

# Experimentally informed site-specific substitution models deepen phylogenetic estimates of the divergence of viral lineages

Sarah K. Hilton<sup>1,2</sup> and Jesse D. Bloom<sup>1,2</sup>

<sup>1</sup>Division of Basic Sciences and Computational Biology Program,  
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

<sup>2</sup>Department of Genome Sciences, University of Washington, Seattle, WA  
E-mail: jdbloom@fredhutch.org.

## Abstract

[[≤ 250 words, currently 217](#)] Molecular phylogenetics is often used to estimate the time since the divergence of modern gene sequences. Such phylogenetic techniques often estimate substantially shallower divergence times than other methods. For instance, in the case of viruses there is independent evidence that molecular phylogenetics can underestimate deep divergence times. This discrepancy is thought to be caused in part by inadequate models of purifying selection leading to branch-length underestimation. Here we show that substitution models informed by experimental measurements of site-specific purifying selection lengthen the branches on phylogenies of influenza virus hemagglutinin. This deepening of branch lengths is due to better modeling how site-specific amino-acid preferences affect the stationary state of the substitution models, and is independent of the branch-length-extension due to modeling site-to-site variation in substitution rate. We show that the deepening of branch lengths from experimentally informed site-specific substitution models is similar to that achieved by other approaches that allow the stationary state to vary across sites. However, the improvements from these site-specific models are limited by the inherent tension between the enhanced accuracy of accounting for site-specific amino-acid preferences and the fact that these preferences shift over long evolutionary times. Overall, our work underscores the importance of modeling how site-specific purifying selection affects the stationary state as well as the substitution rate when estimating deep divergence times.

## Introduction

[from JDB: what is the "age" of a virus? Maybe "divergence time of viral lineages"] skhcommentfrom JDB: what is the less than a million actually? "Old" is not the right phrase. Estimating the divergence time of viral lineages of a virus is essential to understanding its evolutionary history, including its emergence, spread, and past zoonoses. This estimation is commonly done using the concept a "molecular clock" to transform the branch lengths of the viral phylogenetic tree into age in years. However, this molecular dating technique often underestimates the age of many viruses, including measles, foamy virus, and ebola [(citations)], compared to other methods which are independent of the viral phylogeny. For example, SIV (the original source of HIV) is estimated to be less than a million years old based on the viral phylogeny (Sharp et al., 2000; Wertheim and Worobey, 2009; Worobey et al., 2010) but estimated to be several million years old based on the host tree or endogenous retroviral elements (Compton et al., 2013) [(other citations)]. Overall, there is a systematic and substantially large underestimation of of branch length on viral phylogenies. [long branches]

Branch length underestimation is due, in part, to strong purifying selection masking the evolutionary signal in the observed sequences. Purifying selection can lead to mutational saturation, where multiple unobserved, substitutions occur at a single site along a long branch and erase the divergence signal (Holmes, 2003). Furthermore, proteins do not have equal preference for all amino acids at all sites, this evident by a simple visual inspection of a multiple sequence alignment. How many and which amino acids tolerated at each site of the protein generate a site-specific expected rate of change. Failing to account for these site-specific constraints will lead to branch length underestimation. [you will have mutational saturation no matter what - this is a separate, addressable issue?] [talk about the high mutation rate in viruses?]

Substitution models that incorporate site-to-site rate variation have been developed to decrease the bias in long branch estimation. The most common strategy is to allow a single rate-controlling parameter to vary according to some statistical distribution, such as a  $\Gamma$ -distributed  $\omega$  (dN/dS) (Yang et al., 2000). This flexibility in the value of  $\omega$  accounts for the site-to-site rate variation by allow some sites to have a higher dN/dS value than others. While this modification is simple and only requires the addition of one extra parameter, it does not describe site-specificity in its stationary state. That is, at evolutionary equilibrium, this model still assumes that each site in the protein evolves identically.

An alternative approach is to model the site-specific amino-acid frequencies explicitly, such as those models in the mutation-selection family (Halpern and Bruno, 1998). In these models, each amino-acid at each site in the protein is described by its own parameter and these differences are reflected in the stationary state of the model. The rate of change at a given site is controlled by these amino acid profiles and can now vary from site to site, as expected based on observations in nature. Importantly, these rate variations are not constrained to an arbitrary statistical distribution but by parameters with a direct

biological interpretation.

