Structurally informed evolutionary models improve phylogenetic reconstruction for emerging, seasonal, and pandemic influenza viruses

Xueting Qiu[1], Justin Bahl[1,2]

Affiliation:

[1]*The Center for Infectious Disease, The University of Texas School of Public Health, Houston, United States of America*

[2]*Program in Emerging Infectious Diseases, Duke-National University of Singapore Graduate Medical School, 8 College Road, Singapore 169857, Singapore*

Corresponding Authors

Email: Xueting.Qiu@uth.tmc.edu; Justin.Bahl@uth.tmc.edu

Abstract

Precise estimation of genetic substitution patterns is critical for accurate reconstruction of pathogen phylogenies. Few studies of viral evolution account for mutational rate variation across a single gene. This is especially true when considering evolution of segmented viruses where individual segments are short, encoding for few proteins. However, the structural and functional partition of these proteins could provide valuable information for more accurate inference of viral evolution, due to the intense immune selection pressure on different functional domains. In this study we developed and evaluated a structurally informed partitioning scheme combined with an approximate codon substitution model that accounts for rate variation among immunogenic head and stalk domains of the surface protein hemagglutinin (HA) of influenza viruses. We evaluated the model fit with a Bayes factor, using path-sampling (PS) and stepping-stone sampling (SS) approaches to calculate the marginal likelihood estimation. We evaluated and compared 4 different models - HKY85, SRD06 codon, HKY85 plus functional partitioning, SRD06 plus functional partitioning on pandemic H1N1/2009, seasonal H3N2, B-Yamagata and Victoria lineages, and two highly pathogenic avian influenza A viruses H5Nx and H7N9. The Bayes factor tests showed that structurally informed partitioning with SRD06 performed better for all datasets with decisive support. Significantly faster nucleotide substitution rates for head domain, compared to stalk domain was observed and may provide insight for stalk derived universal vaccine design. In summary, we show that integrating a functionally conserved partitioning scheme based on protein structures of immune targets allow for significant improvement of phylogenetic analysis and providing important biological insights.

Introduction

The growing importance of statistical phylogenetic methods to study the epidemiology, evolution and ecology of rapidly evolving viral pathogens has been driven by the growing availability of full genome sequences. Precisely estimating the pattern of genetic variations is critical to reconstruct the accurate pathogen phylogenies. Substitution models have been developed to calculate genetic distances among viral isolates often allow for rate variation between transitions and transversions (Hasegawa, Kishino, & Yano, 1985), or incorporates nucleotide base frequencies with substitution rate parameters (Lanave, Preparata, Sacone, & Serio, 1984; Tavaré, 1986). Segmented viruses, such as Influenza A Virus (IAV) contain short gene segments that encode for few proteins. Consequently, the literature is dominated by the analysis of protein coding regions. Grouping sites into a single partition assumed to have evolve under similar processes is a strategy to approximate more complex codon models or to account for rate variation across genetic domains. (Lanfear, Calcott, Ho, & Guindon, 2012). Approximate codon-models, such as the SRD06 incorporate information about the genetic code allowing for a rate variation between the third codon position and the first and second positions due to rapid accumulation of synonymous mutations in the toggle position of the codon (Shapiro, Rambaut, & Drummond, 2006).

Viral phylodynamic studies have demonstrated how epidemiological, immunological and evolutionary processes can determine genetic variation (Volz, Koelle, & Bedford, 2013). These efforts have shed light on how transmission dynamics impact on viral genetic changes: not only mutation rates and generation times but also selections heavily act on the patterns of viral genetic variation and viral phenotypes. Natural selection of IAV to escape human immunity is manifest as antigenic drift resulting in a rapid turnover of antigenic phenotypes (Koelle, Cobey, Grenfell, & Pascual, 2017). This results in vaccine candidates have to be selected for each epidemic

season in a timely manner with the risk of imprecise match of antigenic phenotypes in the circulating strains. This phenomenon has been clearly observed in influenza viruses (Bedford et al., 2015), where the surface protein hemagglutinin (HA) contains the major antigenic sites recognized by host immune system and has been used to evaluate antigenic drifts.

HA glycoprotein with a spike-shaped structure binds to the receptors on the targeted host cell membrane when the virus begins the infection (Riwilaijaroen & Uzuki, 2012). This protein has two main structure parts – the globular head domain and stalk domain. Functionally different, the globular head domain contains the receptor binding sites, while the stalk domain is a main structure responsible for membrane fusion machinery (Riwilaijaroen & Uzuki, 2012; Wiley & Skehel, 1987). Even though immune selection strongly drives antigenic drift, allowing for accumulation of mutations in the head domain, the stalk domain is functionally conserved across viral subtypes. Current seasonal influenza virus vaccines induce humoral immune responses mainly targeting the immunodominant globular head domain (Nachbagauer et al., 2016; Wang & Palese, 2011), which can have strong specific protection from infecting virus that is similar with vaccine candidates, but have weak cross-reactivity to antigenic drift variants. Conversely, studies in the ferret experimental model demonstrate that the stalk domain was highly conserved across IAV subtypes and stalk specific antibodies could provide cross-reactive protection to a high diversity of IAV subtypes (Krammer et al., 2014; Nachbagauer et al., 2016). Despite the importance of understanding how substitution rates and patterns vary across functionally conserved domains in universal vaccine design (Krammer & Palese, 2013; Subbarao & Matsuoka, 2013), few studies have incorporated protein structure in phylogenetic models.

In this study we propose a novel phylogenetic model that incorporates a protein structure informed partitioning scheme to account for variable evolutionary rates resulting from immune

selection. We aim to evaluate the appropriateness of the novel model to capture the biological realism and reliably to estimate biologically informative parameters from available genetic data. We compare the model performance for Seasonal H3N2, H1N1/2009, Highly Pathogenic Avian Influenza A Virus H5NX and H7N9 virus and the two seasonal Influenza B virus lineages, Yamagata and Victoria. Estimation of viral drift rates for the head and stalk domain will be important for vaccine design. We evaluate and compare 4 different models: HKY85, SRD06 codon, HKY85 plus protein structurally informed partitioning, SRD06 plus protein structurally informed partitioning and determine the minimum dataset size necessary to robustly estimate the parameters of interest. We formally test the hypothesis that rates in each domain are significantly different and estimate the relative selection pressure for each gene region.

