MUSCLE is faster and just as accurate as ClustalW in aligning genetically homogenous influenza outbreaks

Hunter J. Ries

Department of Pathobiological Sciences, University of Wisconsin–Madison, Madison, Wisconsin, USA

Abstract

Deep-sequencing viral genomics has provided high-resolution glimpses into respiratory virus evolution and epidemiology. While this has formed the foundation for modern public health infectious disease surveillance, it proves troublesome for resolving phylogenies in low-diversity outbreaks. During seasonal epidemics of influenza virus, acute infections drive transmission but not evolution within communities. Viruses transmitted from one host to another often share near-identical sequences, suggesting the virus is transmitted before it can accumulate mutations or adapt to its host. Thus, many community sequences are genetically similar but not closely related in their host histories. MUSCLE and ClustalW are standard multiple-sequence alignment tools used to align influenza virus sequences; however, the impact of alignment software on maximum likelihood trees needs to be clarified. In this study, I compare the maximum likelihood trees from ClustalW- and MUSCLE-aligned influenza A virus sequences from three separate, relatively genetically homogenous outbreaks. While MUSCLE is traditionally used for large or highly divergent populations and ClustalW for smaller or similar populations, both demonstrated similar tree topology, length, and likelihood. However, MUSCLE outpaced ClustalW and did not do so at the cost of accuracy. This study positions MUSCLE as rapid and accurate in our datasets, especially those with many samples.

Introduction

Influenza viruses cause significant morbidity and mortality globally (Krammer et al., 2018). Upon infection, the virus replicates and has the potential to introduce mutational errors into its genome. Interestingly, limited diversity is generated throughout this acute infection, but it is not often transmitted between hosts (Amato et al., 2021; Moncla et al., 2016; Sobel Leonard et al., 2016). Although influenza viruses may adapt throughout infection, diversity generated within a host is typically lost during transmission, resulting in poorly-resolved phylogenies dense with polytomies and short branch lengths.

Modern infectious disease genomic surveillance utilizes sequencing tools to assess and investigate outbreaks as they spread. These tools include Oxford Nanopore Technologies, which provides real-time data useful for consensus generation, and Illumina next-generation sequencing, a massively parallel process producing large amounts of high-resolution sequence data. Both have uses in public health interventions and academic investigations; however, the problem of low-diversity influenza outbreaks persists.

In their recent paper, Lauring *et al.* seek to investigate the evolution of influenza virus between hosts using a prospectively-enrolled community cohort based in Ann Arbor, Michigan, USA (McCrone et al., 2018). Using data generated with Illumina next-generation sequencing, they analyze flu evolution across flu seasons and within putative transmission pairs. In their 422 sequences, spanning three influenza A subtypes across three years, they find remarkably little surface protein (hemagglutinin; HA) diversity in influenza virus phylogenies within seasons and only see divergence across seasons. As hemaglutinin is a primary target of host antibodies and, thus, undergoes significant evolutionary pressures, this remained the focus of McCrone *et al.*'s study and the present study(Alymova et al., 2016). Their methods for constructing maximum likelihood trees left much to be desired, but the authors did include the underlying consensus sequences used in their study.

MUSCLE and ClustalW have been widely established as quick and accurate multiple-sequence alignment tools (Edgar, 2004; Thompson et al., 1994). MUSCLE rapidly aligns large datasets, while ClustalW is more suitable for smaller datasets (Sievers & Higgins, 2018). MUSCLE gains its documented speed advantage in most datasets using K-mer-based distance matrices and guide trees to progressively align and refine sequences (Edgar, 2004). ClustalW employs a similar method with final tree iteration, which has the potential to build upon faulty alignments but claims accuracy improvements (Edgar, 2004; Thompson et al., 1994). The application of ClustalW and MUSCLE to the field of influenza virus evolution has clear significance. However, there remains to be a published analysis on which tool to use for low-diversity viral sequences.

