**Automatic detection of rate change in large data sets with
an unsupervised approach: the case of influenza viruses**

Journal:	<i>Genome</i>
Manuscript ID	gen-2015-0163.R1
Manuscript Type:	Article
Date Submitted by the Author:	21-Jan-2016
Complete List of Authors:	Labont��, Kasandra; University of Ottawa, Biology Aris-Brosou, Stephane; University of Ottawa
Keyword:	Influenza, hemagglutinin, neuraminidase, machine learning

Automatic detection of rate change in large data sets with an unsupervised approach: the case of influenza viruses

Kassandra Labonté,^a Stéphane Aris-Brosou^{ab}

^aDepartment of Biology, Centre for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON K1N 6N5, Canada

^bDepartment of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Stéphane Aris-Brosou (e-mail: sarisbro@uottawa.ca).

ABSTRACT

Influenza viruses evolve at such a high rate that vaccine recommendations need to be changed, but not quite on a regular basis. This observation suggests that the rate of evolution of these viruses is not constant through time, which begs the question as to when such rate changes occur, if they do so independently of the host in which they circulate and/or independently of their subtype. To address these outstanding questions, we introduce a novel heuristics, Mclust*, that is based on a two-tier clustering approach in a phylogenetic context to estimate (i) absolute rates of evolution and (ii) when rate change occurs. We employ the novel approach to compare the two influenza surface proteins, hemagglutinin and neuraminidase, that circulated in avian, human and swine hosts between 1960-2014 in two subtypes: H3N2 and H1N1. We show that the algorithm performs well in most conditions, accounting for phylogenetic uncertainty by means of bootstrapping and scales up to analyze very large data sets. Our results show that our approach is robust to the time-dependent artifact of rate estimation, and confirm pervasive punctuated evolution across hosts and subtypes. As such, the novel approach can potentially detect when vaccine composition needs to be updated.

Keywords: Influenza, hemagglutinin, neuraminidase, machine learning

Introduction

Zoonotic infectious diseases such as influenza evolve continuously to evade their host's immune system. As such, they lead to losses in revenue in the farming industry and of human lives, so that it is of paramount importance to understand the pace at which these viruses change through time. Recently, it was shown that influenza viruses circulating in different hosts do evolve at different rates (Worobey et al., 2014; Duchêne et al., 2014), potentially exhibiting a time-dependent artifact where more recent rate estimates appear larger (Duchêne et al., 2014). However, it is also clear that within a given host, rates of evolution can change through time: in humans for instance, antigenicity of H3N2 influenza A viruses tends to change every 2-5 years (Plotkin et al., 2002). These stochastic rate variations may complicate recommendations on vaccine composition, that in any case needs to be changed in an unpredictable manner. As such, a method that could automatically tell us when rates of evolution do change could help us improve the timing of vaccine composition updates.

Sophisticated models that incorporate rate change do exist and have been employed to make specific inference about the evolutionary process. For instance, when analyzing the viral sequences in a HIV-infected patient who is moved to antiretroviral therapy, it becomes possible to fit a time-heterogeneous model and estimate parameters relevant to changes in selective pressure following the onset of treatment (Goode et al., 2008). This approach, based on a maximum likelihood model, was recently implemented in a Bayesian framework, thereby incorporating phylogenetic uncertainty, and extended to allow evolutionary parameters to change across any discrete data types such as hosts or countries (Bielejec et al., 2014). This latter model assumes that, at any time point, all the sequences share the same evolutionary process, but this potential limitation can be relaxed

under structured models (Worobey et al., 2014; Vaughan et al., 2014). Yet, in spite of these extreme sophistications, all current models require or assume prior knowledge about the position of these breakpoints in time, when the pace of the evolutionary process changes. For instance, in the case of influenza viruses, we typically do not know *a priori* when these rate changes occur – or even if they do occur within a particular viral gene of a particular subtype circulating in a given host, even if strong evidence exists supporting the punctuated pace of evolution of these viruses (Bedford et al., 2014).

Here we describe a novel heuristic based on unsupervised machine learning to automatically detect rate change. We set the approach in a phylogenetic context, to specifically detect rate change in sequentially sampled viral sequences, at a fraction of the computational cost required by a full probabilistic approach. We compare the novel approach with Bayesian inference to demonstrate its accuracy, and evaluate the impact of phylogenetic uncertainty by incorporating bootstrap resampling. We finally apply the method to analyze large data sets (upwards of 7,600 aligned sequences and a grand total of 28,668 sequences). We show (i) that the new method is robust to the time-dependent artifact and (ii) that there exists an extreme rate variation through time in the two antigenic determinants of influenza viruses (hemagglutinin and neuraminidase).

Materials and Methods

Full-length coding sequences for hemagglutinin (HA) and neuraminidase (NA) were downloaded from the Influenza Virus Resource (Bao et al., 2008) in May 2014 for each host (avian, human and swine), independently. All sequences collected between 1960-2014 were retrieved. These six DNA data sets were then independently aligned with Muscle (Edgar, 2004) at the protein level using a local version of TranslatorX (Abascal et al., 2010).

Alignments were visually inspected with Jalview (Waterhouse et al., 2009) and rogue sequences were removed. Alignments were further checked by reconstructing phylogenetic trees by maximum likelihood under GTR + Γ (Aris-Brosou and Rodrigue, 2012) with FastTree (Price et al., 2010) to remove extremely divergent sequences. These alignment steps were performed a second time after removing rogue and / or extremely divergent sequences.