Methods

Dataset

To test on the stability of the performance of the proposed structurally informed partitioning model and compare to other models, we included pandemic H1N1/2009, three types of seasonal influenza viruses (H3N2, B-Yamagata and B-Victoria) and two highly pathogenic avian influenza (HPAI) H5Nx and H7N9. Except for HPAI H7N9, we used all other datasets from published research, including H1N1/2009 viruses (Su et al., 2015), seasonal H3N2 (Bedford et al., 2015), HPAI H5N1 (Qiu et al., 2017), B-Yamagata and B-victoria (Vijaykrishna et al., 2015). Specifically, these global H1N1/2009 viruses (2009-2014) is the dateset (n=400) used to generate figure 4a in Su et al., 2015 paper. After removing duplicates, we have 391 isolates for H1N1/2009. HPAI H5N1 contained 1095 isolates from 1996-2017, after removing the vaccine candidates from Qiu et al., 2017 paper. These two B lineages (348 isolates of B-yamagata and 538 isolates of B-Victoria) from Vijaykrishna et al, 2015 paper, were full genome of influenza B viruses sampled during 2002 -2013 in eastern Australia and New Zealand. For B-Victoria, we subsampled down to 50% of the dataset and removed duplicates, resulting in n =214. For B-Yamagata, we had 241 isolates after removing the duplicates. Genetic data and epidemiologic metadata of HPAI H7N9 during 2013-2015 outbreaks were downloaded from Global Initiative on Sharing Avian Influenza Data (GISAID, http://platform.gisaid.org/). After alignments and removed duplicates, this final dataset for HPAI H7N9 contains 365 isolates.

Definition of HA protein functional partitions

HA glycoprotein has been defined as different functional domains, including signal peptide, stalk domain, globular head domain, transmembrane domain and cytoplasmic domain (Riwilaijaroen & Uzuki, 2012; Wiley & Skehel, 1987). Given the similar evolutionary characteristics of conserved functional domains, we combined signal peptide, transmembrane

and cytoplasmic domain as one domain named STC in our modeling. The stalk domain has two gene regions, separated by head domain. The nucleotide number of functional partitions for each influenza subtype included in this study have been listed in supplemental Table 1.

Tree likelihood construction in the structurally informed model

Bayesian frame work has been used for constructing the new model. The general Bayesian inferences are based on the posterior probability of a hypothesis, for example, a given tree τ. So the posterior probability of the tree based on the data can be obtained using Bayes theorem,

$$Pr\ (\tau|x) = \frac{Pr(\mathcal{X}|\tau)\ Pr(\tau)}{\sum Pr(\mathcal{X}|\tau) Pr\ (\tau)}$$

Where Pr(τ) is the prior probability of the tree hypothesis τ based on data x; Pr(x|τ) is the probability of observing the data at a specific site.

Then assuming independence of substitutions across sites, the general likelihood based on a specific phylogenetic model, i.e. the probability of observing the aligned matrix X of all j sequences is,

$$f\ (X|\tau_j, \nu_j, \theta, \phi) = \prod_{i=1}^{c} f\,(x_i|\tau_j, \nu_j, \theta, \phi)$$

Where c is the number of total sites along the sequences; τ represents the tree, ν represents the branch lengths, θ means the substitution model parameters, ϕ represents all other model parameters.

Our proposed structurally informed partitioning model based on HA protein functional domains extended the formula of tree likelihood estimation. Each of the defined partition has its tree likelihood estimated independently. The overall tree likelihood is the products of each partition-specific likelihood.

Phylogenetic models and simulations

Four different Bayesian phylogenetic models were conducted for each dataset – HKY with Gamma invariant distribution, SRD06 codon model, HKY plus structurally informed partitions (STC, stalk and head), and SRD06 codon plus structurally informed partitions. Absolute rates for each partitioning have been coded in these BEAST xml files. These models were conducted in BEAST v1.8.4 with an uncorrelated lognormal relaxed molecular clock that allows for rate variation across lineages. Bayesian Skyride coalescent tree prior was used to incorporate the changes of viral populations. Each model was performed three independent runs of 100 million Markov Chain Monte Carlo (MCMC) generations. To report the substitution rates and phylogenetic trees, three runs for each model were combined after removal of burn-in to achieve an Effective Sample Size (ESS) of >200 as diagnosed in Tracer v 1.5. Visualized violin plots for substitution rates were generated in R package.

Model selection criteria

The goal of model selection is to compare models and identify a model that is sufficiently complex to capture the biological realism and evolutionary processes that have occurred but to avoid overparameterized models with more parameters than can be reliably estimated from the available data (Baele, Lemey, et al., 2012a; Lanfear et al., 2012; Steel, 2005). Previously the harmonic mean estimate (HME) and Akaike's information criterion through MCMC (AICM) (Raftery, Newton, Satagopan, & Krivitsky, 2006) have been commonly used. Recent years, several new approaches to perform model selection in the field of phylogenetics, such as path-sampling (PS) (Lartillot, Philippe, & Lewis, 2006), stepping-stone sampling (SS) (Xie, Lewis, Fan, Kuo, & Chen, 2011) and generalized stepping-stone sampling (Fan, Wu, Chen, Kuo, & Lewis, 2011). Baele et al. (2012) have tested HME, AICM, PS, and SS approaches for

demographic and molecular clock model comparison, which "confirmed that HME systematically overestimates the marginal likelihood and fails to yield reliable model classification, and PS and SS substantially outperform HME estimators" (Baele, Lok, et al., 2012). In this study, we conducted both PS and SS for the model selection procedure.

The marginal likelihood is the probability of the data (that is, likelihood) given the model type, not assuming any particular model parameters (Baele, Lemey, et al., 2012b). To compare the superiority of each model, we used path sampling (PS) and stepping-stone (SS) sampling to calculate the marginal likelihood estimator, with Beta path step distributions. The marginal likelihood estimation for both PS and SS ran 100 path steps and length of chains as 1 million (Baele, Lemey, et al., 2012a). Marginal likelihood estimates coding referred to *BEAST Model Selection Tutorial* (http://beast.bio.ed.ac.uk/Model-selection). Bayes Factors (BF), the ratio of marginal likelihood estimations from two models, were used to perform model selection. When convert to log scale, BF becomes the difference between two log marginal likelihood estimations. The mathematical conversion is below (Kass & Raftery, 1995):

For a model with parameters Θ, the marginal likelihood for the model M with data X is,

$$p\,(X \mid M) = \int p(X \mid \Theta, M_1) \cdot p(\Theta \mid M)\, d\Theta \quad (1)$$

Then the Bayes factor can be calculated for model M1 against model M2,

$$BF_{12} = \frac{p\,(X \mid M_1)}{p\,(X \mid M_2)} = \frac{\int p(X \mid \Theta_1, M_1) \cdot p(\Theta_1 \mid M_1)\, d\Theta_1}{\int p(X \mid \Theta_2, M_2) \cdot p(\Theta_2 \mid M_2)\, d\Theta_2} \quad (2)$$

We derived the log scale of BF is in (3),

$$logBF_{12} = \log p(X \mid M_1) - \log p(X \mid M_2) = dB \quad (3)$$

With log scale BF as dB, we have the criteria for model selection: dB <0 means the selection supports M2; 0 ≤ dB <3 means no evidence for supporting M1; 3 ≤ dB < 10 means substantial

strength of evidence for supporting M1; 10 ≤ dB <15 means strong evidence of supporting M1; 15 ≤ dB < 20 means very strong evidence of supporting M1; dB ≥ 20 means decisive evidence of supporting M1. The histogram figures were generated in Tableau v.9.3 for the comparisons of BFs.

