Phylogenetic discordance can substantially overestimate genomic reassortment in avian influenza virus

Hugo G. Castelán-Sánchez[1] and Art F. Y. Poon[1,2,3]

[1] Department of Pathology & Laboratory Medicine, Western University, Dental Sciences Building, Rm. 201, London, Ontario N6A 5C1, Canada
[2] Department of Microbiology & Immunology, Western University, 1151 Richmond Street, London, Ontario N6A 3K7, Canada
[3] Department of Computer Science, Western University, Rm. 355, Middlesex College, London N6A 5B7, Canada

Abstract

Recombination plays an important role in the evolution of RNA viruses, as it allows the exchange of genetic material between viral lineages. Reassortment, a form of recombination specific to segmented genomes, involves the exchange of entire segments and has contributed significantly to the adaptation and spread of influenza viruses through novel genomic combinations, *i.e.*, antigenic shifts. It is usually identified by phylogenetic discordance: differences in the topologies of trees reconstructed from different genomic segments. However, phylogenetic discordance can also result from error in reconstructing trees. To characterize the impact of reconstruction error, we curated a database of $n = 11,765$ complete genomes of avian H5Nx influenza A viruses from avian hosts. We found evidence of widespread reassortment as measured by inferred subtree-prune-regraft (SPR) events, consistent with previous studies. Next, we ran replicate simulations of sequence evolution along the reference tree for the segment encoding hemagglutinin (HA), adjusting simulations for the lengths and clock rates of the other segments. These simulations provided a baseline for the expected amount of phylogenetic discordance in the absence of any reassortment. When sampling HA sequences at random from the database to build reference trees, we observed that simulating other segments without reassortment still yielded about 32% as many SPRs as the real segment data on average. The average proportion of SPRs without reassortment was greatly reduced (4%)

if we selected an equivalent number of HA sequences retaining the most genetic diversity, which was consistent with the accuracy of phylogenetic reconstruction being the limiting factor. This implies that measuring reassortment by SPRs may have a high false positive rate, and that previous evidence of extensive reassortment in influenza viruses should be interpreted with caution. In addition, we observed that the SPRs reconstructed on simulated trees had significantly shorter distances between the prune and regraft locations than real trees. These results suggest that down-sampling sequences to maximize evolutionary divergence and filtering out the shortest SPRs may be effective measures against false positives.

Introduction

Recombination is an important mechanism of genomic evolution in RNA viruses (Simon-Loriere and Holmes 2011). This process involves the exchange of genetic material between different genomic backgrounds, which is facilitated by the coinfection of a host by two or more distinct virus lineages. Recombination uncouples the evolutionary fate of a mutation from the rest of the genome, and under certain conditions (Otto and Barton 1997), recombination will increase the diversity of genotypes in the population (Pérez-Losada et al. 2015; Wang et al. 2022). Many RNA viruses have segmented genomes, including members of the *Reoviridae* and *Orthomyxoviridae* families (McDonald et al. 2016). In these viruses, recombination can occur by reassortment, in which the genomic segments packaged into a virion are derived from different parental genomes. Additionally, many other viruses have multipartite genomes where segments are packaged into separate virus particles. These viruses can reassort segments from different infected cells or even from different hosts (Varsani et al. 2018). Reassortment involves the wholesale exchange of entire segments between genomes, which can accelerate the acquisition of combinations of mutations that are required to overcome host barriers more rapidly than the progressive accumulation of individual mutations (Vijaykrishna et al. 2015).

Influenza A virus (IAV; family *Orthomyxoviridae*, genus *Alphainfluenza*) provides several im-

portant examples of the possible impacts of genetic reassortment on the health of human and non-human animal populations. This virus has been responsible for numerous and severe epidemics and pandemics (Sutton 2018), several of which are associated with a reassortant virus (Lowen 2017). IAV has a negative-sense RNA genome that consists of eight genomic segments that encode 10 major proteins and other accessory gene products (Pinto et al. 2021). This number of segments implies that about 99% virus particles produced by a co-infected host cell will contain a reassortant genome. The co-circulation of multiple IAV subtypes and clades in domesticated animal host species provides abundant opportunities for reassortment between divergent lineages (Ryt-Hansen et al. 2021; Li et al. 2024). Thus, reassortment can generate new chimeric IAV genomes that potentially increase the capacity of this virus to be transmitted to a new host species and establish a productive infection (Bouvier and Palese 2008; Ma et al. 2016).