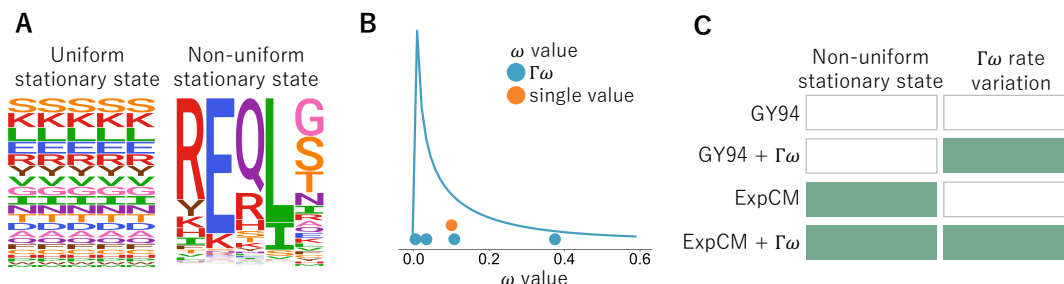
Mutation-selection models are presumably more biologically relevant but pose more practical challenges than the  $\Gamma\omega$  models. These models are highly parametrized with 19 free parameters (the 20 amino acid preferences are constrained to sum to one) per site leading to thousands of parameters for the length of a normal protein. One way to avoid overfitting is to implement the model as a mixture model in either a bayesian (Lartillot and Philippe, 2004) or maximum likelihood framework (Si Quang et al., 2008).

Alternatively, you can reduce the parameter space by defining the amino-acid frequencies *a priori*. We have shown previously that we can define an Experimentally Informed Codon Model (ExpCM) (Bloom, 2014a,b) from the mutation-selection family using measurements from deep mutational scanning (Fowler and Fields, 2014), a high-throughput functional assay. ExpCM are therefore defined by amino-acid preferences measured in a *single* genetic background and do not reflect any epistatic changes which may have occurred over the virus's evolutionary history. But they contain no more parameters than the traditional codon models while maintaining a site-specific stationary state. We hypothesize that the ExpCM will estimate longer branches than the traditional models due to the protein-specific description of purifying selection. [CAT model has been shown to work well (better) on saturated data.]

In order to test this hypothesis, we compared the branch lengths of a influenza virus HA phylogenetic trees optimized by different substitution models. We found that the ExpCM did extend the length of branches from the focal sequence on the tree [define focal] and that this extension was seen even in the context of  $\Gamma$ -distributed rate variation. Furthermore, we found this extension occurred even in the presence of  $\Gamma$ -distributed  $\omega$ , indicating that they are both important for modeling purifying selection. This supports the conclusion that modeling purifying selection, especially in a model with a non-uniform stationary state, is important to estimating the branch lengths on phylogenetic trees.

# Results and Discussion

## Substitution models



**Figure 1: Comparison of substitution model features.** (A) Substitution model stationary states can either be identical at every site in the protein or allow to vary from site to site. (B) The relative rate of non-synonymous change,  $\omega$ , can be defined as one, gene-wide average or allowed to vary following some statistical distribution. To ensure computational tractability,  $\omega$  takes on the mean value of  $K$  bins ( $K = 4$  shown). (C) Substitution models can both, one, or neither of these features. We used models from the GY94 and the ExpCM families.

It is well known that protein sites evolve under different pressures. Some sites appear to tolerate many amino acids while other sites appear to tolerate only a select few. Therefore, it is important to use substitution models which are able to describe this purifying selection, and the resulting site-to-site rate variation, when modeling protein evolution.

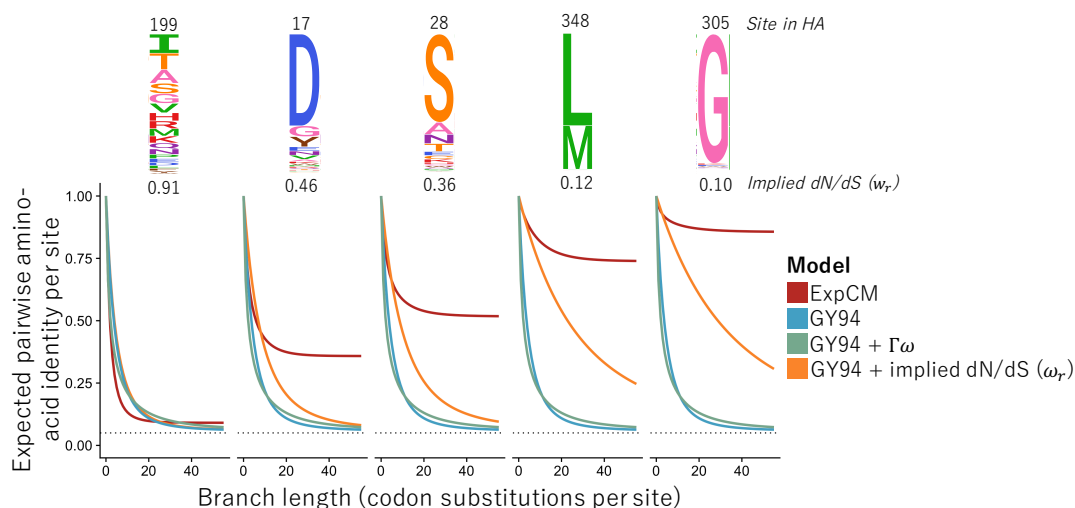
The standard strategy to account for purifying selection is to allow the rate of non-synonymous change to vary among sites following some statistical distribution. This variation is commonly implemented by allowing the rate parameter,  $\omega$ , to follow a  $\Gamma$ -distribution ( $\Gamma\omega$ ). In order to achieve computational tractability, the distribution is discretized into  $K$  bins and  $\omega$  is assigned to the mean of each bin and the site likelihood is the average likelihood given each of the  $\omega$  values. Under this model, some sites would have a higher rate of non-synonymous change than others but this rate would be agnostic to the amino-acid identities themselves. All non-synonymous changes would be treated as equally.