Sensitivity analysis

To evaluate how many data is sufficient to use the structurally informed partitioning model, we conducted sensitivity analysis on HPAI H5Nx from 1996-2017, which is the largest dataset with 1095 isolates. We subsampled the dataset to keep 80%, 60%, 40% and 20% of the dataset, respectively. We conducted 3 runs of 100 million MCMC chain length of the four models for the full dataset and these subsampled datasets with marginal likelihood estimations. Tree height parameters, substitution rates of each functional partition and model selection BFs were used to evaluate the data sufficiency for the stability of analysis results.

Results

Model selection

To compare the superiority of each model, we used the log scale Bayes Factor, which is the difference of two log marginal likelihood estimation for two models. Based on the model selection procedure for all subtypes (Figure 1), we found neither HKY85 nor HKY85 with structurally informed partitioning has decisively significant advantages than SRD06 codon model or structurally informed partitioning model with SRD06 codon model. Furthermore, the new model with mixed partitioning of SRD06 plus structurally informed model showed superiority to other models in all datasets. BFs for SRD06 codon plus structurally informed partitioning model compared to SRD06 only showed strong support for mixed partitioning, or at least not significantly favors SRD06 codon only model (only B-Yamagata showed no significant difference between these two models. The structurally informed partitioning with SRD06 model was more favored in large dataset with a decisive BF support, for example H5Nx.

Tree root height and mutation rate comparison

To evaluate the accuracy and stability of model performances, tree parameters and estimated substitution rates were used to compare different models. Results from four models showed very similar tree root height and 95% highest posterior density (HPD) for each influenza subtypes (Table 1). But we did observe one exception is that for B-Yamagata, SRD06 codon model report a larger root height compared to other three models.

Though the overall mean substitution rates (Figure 2) are similarly estimated by four models for each dataset, HKY85 plus structurally informed partitioning model and SRD06 codon plus structurally informed partitioning model provided detailed substitution rates for STC, stalk and head domains, respectively. However, the rates reported by HKY85 plus structurally

informed partitioning model were underestimated, compared to SRD06 codon plus structurally informed partitioning. One thing to note, STC domain with very small amount of data information resulted in unprecise estimation of substitution rates, showing long tails in violin plots.

Substitution rates of stalk and head domains

Structurally informed partitioning with SRD06 codon model provided separate substitution rates for stalk and head domains. Absolute rates were calculated directly with embedding in related codes into the xml file. From the violin plots in Figure 2 showing the substitution rates for stalk and head domains from each step of the Bayesian simulations, we observed that head domain has a significantly higher mutation rate than stalk domain for all influenza types, under the unit of substitutions per site per year.

Approximate dS/dN ratio

As the new model with SRD06 codon position partitioning, it allowed us to estimate the approximate ratio of synonymous substitutions (codon position 3) over non-synonymous substitutions (codon positions 1 and 2), that is, dS/dN for stalk domain and head domain separately. As shown in Table 2, we found the range of dS/dN is 4.16-9.62 for stalk domain and 1.15 – 4.00 for head domain, respectively. We observed that stalk domain has a higher dS/dN than head domain for each influenza subtypes.

Evaluating data sufficiency for structurally informed partitioning model

With our largest dataset H5Nx, we performed sensitivity analysis to estimate the stability of model performance and requirement of data amount. We used five datasets of H5Nx – full, 80%, 60%, 40% and 20% of the dataset. The model selection procedure (Figure 3) showed that all these five datasets had decisive BFs to favor the structurally informed partitioning model with

SRD06. Tree root height estimation (Table 3) from these five datasets showed similar results and 95% HPD, but they have been slightly overestimated 20% of H5Nx dataset. The violin plots of substitution rates (Figure 4) showed underestimation with 20% H5Nx dataset, giving us the hint that the 20% dataset would not be able to tell the precise evolution story of H5Nx with our new model.

Discussion

In our study, the model selection and comparison procedure supported the superiority of the new proposed hierarchical model – structurally informed partitioning with SRD06 codon model. The advantages were demonstrated by both the supportive Bayes Factors and the stability of model performance on phylogenetic tree root height and substitution rates. The substitution rates of stalk and head domains of HA protein are different, with a higher rate in head domains for all the tested influenza types. With the most complete dataset of HPAI H5N1 and its randomly subsampled datasets, we showed that the data sufficiency for applying the mixed partitioning model is no less than xx% of data with a 21-year time span.

When developing the new model, we tried to capture as much information of the structure and function of HA protein as we can. One solution to do so is to compare all possible partitioning schemes for a given dataset. For example, we could do a more detailed functional partitioning with separating all possible functional sites. However, this approach is usually computationally intractable because the number of possible partitioning schemes is tremendous even for relatively small numbers of data blocks (Li et al. 2008). As a result, most studies either choose a single partitioning scheme a priori or select the best-fit scheme from a handful of candidate schemes (Brandley et al. 2005; McGuire et al. 2007). Thus, despite significant advances in phylogenetic methods in recent years, the accuracy of the inferences we can make from partitioned phylogenetic analyses remains limited by our ability to select appropriate partitioning (Lanfear et al., 2012). Therefore, we combined some functional domains of HA and only explored three different domains of HA protein, with conducting a formal model selection procedure to quantitatively evaluate its superiority (Baele, Lok, et al., 2012).