In this study, I compare the maximum likelihood trees from ClustalW- and MUSCLE-aligned influenza A virus sequences from three separate, relatively genetically homogenous outbreaks. Using the Lauring *et al.* dataset of Hong Kong, Perth, and California '09 influenza A virus outbreak sequences, we demonstrate that MUSCLE rapidly outpaces ClustalW, with approximately eight sequences per second in small datasets and six sequences per second in larger datasets, compared to .5 and .1 sequences per second, respectively, for ClustalW The Hong Kong subtype was the most extensive dataset with 309 sequences; Cali '09 had 53 sequences, and Perth had 57 sequences. This variance in sample number allowed us to assess numerous metrics of accuracy and speed across both subtypes and sample sizes. MUSCLE and ClustalW provided similar RAxML-NG maximum likelihood tree topologies in small datasets and nearly identical tree lengths and likelihoods for all datasets. Together, the results of my analyses suggest MUSCLE outpaces ClustalW and does not do so at the expense of accuracy, as ClustalW suggests.

Methods

Data collection

Influenza virus A fasta consensus sequences were retrieved from the McCrone et al. 2018 GitHub repository (https://github.com/lauringlab/Host_level_IAV_evolution; (McCrone et al., 2018). For example, fasta consensus files for one component of the Hong Kong subtype are located in ~/Host_level_IAV_evolution/data/processed/HK_1/parsed_fa. The Perth and California '09 data are in sub-folders within ~/Host_level_IAV_evolution/data/processed/. The study present did not analyze influenza B virus fasta consensus sequences (e.g., Victoria). All consensus files were transferred into subtype-specific folders, where Hemagglutinin (HA) gene sequences were extracted from the whole-genome consensus sequences. Sequence names were annotated with file names to preserve data integrity throughout processing. All sequences with directory or GenBank information were altered to file name annotations. Sequences were otherwise not altered.

ClustalW multiple sequence alignment

ClustalW (version 2.1) aligned all sequences by subtype using standard, default parameters (Thompson et al., 1994). These parameters include unrooted Neighbor-Joining trees, IUB nucleotide sequence weight matrix, and the final tree iteration. The ClustalW algorithm calculates pairwise distance matrices to construct an unrooted, additive-distance Neighbor-Joining guide tree. The sequences are then aligned pairwise, following the guide tree. Most intriguingly, ClustalW assigns weights to sequences according to branch length, meaning similar sequences are given less weight than more divergent sequences. This could increase the program's efficiency but is likely not ideal for our dataset of genetically homogenous sequences. ClustalW was chosen for this dataset because it is commonly used in numerous publications assessing influenza A virus phylogenies (Deem & Pan, 2009; Lee et al., 2014; Saxena et al., 2010). The main weakness of this program is its weighting algorithm, as previously mentioned, which may misalign sequences with low divergence. ClustalW outputs one optimum pairwise alignment rather than a collection of high-scoring alignments. In the present study, we did not assess the impact of iterations of trees or alignments. ClustalW fasta sequence alignment was the only desired output of the program.

MUSCLE multiple sequence alignment

MUSCLE (version 3.8.1551) aligned all sequences by subtype using standard, default parameters (Edgar, 2004). These parameters include unrooted UPGMA trees, the k-mer-based distance matrix, and the final tree refinement. MUSCLE fasta sequence alignment was the only desired output of the program. The MUSCLE alignment algorithm, in contrast to the ClustalW algorithm, undergoes successive guide tree, distance matrix, and alignment calculations. K-mer distance matrices construct the UPGMA tree, which determines the order of progressive alignment. In contrast to ClustalW, which uses Neighbor-Joining, MUSCLE uses UPGMA, which constructs a rooted, ultrametric phylogenetic tree. The formation of numerous ultrametric trees, rather than a single additive-distance tree, likely contributes to the speed reported for MUSCLE, although not at the cost of accuracy compared to Neighbor-Joining additive trees (Edgar, 2004). Lastly, the multiple sequence alignment is performed by re-estimating pairwise alignments of internal nodes, where new trees are compared to old trees and reassessed until convergence. The significant MUSCLE weaknesses lie in its rapid k-mer-based distance matrix calculations, which reduce the resolution of sequence data but vastly improve speed. This program was chosen to compare against ClustalW, as it is also used in aligning influenza A sequences, but I have not observed it used as often as ClustalW (Boni et al., 2008; ElHefnawi et al., 2011; Sahini et al., 2010).

