SUPPLEMENTARY INFORMATION

Supplementary Table 1. HAPMs and HP clusters inferred to be evolving under positive Selection.

Supplementary Figure 1. Large-sale phylogenetic trees for the HA genome segments of the H7NX and H5NX viruses.

Supplementary Figure 2. Trees for the PB2, PB1, PA and NP viral genome segments for both virus subtypes.

Supplementary Figure 3. Trees for the M1 and NS1 viral genome segments for both virus subtypes.

Supplementary Figure 4. Trees for the NA genome segments (N1-N9 subtypes) of H7NX and H5NX viruses.

Supplementary Figure 5. Full time-scaled MCC tree for H5 with the ancestral reconstructions for site 154.

Supplementary Figure 6. Full time-scaled MCC tree for H5 PB2 with the ancestral reconstructions for site 627.

Supplementary Figure 7. Full time-scaled MCC tree for H7 with the ancestral reconstructions for site 143.

Supplementary Figure 8. Full time-scaled MCC tree for H7 PB2 with the ancestral reconstructions for site 335.

Supplementary Text 5. Full mathematical description of the model and

simulations.

Supplementary Table 1. HAPMs and HP clusters inferred to be evolving under positive selection

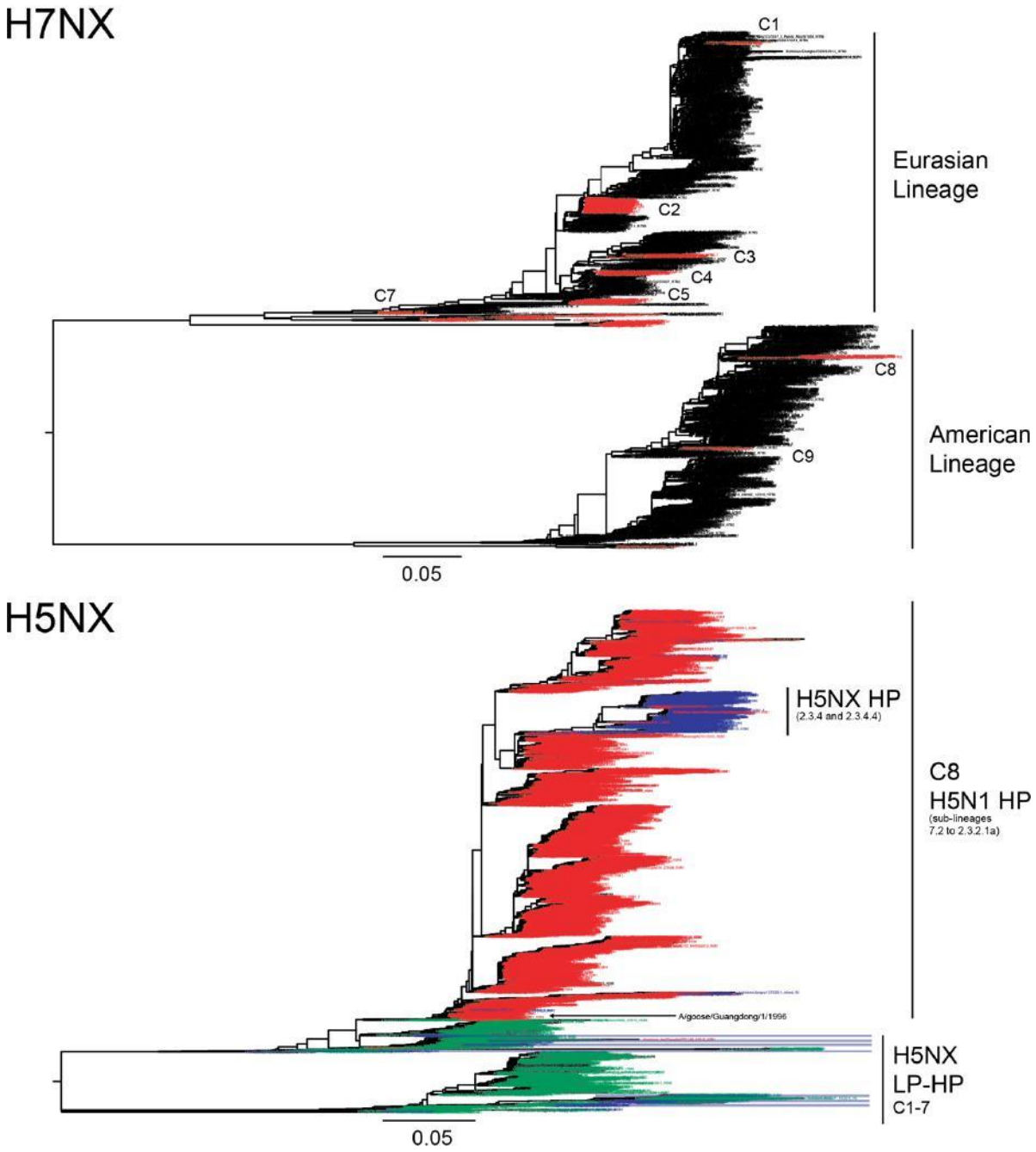
H7NX	Site models (pervasive selection)				Site models (episodic selection)		Branch-site models (episodic selection)						Directional Selection (DEPS)
Gene*	SLAC ≤ 0.05	SITE*	FEL ≤ 0.05	SITE*	MEME ≤ 0.05	SITE*	BSA ≤ 0.05 ‡	CLUSTER	SITE*§	bsREL ≤ 0.05	CLUSTER	SITE*	DEPS: Mutation/EBF value‖
PB2	NO	–	NO	--	YE	--	NO	C6,	**35**	NO	--	--	--
PB1	NO	–	YES	**152**	S	--	NO	C7	**5**	NO	--	--	--
PA	NO	–	NO	--	YE	--	NO	C9 -	**15**	NO	--	--	--
H7	NO	–	YES	**143**	S	--	YES	- C4	**2** --	YE	C4, C7	--	**143T/1139**
NP	NO	–	NO	--	NO	--	NO	--	-- -	S	--	--	--
M	NO	--	NO	--	YE	--	NO	--	-	NO	--	--	--
NS	NO	--	YES	**180,** 209	S YE	**180,** 209	NO	C8	--	NO	--	--	209N/10E5
H5NX	SLAC ≤ 0.05	SITE*	FEL ≤ 0.05	SITE*	MEME ≤ 0.05	SITE*	BSA ≤ 0.05 ‡	CLUSTER	SITE *§	bsREL ≤ 0.05	CLUSTER	SITE*	DEPS: Mutation/EBF value‖
PB2	NO	--	NO	--	YES	--	NO	--	--	YES	C1	--	--
PB1	NO	--	NO	--	NO	--	NO	--	--	NO	--	--	--
PA	NO	--	NO	--	NO	--	NO	--	--	YES	C4	--	--
H5	YES	--	YES	**154**	YES	**127, 154**	NO	--	--	NO	--	--	**S145L/9814**, **N154I/101**, **S157P/>105**, **I167T/>105**
NP	NO	--	NO	--	NO	--	NO	--	--	NO	--	--	--
M	NO	--	NO	--	YES	--	NO	--	--	NO	--	--	--
NS	YES	--	YES	213	YES	213	YES	C2	--	NO	--	--	**P213S/>105**