Each tree was then rooted with the oldest sequence, and patristic distances (*i.e.*, on the estimated tree) were computed between the root and each sequence. To account for rate change through time, an unsupervised clustering algorithm based on fitting mixtures of normal distributions by expectation-maximization, Mclust (Fraley et al., 2012), was employed to group sequences according to their root-to-tip distance and sampling time. The number K of components in this bivariate clustering was estimated based on BIC by varying $K \in [1, 20]$. In addition, we specified a conjugate normal prior on the mean, conditional on the covariance matrix, and an inverse Wishart prior on the covariance matrix. The identified clusters were then filtered to exclude transient trends, thereby keeping clusters that persisted for > 4 years. Robust regressions (Yohai, 1987) were then used to estimate cluster-specific rates of evolutions (slope of root-to-tip distance *vs.* sampling date). Only persistent clusters with significant slopes ($\alpha = 0.01$) were kept for downstream analyses.

Because the above algorithm is ignorant of the phylogeny, it is possible that sequences from different phylogenetic clades end up in the same rate cluster (see Fig.S1 for an example). To help alleviate this potential issue, a third variable accounting for clade identity was included in Mclust*. Clade identity was obtained based on node support and maximum genetic distance within a phylogenetic tree as implemented in Cluster Picker (Ragonnet-Cronin et al., 2013): support, estimated by the approximated likelihood ratio

test (Anisimova and Gascuel, 2006), was set to 90% and distance to the mean branch length multiplied by 25. Singleton sequences, *i.e.* those that were not included in any clade by this approach, were assigned their own clade identifier.

The robustness of the procedure, which we name Mclust* (the star telling it apart from Mclust, the second clustering tier of our algorithm), was assessed with respect to two key steps of the algorithm: the clustering and the robust regressions. First, because the conjugate priors in the clustering algorithm can lead to over-splitting the data, three levels of shrinkage were evaluated: $\tau = 0, 1$ or 5 . Second, because our definition of persistent clusters (circulating for > 4 years) could be too restrictive, more transient clusters (circulating > 2 years) were permitted in a second set of analyses. Shorter time intervals could not be set as we only retrieved the sampling year of each sequence, so as to maximize sample size (collection date of some of the older sequences is often limited to sampling year, excluding month and day information). As such, including two consecutive years is the finest time resolution that we can use to fit a linear model (through two x coordinates). In all, this led to a 3 (hosts) $\times 2$ (subtypes) $\times 2$ (segments) $\times 3$ (τ) $\times 2$ (minimum durations) experimental design.

Two steps were taken to validate the Mclust* algorithm. First, cluster-specific sequences were extracted for a small data set (HA of subtype H1N1 circulating in avian hosts; 240 sequences) and were subjected to a Bayesian relaxed clock analysis (Drummond and Rambaut, 2007) for sequentially sampled data (Rambaut, 2000) with BEAST v1.8.0. The same substitution model as above (GTR + Γ) was used. A lognormal uncorrelated clock and a constant-size coalescent prior were placed on times and rates, respectively. Each analysis was run twice to check convergence. Each Markov chain Monte Carlo (MCMC) sampler was run for 100 million steps with a thinning (subsampling) of 5,000 steps. In all cases, a conservative burn-in representing 25% of chain lengths proved suf-

ficient to reach target distributions; burn-ins removed, MCMC logs were then combined across each duplicate run. Second, this same data set (240 sequences; no subsetting) was subjected to a relaxed molecular clock analysis to check the position of the root. The MCMC samplers were run with the same specifications as on the cluster-specific subsets above. Where relevant, trees were compared with the SH test (Shimodaira and Hasegawa, 1999) as implemented in Phangorn (Schliep, 2011) with 10,000 replicates.

To test for the effect on rate estimation of the duration of ‘persistent clusters’ (> 2 or 4 years) and the effect of the upper bound on K , an ANOVA was conducted. A second ANOVA was conducted to test for evidence of any rate increase with time and/or with cluster duration; mean cluster date was used as a proxy for time; a nested design was adopted (time and duration within gene within hosts within subtypes). Rate estimates were $\log_{10}$ -transformed and all statistical analyses were carried out in R v3.1.1 (R Development Core Team, 2010).

For each data set, 100 bootstrap alignments were generated with R and FastTree was once more employed to estimate each bootstrap tree. Rate estimates are presented based on these bootstrap distributions, which therefore incorporate phylogenetic uncertainty. For each replicate, rate homogeneity was tested by means of an ANCOVA fitting a model predicting root-to-tip distances as a function of dates and a factor encoding cluster identity (the covariate). Scripts and alignments used in this study are available at <https://github.com/sarisbro>.

To assess the existence of a match between rate change and antigenic change, we compiled data from antigenic maps (Smith et al., 2004). These maps visualize in two dimensions the results of hemagglutination inhibition assays, where virus isolates are tested against a large number of strain-specific antisera. Such data are currently only available for H3N2, in humans, from 1968 to 2002 (Smith et al., 2004) and up to 2011

(Fonville et al., 2016).