Maximum likelihood tree generation

ClustalW or MUSCLE fasta alignments were input for RAxML-NG (version 1.1) analyses (Kozlov et al., 2019). Firstly, alignments were checked for single tree inference on alignment with the GTR model, ML estimates of substitution rate and nucleotide frequencies, and a four-category discrete gamma model of rate heterogeneity. Corrected .phy files were parsed for alignment before the tree was inferred. Tree inference was performed using the abovementioned parameters and seed 920 for all corrected .phy alignments without bootstrap replicates. RAxML-NG creates maximum likelihood trees by iteratively testing and altering trees to find the tree with the highest likelihood. This study uses the GTR model employed by many other influenza A virus phylogenetic studies (Forster et al., 2020; M. Nelson et al., 2009; M. I. Nelson et al., 2012). The relatively uncomplicated model allows for unequal base transition and transversion rates. While the GTR+G parameter may introduce bias by assuming that substitution processes are time-reversible and, most critically, stationary, it is a bias that can largely be offset by software parameter alterations (Barba-Montoya et al., 2020).

Tree rooting and analysis

RAxML-ng bestTree files were analyzed in RStudio (version "Spotted Wakerobin") and R (version 4.2.1 “Funny-Looking Kid”) using APE (Paradis et al., 2004), adegenet (Jombart, 2008), phangorn (Schliep, 2011), phytools (Revell, 2012), ggplot (Wickham, 2016), and ggtree (Yu et al., 2017). Briefly, trees were imported with read.tree, tips were cleaned, roots were resolved with midpoint.root, and the likelihood was calculated with pml. Maximum likelihood trees were compared between ClustalW and MUSCLE with cophylo. Maximum likelihood trees were individually plotted using ggtree. Midpoint rooting was chosen to visualize differences in tree topology, where subtle differences in branch lengths alter large-scale tree topology.

Data availability

All analyses outlined in this study are available at https://github.com/RiesHunter/myProject.

Results

ClustalW and MUSCLE produce similar tree topologies through RAxML-NG.

Maximum likelihood trees from ClustalW or MUSCLE alignments across subtypes show considerable congruence in topologies following midpoint rooting (**Figure 1**). Although limited influenza A virus diversity occurs within-season, as predicted, maximum likelihood phylogenies are relatively consistent between alignment software. Most notably, in the Hong Kong subtype, large polytomies are formed. This is likely due to sequence identicality rather than alignment issues. Still, ClustalW and MUSCLE appear to group sequences incredibly similarly.