Another strategy to account for purifying selection is to define a substitution model with a site-specific stationary state. Such models describe the evolution at each site uniquely and independently of the other sites. For example, models from the Halpern-Bruno mutation-selection family explicitly describe the frequency of each amino acid at each site. However, these more mechanistic models, without an arbitrary statistical distribution, come at a cost in the form of an increased number of parameters. When using a mutation-selection model is crucial to invoke a strategy to avoiding overfitting the data. We chose to define the amino-acid frequency parameters *a priori* using measurements from a high-throughput functional assay called deep mutational scanning, Experimentally Informed Codon Models

(ExpCMs).

These two strategies,  $\Gamma\omega$  rate variation and non-uniform stationary states, are not truly distinct. Differences in stationary state manifest as differences in non-synonymous rates across sites (Spielman and Wilke, 2015). Moreover, these two strategies are not mutually exclusive. Technically, substitution models can have a non-uniform stationary state and  $\Gamma\omega$  rate variation and biologically, these models can account for both strong purifying and strong positive selection.

## Effect of stationary state and rate variation on branch length estimation



**Figure 2: Effect of stationary state and rate variation on long branch estimation.** The expected pairwise amino-acid identity for five sites in influenza hemagglutinin (HA) for four different substitution models. The logoplots show the HA amino-acid preferences from a deep mutational scan performed by Doud and Bloom (2016). The implied dN/dS value was calculated from the ExpCM following Spielman and Wilke (2015).

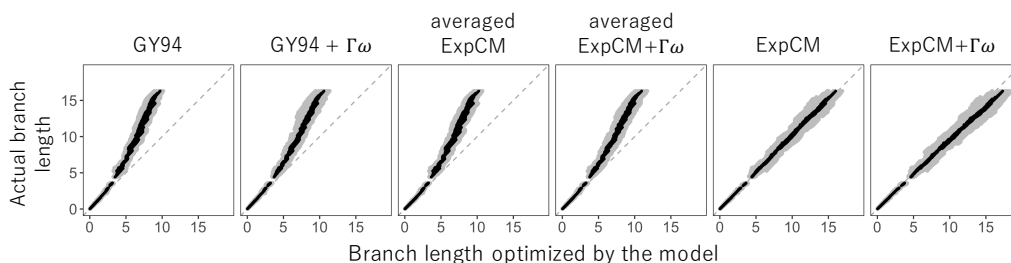
Along a single branch, a substitution model transforms sequence divergence to branch length, which under a molecular clock assumption is proportional to time. However, every substitution model has a stationary state, the expected sequence divergence given a very long time, at which the relationship between sequence divergence and time is lost. The stationary state is represented by the long “tails” in Figure 2, where the expected sequence divergence remains constant as time increases.

A substitution model’s stationary state is defined by the parameters in the model and can vary from model to model. Some model features affect how long it takes a model to reach its stationary state, whereas other features affect the stationary state itself. Adding  $\Gamma\omega$  rate variation affects the “path” to the stationary state but does not affect the stationary state itself. In addition, while  $\Gamma\omega$  accounts for site-to-site rate variation, the resulting stationary state is identical at each site in the protein. Under these

models with uniform stationary states, a mutationally constrained site and a site which tolerates many mutations will have the sequence divergence given a very long time. Models with non-uniform stationary states, by definition, acknowledge that differences in sites will persist even at a very long period of time. The difference in branch length estimation will be most dramatic at constrained sites in the protein. At a constrained site, the ExpCM will estimate a longer branch than the GY94 model because the overall sequence divergence at that site is expected to be small.