Results

Our sampling design led to twelve large data sets containing an average 2,389 sequences and a grand total of 28,668 sequences (Table 1). The rooted maximum likelihood estimates of the corresponding phylogenies¹ (Fig. S2) were used to estimate patristic distances between each tip and the root, for each tree. The procedure then clustered these patristic distances and used phylogenetic information in a second clustering step to determine automatically the breakpoints at which rate changes occur. The Mclust* procedure is illustrated in figures S3-S4, for HA and NA respectively, under the specifications that (i) cluster persistence is at least 4 years and (ii) at most nine clusters are inferred. Our robustness analyses (figures S5-S18) on the Mclust* point estimates (tables S1-S8) showed no significant effect of our definition of cluster persistence ($P = 0.929$) or of the upper bound on K in the second clustering algorithm ($P = 0.474$). Therefore, the results presented below focus on clusters circulating for at least 4 years and a maximum number of clusters set to $K = 9$.

To validate our rate estimation procedure, we first tested the correctness of the choice of the root by running a standard analysis on a small data set (HA/H1N1/avian; 240 sequences). As our procedure relies on root-to-tip distances to estimate rates, it is critical that these distances be estimated accurately. As expected, the Bayesian procedure exhibited a perfect relaxed molecular clock (Fig. S18a), and the maximum likelihood estimates were also consistent with a clock (Fig. S19b; robust regression: $P < 0.0001$). Critically, the maximum *a posteriori* tree was not significantly different from the maximum likelihood tree (SH test: $P = 0.4966$). These results support that both root placement and

root-to-tip distances were correctly estimated by our procedure.

To further validate the procedure, we focused on rate estimates. For this same data set (HA/H1N1/avian), Mclust* estimated five rate clusters (Table 1), arbitrarily labelled 2, 3, 4, 5 and 7 (Fig. S20a). For clusters 3, 5 and 7, the 95% highest posterior density (HPD) of the rate based on the analysis of the full data set contained the Mclust* estimate, and overlapped with the Mclust* rate distribution obtained by bootstrap; clusters 2 and 4 did not. In the case of cluster 4 however, both the 95% HPD and bootstrap distributions overlap, showing that the Mclust* point estimate can be affected by phylogenetic uncertainty. In the case of cluster 2, subsetting the Bayesian analysis to sequences in cluster 2 led to a 95% HPD farther *away* from the Mclust* bootstrap distribution (Fig. S20b), but further removal of outliers (Fig. S21) led to overlapping Bayesian and bootstrap distributions, thereby showing that the Bayesian relaxed clock is sensitive to which sequences are analyzed. In contrast, the robust regressions were completely immune to sequence subsetting, returning exactly the same rate estimate for this cluster ($\hat{r}_2 = 0.0018$; $P = 1$). These results therefore show that while robust linear rate point estimates may be affected by phylogenetic uncertainty, bootstrap distributions provide us with accurate and robust estimates. The Mclust* rate estimates should therefore be confirmed by a bootstrap analysis.

Now that we understand the performance of the novel rate estimates better, we can address the broader biological question of rate variation in HA and NA within influenza viruses circulating in different hosts and different subtypes through 54 years of data. Table 1 suggests that there might be significant (at the 1% level) variation within all subtypes within all hosts for both HA and NA, except maybe in the case of HA (H1N1 and H3N2) and NA/H1N1 circulating in birds. To assess the robustness of these results with respect to phylogenetic uncertainty, we conducted a series of bootstrap analyses. For

each replicate, the P -value of the ANCOVA testing for among-cluster rate homogeneity was recorded (figures S22-S33). The proportion of replicates with P -values smaller than the significance level ($Pr(P < 0.01)$ in Fig. 1-3) is in broad agreement with the results shown in Table 1: the majority-rule probability $Pr(P < 0.01) > 0.50$ is consistent with significant P -values (Table 1).

We further tested if there is any evidence for time-dependency of rate estimates. The results of the nested ANOVA suggested a marginal effect of cluster duration ($F_{11} = 2.50$, $P = 0.0471$; clusters that last longer would tend to have higher rates), but no effect of time ($F_9 = 0.53$, $P = 0.8307$). The Mclust* algorithm is therefore robust to the time-dependent rate estimation artifact, at least on this data set comprising 28,668 influenza sequences. Altogether, these results show evidence for widespread rate heterogeneity both in HA and NA within hosts and subtypes, and hence widespread punctuated rate change in influenza viruses.

We finally assessed the connection between rate change and antigenic cluster change in the case of H3N2 viruses circulating in humans. Our sampling period, from 1968 to 2014, includes 15 antigenic cluster changes. Figures 2 and 4 show that we detect fewer rate change events than antigenic cluster changes, but that the dates match exactly in 1968, 2002, 2007 and 2011 for HA (Fig. 4a) and in 1968, 1997 and 2007 for NA (Fig. 4b).

Discussion

Although great strides have recently been made in both statistical modeling and statistical computing in the context of phylogenetics (Suchard and Rambaut, 2009), such progress hardly keeps up with the ever-increasing size of the data sets available. As a result, data sets with 10,000+ of sequences generally preclude the use of proper statistical models.

The heuristic presented here aims directly at addressing this issue, and is used to shed some light on the mode of evolution of the two antigenic determinants of influenza viruses.

Our estimation procedure takes root in the very definition of a rate: the expected amount of change at a site per unit of time (Drummond et al., 2012). A rate is therefore the slope of the corresponding regression. Here we use a phylogeny estimated by maximum likelihood, whose branch lengths are in unit of expected number of changes per site, to extract root-to-tip distances. Note that the algorithm used for this estimation step (Price et al., 2010) is extremely quick (analyzing 5,000 sequences takes ~ 5 min on a 1.8GHz/i7 laptop) and accurate (Liu et al., 2011). The Mclust* procedure then relies on the robust estimation of the slope of the linear regression of root-to-tip distances *vs.* sampling dates of viral sequences. As such, it is similar to an approach reconstructing rate trajectories (Squires et al., 2014), while our ultimate goal was to estimate absolute rates of evolution. One unique feature of our approach is its robustness to the time-dependency of rate estimates, common in simple relaxed molecular clocks (Duchêne et al., 2014). However, these results are based on the analyses of empirical data sets, so that simulations would be warranted to confirm both accurate rate estimation and robustness to artifacts such as time-dependency.