The comparative analysis of IAV genome sequences plays an important role in measuring the frequency of reassortment and detecting the emergence of new reassortant viruses (Kong et al. 2015; Taylor et al. 2023). To this end, a number of different methods have been developed to identify reassortment events. In the absence of recombination, the evolutionary history of a sample population of genome sequences may be represented by a single phylogenetic tree. Reassortment uncouples the evolutionary histories of the affected genomic segments. Consequently, a common technique to detect reassortment is to reconstruct a phylogeny for each segment, and then to identify discrepancies in the topology and/or labels among the resulting trees. Discordant trees are often visualized as tanglegrams (Page 2003), in which two trees with a common set of labels are drawn in opposing orientations, with subtrees rotated to maximize the concordance of tip labels between trees. Tanglegrams can be used to visually assess a small number of reassortment events. In addition, the optimization algorithm for untangling pairs of trees provides a rough measure of the extent of discordance that may be due to recombination, *i.e.*, entanglement. However, tanglegrams do not readily scale to resolving large numbers of reassortment events. Evolutionary events that produce discordant trees can be explicitly modeled using an ancestral recombination graph

(ARG; Wong et al. 2024). Although there has been a recent resurgence of interest in ARGs (Frost et al. 2015), they remain too computationally complex for large-scale application.

A compromise between these two extremes is to infer the frequency of reassortment events by the number of tree rearrangement operations required to convert one segment tree to another (Svinti et al. 2013), which can be interpreted as an edit distance. Screening for tips that transition between clusters of labels in the respective trees is an equivalent approach Silva et al. 2012; Gong et al. 2021. There are many edit distances available for this purpose, including the Robinson-Foulds distance (RF; Briand et al. 2020), tree bisection and reconnection (TBR; Kelk and Linz 2020), nearest neighbor interchange (NNI; Collienne and Gavryushkin 2021), and subtree prune and regraft (SPR; Wu 2009). Among these, the SPR distance is frequently used in studies of lateral gene transfer and recombination (Whidden et al. 2014), possibly because the SPR most closely resembles the effect of recombination on the topologies of trees on either side of the breakpoint.

In this work, we performed a phylogenetic analysis starting from nearly twelve thousand full genome sequences for influenza A H5Nx viruses to identify putative rearrangement events by reconstructing SPR rearrangements. Our main objective was to evaluate the reliability of this approach given the uncertainty in reconstructing accurate phylogenies under realistic conditions. To this end, we reconstructed phylogenies for each IAV H5Nx segment and then selected one of these trees as the template on which we simulated sequence evolution for the other segments in the absence of reassortment. We compared the entanglement values and the SPR distances between the reference tree for segment 4 (encoding hemagglutinin, HA) and trees for all other segments for both real and simulated data sets to assess the specificity of these methods for detecting reassortment. In addition, we characterized the SPR operations reconstructed from these data in terms of the size of the subtree and the distance traversed in the tree.

Methods

Data collection

We retrieved influenza A virus H5Nx genome sequences isolated from avian hosts from the NCBI Genbank and GISAID databases (accessed 5 May 2024). The GISAID query returned 5,994 to 6,000 sequences per segment, and the Genbank query returned 6,157 to 6,170 sequences per segment. We combined these sequence data sets by genomic segment. There was only a small proportion of sequences that were exact duplicates between databases. For example, out of 12,167 neuraminidase (NA) sequences, there were 8,347 unique sequences, of which 1,417 (17%) had two or more instances — only 171 (12%) of these were represented in both databases, involving a total of 673 sequence records. This low amount of overlap is largely due to a transition away from depositing sequences in Genbank to GISAID around 2020, coinciding with recent outbreaks of avian H5N1 IAV (Supplementary Figure S1). Sequence identity was not sufficient to determine whether the sequences are derived from the same isolate. Consequently, we merged these datasets by harmonizing the sequence labels into the World Health Organization (1980) nomenclature format, including host species, country of sampling, isolate name and year of sampling, and then filtering sequences with duplicate labels. This reduced the combined data to 11,795 sequences per segment. We generated alignments for each segment using MAFFT (v7.490; Katoh and Standley 2013), removed duplicate sequences, and then manually inspected and cleaned each segment alignment with trimAl (Capella-Gutiérrez et al. 2009, p. v1.5.0).

We reconstructed maximum likelihood phylogenies using the standard generalized time-reversible (GTR) model with gamma-distributed rate variation and empirical base frequencies IQ-TREE2 (Minh et al. 2020). To reduce the number of genomes, we applied two subsampling strategies to obtain subsets of sequences. First, we tested the prune function of the Biopython Phylo module to progressively remove terminal branches in the HA tree whose length exceeded a cutoff. After experimenting with different thresholds, we selected a cutoff of 0.08 expected substitutions per site, resulting in a subset of 213 HA sequences. In addition, we generated five

independent random samples of 213 sequences per segment from the entire dataset. We used the labels of the sampled HA sequences to retrieve the corresponding sequences for the other seven segments and reconstruct trees. These two sampling strategies provided a means of evaluating the effect of phylogenetic reconstruction error on levels of phylogenetic discordance.