* Genes/sites shown in this table correspond only to those where candidate HAPMs were detected by reconstruction of ancestral states. Highlighted in black are the HP-positively associated HAPMs.

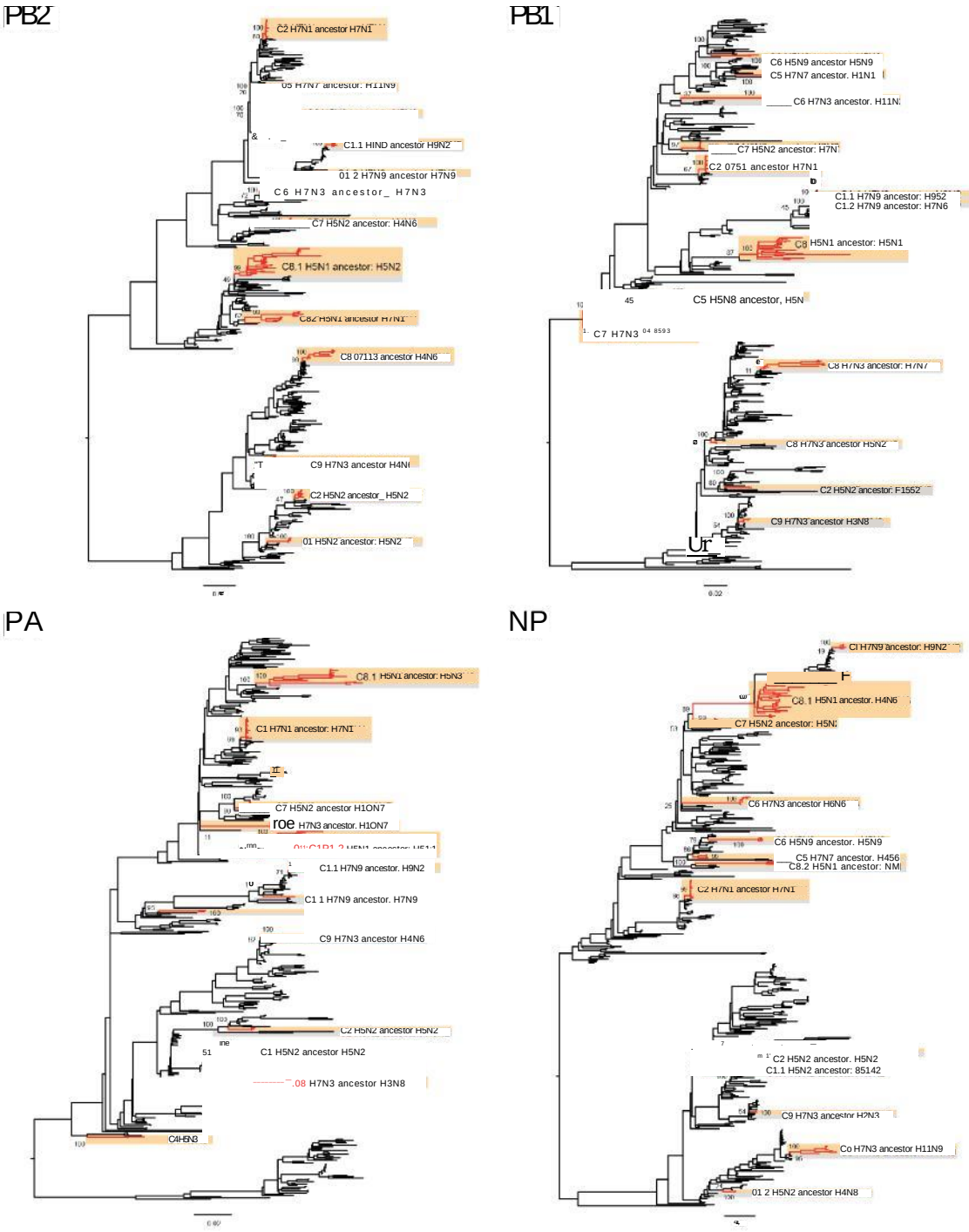
‡ Although the significance for the BSA test can be negative due to very small global proportion of PSS, sites can still be significantly scored under BEB.

