Experimentally Informed Site-Specific Substitution Models Substantially Deepen Viral Divergence Estimates

Sarah K. Hilton[1,2] and Jesse D. Bloom[1,2,]

[1]Division of Basic Sciences and Computational Biology Program,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA
E-mail: jbloom@fredhutch.org.

Abstract

≤ 250 words, currently 159

Molecular dating techniques have been used to estimate the divergence date of many viruses. However, these estimates are consistently substantially younger than estimates from methods which are independent of branch length estimation. This discrepancy is caused, in part, by inadequate modeling of purifying selection leading to branch length underestimation. Here, we show that substitution models informed by empirical measurements of mutational constraint better than traditional models and extend branch lengths. We used models informed by deep mutational scanning experiments performed in two, highly diverged influenza virus hemagglutinin homologs to optimize the branch lengths of a phylogenetic tree. For each experimentally informed model, we observed extension in branch length from the experiment's focal sequence. This extension in branch length due to explicit modeling of site-specific purifying selection is observed in the presence and absence of standard methods for modeling site-to-site variation. Overall, this study underscores the importance of modeling purifying selection when estimating branch lengths and, by extension, divergence dates.

Introduction

Estimating the age 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 which address 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 Γ-distributed ω (dN/dS) (Yang et al., 2000). This flexibility in the value of ω 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 assumably 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 do contain as a few parameters as 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 Γ-distributed rate variation. Furthermore, we found this extension occurred even in the presence of Γ-distributed ω, 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

Effect of stationary state and rate variation on long branch estimation

What are the points for this section?

1. Different substitution models have different definitions of rate variation and this allows us to compare the effects of different attributes.

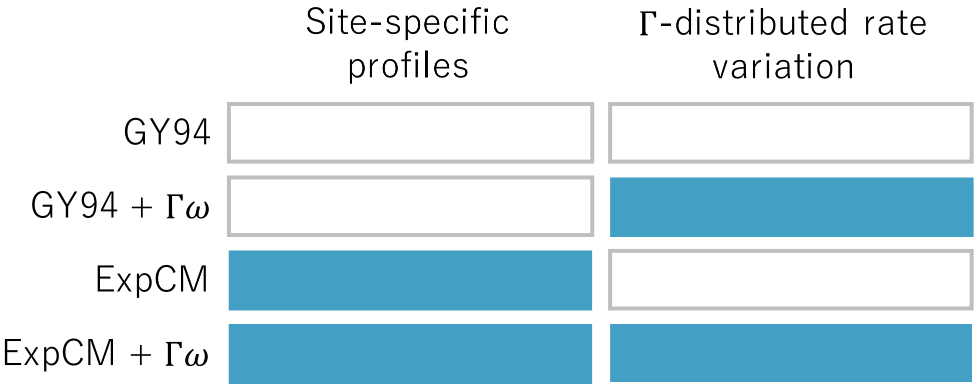
2. GY94 has a uniform stationary state and no rate variation - it estimates the shortest branches and reaches equilibrium right around 0.05.

3. The addition of rate variation for a model with a uniform stationary state does increase the branch length.

4. Using site-specific models allows you to model site-specific constraint. The more constrained the site the bigger the difference between non-uniform and uniform substitution models

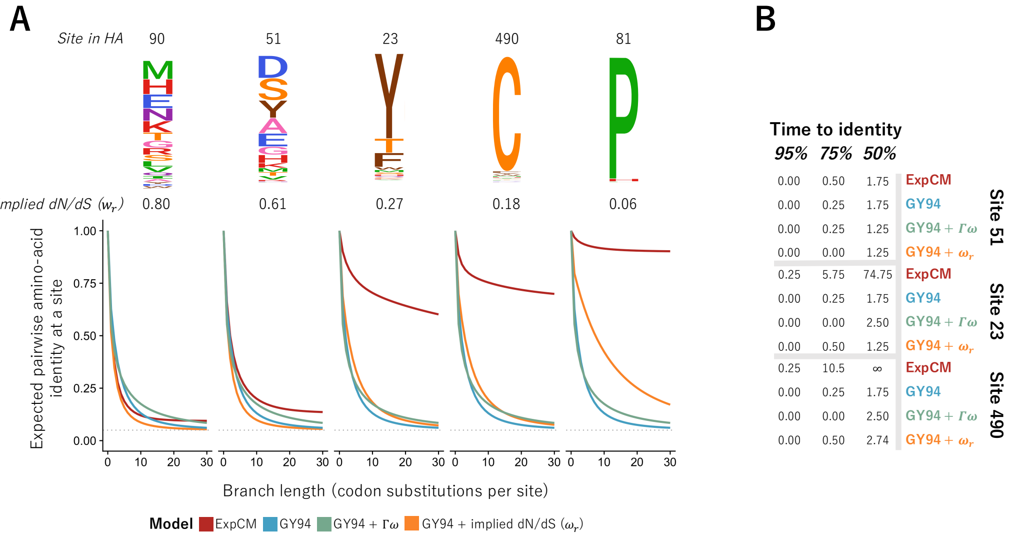
5. You can infer an dN/dS value from mutation-selection models. When you add this value to the base model you see the line come up but not as much as the true ExpCM.