The new model proposed in this study is significant in several perspectives. This is so far the first time to develop a model for influenza viruses with incorporating biological functional partitioning inside one protein to introduce a much closer look of the evolutionary process. Partitioning is a great approach for capturing similarity inside one subgroup of data and compare different subgroups to generate new insight of questions, which balances model complexity and number of parameters (Lanfear et al., 2012). For example, SRD06 codon model categories the first and second codon position together since they behaved more similarly than the third codon position (Lanfear et al., 2012; Shapiro et al., 2006). Some model did partitioning on multiple proteins to be able to run one model and evaluate the co-evolution and separate rates for each protein (Mintaev, Alexeevski, & Kordyukova, 2014). These models function well for some purposes of observing the evolutionary process, however, it did not include the biological structure and function of each protein. Molecular mechanism of influenza viruses, for example, virus entry or fusion or antigenic-immune stimulation, is determined by protein structure and function translated from genetic information but the genetic information also reflects different changes of these functional partitions with different selection pressure (Kordyukova, 2017; Lu et al., 2012; Skehel & Wiley, 2000). Our new functional partitioning model considered the specific functions of head domain to mainly bind to host-cell receptors and the stalk domain to perform membrane fusion. Our model did capture the different evolutionary process of these two domains, which showed head domain has a higher substitution rates, probably indicating a higher immune selection pressure on head domain. Another thing, the model selection procedure showed the models with SRD06 significantly ran over the HKY85 based model. We observed more accurate substitution rates with the new model, compared to HKY85 plus functional partitioning, which means SRD06 is a necessary component to be retained in the new model. Therefore, the new

model with combing SRD06 and functional partitioning is the best to be absolutely supported by BFs and meanwhile provide the most accurate estimation of mutation rates for each domain. Furthermore, even we added in new partitioning parameters, the running time of our model did not require longer time or computing power than SRD06 codon model only under the same computing environment. And the sensitivity analyses confirmed that our model requires a reasonable amount of data to be conducted, i.e., did not request extra efforts on data sampling.

Our new model with the superior characteristics above provides insight for universal vaccine design. The hierarchical evolutionary model with incorporating protein structure and function provides a more accurate estimation of nucleotide substitution rates for stalk and head domains. The higher rates observed in head domain are probably due to strong immune selection targeted on the globular head, where some sites are mostly targeted by the immunity induced by seasonal influenza vaccine (Nachbagauer et al., 2016). Proven by many studies, stalk domain or more conserved sites on head domain could be used to induce more broad-reactive vaccines (Krammer et al., 2014; Nachbagauer et al., 2016; Subbarao & Matsuoka, 2013). Therefore, our model could be further developed to identify these conserved regions by estimating region specific rates and substitution patterns under a shared viral phylogeny. This model can be used to find out all common conserved regions for all viral types in influenza group 1 (H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, and H17) and group 2 (H3, H4, H7, H10, H14 and H15) (Lu et al., 2012), which probably is the most promising approach to achieve universal vaccine design. But for sure, the results from the computational design should be tested on animal models to record the immune-boosting capabilities.

Several limitations existed. We did not compare with other molecular clock and coalescent models to select the most BF supported one out of all possible combinations, which

needs too much efforts and barely possible to test all. Our reasoning on not doing all selections is that we used the lognormal uncorrelated relax clock and skyride coalescent model, which superiority has been supported in previous model selection study (Baele, Lemey, et al., 2012a). Furthermore, our results of the most common ancestor and overall substitution rates are similar with the original studies where our datasets were from. Another limitation is that though PS and SS sampling procedure for marginal likelihood estimate is only determined by model type with free assumptions on any particular model parameters, model selection can be affected by the run quality. We could throw away some obviously bad simulations, but it is really hard to have a set of quantitative criteria. The way we tried to deal with this issue is to run more than 3 simulations at the same time, to give a choice of better runs and guarantee the number of qualified runs.

Future perspectives on this study could be two main directions: 1) refine and improve the model when we have a deeper understanding on HA protein functions; 2) apply this model to other viruses. We keep updating our knowledge on the surface protein of influenza viruses, with rapid new findings reported nowadays. Genetic information and modeling has its limitation on interpreting mechanisms, therefore, it is necessary to incorporate and update with more advanced functional information into our model. As we discussed, common conserved regions of the high diverse viruses could be nicely identified by our hierarchical partitioning model to help with universal vaccine design. This model is developed not only for influenza viruses, but also for other similar or very different viruses. For example, in respiratory syncytial virus (RSV) the fusion protein functions similar as HA protein in influenza virus and evolves under similar substitution rates and pattern (Borchers, Chang, Gershwin, & Gershwin, 2013). For very different viruses, for example, HIV viruses have a significant phenomenon that within host evolution and between host evolution are tremendously different. Will this model still fit? We

cannot answer it. But even our model did not function well on HIV, trying our model on HIV probably leads to developing a new model that could incorporate the different evolutionary processes within and among hosts. For this situation, our model could be a stepping stone to reach a more ideal solution.

Acknowledgements

This study was partially funded by NIAID Centers of Excellence for Influenza Virus Research and Surveillance (CEIRS, HHSN272201400006C). These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References:

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012a). Improving the accuracy of demographic and molecular clock model comparison while accomodating phylogenetic uncertainty. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msl161

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012b). Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty Research article, *29*(9), 2157–2167. https://doi.org/10.1093/molbev/mss084

Baele, G., Lok, W., Li, S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2012). Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics Letter Fast Track, *30*(2), 239–243. https://doi.org/10.1093/molbev/mss243

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., … Russell, C. A. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, *523*, 217. Retrieved from http://dx.doi.org/10.1038/nature14460

Borchers, A. T., Chang, C., Gershwin, M. E., & Gershwin, L. J. (2013). Respiratory Syncytial Virus—A Comprehensive Review. *Clinical Reviews in Allergy & Immunology*, *45*(3), 331–379. https://doi.org/10.1007/s12016-013-8368-9

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., & Lewis, P. O. (2011). Choosing among Partition Models in Bayesian Phylogenetics. *Molecular Biology and Evolution*, *28*(1), 523–532. Retrieved from http://dx.doi.org/10.1093/molbev/msq224

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. https://doi.org/10.1007/BF02101694

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Koelle, K., Cobey, S., Grenfell, B., & Pascual, M. (2017). Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans, *314*(5807), 1898–1903.

Kordyukova, L. (2017). Structural and functional specificity of Influenza virus haemagglutinin and paramyxovirus fusion protein anchoring peptides. *Virus Research*, *227*(Supplement C), 183–199. https://doi.org/https://doi.org/10.1016/j.virusres.2016.09.014

Krammer, F., Hai, R., Yondola, M., Tan, G. S., Leyva-Grado, V. H., Ryder, A. B., … Albrecht, R. A. (2014). Assessment of influenza virus hemagglutinin stalk-based immunity in ferrets. *Journal of Virology*, *88*(6), 3432–42. https://doi.org/10.1128/JVI.03004-13

Krammer, F., & Palese, P. (2013). Influenza virus hemagglutinin stalk-based antibodies and vaccines. *Current Opinion in Virology*, *3*(5), 521–530. https://doi.org/https://doi.org/10.1016/j.coviro.2013.07.007

Lanave, C., Preparata, G., Sacone, C., & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, *20*(1), 86–93. https://doi.org/10.1007/BF02101990

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701. Retrieved from http://dx.doi.org/10.1093/molbev/mss020

Lartillot, N., Philippe, H., & Lewis, P. (2006). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, *55*(2), 195–207. Retrieved from http://dx.doi.org/10.1080/10635150500433722