The major novelty introduced here essentially reside in the use of a two-tier clustering algorithm that permits rate variation through time without specifying the time points when these changes occur. As recently re-emphasized (Bielejec et al., 2014), appropriate modeling is possible, for instance by discretizing time according to a fine grid of intervals (Rodrigo et al., 2008). This however would not only lead to a computationally expensive solution (Bielejec et al., 2014), but also to one where co-circulation of different clades evolving at different rates would not be permitted. Implementing these ideas in a structured coalescent (Vaughan et al., 2014) would solve this latter limitation, but the

computational costs would still remain too high for any practical use. Even if not entirely satisfactory from a modeling standpoint, it may be that the future of phylogenetics lies in hybrid algorithms. These would implement sound statistical models that, instead of integrating over all uncertainties, would resort to machine learning, either unsupervised as here, or supervised (Quang et al., 2008; Le et al., 2008), in order to alleviate some of the computational costs.

One problem here is our use of frequentist statistics (ANCOVA) to compare regressions based on thousands of data points. For instance, in the case of Human HA/H3N2, the bootstrapped distributions show wide overlap (Fig. 2c) when the ANCOVA shows that slopes are very significantly different (Table 1: $P = 5.7 \times 10^{-146}$). The use of recent developments in large-scale inference (Efron, 2010) should address this issue. Future work should also expand this analysis to the six other segments of the influenza/A genome in order to test if there are differences between the external (HA and NA) and the internal (PB2, PB1, PA, NP, M and NS) segments. Yet, the major limitation of our approach probably lies in the *a priori* requirement of setting a minimal cluster span (2 or 4 years), which limits the resolution of our ability to detect rate change. In the case of HA/H1N1 for instance, rates cluster on a year to year basis since the 2009 pandemic, which leads us to discard information on all these years of sampling. Although one could argue that this represents a conservative feature, incorporating a cluster-merge process (Ding and He, 2002) would ensure that data are not discarded that quickly. However, by decreasing the number of rate clusters, this additional cluster-merge step is unlikely to improve the match between rate change and antigenic cluster change (Fig. 4): this might be because a change in antigenic properties can be caused by rate change (*e.g.*, following a reassortment event), but not necessarily so. Note also that we only considered collection year in the metadata associated with each sequence: this was done to maximize sample size, as all

sequences do not have a full collection date (year/month/day). However, full collection dates would allow us to use much shorter time intervals for cluster persistence: instead of being based on 4 years, just a few weeks of data could be used to obtain a much finer resolution, compatible with an early warning system in influenza surveillance and vaccine update recommendation.

Our biological finding of punctuated evolution of HA and NA echoes recent findings (Bedford et al., 2014) and should be confirmed for the six other genes of the influenza/A genome, expanded and extended to include more hosts and more subtypes (Worobey et al., 2014), in order to test whether natural hosts have indeed more constant rates of evolution – or not. Reciprocally, our results also mean that in the presence of rate heterogeneity, estimates of divergence times are likely to be impacted. The extend to which current divergence estimates are affected remains to be evaluated, but the Mclust* algorithm can easily be implemented in a local molecular clock framework (Yang, 2004; Aris-Brosou, 2007). By incorporating a finer granularity on the collection dates of influenza sequences, our approach can potentially be used to detect rate change almost in real time, and hence be used as a further indicator as to when vaccine composition should be updated.

Footnotes

¹Supplementary data are available with the article through the journal Web site at <http://nrcresearchpress.com/doi/suppl/10.1139/gen-2015-0163>.

Acknowledgements

We thank two anonymous reviewers whose comments helped improve this article. This work was supported by the Natural Sciences Research Council of Canada and by the Canada Foundation for Innovation (SAB).

References

- Abascal, F., Zardoya, R., and Telford, M.J. 2010. Translatorex: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**(Web Server issue): W7–13. doi:10.1093/nar/gkq291.
- Anisimova, M. and Gascuel, O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**(4): 539–52. doi:10.1080/10635150600755453.
- Aris-Brosou, S. 2007. Dating phylogenies with hybrid local molecular clocks. *PLoS One* **2**(9): e879. doi:10.1371/journal.pone.0000879.
- Aris-Brosou, S. and Rodrigue, N. 2012. The essentials of computational molecular evolution. *Methods Mol Biol* **855**: 111–52. doi:10.1007/978-1-61779-582-4_4.
- Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. 2008. The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**(2): 596–601. doi:10.1128/JVI.02005-07.
- Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley, J.W., Russell, C.A., Smith, D.J., and Rambaut, A. 2014. Integrating influenza antigenic dynamics with molecular evolution. *Elife* **3**: e01914. doi:10.7554/eLife.01914.
- Bielejec, F., Lemey, P., Baele, G., Rambaut, A., and Suchard, M.A. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Syst Biol* **63**(4): 493–504. doi:10.1093/sysbio/syu015.
- Ding, C. and He, X. 2002. Cluster merging and splitting in hierarchical clustering algorithms. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, pp. 139–146.