Data simulation

To quantify the discordance among phylogenies reconstructed from sequences that have evolved in the complete absence of reassortment, we simulated sequence alignments along the tree inferred from the real HA alignment (T_{HA}) for every other segment. Since every segment would evolve along exactly the same tree topology, any reassortment event inferred for the simulated trees could be interpreted as a false positive. We proportionally rescaled branch lengths in T_{HA} so that the total tree length corresponded to the length of the tree reconstructed from actual sequences of that segment. Our motivation for adjusting tree lengths in this manner was to accommodate the substantial variation in rates of evolution among segments, which could otherwise confound our assessment of false positive rates. For each rescaled tree, we used the Python module *Pyvolve* (Spielman and Wilke 2015) to simulate sequence evolution under the default Goldman-Yang codon substitution model with the dN/dS ratio set to 0.5 and sequence length matched to each genomic segment.

Phylogenetic analysis

We used two approaches to infer reassortment events. First we used a tanglegram, which is a graphical comparison of two trees with a common set of tip labels. The trees are rotated and optionally re-rooted to minimize the discordance in the vertical location of labels, as quantified by the entanglement statistic. Our working assumption is that increasing amounts of reassortment will induce greater amounts of entanglement between trees of different segments. We used the *step2side* method implemented in the Python module *tanglegram* (Nguyen et al. 2022) to disentangle the HA tree and trees from every other segment. We further visualized the tanglegram using the cophylo function from the phytools package in R, which implements a closest-neighbor (NN) tanglegram

approach by heuristically rotating the trees to minimize link crossings between matched tip labels (Revell 2024).

Second, we used the *Espalier* library (Rasmussen and Guo 2023) in Python to efficiently calculate the approximate number of SPR (subtree prune and regraft) operations required to convert one phylogenetic tree into another. This number provides another means of estimating the number of reassortment events between phylogenetic trees. We obtained SPR distances between the HA tree and the trees for every other segment for the actual IAV genome sequence data, and we made the same comparisons for trees reconstructed from simulated data.

We hypothesized that false positive reassortment events due to phylogenetic uncertainty would tend to involve SPR moves separated by shorter distances in the respective trees. To evaluate this hypothesis, we quantified the total path length separating the inferred graft points for each SPR operation. Let the tree reconstructed from real sequences for the i-th segment be denoted by T_i. Moreover, let the tree reconstructed from simulated sequences that were evolved along the real HA tree T_4 be denoted by T_i^*. For each subtree S_j in the maximum agreement forest between trees T_4 and T_i for $i \neq 4$, we first locate S_j in the reference (segment 4) tree T_4 by searching for the internal node with the fewest descendants that match all tip labels in S_j. We call this node the root of subtree j in T_4 and denote it as $r_4^{(j)}$. Next, we located the internal node in T_4 with the same direct common ancestor as $r_4^{(j)}$, which we refer to as its sibling node, $s_4^{(j)}$. Switching to the other segment tree T_i, we locate the same subtree S_j to identify $r_i^{(j)}$ by matching tip labels. Finally, we locate the internal node in T_i with the fewest descendants that match all tip labels associated with $s_4^{(j)}$, which we denote as $s_i^{(j)}$. The sibling distance is then defined as the total path length (*i.e.*, sum of branch lengths) from $s_i^{(j)}$ to $r_i^{(j)}$. Put another way, $s_i^{(j)}$ is the approximate location from which the SPR operation pruned the subtree, and $r_i^{(j)}$ is the location where the subtree was grafted to. A diagram summarizing this procedure is provided as Supplementary Figure S2. Following this method, we calculated the sibling distances for all SPR events between T_4 and both real (T_i) and simulated (T_i^*) trees for other segments.

Results

We obtained a total of 11,795 H5Nx influenza A virus genome sequences from the NCBI Genbank and GISAID databases. These sequences represented a variety of avian hosts, with the most frequent annotations being 'chicken' ($n = 2,968$), 'duck' ($n = 1,774$) and 'turkey' ($n = 927$). There was a total of 527 unique annotations for host organisms at varying taxonomic levels; for instance, mallard duck ($n = 47$) and Muscovy duck ($n = 199$) were both sub-classifications below 'duck'. Similarly, there were 621 unique geolocation annotations, with the three most common being 'Netherlands' ($n = 845$), 'Minnesota' ($n = 569$), and 'Vietnam' ($n = 557$). Many records were labeled by state instead of country, obscuring a relative abundance of samples from China and the United States, for example. We reconstructed a maximum likelihood phylogeny relating all HA (segment 4) sequences, and then we generated a sub-sample of $n = 213$ (1.8%) sequences maximizing the total tree length by progressively removing the shortest terminal branches. Next, we extracted the corresponding sequences for the other segments (*i.e.*, from the same virus isolates as the HA sample) and regenerated maximum likelihood phylogenies for each segment. These seven datasets are intended to minimize the error in phylogenetic reconstruction for a given sample size, and will be referred to as the *control samples*. In addition, we generated five random samples of the same number of HA sequences, extracted sequences for the other segments, and reconstructed phylogenies under the same workflow. These random samples are meant to be more consistent with a conventional phylogenetic study.