§ PSS scored under BEB (PP ≥ 90%).

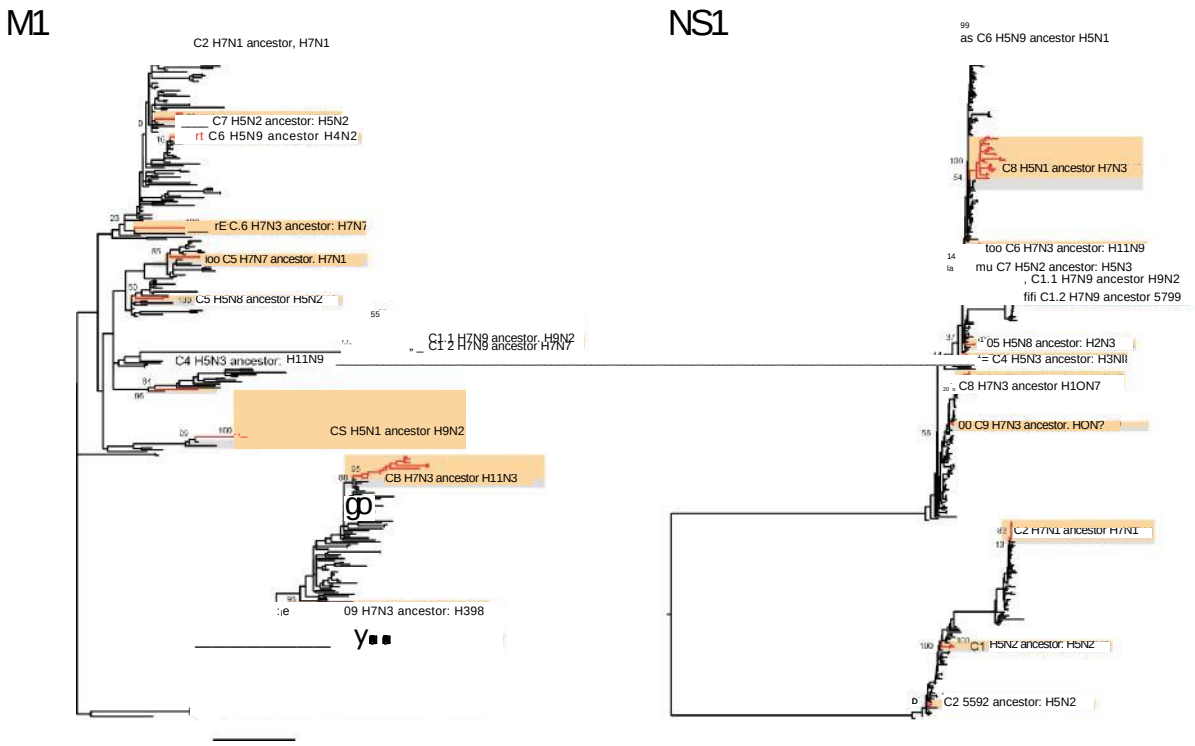
‖ Empirical Bayes Factor (EBF) for convergent evolution using Directional Evolution of Protein Sequences (DEPS).



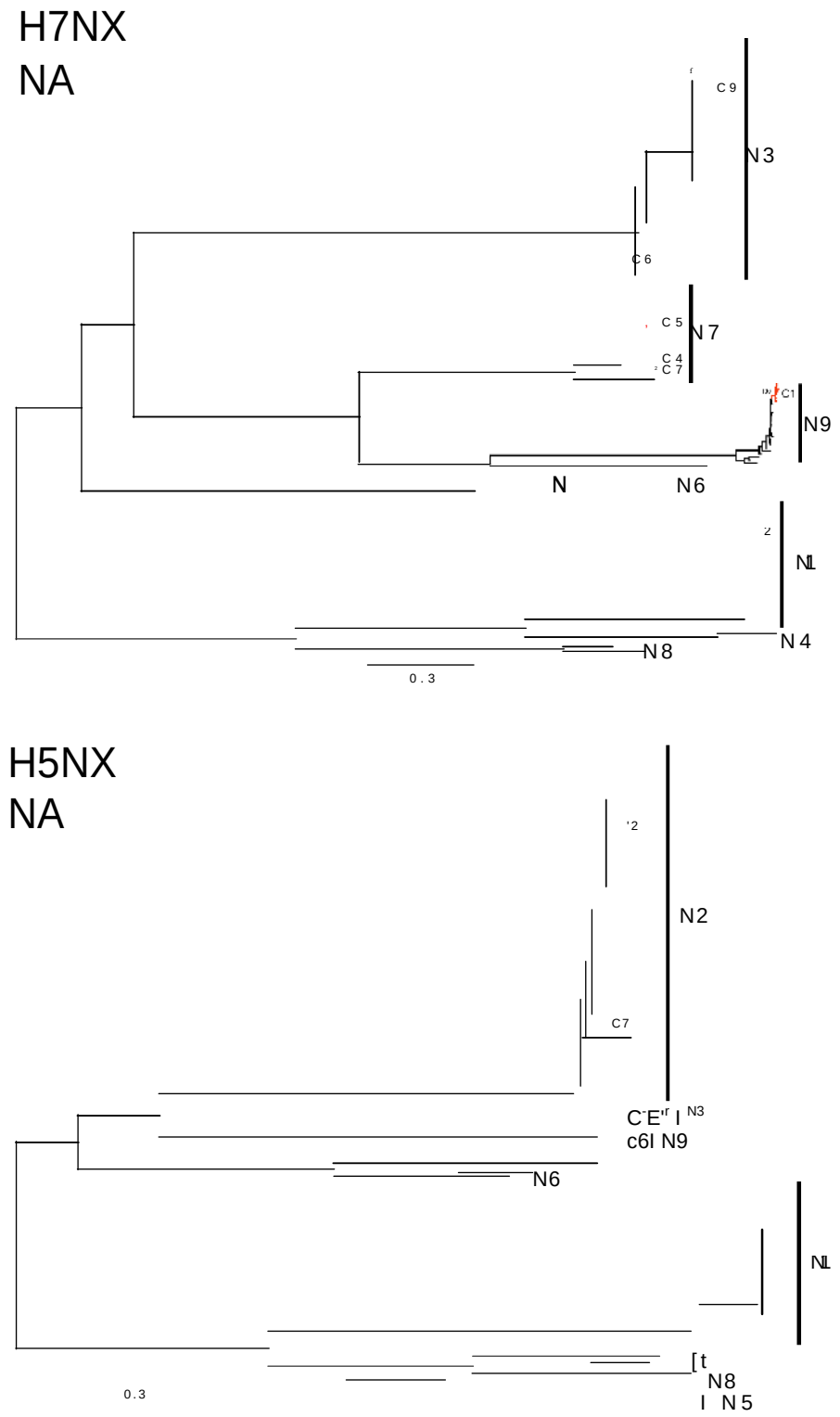
Supplementary Figure 1. Large-sale phylogenetic trees for HA genome segments of H7NX and H5NX viruses