- Drummond, A.J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214. doi:10.1186/1471-2148-7-214.
- Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**(8): 1969–73. doi:10.1093/molbev/mss075.
- Duchêne, S., Holmes, E.C., and Ho, S.Y.W. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc Biol Sci* **281**(1786). doi:10.1098/rspb.2014.0732.
- Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5): 1792–7. doi:10.1093/nar/gkh340.
- Efron, B. 2010. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, *Institute of mathematical statistics monographs*, volume 1. Cambridge University Press, Cambridge.
- Fonville, J.M., Fraaij, P.L.A., de Mutsert, G., Wilks, S.H., van Beek, R., Fouchier, R.A.M., and Rimmelzwaan, G.F. 2016. Antigenic maps of influenza a(h3n2) produced with human antisera obtained after primary infection. *J Infect Dis* **213**(1): 31–8. doi:10.1093/infdis/jiv367.
- Fraley, C., Raftery, A.E., Murphy, T.B., and Scrucca, L. 2012. mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington.
- Goode, M., Guindon, S., and Rodrigo, A. 2008. Modelling the evolution of protein coding sequences sampled from Measurably Evolving Populations. *Genome Inform* **21**: 150–64.
- Le, S.Q., Lartillot, N., and Gascuel, O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc Lond B Biol Sci* **363**(1512): 3965–76. doi:10.1098/rstb.2008.0180.
- Liu, K., Linder, C.R., and Warnow, T. 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One* **6**(11): e27731. doi:10.1371/journal.pone.0027731.
- Plotkin, J.B., Dushoff, J., and Levin, S.A. 2002. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc Natl Acad Sci U S A* **99**(9): 6263–8. doi:10.1073/pnas.082110799.
- 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. doi:10.1371/journal.pone.0009490.

- Quang, L.S., Gascuel, O., and Lartillot, N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**(20): 2317–23. doi:10.1093/bioinformatics/btn445.
- R Development Core Team 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A.J.L., Lycett, S., and UK HIV Drug Resistance Database 2013. Automated analysis of phylogenetic clusters. *BMC Bioinformatics* **14**: 317. doi:10.1186/1471-2105-14-317.
- Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**(4): 395–9.
- Rodrigo, A., Bertels, F., Heled, J., Noder, R., Shearman, H., and Tsai, P. 2008. The perils of plenty: what are we going to do with all these genes? *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1512): 3893–3902.
- Schliep, K.P. 2011. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**(4): 592–593.
- Shimodaira, H. and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution* **16**: 1114–1116.
- Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**(5682): 371–376. doi:10.1126/science.1097211.
- Squires, R.B., Pickett, B.E., Das, S., and Scheuermann, R.H. 2014. Toward a method for tracking virus evolutionary trajectory applied to the pandemic h1n1 2009 influenza virus. *Infect Genet Evol* doi:10.1016/j.meegid.2014.07.015.
- Suchard, M.A. and Rambaut, A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**(11): 1370–6. doi:10.1093/bioinformatics/btp244.
- Vaughan, T.G., Kühnert, D., Poppinga, A., Welch, D., and Drummond, A.J. 2014. Efficient bayesian inference under the structured coalescent. *Bioinformatics* doi:10.1093/bioinformatics/btu201.
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. 2009. Jalview version 2: a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9): 1189–91. doi:10.1093/bioinformatics/btp033.
- Worobey, M., Han, G.Z., and Rambaut, A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* **508**(7495): 254–7. doi:10.1038/nature13016.

- Yang, Z. 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. *Acta Zool Sinica* **50**(4): 645–656.
- Yohai, V.J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics* **15**(2): 642–656.

Draft

Table

Table 1. Summary statistics of the data used in this study. Three hosts, two subtypes and two segments were used. Sample sizes (n), estimated number of rate clusters ($\hat{K}$) and P -values of the ANCOVAs testing for equality of rates (P_{ANCOVA} ; $K \in [1, 9]$) are shown.

Segment	Host	Subtype	n	$\hat{K}$	P_{ANCOVA}
HA	Avian	H1N1	240	5	0.4410
		H3N2	243	5	0.7539
	Human	H1N1	7618	1	—
		H3N2	5251	2	5.7×10^{-146}
	Swine	H1N1	1597	7	2.1×10^{-59}
		H3N2	1552	6	0.0046
NA	Avian	H1N1	182	3	0.0262
		H3N2	184	2	1.5×10^{-6}
	Human	H1N1	5193	3	2.1×10^{-44}
		H3N2	4971	3	8.3×10^{-77}
	Swine	H1N1	834	3	1.3×10^{-16}
		H3N2	803	6	1.9×10^{-5}

Figure captions

Fig. 1. Mclust* rate estimation in influenza viruses circulating in avian hosts. Results are shown for HA in H1N1 ((a)-(b)) and H3N2 ((c)-(d)) and for NA in H1N1 ((e)-(f)) and H3N2 ((g)-(h)). For each data set, the root-to-tip distance *vs.* sampling dates is first shown, followed by bootstrap distributions of the Mclust* rate estimates. Only significant clusters are labeled; they are consistently color-coded across both types of graphs. The cluster mapping is shown on each phylogenetic tree as an inset to bootstrap distributions. Also shown at the top left corner of each bootstrap panel, $Pr(P < 0.01)$, the proportion of times across the bootstrap replicates that the ANCOVA shows significant differences at the 1% level among cluster-specific rates – a measure of rate heterogeneity.