To visually assess the potential extent of reassortment in the evolutionary history of avian H5Nx genomes, we generated tanglegrams for the HA tree against the trees of every other segment. A tanglegram is a graphical method that re-roots and rotates branches in two trees with matching labels to minimize the discordance in the vertical positions of their labels. Discordance can be quantified by an entanglement statistic that sums the difference in the vertical positions of each label, and normalizes this total such that the statistic ranges from 0 (completely congruent) to 1 (worst case scenario). For example, the tanglegram between trees relating HA and NA segments

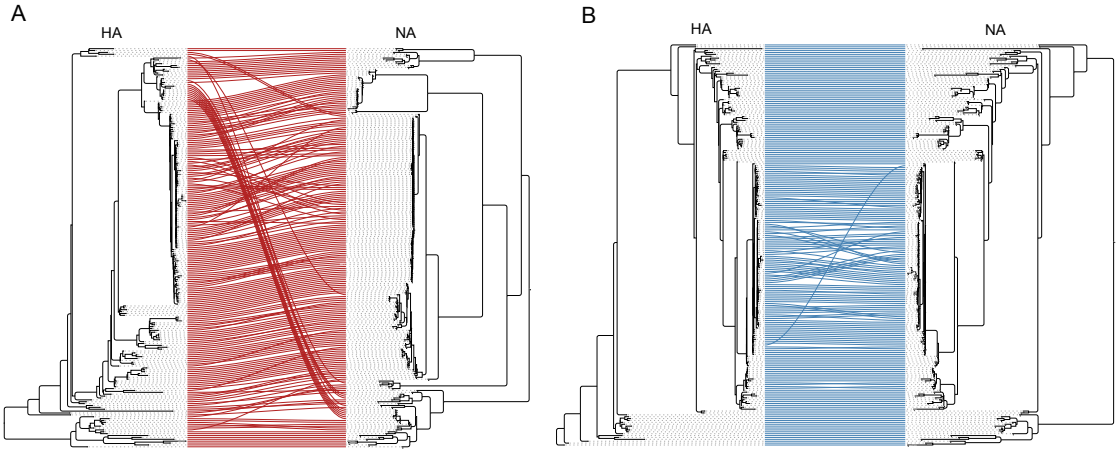

Figure 1: Tanglegrams between a random sample of $n = 213$ influenza A virus H5Nx sequences for segments 4 (HA) and 6 (NA). The NA tree (right side) was reconstructed either from actual segment 6 sequences from the same isolates (A), or generated by simulating sequence evolution along the rescaled HA tree. The entanglement values associated with these pairs of trees were 0.13 (actual) and 0.07 (simulated), respectively. An alternate version using the Python module *baltic* is provided as Supplementary Figure S3.

from a random sample of genomes demonstrates a substantial level of entanglement (0.13; Figure 1A), which implies multiple reassortment events have occurred in the evolutionary history of these sequences. However, these results do not indicate how much discordance can be attributed to inaccuracies in reconstructing the respective phylogenies. To illustrate, Figure 1B displays a tanglegram between the same HA tree and a tree relating sequences simulated along the HA tree adjusted to the sequence length and molecular clock of the segment encoding NA. Since the underlying phylogenies are congruent, this simulations provide an estimate of the expected amount of phylogenetic discordance in the absence of reassortment. Entanglement values were significantly lower for simulated trees (-0.12, 95% confidence interval $-0.14, -0.10$; Figure 2). Put another way, a tree reconstructed from sequences simulated without reassortment was about 48% less 'entangled' than a tree from actual sequences. We also observed significantly lower entanglement on trees generated on random samples (-0.04, 95% CI $-0.07, -0.01$) relative to control samples. However, entanglement is not a reliable statistic for measuring phylogenetic discordance (De Vienne 2019), and is difficult to interpret in terms of how much reassortment may have occurred

between segments.