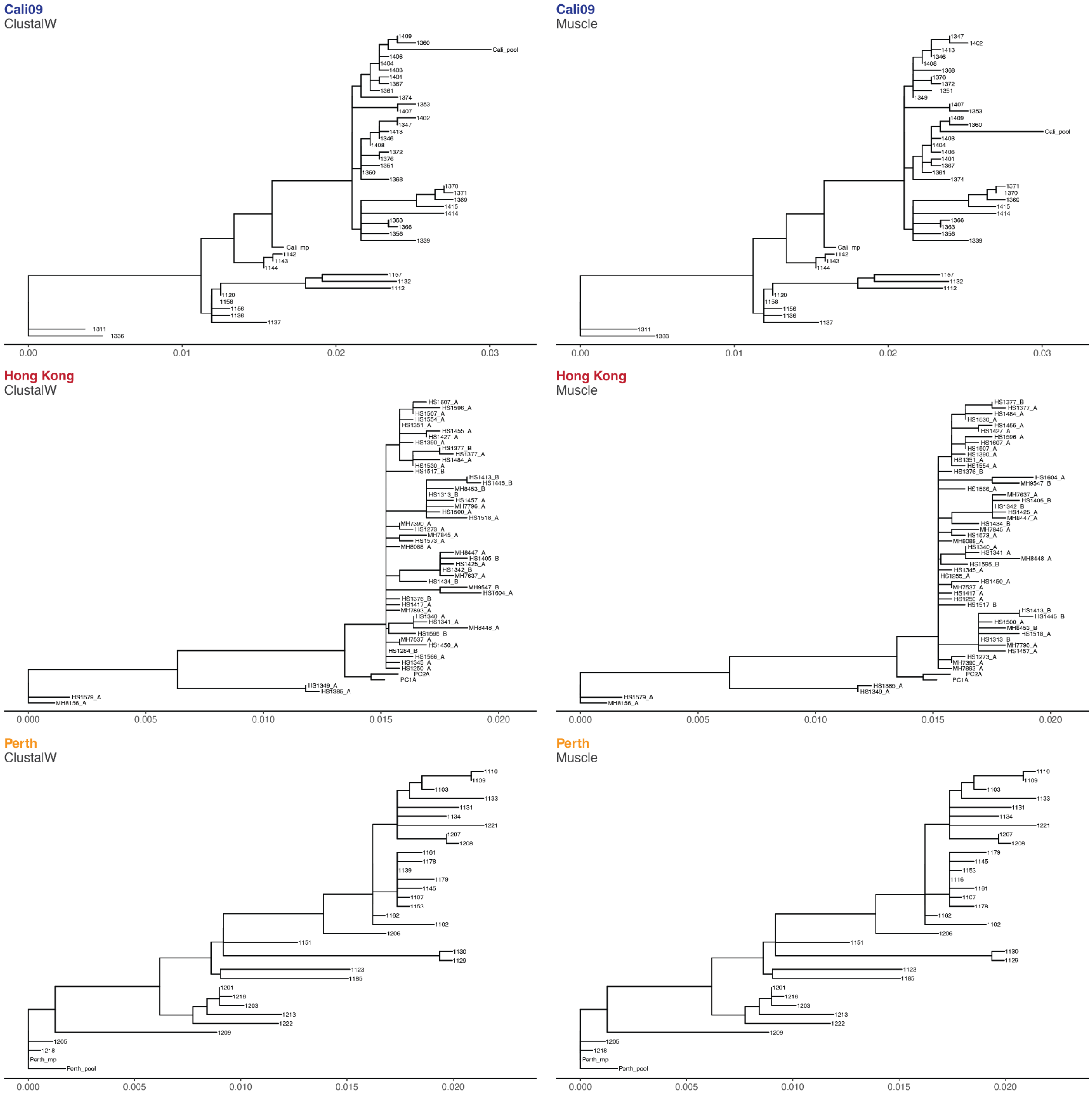

Figure 1. Maximum likelihood trees for each influenza subtype aligned by either ClustalW (left) or MUSCLE (right). California '09 (blue), Hong Kong (red), and Perth (yellow) maximum likelihood trees are shown in the rows. Trees were created in RStudio using ggtree, where nucleotide substitutions per site were measured on the x-axis. Tip labels are the anonymized sample name from the original consensus sequences.

ClustalW and MUSCLE yield nearly identical maximum likelihood tree topologies in small, low-diversity datasets.

To visualize tree congruence between ClustalW- and MUSCLE-aligned RAxML-NG trees, I plotted maximum likelihood trees in cophylo, through the exact midpoint rooting as previously (**Figure 2**). In Cali '09, nearly all clades are identical, with a minor difference in branch lengths between samples 1403 and 1406. The Hong Kong subtype showed the largest incongruence between trees; although this was the most extensive dataset at 309 sequences, Cali '09 had 53 sequences, and Perth had 57 sequences. This was expected, as large polytomies were present within the **Figure 1** Hong Kong trees, and many are likely due to minor branch length differences. Interestingly, in both the Perth and Hong Kong subtypes, a tip appears missing from the phylogeny: 1139/1116 and HS1284_B/HS1255_A, respectively. I am unaware of why this tip is missing, although each alignment software appears to have discarded a sample from each group.

