). Specifically, the standard branch length for Markov models of evolution is defined as $B(\cdot) = \sum_{i \neq j} q_{ij} \times \pi_j$, where $(\cdot)$ indicates dependence on estimable model parameters, i, j enumerate sense codons, and π_j are target codon equilibrium frequencies. We define $B_{2H}(\cdot) = \sum_{i \neq j, \delta(i,j)=2} q_{ij} \times \pi_j$, $B_{3H}(\cdot) = \sum_{i \neq j, \delta(i,j)=3} q_{ij} \times \pi_j$, where sums are taken over only those codons that differ at 2 (or 3) nucleotide positions, and use those to estimate the fractions of total branch lengths attributable to 2H as B_{2H}/B and to 3H as B_{3H}/B. B, B_{2H}, and B_{3H} are polynomial expressions in terms of estimated model parameters (π, θ, ω, etc). An example of how 2H and 3H substitution fractions behave for some fixed values of other model parameters for 4-taxon simulation trees are shown in supplementary fig. S3, Supplementary Material online. These quantities are biologically meaningful and comparable with literature estimates for similar MNM-aware studies, for example, Schrider et al. (2011), De Maio et al. (2013).

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.

A notable limiting assumption of the model, made for computational tractability, is that MH substitutions spanning codon boundaries, for example, a substitution in the third position in a codon coupled with a substitution in the first position of the subsequent codon, are not

accounted for. Doing so in a systematic fashion breaks the assumption of site independence, making computation largely intractable. One approximate solution is a mean-field type approximation (Whelan and Goldman 2004), but even this approach is computationally quite expensive. Not accounting for "codon-spanning" multiple hits is a conservative approach, since it will *miss* a subset of MH substitutions, treating them instead as several single-nucleotide substitutions occurring independently.

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^2_0 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 datasets can be downloaded from https://data.hyphy.org/web/busteds-mh/. Additional information is present in the README.md file for this repository, including details of how to generate alignments under the +S and +S+MH models.

All 4-taxon alignments were generated were 800 codons long, used equal (0.25) positional nucleotide frequencies to parameterize the equilibrium codon distribution, and nucleotide bias parameters corresponding to the HKY85 model with the transition/transversion ratio of $\kappa = 2$.

Implementation

All analyses were performed in HyPhy version 2.5.42 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, $P(\omega_3^{bs} > 1|D_s) = P(D_s|\omega_3^{bs} > 1) \times P(\omega_3^{bs} > 1)/P(D_s)$, 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 $\frac{P(\omega_3^{bs}>1|D_s)/(1-P(\omega_3^{bs}>1|D_s))}{P(\omega_3^{bs}>1)/(1-P(\omega_3^{bs}>1))}$. As discussed in Murrell et al. (2012), these empirical estimates are quite noisy and should only be used for exploratory purposes, for example, to look for "hot-spots" in a tree (cf. fig. 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 (table 7).

BUSTED ModelTesting

We recommend our BUSTED model testing and averaging procedure (see main text) in order to select

Table 7. Model Run Times for the Enard Dataset.

Model	Median (s)	Mean (s)	Relative
BUSTED	**50**	**60**	**–**
+S	**174**	**219.1**	**3.3**
+MH	**279**	**379.1**	**6.1**
+S+MH	**1,405**	**1,788.5**	**28.6**

NOTE.—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.

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 EDS 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 EDS. 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).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank members of the HyPhy and Datamonkey teams for their contribution to this work, as well as two anonymous reviewers whose meticulous reading of the manuscript and detailed critiques allowed us to significantly improve the manuscript. This research was supported in part by grants GM144468 (NIH/NIGMS), AI140970 (NIH/NIAID), AI134384 (NIH/NIAID), and NIH NIGMS R35GM142677 (Enard).

Data Availability

All data is available at https://data.hyphy.org/web/busteds-mh/. We also cite the original sources of empirical datasets.

References

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

Anisimova M, Yang Z. 2004. Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? *J Mol Evol*. **59**(6):815–826.

Arana ME, Seki M, Wood RD, Rogozin IB, Kunkel TA. 2008. Low-fidelity DNA synthesis by human DNA polymerase theta. *Nucleic Acids Res*. **36**(11):3847–3856.