6.



	Site-specific profiles	Γ-distributed rate variation
GY94	□	□
GY94 + $\Gamma\omega$	□	■
ExpCM	■	□
ExpCM + $\Gamma\omega$	■	■

Figure 1: Comparison of substitution model features. Site-specific amino-acid profiles and Γ-distributed rate variation are both substitution model features which have been shown or theorized to lengthen branches. The models YNGKP M0, YNGKP M5, ExpCM, and ExpCM+$\Gamma\omega$ represent all possible combinations of these two features. Blue indicates presence and white indicates absence of a feature.



Time to identity

95%	75%	50%		
0.00	0.50	1.75	ExpCM	Site 51
0.00	0.25	1.75	GY94	
0.00	0.25	1.25	GY94 + $\Gamma\omega$	
0.00	0.00	1.25	GY94 + ω_r	
0.25	5.75	74.75	ExpCM	Site 23
0.00	0.25	1.75	GY94	
0.00	0.00	2.50	GY94 + $\Gamma\omega$	
0.00	0.50	1.25	GY94 + ω_r	
0.25	10.5	∞	ExpCM	Site 490
0.00	0.25	1.75	GY94	
0.00	0.00	2.50	GY94 + $\Gamma\omega$	
0.00	0.50	2.74	GY94 + ω_r	

Figure 2: Effect of stationary state and rate variation on long branch estimation. The expected pairwise identity trajectories were calculated using Equation 6 and models described in Table 2. The trajectories of the YNKGP M0 (blue) and YNGKP M5 (green) do not vary from panel to panel because neither model is site-specific. The deviation in trajectory of the ExpCM (red) from the YNGKP M0 (blue) increases from left to right as the mutational constraint of the amino-acid profile increases (logoplots, above). The deviation in trajectory of the YNGKP model with a site-specific ω value inferred from the ExpCM (yellow, Equation 5) is also positively correlated with the constraint of the site-specific amino-acid profiles but the effect size is smaller.

Model performance under simulated, site-specific data.

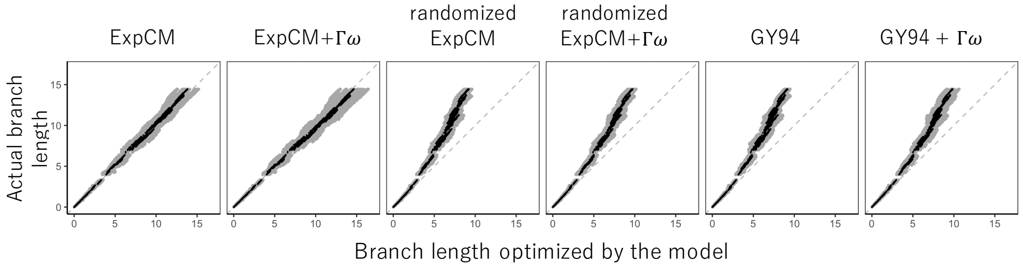


Figure 3: Model performance under simulated, site-specific data. Alignments were simulated under an ExpCM (**??**) along an HA tree and the branches were re-optimized by a model from the ExpCM or YNGKP family. The randomized ExpCMs have amino-acid profiles shuffled among the sites These randomized models 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.