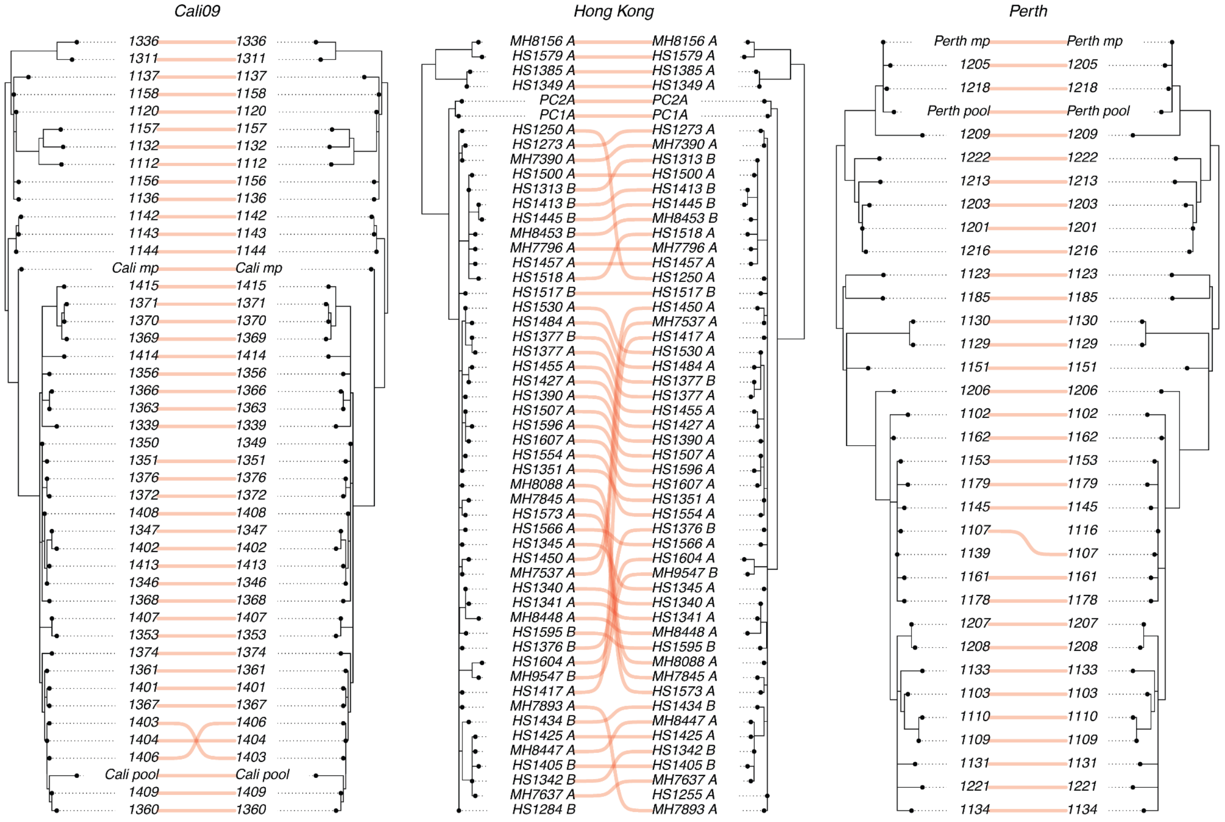

Figure 2. Comparison of ClustalW (left) MUSCLE (right) for each influenza subtype, Cali09, Hong Kong, and Perth. Trees were created in RStudio using cophylo, with tip labels connected with orange lines.

MUSCLE outpaces ClustalW with nearly identical likelihood and tree lengths.

To assess the efficiency of ClustalW and MUSCLE against the two small datasets (Cali ‘09 and Perth) and the larger dataset (Hong Kong; HK), a seconds counter was utilized in the script for both run_clustalw.sh and run_muscle.sh (**Figure 3**). As predicted, MUSCLE outpaced ClustalW in every subtype group. Strikingly, MUSCLE processed approximately six sequences per second in the large data set, while ClustalW only processed .1 sequence per second. The inverse: MUSCLE processed a sequence every 0.16 seconds, and ClustalW processed one sequence every 10.3 seconds. With significant speed efficiency differentials, it is clear that MUSCLE is the better choice for massive datasets.