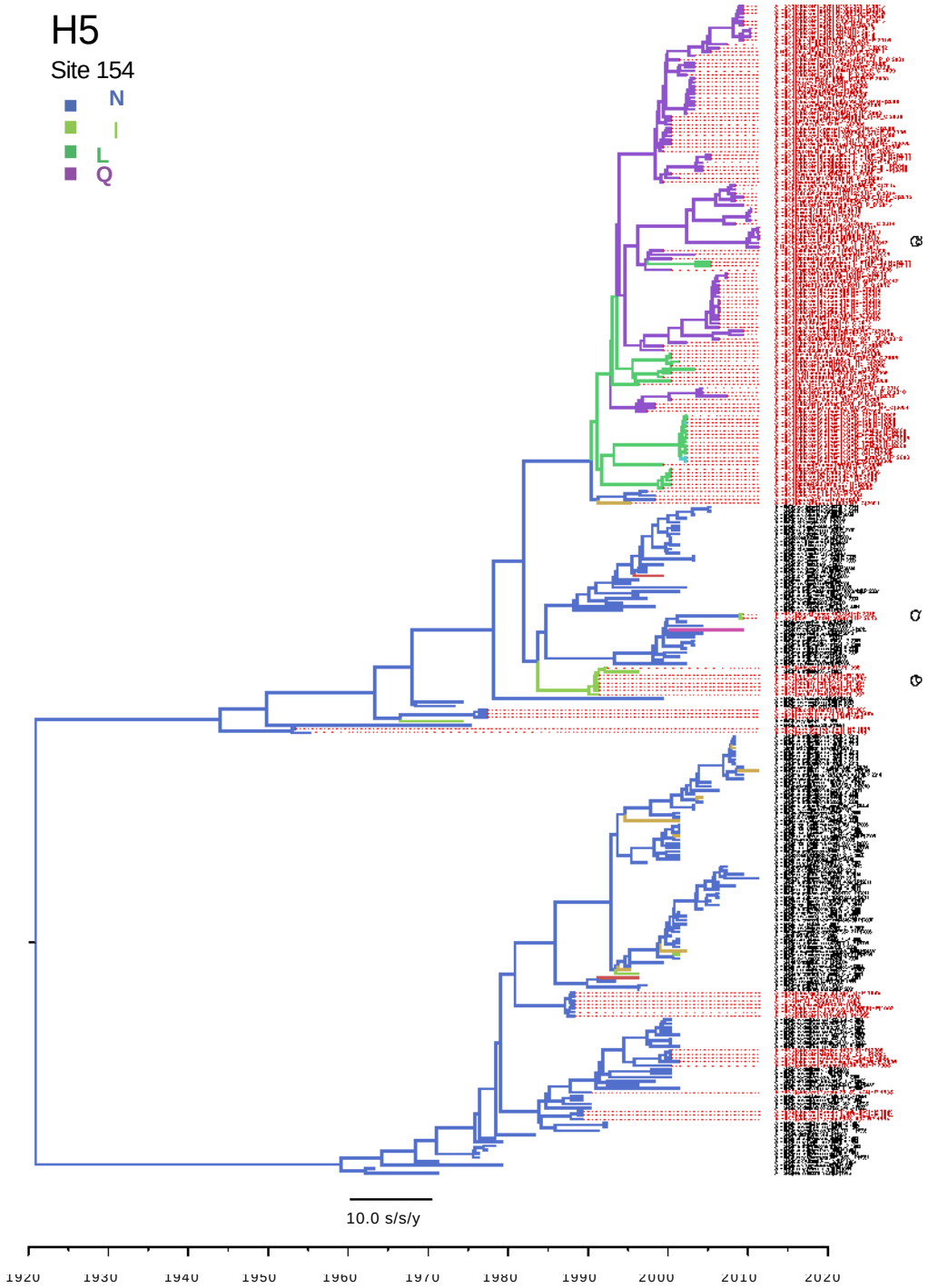
Supplementary Figure 2. Maximum likelihood trees for the internal genome segments of H7NX viruses