This difference in branch length estimation cannot be recapitulated by more complex modeling of the  $\omega$  parameter. We transformed the GY94 model into a site-specific model by inferring a unique  $\omega$  value from the ExpCM, following the work of [Spielman and Wilke \(2015\)](#), for each site in the protein. This model takes longer to reach stationary state than the GY94 model ultimately has the same divergence level at stationary state. As this model shows, no matter how “well” a model accounts for site-to-site rate variation, it will underestimate long branches with a uniform stationary state model.

### Failure to account for site-specificity leads to branch length underestimation.



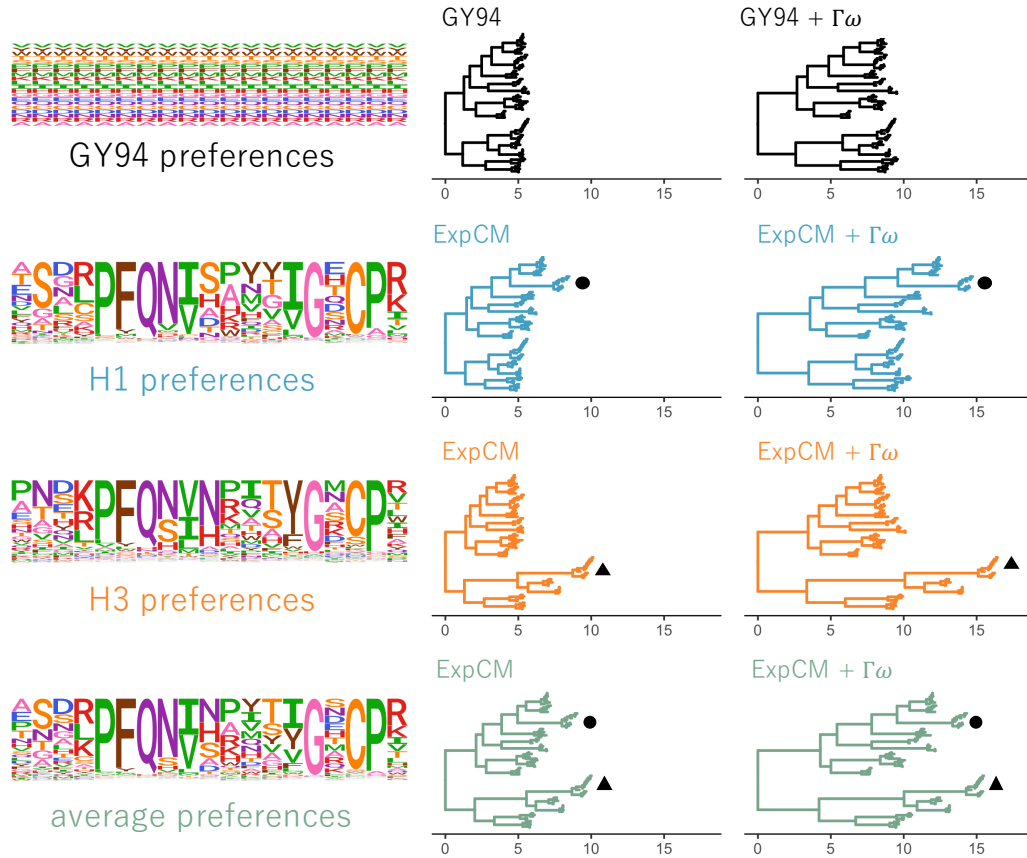
**Figure 3: Model performance under simulated, site-specific data.** [Fonts are small, and tick labels definitely too small. I would show averaged.] Alignments were simulated under an ExpCM (??) along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The randomized ExpCMs have amino-acid profiles shuffled among the sites. These randomized models are still site-specific but the relationship between the site and the experimental data is broken. Grey points represent the length of one branch and the black points are the mean branch lengths over eight simulations. The grey, dashed line is the reference line  $y = x$ , depicting the behavior of a model which is an unbiased estimator of the simulated branch length.

[\[Section Outline:\]](#)

- To test the effect of substitution model on data with site-specific amino acid frequencies, we simulated sequences under an ExpCM and re-inferred the branch lengths using different substitution models.
- Uniform stationary state models underestimate long branches

- Both the GY94 and the GY94+ $\Gamma\omega$  underestimate long branches. The  $\Gamma\omega$  rate variation scheme is not enough to prevent this underestimation at every long branches.
- Another uniform stationary state model, the averaged ExpCM, also underestimates long branches. This model derives some information from the experiments but lacks site-specificity.
- Non-uniform stationary state models accurately estimate long branches (both the ExpCM and ExpCM+ $\Gamma\omega$ )
- If sequences have site-specific amino acid frequencies then non-uniform stationary state models, even with  $\Gamma\omega$ , will underestimate long branches.