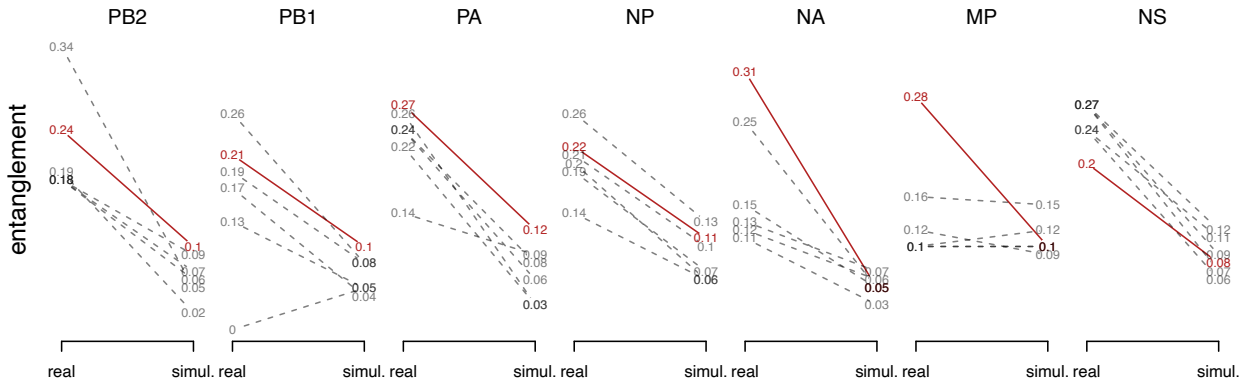

Figure 2: Slopegraphs summarizing entanglement values for real and simulated trees. Solid (red) lines indicate entanglement values obtained for the control (maximum diversity) sample of HA sequences, and dashed lines indicate random samples. 'Real' values were obtained for trees reconstructed from actual sequences for the other segments. Simulated ('simul.') values were obtained for trees reconstructed from sequences simulated along the rescaled HA tree.

A more robust and informative method for quantifying the discordance between phylogenies is the minimum number of subtree-prune-regraft (SPR) rearrangements required to transform one tree topology into another. This number defines an edit distance between the trees. A useful characteristic of SPR is that this rearrangement is similar to the effect of reassortment on trees Svinti et al. 2013. We calculated the SPR distances between the same pairwise comparisons of trees as we used for entanglement. As expected, the effects of reassortment and phylogenetic reconstruction error were much clearer using SPR counts (Figure 3). SPR counts were significantly lower for simulated trees without reassortment (Poisson regression, $P < 10^{-12}$, 95% CI $-1.59, -1.45$). For random samples of HA sequences, trees reconstructed from other segments simulated without reassortment had about 32.0% (interquartile range, IQR $= 17.8\% - 36.7\%$) of SPRs as the real trees, on average. Specifically, the median SPR distance was 98 (IQR 77.5, 103) for real trees and 22 (16, 27.5) for simulated trees ($n = 35$). These proportions were much lower for the control samples (3.8%, range $1.4\% - 5.5\%$), supporting the hypothesis that phylogenetic uncertainty is the main driver of false positives. For these $n = 7$ samples, the median SPR distances were 143 (range 127, 150) for real and 5 (2, 7) for simulated trees. In addition, we obtained consistently

higher SPR counts across segments for control samples (mean 140.7, range 127 − 150; Figure 3), which was likely driven by more complete coverage of the underlying evolutionary history of avian H5Nx isolates.

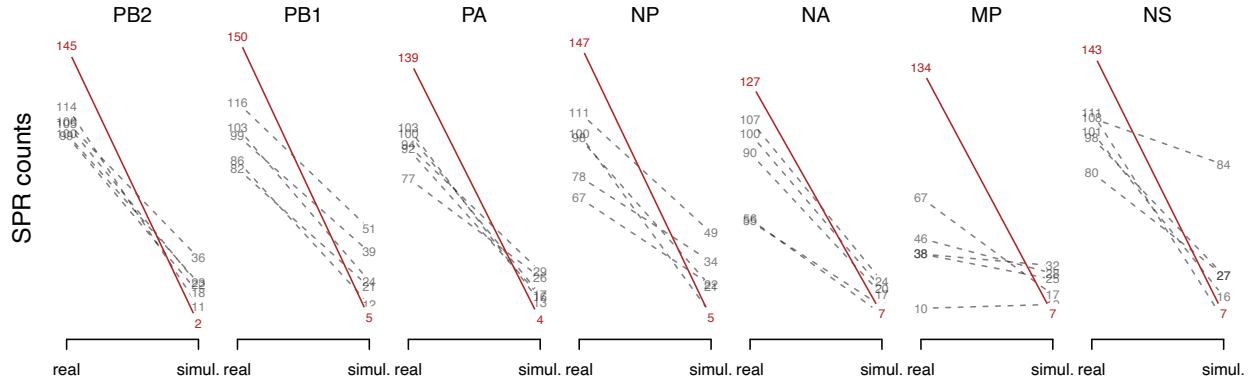

Figure 3: Slopegraphs summarizing subtree-prune regraft distances (minimum SPR counts) between real and simulated trees. Solid (red) lines indicate SPR counts obtained for the control sample of HA sequences, and dashed lines indicate random samples. 'Real' counts were obtained for trees reconstructed from actual sequences for other segments. Simulated ('simul.') counts were obtained for trees reconstructed from sequences simulated along the rescaled HA tree.