Fig. 2. Mclust* rate estimation in influenza viruses circulating in humans. See caption to figure 1 for details. Dates of antigenic cluster changes, as derived from antigenic maps, are shown for H3N2 by vertical broken lines (orange).

Fig. 3. Mclust* rate estimation in influenza viruses circulating in swines. See caption to figure 1 for details.

Fig. 4. Match between timing of rate change and antigenic cluster change. Timing of rate change is inferred by the Mclust* algorithm, while dates of antigenic cluster changes come from the literature. Results are shown for HA (in (a)) and for NA (in (b)). The red line indicates matchings ($y = x$).

(a)

Timing of rate changes

(b)

Timing of rate changes

1970 1980 1990 2000 2010

Automatic detection of rate change in large data sets with an unsupervised approach: the case of influenza viruses

Kassandra Labonté,^a Stéphane Aris-Brosou^{ab}

^aDepartment of Biology, Centre for Advanced Research in Environmental Genomics,
University of Ottawa, Ottawa, ON K1N 6N5, Canada

^bDepartment of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5,
Canada

Corresponding author: Stéphane Aris-Brosou (e-mail: sarisbro@uottawa.ca).

Supplementary figures and tables

Impact of including clade information into Mclust*	2
Maximum likelihood trees	3
Rate estimation under Mclust*	4
Point rate estimates and ANCOVA results	20
Comparison with Bayesian relaxed clock analyses (HA / H1N1 / avian)	28
Bootstrap distributions of the ANCOVA <i>P</i> -values	31

Figure S1: **Impact of including clade information into Mclust***. The top row shows that no including clade information leads to clusters than can group sequences present in distinct clades. The bottom row shows that after including clade information into Mclust*, clusters are limited to unique clades.

Figure S2: **Maximum likelihood trees.** These trees were estimated under the GTR + Γ model of evolution. For HA: subtype H1N1 in (a) avian, (b) human and (c) swine hosts; subtype H3N2 in (d) avian, (e) human and (f) swine hosts. For NA: subtype H1N1 in (g) avian, (h) human and (i) swine hosts; subtype H3N2 in (j) avian, (k) human and (l) swine hosts.

Figure S3: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 9]$. These are the results presented in the main text.

Figure S4: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 9]$. These are the results presented in the main text.

Figure S5: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 9]$.

Figure S6: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 9]$.

Figure S7: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 6]$. These are the results presented in the main text.

Figure S8: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 6]$. These are the results presented in the main text.

Figure S9: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 6]$.

Figure S10: **Rate estimation under Mclust*** for NA. Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 6]$.

Figure S11: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 8]$. These are the results presented in the main text.

Figure S12: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 8]$. These are the results presented in the main text.

Figure S13: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 8]$.

Figure S14: **Rate estimation under Mclust*** for NA. Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 8]$.

Figure S15: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 20]$. These are the results presented in the main text.

Figure S16: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 4 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 20]$. These are the results presented in the main text.

Figure S17: **Rate estimation under Mclust* for HA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 20]$.

Figure S18: **Rate estimation under Mclust* for NA.** Persistent clusters are those circulating for > 2 years. The number K of components in Mclust was estimated based on BIC by varying $K \in [1, 20]$.

Table S1: Point rate estimates under the Mclust* algorithm for $K \in [1, 6]$ and clusters circulating for > 4 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates					<i>P</i> -value
HA						
av_h1	0.0036	0.0031	0.0114	0.0035		9.41E-15
av_h3	0.0015	0.0025	0.0021			3.03E-12
hu_h1	0.0031	0.0049				1.20E-37
hu_h3	0.0043	0.0050				3.65E-197
sw_h1	0.0042	0.0058	0.0032	0.0032	0.0018	2.56E-220
sw_h3	0.0039	0.0052	0.0037			2.10E-03
NA						
av_n1	0.0030	0.0032	0.0024			1.28E-04
av_n3	0.0018	0.0035				1.47E-06
hu_n1	0.0036	0.0049				1.85E-55
hu_n3	0.0033	0.0043	0.0020			0
sw_n1	0.0037	0.0033	0.0041	0.0018	0.0032	7.70E-06
sw_n3	0.0034	0.0029	0.0039	0.0022		1.46E-02

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S2: Point rate estimates under the Mclust* algorithm for $k \in [1, 9]$ and clusters circulating for > 4 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates	P -value
HA		
G8new4 av_h1	0.0018 0.0036 0.0036	3.10E-04
G8new4 av_h3	0.0025 0.0021 0.0028	0.7539
G8new4 hu_h3	0.0031 0.0090 0.0041 0.0047	1.08E-206
G8new4 sw_h1	0.0039 0.0058 0.0037 0.0043 0.0032 0.0040	3.27E-73
G8new4 sw_h3	0.0044 0.0033 0.0037	4.55E-03
NA		
G8new4 av_n1	0.0020 0.0029 0.0030 0.0036 0.0022	6.37E-03
G8new4 av_n3	0.0018 0.0035	1.47E-06
G8new4 hu_n1	0.0013 0.0029	2.12E-44
G8new4 hu_n3	0.0033 0.0043 0.0020	0
G8new4 sw_n1	0.0037 0.0031 0.0041 0.0031 0.0027 0.0041	1.33E-16
G8new4 sw_n3	0.0029 0.0029 0.0030 0.0024 0.0039 0.0022	2.07E-03