Actual Data

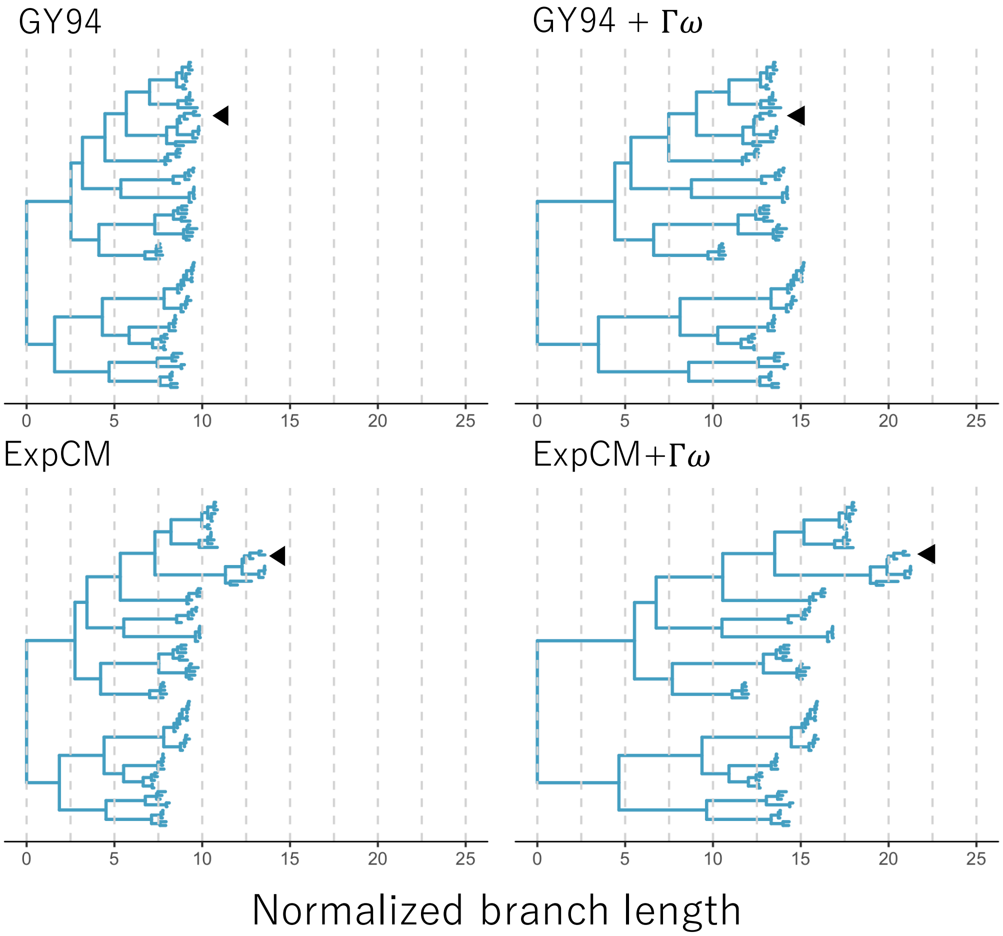


Figure 4: Trees optimized with an ExpCM defined by H1 preferences lengthen branches from the focal H1 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by H1 preferences, **(B)** an ExpCM+$\Gamma\omega$ defined by H1 preferences, **(C)** YNKGP 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).