Even though we expect less phylogenetic discordance in the absence of reassortment, the number of false positives is unacceptably high ($\sim 30\%$) if the sequences have not been down-sampled to maximize evolutionary divergence. Therefore, it would be helpful to be able to distinguish between real and spurious SPR events. We investigated two characteristics of SPR events to determine whether SPRs caused by error in phylogenetic reconstruction were distinguishable from those caused by reassortment. First, we examined the number of terminal nodes (tips) affected by the SPR, *i.e.*, subtree size. An unexpected result is that SPRs involving simulated trees tended to have the largest subtrees (Figure 4), despite there being significantly fewer subtrees extracted from these comparisons. This difference was statistically significant for random samples (paired Wilcoxon rank sum test, $P = 4.16^{-6}$) and marginally significant for control samples ($P = 0.051$). On the other hand, there was no significant difference between real and simulated trees in the proportion of SPRs involving a single branch (binomial regression, $P = 0.76$).

Next, we quantified the distance between the prune and regraft points for all inferred SPR

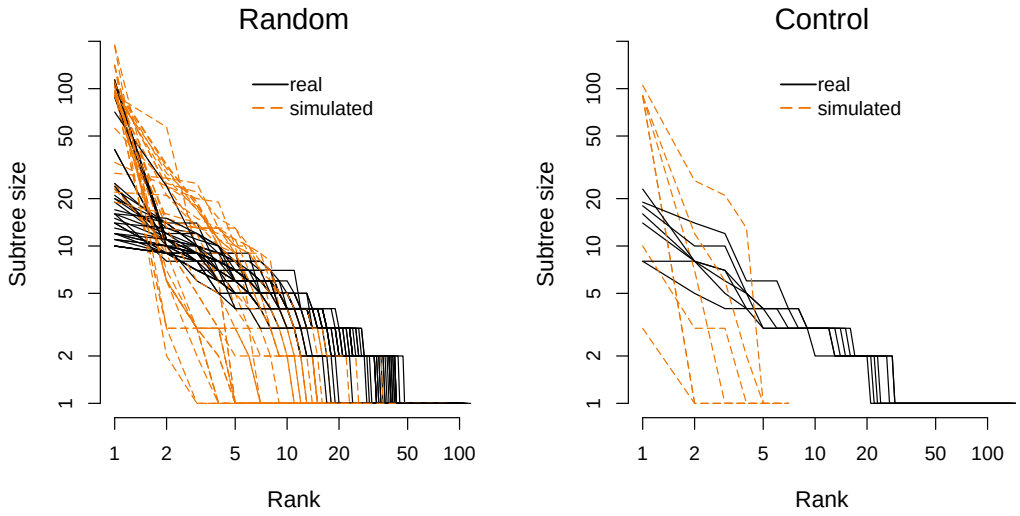

Figure 4: Log-log plots of subtree sizes for all SPR events reconstructed from random and control samples. Each line represents the sequence of subtrees for each segment and replicate (for random samples), sorted in decreasing order of size (number of tips). Solid (black) lines indicate subtrees in trees relating real segments, and dashed (orange) lines indicate simulations. The ranks of the ordered subtrees are plotted along the x-axis (*i.e.*, largest subtree is rank 1), and subtree sizes are plotted along the y-axis.

rearrangements. This task is not straightforward because these points exist on separate trees, so we approximated the path length by locating the smallest subtree containing the same tip labels as the descendants of the sibling node in a second tree, where the sibling node shares the same immediate ancestor as the subtree affected by the SPR. We refer to this path length as the sibling distance for a given SPR. Figure 5 summarizes the distributions of sibling distances between real and simulated trees for random and control samples. Overall, sibling distances tended to be shorter for comparisons to simulated trees, with a median distance of 0.046 for real trees and 0.006 for simulated trees. If we assume that a reassortment event can occur between any two contemporaneous lineages in the population, then a tendency towards shorter sibling distances is more consistent with error in phylogenetic reconstruction. Sibling distances were more variable for comparisons between HA and NA trees (Figure 5), most likely because trees reconstructed from NA sequences were substantially larger than the other segments; conversely, trees relating MP sequences were the shortest. We used a gamma mixed-effects regression model with a log link to test the effect of data