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S3: Point rate estimates under the Mclust* algorithm for $k \in [1, 10]$ and clusters circulating for > 4 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates							P-value
HA								
av_h1	0.0019	0.0033	0.0049	0.0030	0.0030			0.4410
av_h3	0.0025	0.0021	0.0028					0.7539
hu_h3	0.0031	0.0041	0.0047					5.65E-146
sw_h1	0.0039	0.0058	0.0023	0.0041	0.0033	0.0041	0.0040	2.13E-59
sw_h3	0.0044	0.0033	0.0037					4.55E-03
NA								
av_n1	0.0042	0.0029	0.0024	0.0035	0.0033			2.62E-02
av_n3	0.0018	0.0035						1.47E-06
hu_n1	0.0013	0.0029						2.12E-44
hu_n3	0.0033	0.0039	0.0031					8.31E-77
sw_n1	0.0037	0.0031	0.0041	0.0031	0.0027	0.0041		1.33E-16
sw_n3	0.0063	0.0037	0.0030	0.0030	0.0037	0.0022		1.89E-05

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S4: Point rate estimates under the Mclust* algorithm for $k \in [1, 20]$ and clusters circulating for > 4 years. The P -value comes from the ANCOVA testing rate equality.

HA								
av_h1	0.0017	0.0033	0.0039	0.0029	0.0041	0.0065		3.50E-02
av_h3	0.0026	0.0015	0.0037	0.0021	0.0046			4.03E-03
hu_h1	0.0038	0.0023						0.8521
hu_h3	0.0030	0.0023	0.0037	0.0027	0.0018			2.92E-25
sw_h1	0.0044	0.0022	0.0023	0.0036	0.0024	0.0043	0.0040	1.53E-09
	0.0032							
sw_h3	0.0028	0.0051	0.0036					2.34E-04
NA								
av_n1	0.0014	0.0035	0.0029	0.0030	0.0012	0.0025		1.52E-03
av_n3	0.0035	0.0047	0.0021	0.0023	0.0044	0.0066		6.08E-03
hu_n1	0.0034	0.0032						8.67E-03
hu_n3	0.0009	0.0030	0.0035	0.0033	0.0034	0.0026	0.0028	2.01E-22
sw_n1	0.0038	0.0022	0.0039	0.0052	0.0043	0.0018	0.0023	1.01E-02
sw_n3	0.0027	0.0033	0.0037	0.0029	0.0035	0.0038	0.0022	2.01E-02

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S5: Point rate estimates under the Mclust* algorithm for $k \in [1, 6]$ and clusters circulating for > 2 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates	P -value
HA		
av_h1	0.0036 0.0031 0.0114 0.0035	9.41E-15
av_h3	0.0015 0.0025 0.0021	3.03E-12
hu_h1	0.0008 0.0031 0.0049	4.38E-56
hu_h3	0.0043 0.0050	3.65E-197
sw_h1	0.0042 0.0058 0.0032 0.0032 0.0018	2.56E-220
sw_h3	0.0039 0.0052 0.0037	2.10E-03
NA		
av_n1	0.0030 0.0032 0.0024	1.28E-04
av_n3	0.0018 0.0035	1.47E-06
hu_n1	0.0036 0.0049	1.85E-55
hu_n3	0.0033 0.0043 0.0020	0.0000
sw_n1	0.0037 0.0033 0.0041 0.0018 0.0032	7.70E-06
sw_n3	0.0032 0.0034 0.0029 0.0039 0.0022	5.24E-03

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S6: Point rate estimates under the Mclust* algorithm for $k \in [1, 9]$ and clusters circulating for > 2 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates	P -value
HA		
av_h1	0.0018 0.0036 0.0036	3.10E-04
av_h3	0.0025 0.0021 0.0028	0.7539
hu_h3	0.0031 0.0090 0.0041 0.0047	1.08E-206
sw_h1	0.0039 0.0041 0.0058 0.0037 0.0043 0.0032 0.0040	1.32E-112
sw_h3	0.0044 0.0033 0.0037	4.55E-03
NA		
av_n1	0.0020 0.0029 0.0030 0.0036 0.0022	6.37E-03
av_n3	0.0018 0.0035	1.47E-06
hu_n1	0.0013 0.0029	2.12E-44
hu_n3	0.0033 0.0043 0.0020	0.0000
sw_n1	0.0037 0.0031 0.0041 0.0031 0.0027 0.0041	1.33E-16
sw_n3	0.0031 0.0029 0.0029 0.0030 0.0024 0.0039 0.0022	2.03E-04

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S7: Point rate estimates under the Mclust* algorithm for $k \in [1, 10]$ and clusters circulating for > 2 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates								<i>P</i> -value
HA									
av_h1	0.0019	0.0033	0.0049	0.0030	0.0030				0.4410
av_h3	0.0025	0.0021	0.0028						0.7539
hu_h3	0.0031	0.0041	0.0047						5.65E-146
sw_h1	0.0039	0.0042	0.0058	0.0023	0.0041	0.0033	0.0041		4.22E-89
	0.0040								
sw_h3	0.0044	0.0033	0.0037						4.55E-03
NA									
av_n1	0.0042	0.0029	0.0024	0.0035	0.0033				2.62E-02
av_n3	0.0018	0.0035							1.47E-06
hu_n1	0.0013	0.0029							2.12E-44
hu_n3	0.0033	0.0039	0.0031						8.31E-77
sw_n1	0.0037	0.0031	0.0041	0.0031	0.0027	0.0041			1.33E-16
sw_n3	0.0032	0.0063	0.0037	0.0030	0.0030	0.0037	0.0022		1.26E-07

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Table S8: Point rate estimates under the Mclust* algorithm for $k \in [1, 20]$ and clusters circulating for > 2 years. The P -value comes from the ANCOVA testing rate equality.