## empirical Data



**Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to YNGKP models. [Keep branch lengths in substitutions per site. Make a panel A show a prefs snippet, B H1 trees, C H3 trees, D H1+H3. Maybe increase very small font sizes a bit.]** The branch lengths of a base topology inferred using the GTR-CAT model were optimized by (A) an ExpCM defined by H1 preferences, (B) an ExpCM+ $\Gamma\omega$  defined by H1 preferences, (C) YNGKP M0, and (D) YNGKP M5. The branch lengths are normalized to the distance between A/South Carolina/1/1918 and A/Solomon Islands/3/2006 and colored to indicate the distance from the H1 focal sequence (black triangle).

### [Section Outline:]

- We tested the effect substitution model choice on branch length estimation using influenza HA sequences.
  - These sequences presumably evolve with site-specific frequencies but unlike the simulations we do not *know* the model

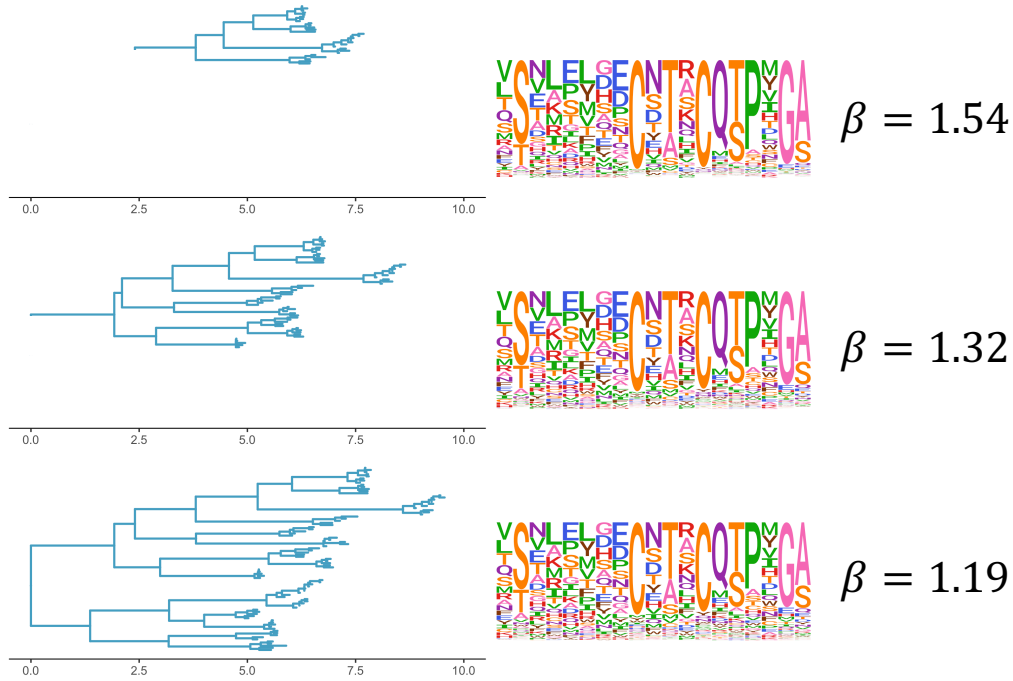


**Table 1:** Model comparison for Fig. [Figure 4](#).

Model	$\Delta AIC$	Log Likelihood	Parameters
ExpCM + $\Gamma\omega$ (average)	0.00	-48751.16	$\omega = (0.19, 0.50, 0.90, 1.86), \beta = 1.70, \kappa = 3.73$
ExpCM (average)	949.70	-49227.01	$\beta = 1.78, \kappa = 3.40, \omega = 0.15$
ExpCM + $\Gamma\omega$ (H1)	1306.44	-49404.38	$\omega = (0.13, 0.44, 0.91, 2.16), \beta = 1.12, \kappa = 3.64$
ExpCM + $\Gamma\omega$ (H3)	1737.40	-49619.86	$\omega = (0.09, 0.33, 0.72, 1.77), \beta = 1.28, \kappa = 3.71$
ExpCM (H1)	2555.58	-50029.95	$\beta = 1.22, \kappa = 3.24, \omega = 0.13$
ExpCM (H3)	3196.50	-50350.41	$\beta = 1.45, \kappa = 3.38, \omega = 0.12$
GY94 + $\Gamma\omega$	4719.34	-51105.83	$\omega = (0.00, 0.03, 0.08, 0.26), \kappa = 3.18$
GY94	7624.64	-52559.48	$\kappa = 2.89, \omega = 0.07$

- These sequences are from fairly diverged proteins. Some sequences share only 40% identity on the amino acid level
- The HA tree has closely related subtypes with long branches between subtypes.
- Both  $\Gamma\omega$  and non-uniform stationary states result in an extension of branch lengths.
  - The addition of  $\Gamma\omega$  extends branches independent of preferences.
  - The addition of preferences extends branches *from the focal sequence* independent of  $\Gamma\omega$ .
- The extension of branch length from the focal sequence shows that the ExpCM does not have equal "relevance" across the whole tree.
  - This result is not entirely surprising. DMS measure the effects of single amino-acid changes in the context of a single genetic background.
  - It is expected that there would be epistatic changes across such a highly diverged phylogenetic tree. "Shifting preferences".