type (real or simulated) on sibling distance for random samples, including segment as a random effect. This model obtained a significant negative fixed effect of simulated data on sibling distance for random samples ($P = 3.1 \times 10^{-12}$, 95% CI $-0.53, -0.31$). There was significant variation among segments (standard deviation 95% CI $0.28, 0.95$). We did not find a significant association between subtree size and sibling distance ($P = 0.13$). Applying the same model to the control samples, we obtained a stronger effect of simulated data ($P < 10^{-12}$, 95% CI $-2.58, -1.83$) and a comparable amount of variation among segments (standard deviation 95% CI $0.29, 1.02$). The median sibling distance for control samples was 0.094 for real trees and 0.0078 for simulated trees. These results imply that an effective means of reducing the number of false positives in a reassortment analysis is to down-sample the data to maximize sequence divergence, and to discard SPRs with short sibling distances.

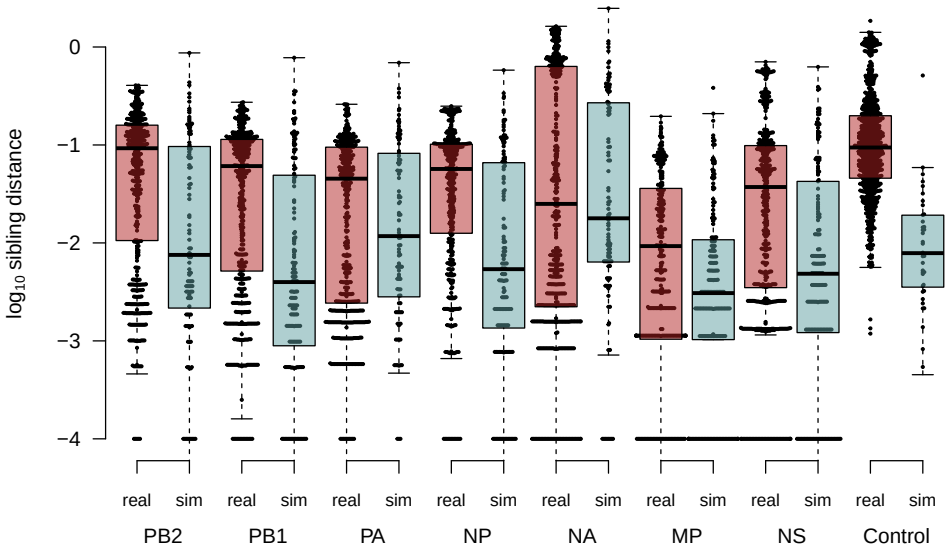

Figure 5: Boxplots showing the distribution of sibling distances ($\log_{10}$-transformed) between prune and regraft locations for SPR events across genome segments (PA, PB1, PB2, MP, NA, NP, and NS) for random samples, and pooled across segments for control samples. We added an arbitrary small amount (10^{-4}) to zero branch lengths to accommodate the log-transformation.

Discussion

Reassortment is undoubtedly an important process in the evolution of influenza A viruses in the form of antigenic shift (McHardy and Adams 2009). Several epidemiologically significant strains such as the H1N1 strain associated with the 2009 swine flu pandemic have originated from reassortment combining segments from divergent lineages (Garten et al. 2009; Rambaut et al. 2008). In addition, a wide diversity of influenza A viruses circulate in numerous avian species (Causey and Edwards 2008). This natural host reservoir represents a significant risk for reassortment to generate new genomic combinations with increased pathogenicity and zoonotic potential (Graziosi et al. 2024; Guo et al. 2017; Pulit-Penaloza et al. 2020). For example, Gu et al. (2012) reported an IAV genome isolated from domestic ducks in China in 2012 that appeared to be a reassortment of H11N2, H7N1 and highly pathogenic avian H5Nx lineages. Consequently, there have been many studies using comparative methods to identify reassortment events in the evolutionary history of AIVs. Lewis et al. (2015) reported extensive reassortment from a tanglegram analysis of 211 complete genomes for a variety of HnNn subtypes of AIVs. Similarly, Dugan et al. (2008) observed a "remarkably high rate of genome reassortment" among 167 complete genomes of AIVs, using the difference in log-likelihoods of tree topologies reconstructed from different segments as a measure of discordance when fitting those trees to the same data. Although this method can be used to reject a null hypothesis of no reassortment, it does not identify specific reassortment events. Lu et al. (2014) used ancestral reconstruction of HA or NA subtypes as discrete traits associated with the other segments from about 3,000 AIV genomes to quantify a range of reassortment rates among subtypes and host species. Using a larger dataset of over 40,000 IAV genomes from a variety of host species, Gong et al. (2021) partitioned the trees for the genomic segments into 493-663 monophyletic clades, such that each genome could be represented by the clade assignments of its segments. They reported nearly 2,000 putative reassortments, where a genome was assumed to be a reassortant if its clade assignments could be produced from a combination of two or more other genomes.