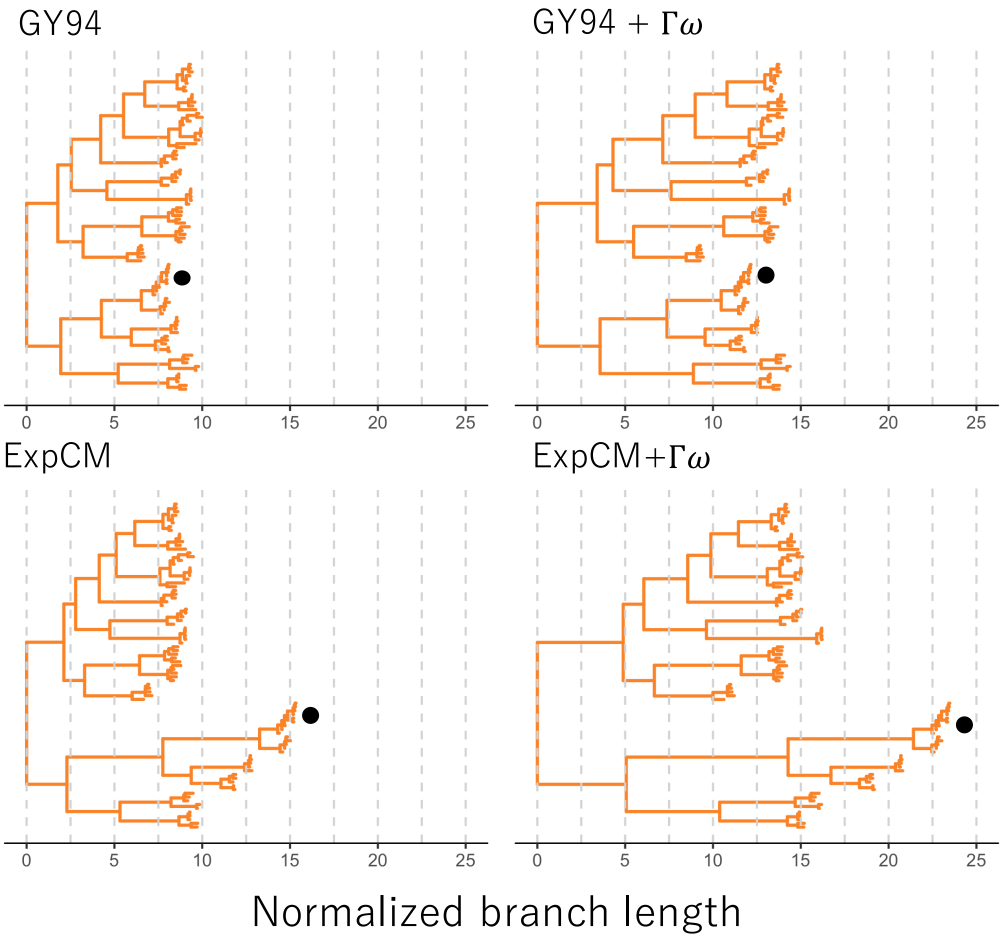


Figure 5: Trees optimized with an ExpCM defined by H3 preferences lengthen branches from the focal H3 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by H3 preferences, **(B)** an ExpCM+Γω defined by H3 preferences, **(C)** YNKGP 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 H3 focal sequence (black circle).