## Competing effects of shifting preferences and long branches.



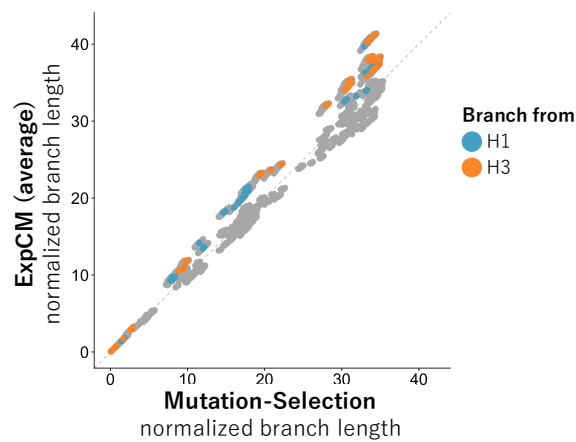
**Figure 5: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A)** An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of the "high" alignment. The colors denote the alignment and the black circle denotes the focal H3 sequence. **(C)** The value of the ExpCM and ExpCM+ $\Gamma\omega$  stringency parameter  $\beta$  decreases as the divergence from the focal H3 sequence increases. **(D)** Comparisons of branch lengths optimized by the four substitution models for the varying degrees of divergence. Black points represent branches from the focal H3 sequence and grey points represent all other branches. The branch lengths are in average number of codon substitutions per site.

### [Section Outline:]

- We investigated the competing effects of "shifting preferences" and long branches by comparing ExpCMs with different "relevances"
- We use the fit stringency parameter  $\beta$  as a measure of relevance.
  - The stringency parameter rescales the preferences. A stringency parameter with a value greater than one indicates that selection in nature prefers the same amino acids as selection in lab but with greater strength.

- There is an inverse relationship between the stringency parameter and overall sequence divergence. As the overall divergence of the tree increases, the stringency parameter decreases.
- [hypothetical data] There is also a positive relationship between the stringency parameter and branch lengths
  - We optimized the same tree with ExpCMs with different stringency parameters.
  - The ExpCMs with higher stringency values estimate longer branches than the ExpCMs with lower stringency values.

## phylobayes



**Figure 6: Comparison of ExpCM and phylobayes**

## Conclusion

1. We don't allow any of the models to vary by lineage.

## Materials and Methods

### Substitution models

#### GY94 models

#### ExpCMs

We recap the **Experimentally Informed Codon Model** (ExpCM) (Bloom, 2014a,b, 2017; Hilton et al., 2017) to introduce nomenclature.

In an ExpCM, rate of substitution  $P_{r,xy}$  of site  $r$  from codon  $x$  to  $y$  is written in mutation-selection form (Halpern and Bruno, 1998; McCandlish and Stoltzfus, 2014; Spielman and Wilke, 2015) as

$$P_{r,xy} = Q_{xy} \times F_{r,xy} \quad (\text{Equation 1})$$

where  $Q_{xy}$  is proportional to the rate of mutation from  $x$  to  $y$ , and  $F_{r,xy}$  is proportional to the probability that this mutation fixes. The rate of mutation  $Q_{xy}$  is assumed to be uniform across sites, and takes an HKY85-like (Hasegawa et al., 1985) form:

$$Q_{xy} = \begin{cases} \phi_w & \text{if } x \text{ and } y \text{ differ by a transversion to nucleotide } w \\ \kappa \phi_w & \text{if } x \text{ and } y \text{ differ by a transition to nucleotide } w \\ 0 & \text{if } x \text{ and } y \text{ differ by } > 1 \text{ nucleotide.} \end{cases} \quad (\text{Equation 2})$$

The  $\kappa$  parameter represents the transition-transversion ratio, and the  $\phi_w$  values give the expected frequency of nucleotide  $w$  in the absence of selection on amino-acid substitutions, and are constrained by  $1 = \sum_w \phi_w$ .