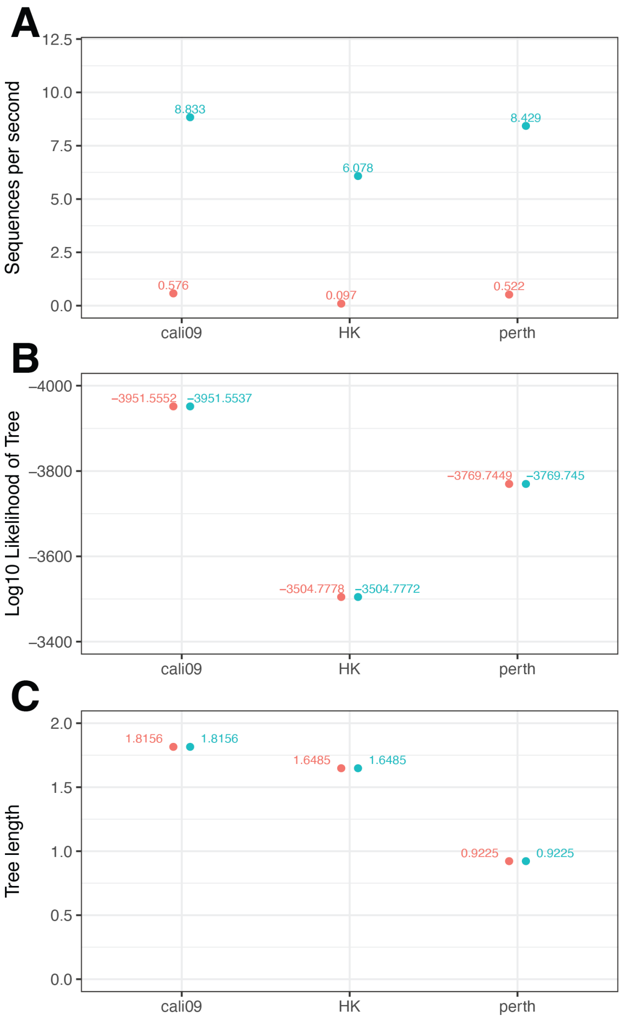

Figure 3. Run statistics of each influenza subtype in each alignment method. **A)** Sequences per second. **B)** Log10 Likelihood of Tree with gamma site rate and GTR substitution model. **C)** Tree length.

Secondly, I assessed the log10 likelihood of the tree and tree length to investigate whether the two programs differed quantitatively in their maximum likelihood accuracy or topology. In all cases, even in the Hong Kong subtype, ClustalW and MUSCLE performed nearly identically. These data suggest that MUSCLE provides comparable accuracy tp ClustalW with massive speed advantages.

Discussion

This study has assessed the impact of multiple sequence alignment software ClustalW and MUSCLE on the tree topology and branch length of RAxML-NG maximum likelihood trees across three different influenza A outbreaks, each of varying sequence numbers. This study found that ClustalW and MUSCLE produce remarkably similar trees in topology, branch length, and likelihood ratio. Unsurprisingly, MUSCLE outpaced ClustalW; however, it did so without the cost of accuracy, as measured in this study. While we concede that there are many other well-vetted metrics for assessing tree accuracy, such as Q and TC scores (Edgar, 2004), our study found striking similarities between ClustalW and MUSCLE among vastly different metrics.

The present study clearly defines the need to further assess phylogenetic software within the context of actual viral disease outbreaks, especially when there is low divergence and high sequence similarity. Vetted test datasets, such as BALiBASE, do not accurately represent actual, highly-homogenous sequence data observed in these outbreaks and thus should be viewed as a measure of the broad scale of algorithm performance (Bahr et al., 2001). As outbreaks continue to occur, and sequencing technologies advance to encourage rapid, real-time sampling, current phylogenetic software will need help to resolve dense phylogenies. This study proposes MUSCLE as an accurate and rapid alternative to ClustalW for aligning densely sampled viral outbreaks.

References

1. Alymova, I. V., York, I. A., Air, G. M., Cipollo, J. F., Gulati, S., Baranovich, T., Kumar, A., Zeng, H., Gansebom, S., & McCullers, J. A. (2016). Glycosylation changes in the globular head of H3N2 influenza hemagglutinin modulate receptor binding without affecting virus virulence. *Scientific Reports*, *6*, 36216. https://doi.org/10.1038/srep36216
2. Amato, K. A., Haddock, L. A., Braun, K. M., Meliopoulos, V., Livingston, B., Honce, R., Schaack, G. A., Boehm, E., Higgins, C. A., Barry, G. L., Koelle, K., Schultz-Cherry, S., Friedrich, T. C., & Mehle, A. (2021). *Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks* (p. 2021.09.28.462198). https://doi.org/10.1101/2021.09.28.462198
3. Bahr, A., Thompson, J. D., Thierry, J. C., & Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research*, *29*(1), 323–326. https://doi.org/10.1093/nar/29.1.323
4. Barba-Montoya, J., Tao, Q., & Kumar, S. (2020). Using a GTR+Γ substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated. *Bioinformatics*, *36*(Supplement_2), i884–i894. https://doi.org/10.1093/bioinformatics/btaa820
5. Boni, M. F., Zhou, Y., Taubenberger, J. K., & Holmes, E. C. (2008). Homologous Recombination Is Very Rare or Absent in Human Influenza A Virus. *Journal of Virology*,