Lu, X., Shi, Y., Gao, F., Xiao, H., Wang, M., Qi, J., & Gao, G. F. (2012). Insights into Avian Influenza Virus Pathogenicity: the Hemagglutinin Precursor HA0 of Subtype H16 Has an Alpha-Helix Structure in Its Cleavage Site with Inefficient HA1/HA2 Cleavage. *Journal of Virology*, *86*(23), 12861–12870. https://doi.org/10.1128/JVI.01606-12

Mintaev, R. R., Alexeevski, A. V, & Kordyukova, L. V. (2014). Co-evolution analysis to predict protein–protein interactions within influenza virus envelope. *Journal of Bioinformatics and Computational Biology*, *12*(2), 1441008. https://doi.org/10.1142/S021972001441008X

Nachbagauer, R., Miller, M. S., Hai, R., Ryder, A. B., Rose, J. K., Palese, P., & García-sastre, A. (2016). Hemagglutinin Stalk Immunity Reduces Influenza Virus Replication and Transmission in Ferrets, *90*(6), 3268–3273. https://doi.org/10.1128/JVI.02481-15.Editor

Qiu, X., Duvvuri, V. R., Gubbay, J. B., Webby, R. J., Kayali, G., & Bahl, J. (2017). specific epitope profiles for HPAI H5 pre- - pandemic vaccine selection and evaluation, (July), 445–456. https://doi.org/10.1111/irv.12466

Raftery, A., Newton, M., Satagopan, J., & Krivitsky, P. (2006). Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. *Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series*. Retrieved from http://biostats.bepress.com/mskccbiostat/paper6

Riwilaijaroen, B. N. S., & Uzuki, Y. S. (2012). Review Molecular basis of the structure and function of H1 hemagglutinin of in fl uenza virus, *88*, 226–249.

Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Molecular Biology and Evolution*, *23*(1), 7–9. Retrieved from http://dx.doi.org/10.1093/molbev/msj021

Skehel, J. J., & Wiley, D. C. (2000). Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annual Review of Biochemistry*, *69*(1), 531–569. https://doi.org/10.1146/annurev.biochem.69.1.531

Steel, M. (2005, June 1). Should phylogenetic models be trying to "fit an elephant"? *Trends in Genetics*. Elsevier Current Trends. https://doi.org/10.1016/j.tig.2005.04.001

Su, Y. C. F., Bahl, J., Joseph, U., Butt, K. M., Peck, H. A., Koay, E. S. C., … Smith, G. J. D. (2015). selection. *Nature Communications*, *6*, 1–13. https://doi.org/10.1038/ncomms8952

Subbarao, K., & Matsuoka, Y. (2013). The prospects and challenges of universal vaccines for influenza. *Trends in Microbiology*, *21*(7), 350–358. https://doi.org/10.1016/j.tim.2013.04.003

Tavaré, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, *17*, 57–86. Retrieved from citeulike-article-id:6732633

Vijaykrishna, D., Holmes, E. C., Joseph, U., Fourment, M., Su, Y. C. F., Halpin, R., … Jennings, L. C. (2015). The contrasting phylodynamics of human influenza B viruses, 1–23. https://doi.org/10.7554/eLife.05055

Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral Phylodynamics, *9*(3). https://doi.org/10.1371/journal.pcbi.1002947

Wang, T. T., & Palese, P. (2011). Catching a Moving Target. *Science*, *333*(6044), 834 LP-835. Retrieved from http://science.sciencemag.org/content/333/6044/834.abstract

Wiley, D. C., & Skehel, J. J. (1987). The Structure and Function of the Hemagglutinin Membrane Glycoprotein of Influenza Virus. *Annual Review of Biochemistry*, *56*(1), 365–394. https://doi.org/10.1146/annurev.bi.56.070187.002053

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M.-H. (2011). Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Systematic Biology*, *60*(2), 150–160. Retrieved from http://dx.doi.org/10.1093/sysbio/syq085

TABLES AND FIGURES

Table 1. Tree root height estimation and its 95% highest posterior density (HPD) from each model for each influenza subtype

Models	H1N1/2009			B-Victoria			B-Yamagata			H7N9			H5Nx-full		
	Height	95%HPD*		Height	95%HPD		Height	95%HPD		Height	95%HPD		Height	95%HPD	
HKY85	5.92	5.75	6.10	11.46	11.10	11.84	13.85	12.88	14.86	3.92	3.47	4.54	21.27	21.08	21.48
SRD06	5.93	5.75	6.12	11.47	11.13	11.87	14.88+	12.85	15.03	3.90	3.47	4.48	21.29	21.08	21.48
p**	5.93	5.75	6.12	11.49	11.13	11.88	13.83	12.87	14.83	3.88	3.43	4.49	21.28	21.08	21.49
pc***	5.92	5.73	6.11	11.47	11.12	11.87	13.84	12.84	14.81	3.95	3.46	4.65	21.28	21.08	21.49

***: 95% HPD: 95% highest posterior density.**

+: The root height estimated by SRD06 model for B-Yamagata is overestimated.

****: p means the HKY85 plus structurally informed partitioning model.**

*****: pc represents SRD06 codon plus structurally informed partitioning model.**

Table 2. Approximate dS/dN for stalk and head domain

Dataset	Substitution rates (subs/per/site)				Approximate dS/dN+	
	Stalk 3*	Stalk 1+2**	Head 3	Head 1+2	Stalk	Head
H1N1/09	6.21E-03	1.34E-03	5.97E-03	2.38E-03	4.63	2.51
B-Victoria	4.04E-03	5.44E-04	4.63E-03	1.51E-03	7.44	3.06
B-Yamagata	4.93E-03	5.12E-04	5.56E-03	1.39E-03	9.62	4.00
H7N9	8.02E-03	2.25E-03	9.26E-03	2.92E-03	3.56	3.17
H5Nx-full	8.00E-03	1.92E-03	7.78E-03	6.77E-03	4.16	1.15

+: dS/dN means the ratio of synonymous substitution to non-synonymous substitution.

***: 3 means the codon position 3.**

****: 1+2 means the codon position 1 and 2.**

Table 3. Tree root height estimation and its 95% highest posterior density (HPD) from each model for all full H5Nx and its subsampled datasets

Models	H5Nx-full			H5Nx-80%			H5Nx-60%			H5Nx-40%			H5Nx-20%		
	Height	95%HPD*		Height	95%HPD		Height	95%HPD		Height	95%HPD		Height	95%HPD	
HKY85	21.27	21.08	21.48	21.32	21.08	21.57	21.34	21.07	21.61	21.41	21.07	21.79	21.66	21.06	22.24
SRD06	21.29	21.08	21.48	21.30	21.08	21.57	21.33	21.07	21.62	21.41	21.07	21.76	21.65	21.09	22.25
p**	21.28	21.08	21.49	21.32	21.08	21.58	21.32	21.07	21.59	21.41	21.07	21.78	21.67	21.08	22.23
pc***	21.28	21.08	21.49	21.33	21.08	21.58	21.35	21.07	21.64	21.39	21.07	21.79	21.67	21.08	22.23

***: 95% HPD: 95% highest posterior density.**

****: p means the HKY85 plus structurally informed partitioning model.**

*****: pc represents SRD06 codon plus structurally informed partitioning model.**

Figure 1. Log scale Bayes Factors for model selection procedure for each influenza subtype. All these BFs were calculated to compare with HKY. The asterisk * means a significant supportive BF for the new structurally informed model, compared to SRD06 codon only model. In the figure, p means the HKY85 plus structurally informed partitioning model; cp represents SRD06 codon plus structurally informed partitioning model.