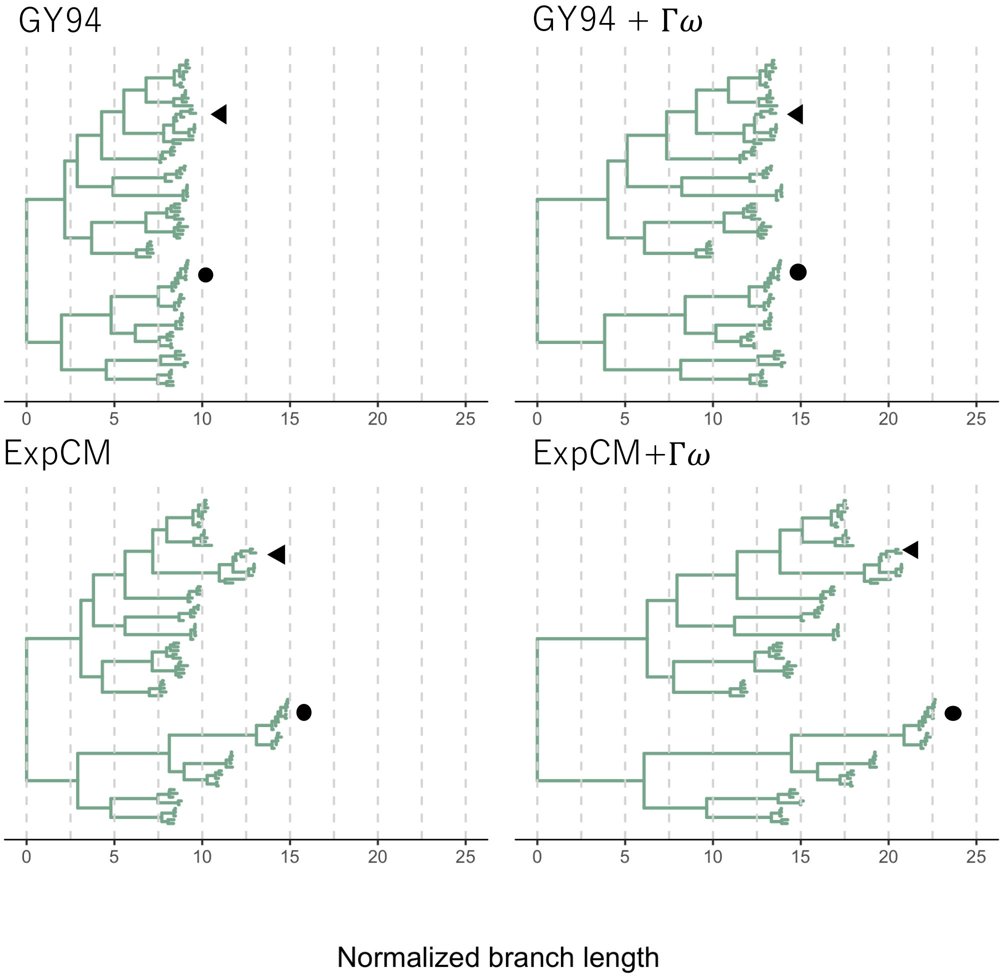


Figure 6: Trees optimized with an ExpCM defined by the average of H1 and H3 preferences lengthen branches from both the focal H3 sequence and the focal H1 sequence compared to YNGKP models. The branch lengths of a base topology inferred using the GTR-CAT model were optimized by **(A)** an ExpCM defined by the average preferences, **(B)** an ExpCM+Γω defined by the average preferences, **(C)** YNKGP 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. The black triangle indicates the H1 focal sequence and the black circle indicates the focal sequence.