The deep mutational scanning data are incorporated into the ExpCM via the  $F_{r,xy}$  terms. The experiments measure the preference  $\pi_{r,a}$  of every site  $r$  for every amino-acid  $a$ . The  $F_{r,xy}$  terms are defined in terms of these experimentally measured amino-acid preferences as

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \\ \omega \times \frac{\ln[(\pi_{r,\mathcal{A}(y)}/\pi_{r,\mathcal{A}(x)})^\beta]}{1 - (\pi_{r,\mathcal{A}(x)}/\pi_{r,\mathcal{A}(y)})^\beta} & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y) \end{cases} \quad (\text{Equation 3})$$

where  $\mathcal{A}(x)$  is the amino-acid encoded by codon  $x$ ,  $\beta$  is the stringency parameter, and  $\omega$  is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences. The ExpCM has six free parameters (three  $\phi_w$  values,  $\kappa$ ,  $\beta$ , and  $\omega$ ). The preferences  $\pi_{r,a}$  are *not* free parameters since they are determined by an experiment independent of the sequence alignment being analyzed.

The ExpCM stationary state frequency  $p_{r,x}$  of codon  $x$  at site  $r$  is (Bloom, 2017)

$$p_{r,x} = \frac{(\pi_{r,A(x)})^\beta \phi_{x_0} \phi_{x_1} \phi_{x_2}}{\sum_z (\pi_{r,A(z)})^\beta \phi_{z_0} \phi_{z_1} \phi_{z_2}}, \quad (\text{Equation 4})$$

## Theoretical effect of model choice on branch length

### Effect of model choice on natural sequences

#### ExpCM + $\Gamma\omega$ and YNGKP M5

#### Spielman $\omega_r$ values inferred from the ExpCM

We inferred the average nonsynonymous fixation rate from the ExpCM following Spielman and Wilke (2015) as

$$\omega_r = \frac{\sum_x \sum_{y \in N_x} p_{r,x} \times P_{r,xy}}{\sum_x \sum_{y \in N_x} p_{r,x} \times Q_{xy}} \quad (\text{Equation 5})$$

where  $p_{r,x}$  is the stationary state of the ExpCM at site  $r$  and codon  $x$ ,  $P_{r,xy}$  is the substitution rate from codon  $x$  to codon  $y$  at site  $r$ ,  $Q_{xy}$  is the mutation rate from codon  $x$  to codon  $y$ , and  $N_x$  is the set of codons that are nonsynonymous to codon  $x$  and differ from codon  $x$  by only one nucleotide.

### Expected pairwise amino-acid identity

*Do I need to talk about the branchScale scaling I used?* The expected pairwise amino-acid identity at a site  $r$  over time  $t$  for a given model is

$$\sum_a \sum_{x \in a} p_{r,x} \sum_{y \in a} [M_r(t)]_{xy} \quad (\text{Equation 6})$$

where  $a$  is all amino acids,  $p_{r,x}$  is the stationary state of the model at site  $r$  and codon  $x$ , and  $[M_r(t)]_{xy}$  is the transition rate from codon  $x$  to codon  $y$  at site  $r$  given time  $t$ .



**Table 2:** ExpCM parameters used to simulate sequences in Fig. ??.

Parameter	Value
$\beta$	1.54
$\kappa$	3.60
$\omega$	0.20
$\phi_A, \phi_C, \phi_G$	0.38, 0.17, 0.23