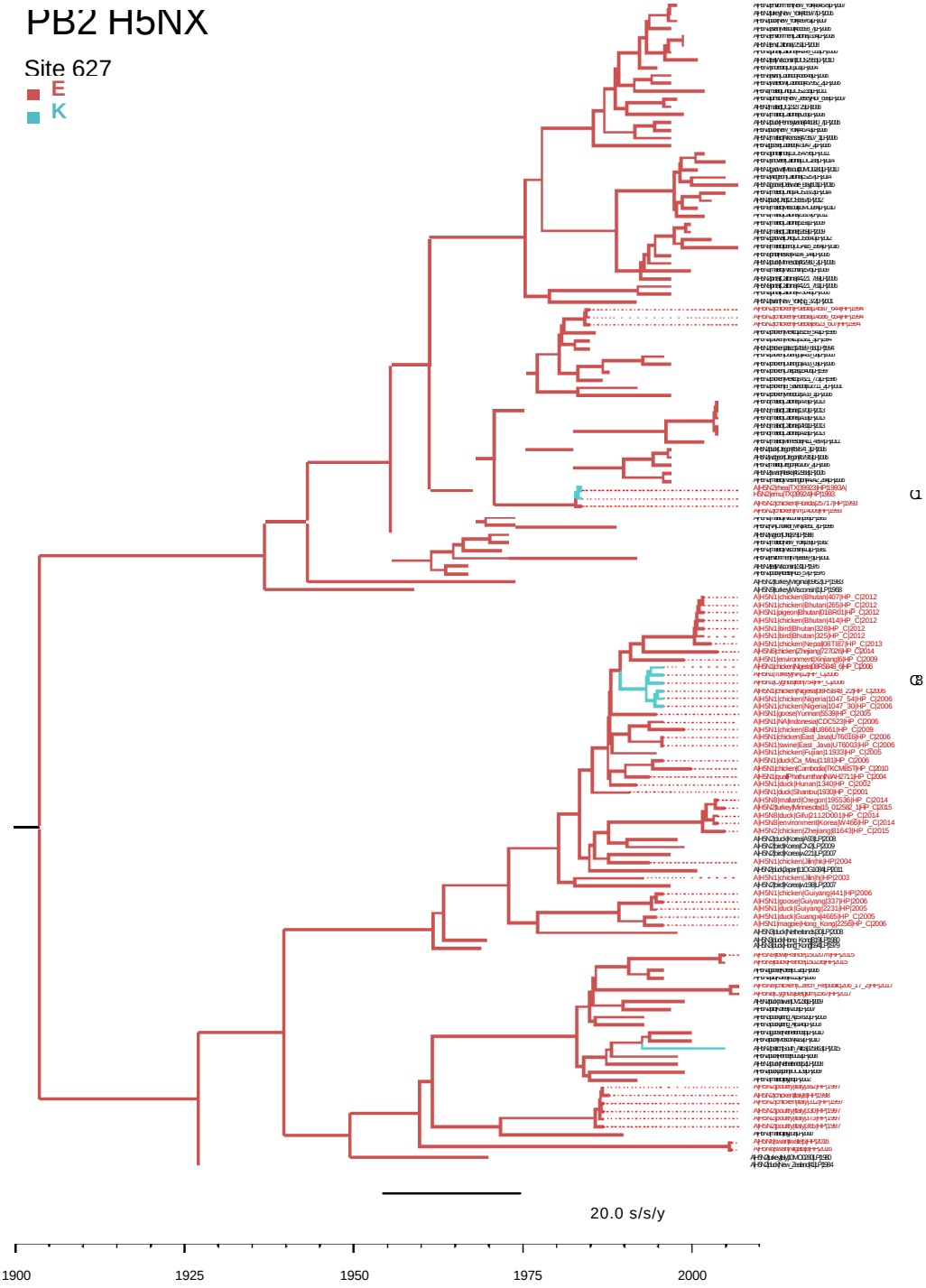
Supplementary Figure 3. Maximum likelihood trees for the internal genome segments of H7NX viruses