0.1 Competing effects of shifting preferences and long branches.

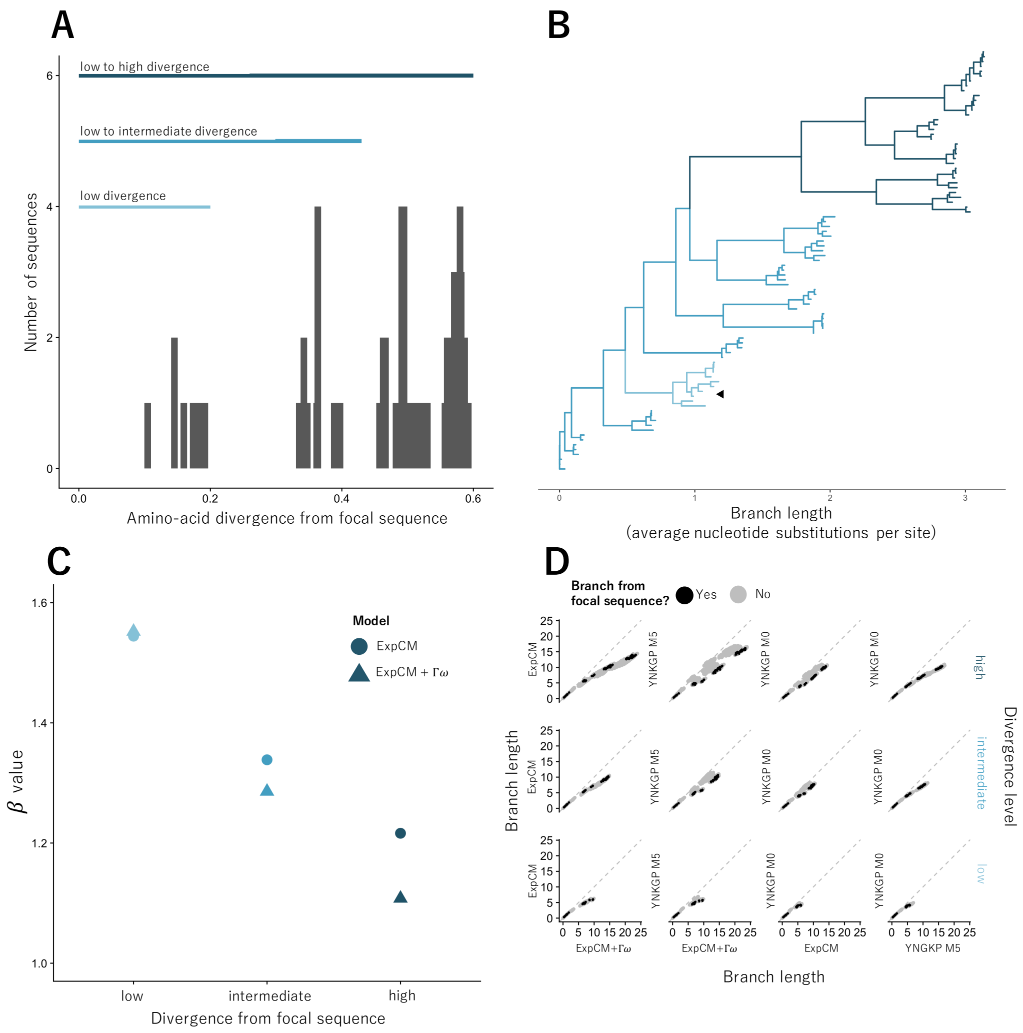


Figure 7: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H1 sequence, referred to as ”low”, ”intermediate”, and ”high”. **(B)** A phylogenetic tree of the ”high” alignment was constructed using the GTR-CAT model. The colors denote the alignment and the black circle denotes the focal H3 sequence. **(C)** The value of the ExpCM and ExpCM+Γω stringency parameter β decreases as the divergence from the focal H1 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.

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 **Exp**erimentally Informed **C**odon **M**odel (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

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

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:

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

The κ parameter represents the transition-transversion ratio, and the ϕ_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

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

where $\mathcal{A}\left(x\right)$ is the amino-acid encoded by codon x, β is the stringency parameter, and ω is the relative rate of nonsynonymous to synonymous substitutions after accounting for the amino-acid preferences. The ExpCM has six free parameters (three ϕ_w values, κ, β, and ω). 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)

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

Theoretical effect of model choice on branch length

Effect of model choice on natural sequences

ExpCM + $\Gamma\omega$ and YNGKP M5

Spielman ω_r values inferred from the ExpCM

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

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

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

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

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

Supplemental Information

Model Parameters for the simulations

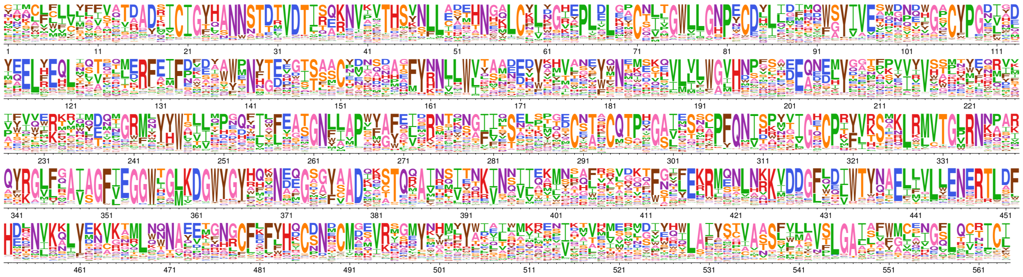


Figure 8: H1 preferences measured by Doud and Bloom (2016) rescaled with the ExpCM stringency parameter optimized in Figure 4A ($\beta = 1.21$)

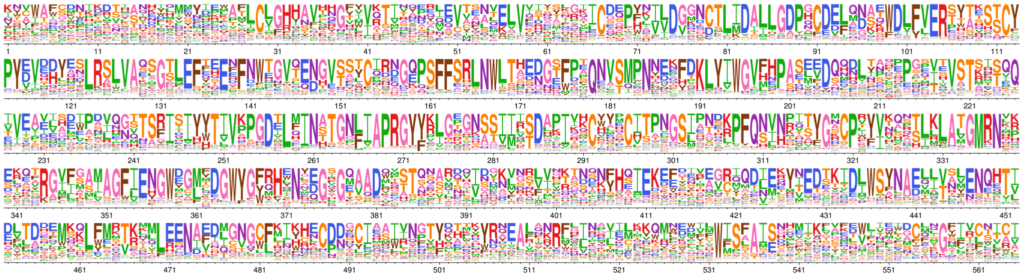


Figure 9: H3 preferences measured by *lee* rescaled with the ExpCM stringency parameter optimized in Figure 5A ($\beta = 1.46$)