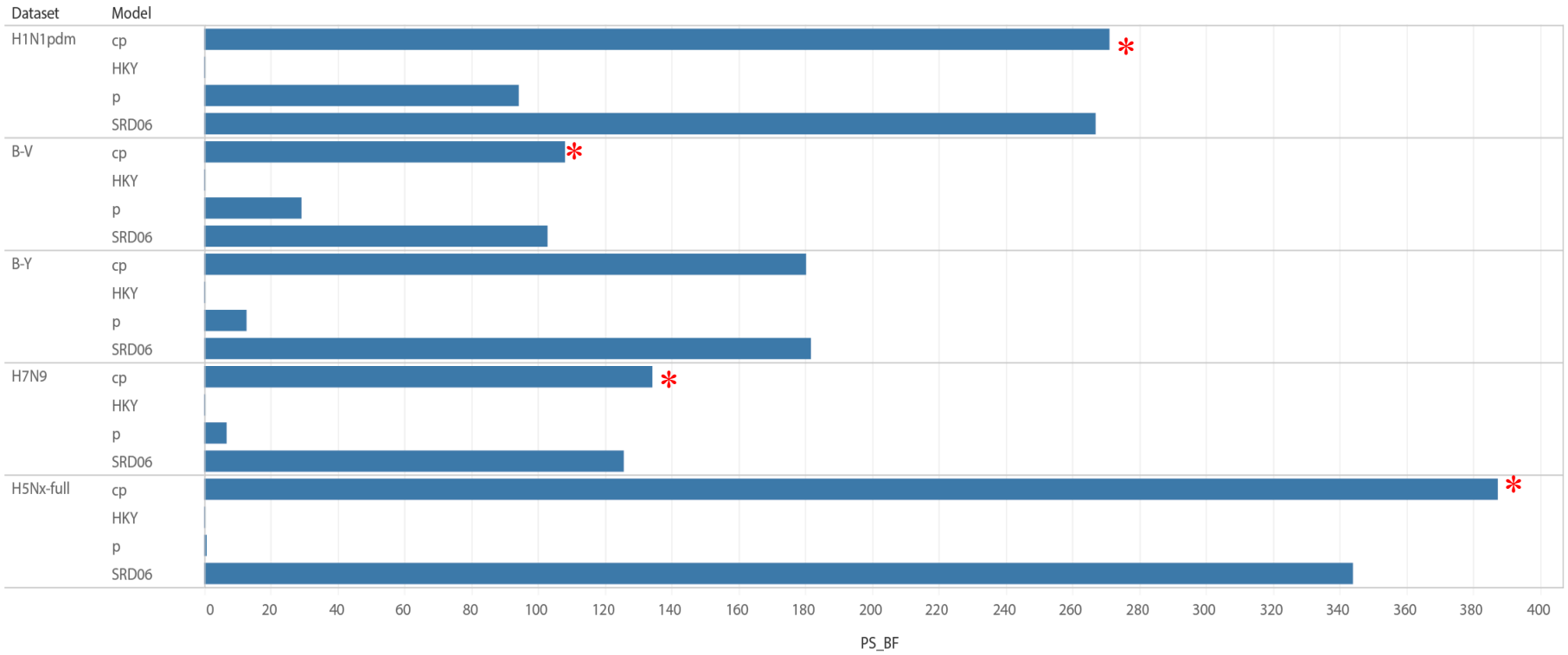
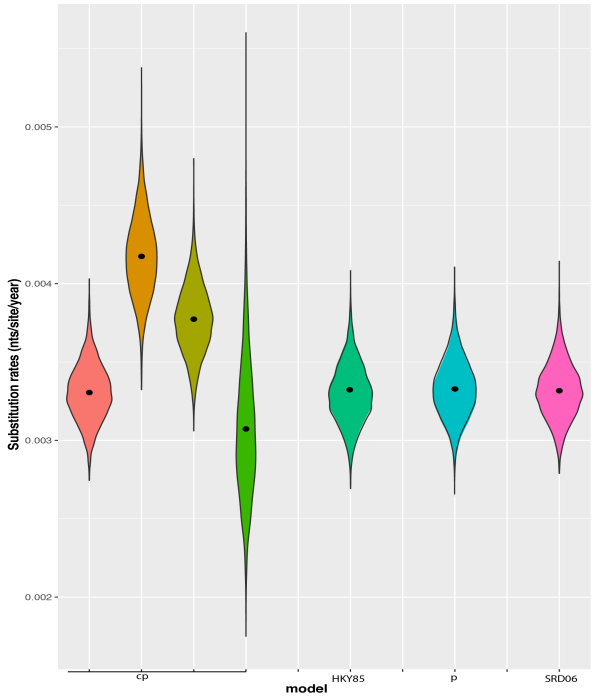
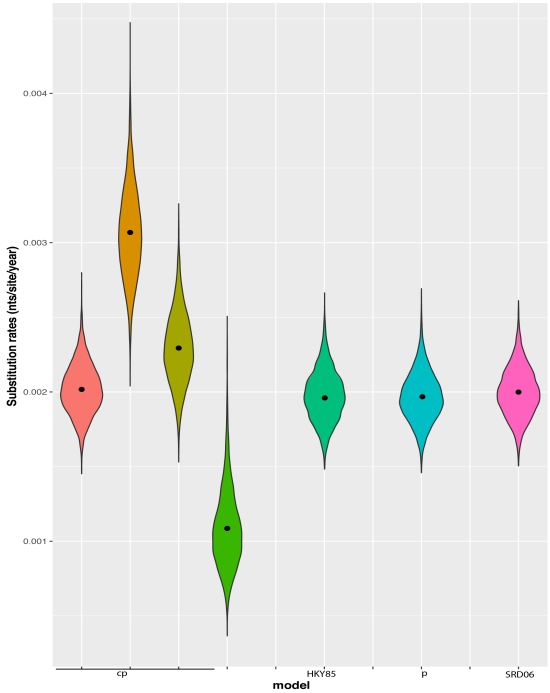


Figure 2. Overall mean substitution rates estimated from four models and domain-specific substitution rates from structurally informed partitioning model. The results showed overall mean substitution rates are similar, but head domain has a faster rate than that of stalk domain.