Assaf ZJ, Tilk S, Park J, Siegal ML, Petrov DA. 2017. Deep sequencing of natural and experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new mutations. *Genome Res*. **27**(12):1988–2000.

Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, et al. 2016. Multi-nucleotide de novo mutations in humans. *PLoS Genet*. **12**(11):e1006315.

Bloom JD. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol*. **31**(8):1956–1978.

Box GEP. 1976. Science and statistics. *J Am Stat Assoc*. **71**(356):791–799.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*. **16**(11):1457–1465.

Chen J-M, Cooper DN, Férec C. 2014. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: ∼0.4% of the single-nucleotide substitution mutation rate. *Hum Mutat*. **35**(3):392–394.

Chen J-M, Férec C, Cooper DN. 2015. Complex multiple-nucleotide substitution mutations causing human inherited disease reveal novel insights into the action of translesion synthesis dna polymerases. *Hum Mutat*. **36**(11):1034–1038.

Cohen ZP, Brevik K, Chen YH, Hawthorne DJ, Weibel BD, Schoville SD. 2021. Elevated rates of positive selection drive the evolution of pestiferousness in the Colorado potato beetle (*Leptinotarsa decemlineata*, Say). *Mol Ecol*. **30**(1):237–254. doi:10.1111/mec.15703

Davydov II, Salamin N, Robinson-Rechavi M. 2019. Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. *Mol Biol Evol*. **36**(6):1316–1332.

Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, Kosakovsky Pond SL. 2010. Codontest: modeling amino acid substitution preferences in coding sequences. *PLoS Comput Biol*. **6**(8):e1000885.

Delport W, Scheffler K, Gravenor MB, Muse SV, Kosakovsky Pond S. 2010. Benchmarking multi-rate codon models. *PLoS ONE* **5**(7):e11587.

De Maio N, Holmes I, Schlötterer C, Kosiol C. 2013. Estimating empirical codon hidden Markov models. *Mol Biol Evol*. **30**(3):725–736.

Dunn KA, Kenney T, Gu H, Bielawski JP. 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, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**:e12469.

Freitas L, Nery MF. 2022. Positive selection in multiple salivary gland proteins of Anophelinae reveals potential targets for vector control. *Infect Genet Evol*. **100**:105271.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. **11**(5):725–736.

Harris K, Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res*. **24**(9):1445–1454.

Hensley NM, Ellis EA, Leung NY, Coupart J, Mikhailovsky A, Taketa DA, Tessler M, Gruber DF, De Tomaso AW, Mitani Y, et al. 2021. Selection, drift, and constraint in cypridinid luciferases and the diversification of bioluminescent signals in sea fireflies. *Mol Ecol*. **30**(8):1864–1879. doi:10.1111/mec.15673

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. **12**(11):756–766.

Huang CJ. 2014. A biocompatible open-surface droplet manipulation platform for detection of multi-nucleotide polymorphism. *Lab on a Chip* **14**(12):2057–2062.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**(1):153–159.