Data	Cluster rates	P -value
HA		
av_h1	0.0017 0.0033 0.0039 0.0029 0.0041 0.0065	3.50E-02
av_h3	0.0026 0.0015 0.0037 0.0021 0.0046	4.03E-03
hu_h1	0.0038 0.0029 0.0044 0.0023	5.83E-05
hu_h3	0.0033 0.0030 0.0038 0.0034 0.0043 0.0040 0.0023 0.0038 0.0033 0.0037 0.0027 0.0018 0.0023	7.53E-118
sw_h1	0.0044 0.0022 0.0023 0.0036 0.0024 0.0043 0.0033 0.0040 0.0026 0.0032	7.68E-09
sw_h3	0.0028 0.0051 0.0046 0.0040 0.0035 0.0036	7.92E-04
NA		
av_n1	0.0014 0.0035 0.0029 0.0015 0.0030 0.0012 0.0044 0.0025	3.28E-04
av_n3	0.0035 0.0047 0.0021 0.0023 0.0044 0.0066	6.08E-03
hu_n1	0.0034 0.0032 0.0044 0.0020	5.99E-08
hu_n3	0.0034 0.0009 0.0030 0.0035 0.0033 0.0034 0.0026 0.0028	1.77E-25
sw_n1	0.0038 0.0022 0.0039 0.0052 0.0043 0.0042 0.0036 0.0030 0.0018 0.0023	1.72E-02
sw_n3	0.0032 0.0027 0.0033 0.0048 0.0033 0.0037 0.0029 0.0035 0.0038 0.0022	1.96E-03

Notes—The Data column shows the host (av: avian; hu: human; sw: swine), segment type (h: HA; n: NA) and subtype (1: H1N1; 3: H3N2).

Figure S19: **Estimated tip dates vs. actual sampling dates.** In both panels, tip dates are represented as the patristic distance from each tip to the estimated root of the tree, the root-to-tip distance. (a) Root-to-tip distance under the Bayesian relaxed molecular clock for the maximum *a posteriori* (MAP) tree as a function of actual sampling dates. (b) Root-to-tip distance under the Bayesian relaxed molecular clock for the MAP tree as a function of root-to-tip distance under the maximum likelihood tree. The result of the robust linear regression is shown (solid line). All dates are at the 1-year resolution.

Figure S20: Rate estimates under our approach *vs.* the Bayesian analyses in the case of HA/H1N1 in avian hosts. (a) Root-to-tip distances estimated under our procedure (with Mclust*) as a function of sampling dates (in years). Significant clusters, for which the slope of the robust regression has $P < 0.01$, are indicated: here, clusters 2, 3, 4, 5 and 7. Numbering is arbitrary. (b-f) The estimated rate (slope of the root-to-tip robust regression) is represented for each of the five clusters (vertical red line). The distributions show of these estimates over 100 bootstrap replicates is shown (solid density in black). Dotted densities show the posterior distribution of the rate when only the in-cluster sequences are analyzed. The orange density, which is the same in all five panels, shows the the posterior distribution of the rate for the entire data set (240 sequences). In (b), posterior “a” shows the results when all the in-cluster sequences are analyzed; posterior “b” shows the results when all the most distant sequences are removed; posterior “c” shows the results when all the least distant sequence is further removed; see figure S20 for identity of removed sequences.

Figure S21: **Identity of sequences removed from cluster 2 in the Bayesian analyses.** Root-to-tip distances as a function of sampling dates are shown for all the sequences belonging to cluster 2 (red squares). The most distant sequences, at the top of the graph, were removed in analyses “b” and “c”; the least distant sequence, at the bottom of the graph, was removed only in analysis “c”.

Figure S22: **Bootstrap distributions of the ANCOVA P -values for HA/H1N1 in avian viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S23: **Bootstrap distributions of the ANCOVA P -values for HA/H3N2 in avian viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S24: **Bootstrap distributions of the ANCOVA P -values for NA/H1N1 in avian viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S25: **Bootstrap distributions of the ANCOVA P -values for NA/H3N2 in avian viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S26: **Bootstrap distributions of the ANCOVA P -values for HA/H1N1 in human viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S27: **Bootstrap distributions of the ANCOVA P -values for HA/H3N2 in human viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S28: **Bootstrap distributions of the ANCOVA P -values for NA/H1N1 in human viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S29: **Bootstrap distributions of the ANCOVA P -values for NA/H3N2 in human viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S30: **Bootstrap distributions of the ANCOVA P -values for HA/H1N1 in swine viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S31: **Bootstrap distributions of the ANCOVA P -values for HA/H3N2 in swine viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S32: **Bootstrap distributions of the ANCOVA P -values for NA/H1N1 in swine viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.

Figure S33: **Bootstrap distributions of the ANCOVA P -values for NA/H3N2 in swine viruses.** One hundred replicates were performed to obtain this distribution; for each replicate, an ANCOVA compared a model constraining the slopes to be equal *vs.* and model that assumed independent slopes. The P -values were $\log_{10}$ -transformed for clarity; the red vertical dashed line represents the 1% (10^{-2}) significance level. The proportion of P -values under the significance threshold ($Pr(P < 0.01)$) can be seen in figure 1.