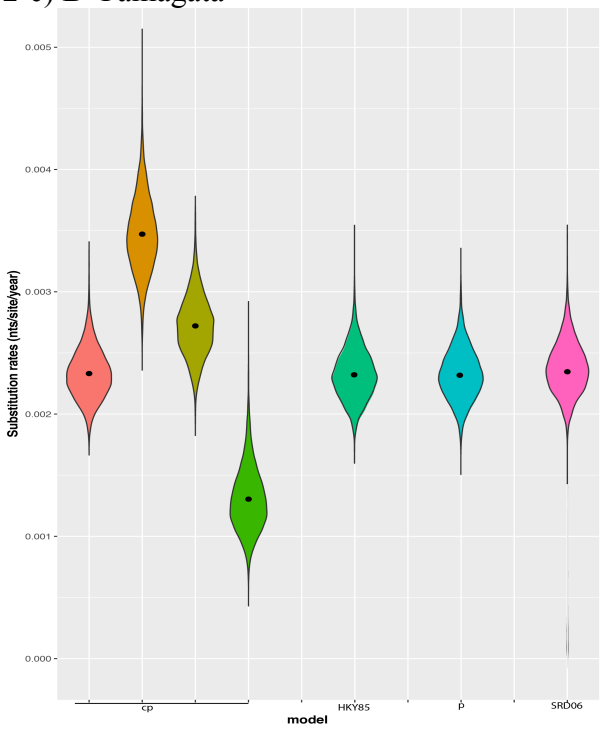
2-a) H1N1/09



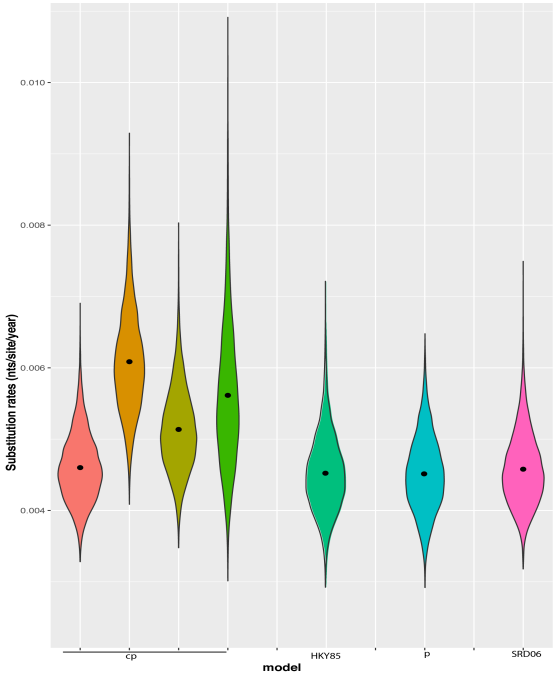
2-b) B-Victoria



2-c) B-Yamagata



2-d) H7N9



2-e) H5Nx-full

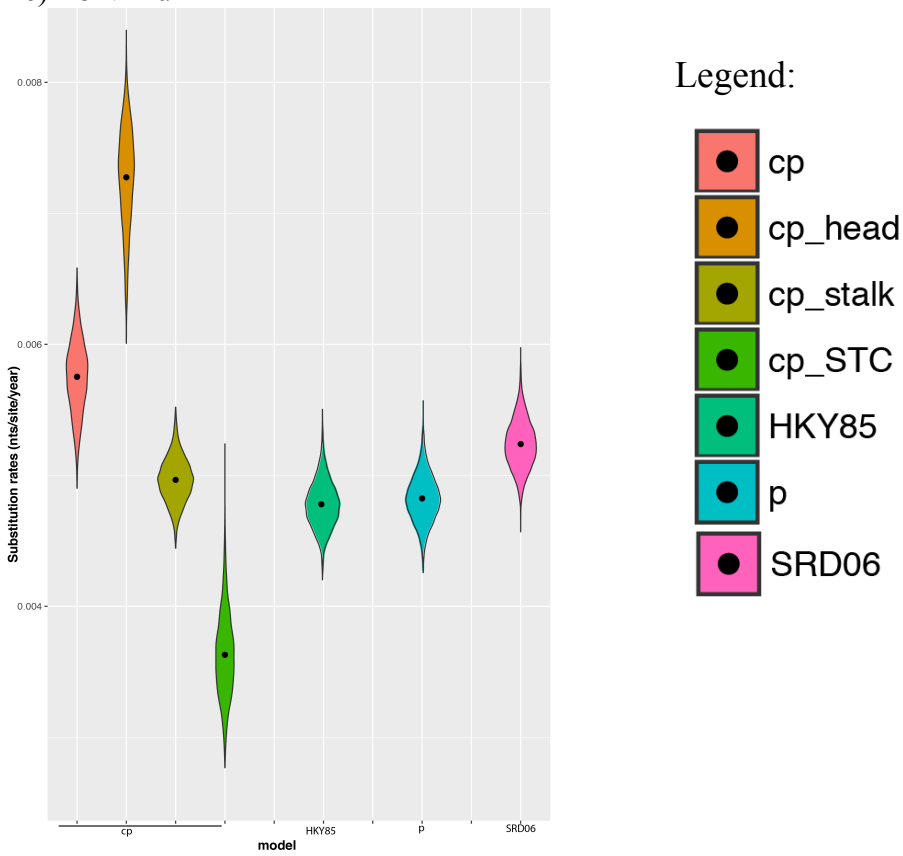

0.008
0.006
0.004
Substitution rates (nts/site/year)
cp
HKY85
p
SRD06
model
Legend:
cp
cp_head
cp_stalk
cp_STC
HKY85
p
SRD06

Figure 3. Log scale Bayes Factors for sensitivity analysis on three HPAI H5Nx dataset. All these BFs were calculated to compare with HKY. The asterisk * means a significant supportive BF for the new structurally informed model, compared to SRD06 codon only model. In the figure, p means the HKY85 plus structurally informed partitioning model; cp represents SRD06 codon plus structurally informed partitioning model.