82(10), 4807–4811. https://doi.org/10.1128/JVI.02683-07

6. Deem, M. W., & Pan, K. (2009). The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Engineering, Design and Selection*, *22*(9), 543–546. https://doi.org/10.1093/protein/gzp027
7. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
8. ElHefnawi, M., AlAidi, O., Mohamed, N., Kamar, M., El-Azab, I., Zada, S., & Siam, R. (2011). Identification of novel conserved functional motifs across most Influenza A viral strains. *Virology Journal*, *8*(1), 44. https://doi.org/10.1186/1743-422X-8-44
9. Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9241–9243. https://doi.org/10.1073/pnas.2004999117
10. Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129
11. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305
12. Krammer, F., Smith, G. J. D., Fouchier, R. A. M., Peiris, M., Kedzierska, K., Doherty, P. C., Palese, P., Shaw, M. L., Treanor, J., Webster, R. G., & García-Sastre, A. (2018). Influenza. *Nature Reviews. Disease Primers*, *4*(1), 3. https://doi.org/10.1038/s41572-018-0002-y
13. Lee, Y.-J., Kang, H.-M., Lee, E.-K., Song, B.-M., Jeong, J., Kwon, Y.-K., Kim, H.-R., Lee, K.-J., Hong, M.-S., Jang, I., Choi, K.-S., Kim, J.-Y., Lee, H.-J., Kang, M.-S., Jeong, O.-M., Baek, J.-H., Joo, Y.-S., Park, Y. H., & Lee, H.-S. (2014). Novel Reassortant Influenza A(H5N8) Viruses, South Korea, 2014. *Emerging Infectious Diseases*, *20*(6), 1086–1089. https://doi.org/10.3201/eid2006.140233
14. McCrone, J. T., Woods, R. J., Martin, E. T., Malosh, R. E., Monto, A. S., & Lauring, A. S. (2018). Stochastic processes constrain the within and between host evolution of influenza virus. *ELife*, *7*, e35962. https://doi.org/10.7554/eLife.35962
15. Moncla, L. H., Zhong, G., Nelson, C. W., Dinis, J. M., Mutschler, J., Hughes, A. L., Watanabe, T., Kawaoka, Y., & Friedrich, T. C. (2016). Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian Adaptation of a 1918-like Avian Influenza

Virus. *Cell Host & Microbe*, *19*(2), 169–180. https://doi.org/10.1016/j.chom.2016.01.011

16. Nelson, M. I., Detmer, S. E., Wentworth, D. E., Tan, Y., Schwartzbard, A., Halpin, R. A., Stockwell, T. B., Lin, X., Vincent, A. L., Gramer, M. R., & Holmes, E. C. (2012). Genomic reassortment of influenza A virus in North American swine, 1998–2011. *The Journal of General Virology*, *93*(Pt 12), 2584–2589. https://doi.org/10.1099/vir.0.045930-0
17. Nelson, M., Spiro, D., Wentworth, D., Fan, J., Beck, E., St. George, K., Ghedin, E., Halpin, R., Bera, J., Hine, E., Proudfoot, K., Stockwell, T., Lin, X., Griesemer, S., Bose, M., Jurgens, L., Kumar, S., Viboud, C., Holmes, E., & Henrickson, K. (2009). The early diversification of influenza A/H1N1pdm. *PLoS Currents*, *1*, RRN1126. https://doi.org/10.1371/currents.RRN1126
18. Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412
19. Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x
20. Sahini, L., Tempczyk-Russell, A., & Agarwal, R. (2010). Large-Scale Sequence Analysis of Hemagglutinin of Influenza A Virus Identifies Conserved Regions Suitable for Targeting an Anti-Viral Response. *PLOS ONE*, *5*(2), e9268. https://doi.org/10.1371/journal.pone.0009268
21. Saxena, S. K., Mishra, N., Saxena, R., Swamy, M. A., Sahgal, P., Saxena, S., Tiwari, S., Mathur, A., & Nair, M. P. (2010). Structural and antigenic variance between novel influenza A/H1N1/2009 and influenza A/H1N1/2008 viruses. *The Journal of Infection in Developing Countries*, *4*(01), Article 01. https://doi.org/10.3855/jidc.546
22. Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706
23. Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290
24. Sobel Leonard, A., McClain, M. T., Smith, G. J. D., Wentworth, D. E., Halpin, R. A., Lin, X., Ransier, A., Stockwell, T. B., Das, S. R., Gilbert, A. S., Lambkin-Williams, R., Ginsburg, G. S., Woods, C. W., & Koelle, K. (2016). Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic

Diversification. *Journal of Virology*, *90*(24), 11247–11258. https://doi.org/10.1128/JVI.01657-16

25. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.
26. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Create Elegant Data Visualisations Using the Grammar of Graphics. https://ggplot2.tidyverse.org/
27. Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628