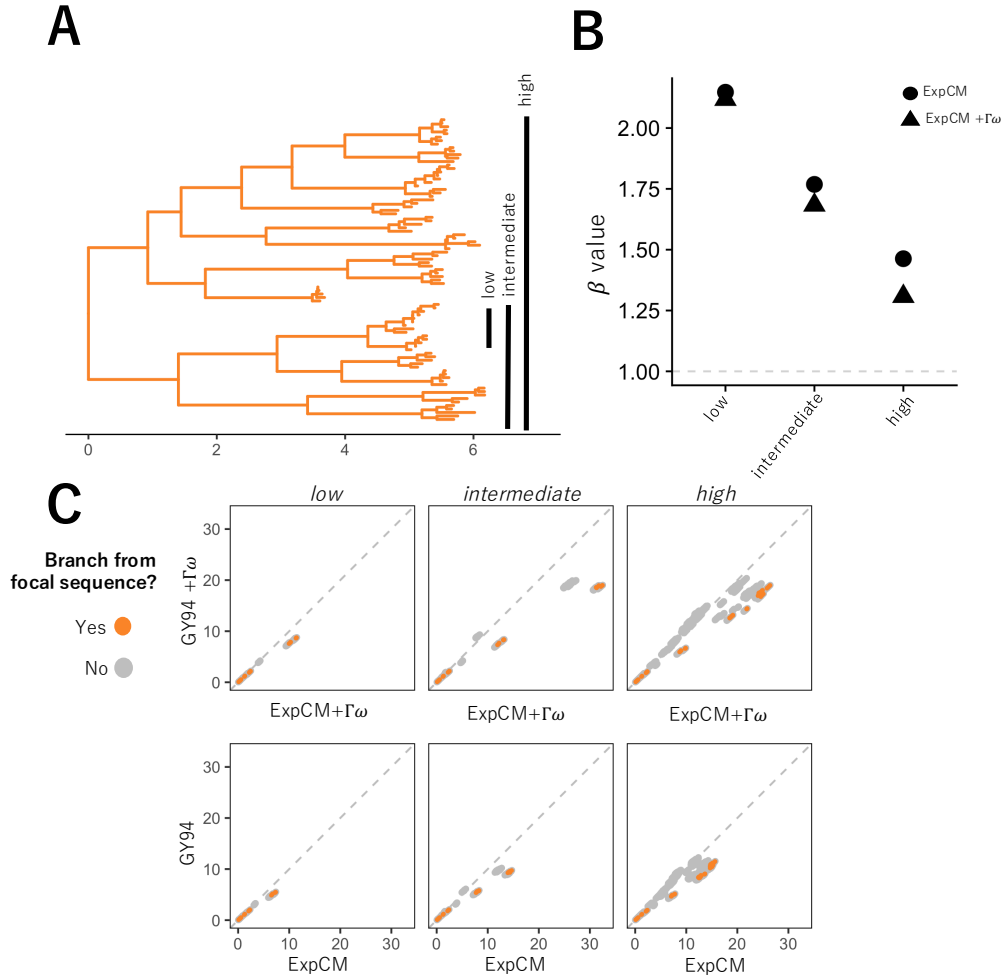
**Table 3:** Model parameters used in Fig. Figure 2.

Model	Parameters
ExpCM	$\beta = 1.54196$ $\kappa = 3.47184$ $\omega = 0.219225$
YNGKP M0	$\kappa = 2.9984$ $\omega = 0.09076$
YNGKP M5	$\kappa = 2.9984$ $\omega = 0.09076$



**Figure 7:** The average of the H1 preferences measured by Doud and Bloom (2016) and the H3 preferences measured by Lee rescaled with the ExpCM stringency parameter optimized in ??A ( $\beta = 1.77$ )

[I need to change the  $\beta$  value when the new phydms results finish running.]



**Supplementary figure 3: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A)** An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as "low", "intermediate", and "high". **(B)** The phylogenetic tree of the "high" alignment. The colors denote the alignment and the black circle denotes the focal H3 sequence. **(C)** The value of the ExpCM and ExpCM+ $\Gamma\omega$  stringency parameter  $\beta$  decreases as the divergence from the focal H3 sequence increases. **(D)** Comparisons of branch lengths optimized by the four substitution models for the varying degrees of divergence. Black points represent branches from the focal H3 sequence and grey points represent all other branches. The branch lengths are in average number of codon substitutions per site.



## References

- Bloom JD. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*. 31:1956–1978.
- Bloom JD. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol. Biol. Evol.* 31:2753–2769.
- Bloom JD. 2017. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*. 12:1.
- Compton AA, Malik HS, Emerman M. 2013. Host gene evolution traces the evolutionary history of ancient primate lentiviruses. *Phil. Trans. R. Soc. B*. 368:20120496.
- Doud MB, Bloom JD. 2016. Accurate measurement of the effects of all amino-acid mutations to influenza hemagglutinin. *Viruses*. 8:155.
- Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nature methods*. 11:801–807.
- Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*. 15:910–917.
- Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*. 22:160–174.
- Hilton SK, Doud MB, Bloom JD. 2017. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*. 5:e3657.
- Holmes EC. 2003. Molecular clocks and the puzzle of rna virus origins. *Journal of virology*. 77:3893–3897.
- Lartillot N, Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*. 21:1095–1109.
- McCandlish DM, Stoltzfus A. 2014. Modeling evolution using the probability of fixation: history and implications. *The Quarterly review of biology*. 89:225–252.
- Sharp P, Bailes E, Gao F, Beer B, Hirsch V, Hahn B. 2000. Origins and evolution of aids viruses: estimating the time-scale.
- Si Quang L, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*. 24:2317–2323.

- Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*. 32:1097–1108.
- Wertheim JO, Worobey M. 2009. Dating the age of the siv lineages that gave rise to hiv-1 and hiv-2. *PLoS computational biology*. 5:e1000377.
- Worobey M, Telfer P, Souquière S, et al. (11 co-authors). 2010. Island biogeography reveals the deep history of siv. *Science*. 329:1487–1487.
- Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.