Jones CT, Youssef N, Susko E, Bielawski JP. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol Biol Evol*. **35**(6):1473–1488.
Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*. **28**(11): 3033–3043.
Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. 2020. Hyphy 2.5-A customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol*. **37**(1):295–299.
Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol*. **24**(7):1464–1479.
Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**(21): 4453–4455.
Kuno G, Chang GJ, Tsuchiya KR, Karabatsos N, Cropp CB. 1998. Phylogeny of the genus *Flavivirus*. *J Virol*. **72**(1):73–83.
Loeb LA, Monnat RJ Jr. 2008. DNA polymerases and human disease. *Nat Rev Genet*. **9**(8):594–604.
Lucaci AG, Wisotsky SR, Shank SD, Weaver S, Kosakovsky Pond SL. 2021. Extra base hits: widespread empirical support for instantaneous multiple-nucleotide changes. *PLoS ONE* **16**(3):e0248337.
MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, Kosakovsky Pond SL, Robertson DL. 2021. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol*. **19**(3): e3001115.
Martin DP, Lytras S, Lucaci AG, Maier W, Grüning B, Shank SD, Weaver S, MacLean OA, Orton RJ, Lemey P, et al. 2022. Selection analysis identifies clusters of unusual mutational changes in omicron lineage BA.1 that likely impact Spike function. *Mol Biol Evol*. **39**(4):msac061.
Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, Lucaci AG, Giandhari J, Naidoo S, Pillay Y, et al. 2021. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**(20):5189–5200.e7.
Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. 2000. Low fidelity DNA synthesis by human DNA polymerase-eta. *Nature* **404**(6781):1011–1013.
Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. 2021. Sustainable data analysis with Snakemake. *F1000Res*. **10**:33.
Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*. **32**(5): 1365–1371.
Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. **8**(7):e1002764.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. **11**(5):715–724.
Pond SK, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS ONE* **5**(7):e11230.
Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. **22**(12):2375–2385.
Posada D, Buckley TR. 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.
Prendergast JGD, Pugh C, Harris SE, Hume DA, Deary IJ, Beveridge A. 2019. Linked mutations at adjacent nucleotides have shaped human population differentiation and protein evolution. *Genome Biol Evol*. **11**(3):759–775.
Rodrigue N, Latrille T, 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.
Saribasak H, Maul RW, Cao Z, Yang WW, Schenten D, Kracker S, Gearhart PJ. 2012. DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J Exp Med*. **209**(6):1075–1081.
Schrider DR, Hourmozdi JN, Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol*. **21**(12):1051–1054.
Seo T-K, Kishino H, Thorne JL. 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):1201–1213.
Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, Duffet R, Zvelebil M, Martinson N, McIntyre J, Morris L, et al. 2007. A model of directional selection applied to the evolution of drug resistance in HIV-1. *Mol Biol Evol*. **24**(4):1025–1031.
Seplyarskiy VB, Bazykin GA, Soldatov RA. 2015. Polymerase ζ activity is linked to replication timing in humans: evidence from mutational signatures. *Mol Biol Evol*. **32**(12):3158–3172.
Shu Y, McCauley J. 2017. Gisaid: global initiative on sharing all influenza data—from vision to reality. *Euro Surveill*. **22**(13):30494.
Spielman SJ, Dawson ET, Wilke CO. 2014. Limited utility of residue masking for positive-selection inference. *Mol Biol Evol*. **31**(9): 2496–2500.
Spielman SJ, Kosakovsky Pond SL. 2018. Relative evolutionary rates in proteins are largely insensitive to the substitution model. *Mol Biol Evol*. **35**(9):2307–2317.
Steward RA, de Jong MA, Oostra V, Wheat CW. 2022. Alternative splicing in seasonal plasticity and the potential for adaptation to environmental change. *Nat Commun*. **13**(1):755.
Stone JE, Lujan SA, Kunkel TA, Kunkel TA. 2012. Dna polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in *Saccharomyces cerevisiae*. *Environ Mol Mutagen*. **53**(9):777–786.
Su C, Nguyen VK, Nei M. 2002. Adaptive evolution of variable region genes encoding an unusual type of immunoglobulin in camelids. *Mol Biol Evol*. **19**(3):205–215.
Tamuri AU, Dos Reis M. 2022. A mutation-selection model of protein evolution under persistent positive selection. *Mol Biol Evol*. **39**(1):msab309.
Tavaré S, Miura RM. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Lectures on mathematics in the life sciences*. Vol. 17. Providence (RI): American Mathematical Society. p. 57–86.
Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol*. **2**(8):1280–1288.
Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, Anyaneji UJ, Bester PA, Boni MF, Chand M, et al. 2022. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**(7902):679–686.
Wagenmakers E-J, 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, Gao C. 2020. Precise, predictable multi-nucleotide deletions in rice and wheat using APOBEC–Cas9. *Nat Biotechnol*. **38**(12):1460–1465.
Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**(4):2027–2043.
Wisotsky SR, Kosakovsky Pond SL, Shank SD, Muse SV. 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):568–573.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol*. **51**(5):423–432.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**(1):431–449.

Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol*. **17**(10):1446–1455.

Yokoyama S, Tada T, Zhang H, 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):13480–13485.