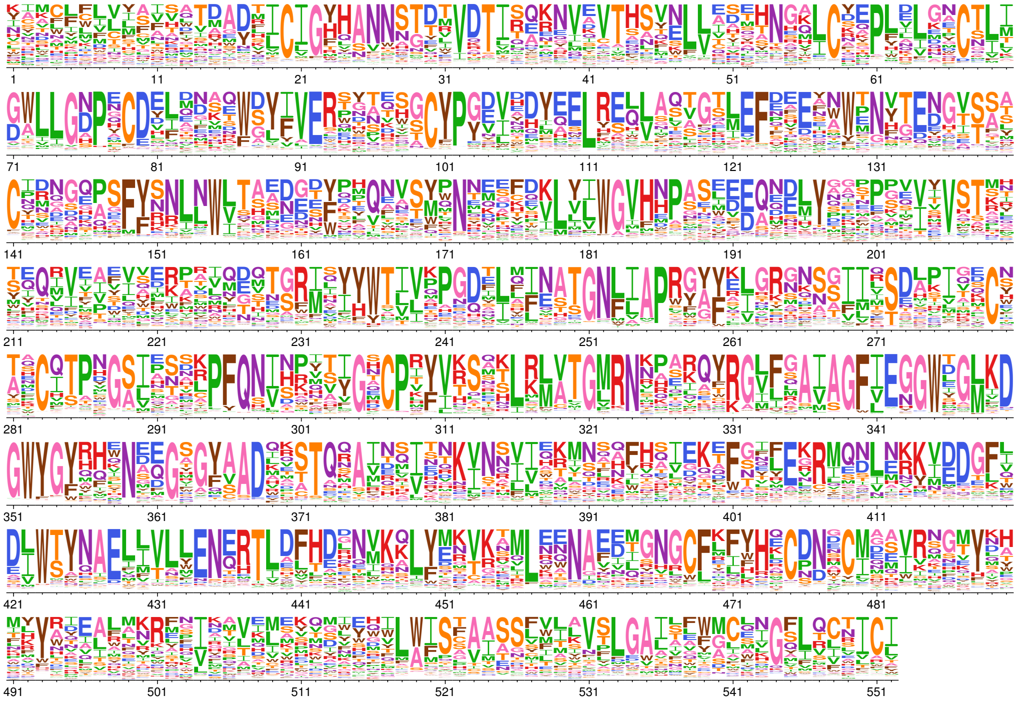


Figure 10: The average of the H1 preferences measured by Doud and Bloom (2016) and the H3 preferences measured by *Lee* rescaled with the ExpCM stringency parameter optimized in Figure 6A ($\beta = 1.82$)