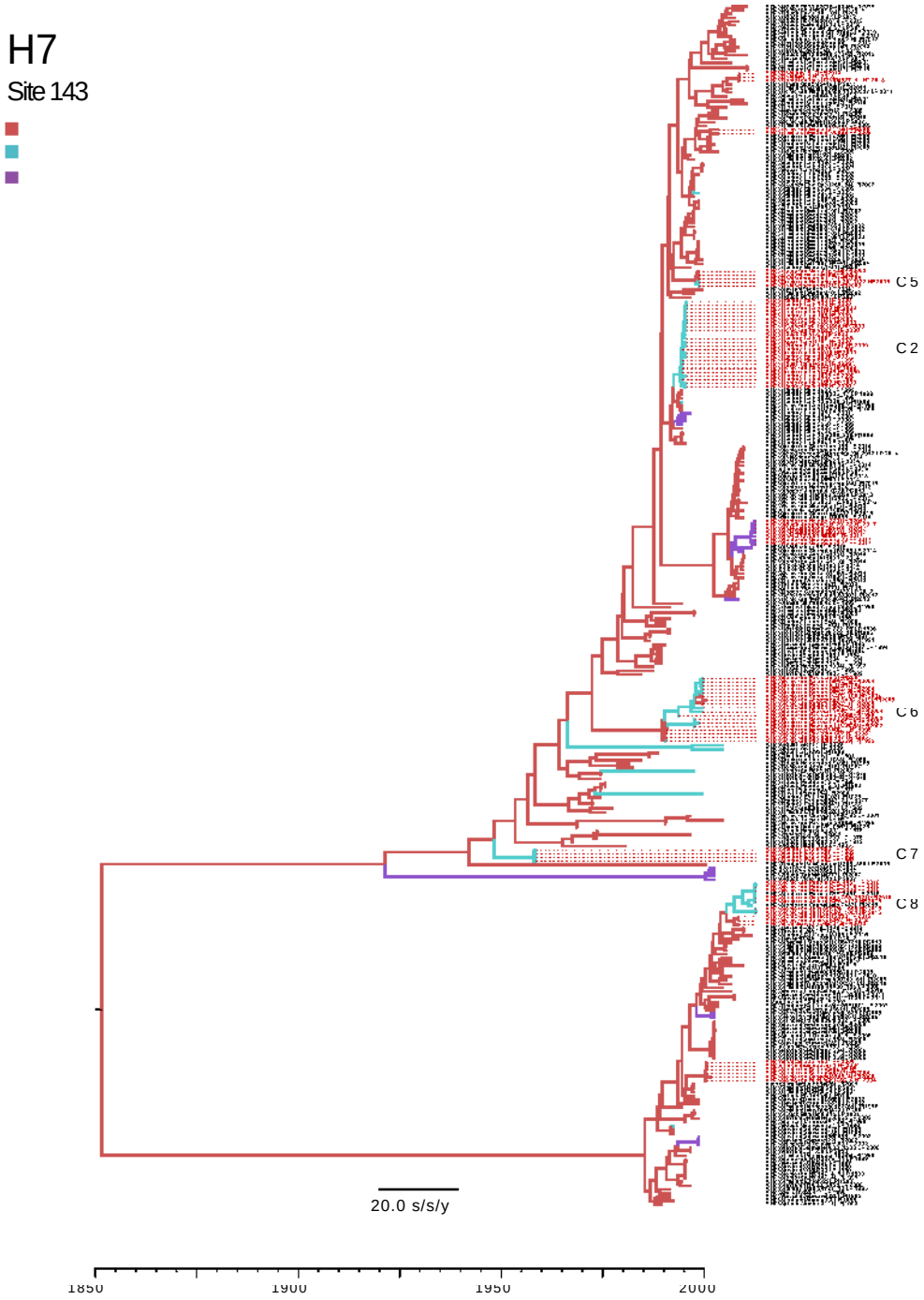
Supplementary Figure 4. Maximum likelihood trees for the NA genome segments (N1-N9 subtypes) of H7NX and H5NX viruses



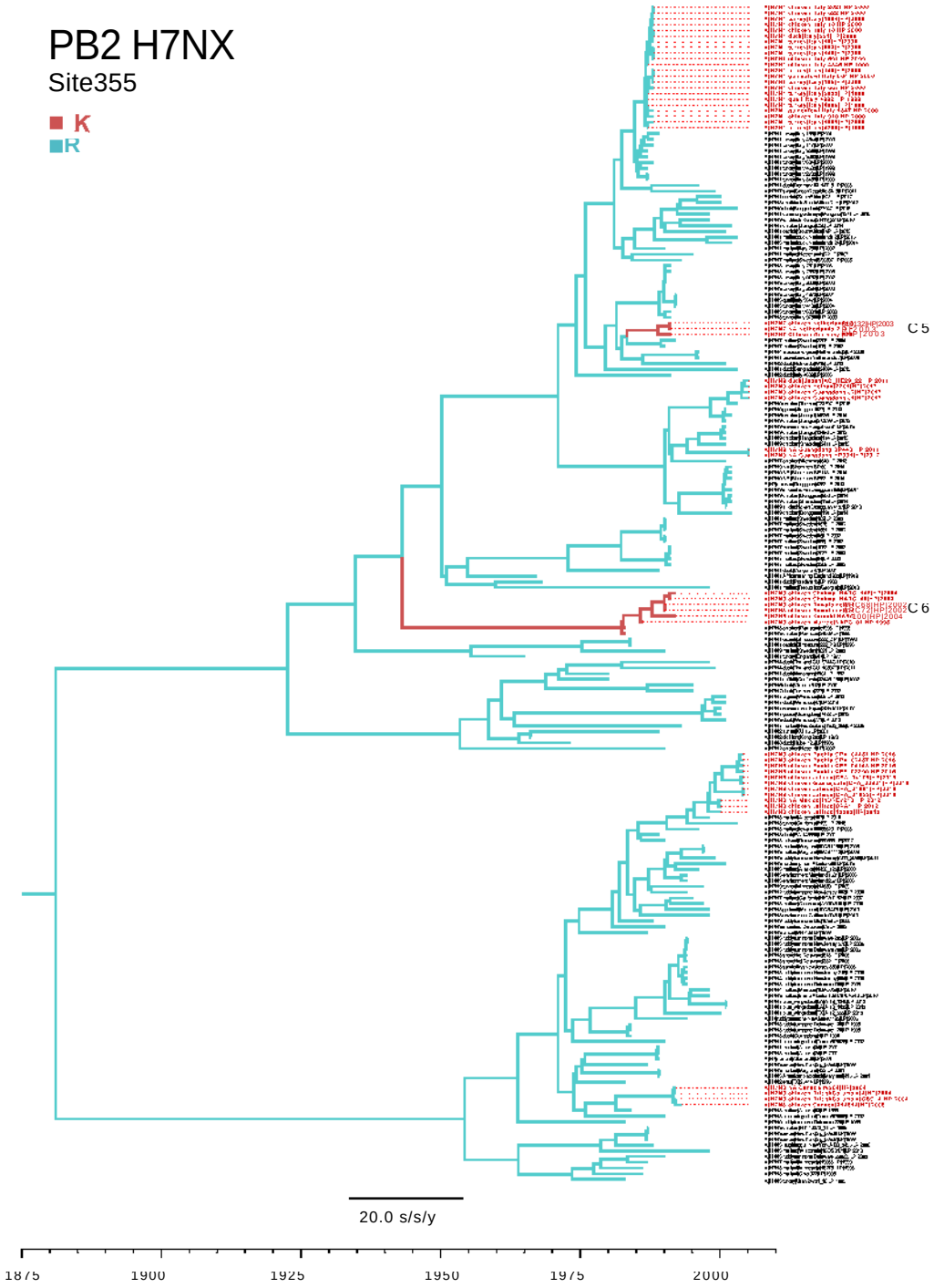
Supplementary Figure 5. Full time-scaled MCC tree for H5 with the ancestral reconstructions for site 154.