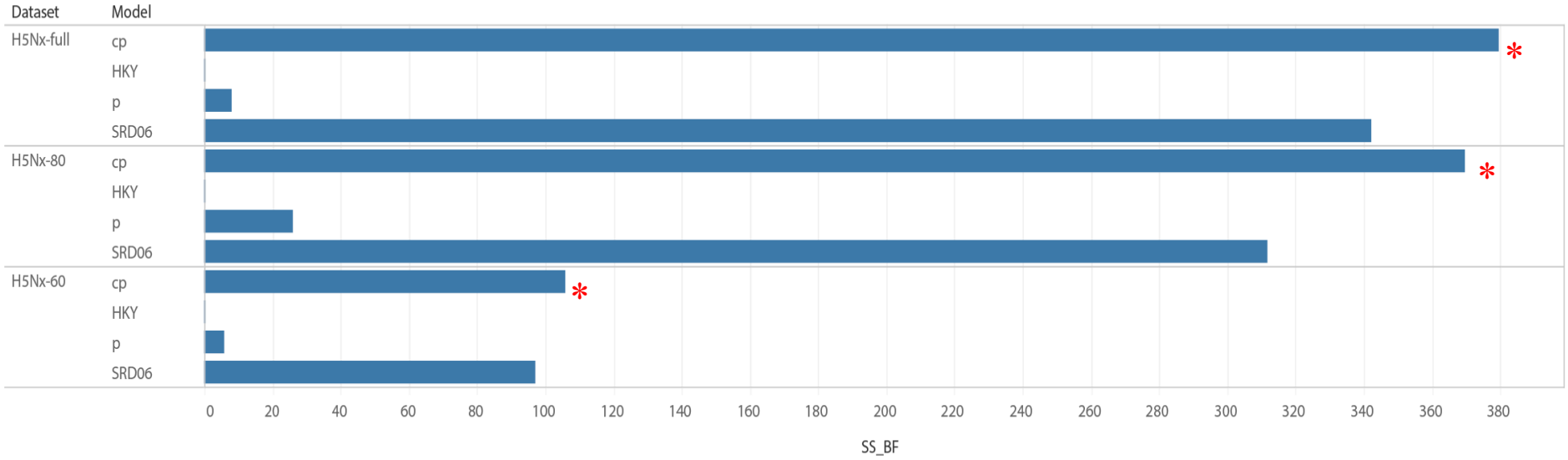
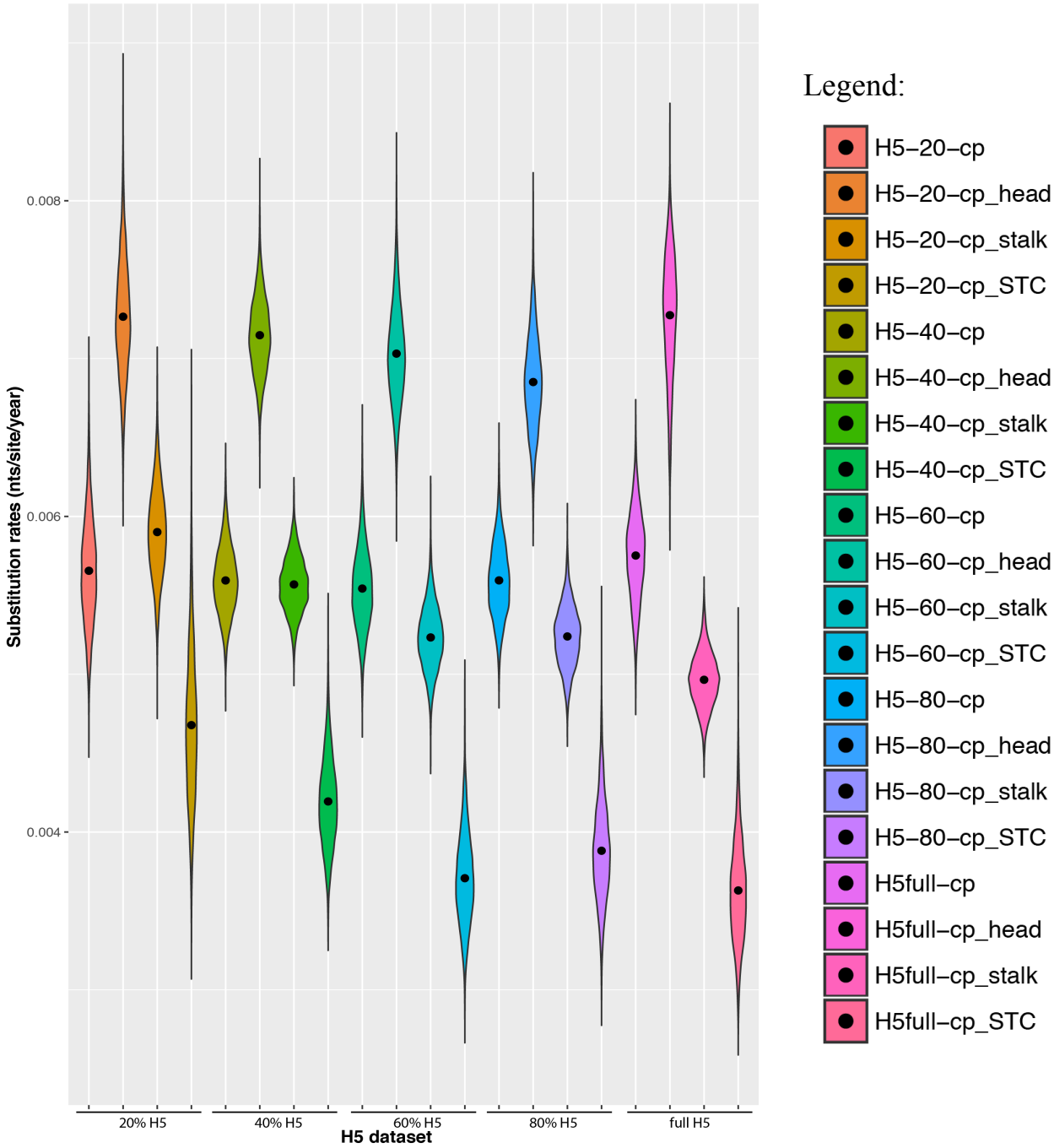


Figure 4. Overall mean substitution rates and domain-specific substitution rates from structurally informed partitioning model for five HPAI H5Nx dataset. The results showed 20% of H5Nx data has an more inaccurate estimation on substitution rates. cp represents SRD06 codon plus structurally informed partitioning model



Supplemental Materials

Supplemental Table 1. The nucleotide number of functional partitions for each influenza subtype included in this study

Influenza subtypes	STC domain	Stalk domain	Head domain
H1N1/2009	1-51, 1591-1701	52-216, 874-1590	217-873
B-Victoria	1-45, 1645-1758	46-171, 922-1644	172-921
B-Yamagata	1-45, 1645-1755	46-171, 922-1644	172-921
H7N9	1-54, 1573-1683	55-216, 856-1572	217-855
HPAI H5Nx	1-48, 1597-1713	49-174, 868-1596	175-867

***note: the positions are corresponding to the aligned HA nucleotide sequences for each influenza subtype.**