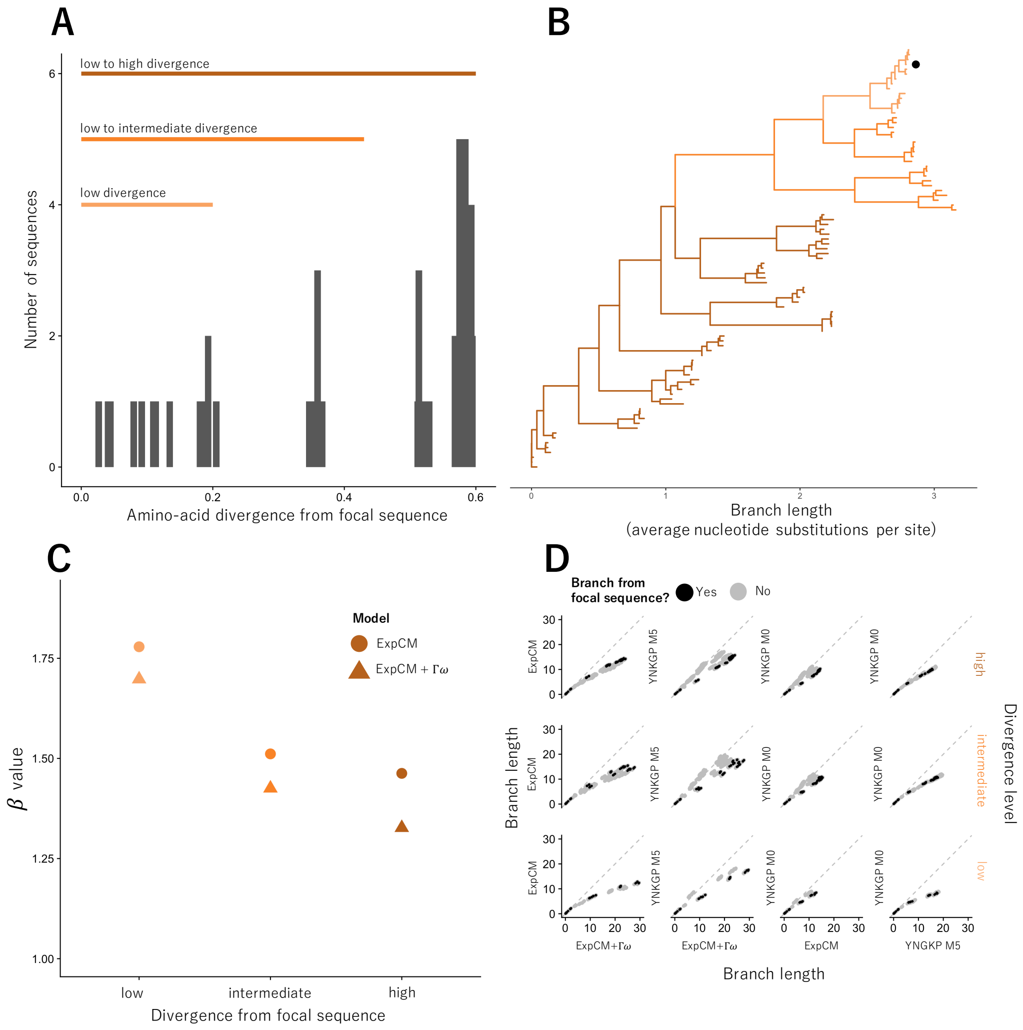


Figure 11: The ExpCM defined by H1 preferences lengthen longer branches on the HA tree. (A) An HA alignment was subsampled to create three smaller alignments with varying degrees of divergence from the focal H3 sequence, referred to as ”low”, ”intermediate”, and ”high”. **(B)** The phylogenetic tree 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 β 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.

Table 1: ExpCM parameters used to simulate sequences in Fig. **??**.

Parameter	Value
β	1.54
κ	3.60
ω	0.20
ϕ_A, ϕ_C, ϕ_G	0.38, 0.17, 0.23

Table 2: Model parameters used in Fig. **??**.

Model	Parameters
ExpCM	$\beta = 2.0$, $\kappa = 1.0$, $\omega = 1.0$, $\phi_A = \phi_C = \phi_T = 0.25$, $\pi_{r,A(X)}$:**?**
YNGKP M0	$\kappa = 1.0$, $\omega = 1.0$, $\phi_{rw} = 0.25$
YNGKP M5	$\kappa = 1.0$, $\alpha_\omega = 0.36$, $\beta_\omega = 1.9$, $\phi_{rw} = 0.25$

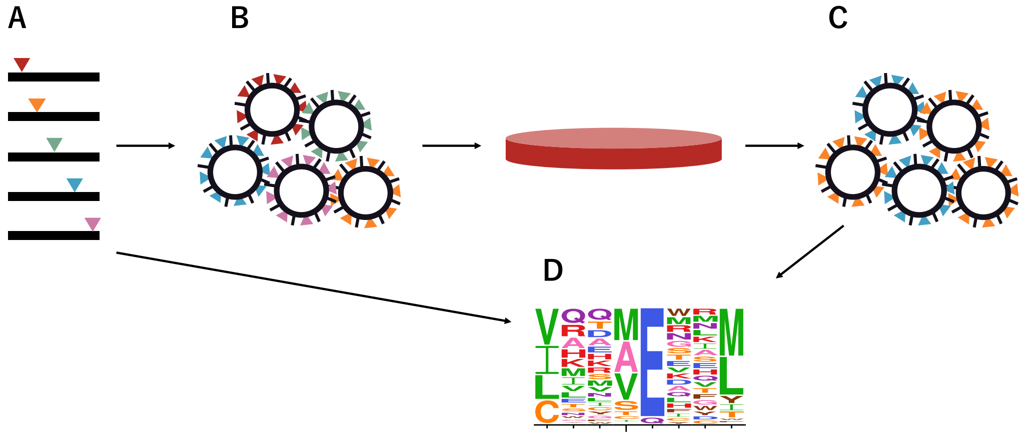


Figure 12: Schematic of deep mutational scanning. (A) All single codon mutations are introduced into the wildtype HA gene. **(B)** Each virus in the mutant virus library contains one HA variant. **(C)** The mutant virus library is passaged in cell culture to select for functional variants. **(D)** Deep sequencing quantifies the frequency of each variant before and after selection. The preference for each amino acid at each site (as quantified by the deep sequencing) is represented as a logoplot.

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.