Supplementary Figure 6. Full time-scaled MCC tree for H5 PB2 with the ancestral reconstructions for site 627.



Supplementary Figure 7. Full time-scaled MCC tree for H7 with the ancestral reconstructions for site 143.



Supplementary Figure 8. Full time-scaled MCC tree for H7 PB2 with the ancestral reconstructions for site 335.

Amino Acid Trait Association Model

Model description

We developed a pair of nested evolutionary models, a null and an alternative model, to test for associations between a target amino acid at an alignment site and a target binary trait. The target amino acid being tested, denoted targetaa, represents one of the 20 possible amino acids. The target trait, denoted T, represents the trait being tested for an association with the target amino acid. The non-target trait is denoted N. The traits, like the amino acids, are assumed to have evolved in an evolutionary manner.

The null model treats the amino acid evolution at a site and the trait evolution as independent of one another, whereas the alternative model treats the target amino acid and the target binary traits as potentially associated. This potential association is introduced into the alternative model via a dependence parameter λ.

For a given amino acid site and set of traits we used maximum likelihood estimation to estimate the parameters of both models and to obtain the maximum likelihood values. The maximum likelihood values were used to compare both models using a likelihood ratio test (LRT) and to calculate a p-value. If the LRT rejects the null model ($p < 0.05$) in favour of the alternative model, this suggests that the target amino acid and target trait are associated.

This association can be a positive association: the target amino acid and the target trait tend to co-occur together, or a negative association: the target amino acid and the target trait tend to actively avoid co-occurring together. When the maximum likelihood estimate for the dependence parameter is larger than one, $\lambda\hat{} > 1$, this suggests a positive association, and when $\lambda\hat{} < 1$, this suggests a negative association.

This test is somewhat analogous to a chi-squared test of association, except it accounts for phylogenetic correlations. A chi-squared test will treat each observation of amino acid and trait at the tips of a phylogeny as independent events, when in reality they are produced by an evolutionary process where the underlying number of events leading to those observations may be small. This is commonly referred to as a founder effect (Bhattacharya *et al.*, 2007), and can result in chi-squared associations whose significance is erroneously inflated. Our test avoids this by explicitly modelling the potential dependence between the amino acid and trait evolutionary processes and testing its significance relative to a model that treats them as independent of one another.

Note that the traits, like the amino acids, are assumed to evolve in an evolutionary manner along the tree, and therefore our test is only appropriate where the trait can be described by an evolutionary process. This is the case for the HP and LP traits because they are a direct function of the presence or absence of an insertion, which is generated by an insertion-deletion evolutionary process along the tree. This test would not be appropriate for a trait such as patient survival, which represents a propensity along the tree rather than propagating in a discrete manner. For traits such as these we recommend the test outlined in Bhattacharya *et al.* (2007).

A formal description of the model is given as follows: the joint evolution of amino acids and traits are modelled using a 40×40 substitution model *Q*, that combines a 20×20 amino substitution model, *A*, and 2×2 by trait model, T. The trait model is a two-state continuous-time Markov model akin to Felsenstein's 1981 DNA substitution model (Felsenstein, 1981). The joint model is given as follows:

$$Q_{ij,mn} = \begin{cases} /1 A_{ij}\lambda & \text{if } i = j \text{ and } j = \text{target aa and } m = n = \\ \mathrm{T} & /1 A_{ij}{}^{1}{}_{\lambda} \text{ if } i = j \text{ and } j = \text{target aa and } m = n = \mathrm{T} \\ & /1 A_{ij} \text{ if } i = j \text{ and } m = n = \mathrm{N} \\ \pi \mathrm{T} \tau\lambda & \text{if } i = j = \text{target aa and } m = \mathrm{N} \text{ and } n = \\ \mathrm{T} & \pi \mathrm{T} \tau^{1}{}_{\lambda} \text{ if } i = j \text{ and } j = \text{target aa and } m = \mathrm{N} \text{ and } n = \\ \mathrm{T} & \pi \mathrm{N} \tau \text{ if } i = j \text{ and } m = \mathrm{T} \text{ and } n = \mathrm{N} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where *i* and *j* represent the initial and end amino acid states, respectively, and *m* and *n* represent the start and end trait states, respectively. This formulation is motivated by the RNA base-pairing model of Muse (1995). Our model shares similarities with other discrete models such as BayesTraits (Pagel, 1994; Pagel and Meade, 2006), however, these models typically only consider two binary traits, whereas our model considers a binary trait and a 20 state amino acid. Furthermore, by following the approach of Muse (1995) our model captures the potential dependence between the trait and the target amino acid in a single parameter rather than requiring a large matrix of free parameters that considers each possible transition separately. This approach simplifies hypothesis testing and implies we can perform maximum likelihood inference as opposed to more computationally time intensive Bayesian inference.

$/1$ is a site-specific amino acid substitution rate, and τ is the trait substitution rate. A is a rate matrix given by the LG2008 substitution model. π^{A} is a vector of 20 amino acid frequencies as specified by the LG2008 model, and $\pi\mathrm{T}$ and $\pi\mathrm{N}$ represents the frequencies of the target (T) and non-target traits (N) states, respectively.

The equilibrium frequencies, π, of the alternative model are given by four separate cases corresponding to the two possible values for traits (T or N), and whether the amino acid

(aa) matches the target amino acid (targetaa) or not:

$$\pi_{aa=targetaa,trait=T} = \kappa^{-1}\pi_{A_{aa}}\pi_T\lambda \quad (2)$$

$$\pi_{aa\neq targetaa,trait=T} = \kappa^{-1}\pi_{A_{aa}}\pi_T \frac{1}{\lambda} \quad (3)$$

$$\pi_{aa=targetaa,trait=N} = \kappa^{-1}\pi_{A_{aa}}\pi_N \quad (4)$$

$$\pi_{aa\neq targetaa,trait=N} = \kappa^{-1}\pi_{A_{aa}}\pi_N \quad (5)$$

Where $\kappa = (\lambda + \frac{1}{\lambda})\pi_T + 2\pi_N$ is a normalising constant.

These equilibrium frequencies provide a intuitive way of understanding the influence of the association parameter λ. It is possible to get a sense of the expected frequencies of particular amino acid and trait associations for given values of λ. Furthermore, they can be used to predict, for a single sequence, the posterior probability of a trait given the amino aa at the target site:

$$p(trait = T \mid aa = targetaa) = \frac{\pi_T \lambda}{\pi_N + \pi_T \lambda} \quad (6)$$

$$p(trait = T \mid aa \neq targetaa) = \frac{\pi_T}{\pi_N \lambda + \pi_T} \quad (7)$$

Also note that the model is time-reversible, and therefore an unrooted tree can be used if the equilibrium probabilities are taken to be the initial probabilities at any rooting of the tree (Felsenstein, 1981).

Simulations

The Influenza H7 ML tree and corresponding HP and LP traits were taken and amino acid alignments were simulated. Note that amino acid sequences were simulated conditioned on the patterns of HP and LP traits observed in the real data. This was done to account for sampling bias. Each alignment consisted of 500 amino acids, with the first twenty sites of each alignment simulated as being associated with the traits, each having a different one of the 20 canonical amino acids as the target amino acid. The remaining 480 amino acid sites were simulated under the LG2008 model and were therefore treated as being independent of the traits. Three different degrees of association were simulated: weak, intermediate, and high, combined with three different rates of trait evolution (2.0, 4.0, and 7.0) - the inferred rate of trait evolution in the H7 ML tree was <4.0 and so this was selected as an intermediate value.

Table 1: Summary of benchmarks results

Simulated association strength	Simulated rate of trait evolution	Recall	Precision
Weak (2.0)	2.0	0.23	0.94
Intermediate (4.0)	2.0	0.35	0.96
Strong (8.0)	2.0	0.50	0.93
Weak (2.0)	4.0	0.26	0.92
Intermediate (4.0)	4.0	0.48	0.92
Strong (8.0)	4.0	0.64	0.96
Weak (2.0)	7.5	0.28	0.97
Intermediate (4.0)	7.5	0.62	0.98
Strong (8.0)	7.5	0.73	0.95

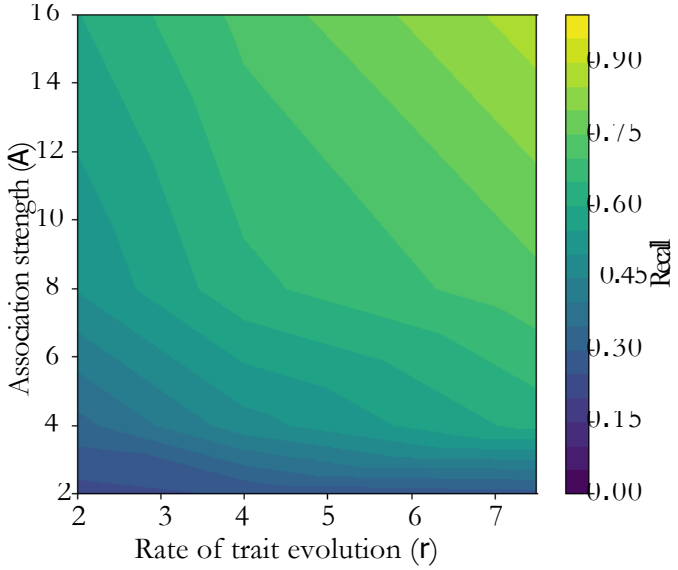


Figure 1: Contour plot of recall (blue-green-yellow colour gradient) as a function of simulated rate of trait evolution (x-axis) and simulated association strength (y-axis).

To account for the potential error introduced during tree inference, an ML tree was inferred using FastTree (Price *et al.*, 2010) for each of the simulated alignments. Potential associations were then estimated using our model and the Bhattacharya method (Bhattacharya *et al.*, 2007) on the first 40 sites of each alignment. The first twenty sites were used to measure the number of true positive and false negative detections, whereas the remaining twenty sites were simulated as independent of the traits and were used to measure the number of true negative and false positive detections. The recall and precision were calculated for each simulation using the number of true positives (TN), false-positives (FP),

and false-negatives (FN). Recall and precision are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

The results in Table 1 indicate that our model has a false-discovery rate (FDR= 100%x[1.0 Precision]) of < 5% across all test conditions which is consistent with our p-value significance threshold of 0.05. The recall of the model increases with the simulated association strength as expected, with stronger associations being more easily detected. Likewise, the recall of the model increases with a higher rate of trait evolution (Table 1 and Figure 1), which is also expected given that higher rates of trait evolution imply a greater number of trait events along the tree and therefore more the power in being able to detect an association.

Software availability

https://github.com/michaelgoldendev/trait-evolution

References

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., *et al.* 2007. Founder effects in the assessment of hiv polymorphisms and hla allele associations. *Science*, 315(5818): 1583–1586.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6): 368–376.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139(3): 1429–1439.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342): 37–45.

Pagel, M. and Meade, A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6): 808–825.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3): e9490.