Although these studies have employed different methods, they share a common premise on utilizing the phylogenetic discordance among genomic segments as evidence of reassortment. It has long been recognized that error in reconstructing phylogenies will also result in discordant trees in the absence of any reassortment (Svinti et al. 2013). Surprisingly, previous studies of reassortment in AIV genomes have seldom checked for these inevitable false positives. We have found that the estimated prevalence of reassortment events in trees simulated in the absence of any reassortment can reach a substantial fraction (about 20% to 40%) of the number of events estimated from trees derived from actual AIV data. This outcome is sensitive to the amount of divergence in the sequence data. For instance, if the sequences are pre-selected to maximize divergence, making it easier to accurately reconstruct the underlying phylogeny, then the number of SPRs in the absence of reassortment is greatly reduced. This sensitivity to sequence divergence is consistent with error in phylogenetic reconstruction being the primary driver of these false positive SPRs. These results suggest that reports of extensive reassortment among influenza virus segments should be interpreted with some caution. We note that the rapid nucleotide substitution rates of AIV should provide a relative abundance of phylogenetically-informative sites in genomic data within a certain timescale. Identifying reassortment events for other viruses with segmented genomes and slower molecular clocks may be an even greater challenge (Varsani et al. 2018).

In this study, we employed two methods to quantify the extent of reassortment in the evolutionary history of real and simulated data: entanglement and the minimum number of subtree-prune-regraft rearrangements, (*i.e.*, the SPR distance). While there are other approaches to quantifying reassortment that have been described in the literature, we submit that entanglement and SPR distances provide a representative selection of methods. Entanglement is generally calculated as a part of generating a tanglegram, a popular method for visualizing phylogenetic discordance (Bansal et al. 2009). It is essentially the total difference in the vertical placement of tip labels. Entanglement can be misleading because the locations of tip labels are not directly proportional to topological congruence. Trees with very different topologies can have a relatively low entanglement, and vice

versa (De Vienne 2019; Briand et al. 2020). Our results are consistent with these findings. Specifically, we observed significantly lower entanglement values for trees simulated in the absence of reassortment (Figure 2). However, this pattern was much more clearly resolved for the same data by SPR distances (Figure 3), which is arguably the most direct measure of reassortment that is readily available to us (Svinti et al. 2013; Lin et al. 2024).

A number of programs have been implemented to measure reassortment from phylogenetic discordance (Suzuki 2010). However, only a few of these programs remain available, particularly for inferring the minimum number of SPR rearrangements between trees. For example, SPRSupertrees (Whidden et al. 2014) is a C++ program distributed under the GNU GPLv3 license that calculates the SPR distance, but its computing time because prohibitively long for phylogenetic trees with hundreds of branches or more. In addition, the website from which this program is distributed was not consistently accessible (last attempt October 15, 2025). Espalier (Rasmussen and Guo 2023) uses a heuristic algorithm to compute the SPR distance in a more computationally efficient manner, retaining only strongly supported rearrangements. However, the algorithm employed by Espalier is approximate and can produce distances that are not symmetric, *i.e.*, $d(A,B) \neq d(B,A)$.

Lin et al. (2024) recently described several other issues in estimating the number and distribution of reassortment events from discordant segment trees. First, they noted that reassortment events may alter branch lengths without modifying the topology of the trees. Reassortments that involve an unsampled lineage cannot be reconstructed by comparing trees. In addition, comparing discordant trees can be affected by a 'multiple hits' problem in which a more recent reassortment event obscures other events that precede it. These issues have the opposite effect compared to phylogenetic uncertainty: they would cause the estimated number of SPR rearrangements to underestimate the actual extent of reassortment. Consequently, measuring the overall rate of reassortment or reconstructing a number of these events by the large-scale comparative analysis of genomes should be considered an open problem for influenza viruses and other segmented viruses. Fortunately, there has been a resurgence of interest in ancestral recombination graphs (*e.g.*, Ras-

mussen and Guo 2023; Barrat-Charlaix et al. 2022; Lewanski et al. 2024) that will likely drive further advances in this area.

Acknowledgments

We thank David Rasmussen for providing advice and insight into the algorithms employed by Espalier, and Sarah Otto for constructive comments on earlier versions of this manuscript.

References

Bansal, Mukul S et al. (2009). "Generalized binary tanglegrams: Algorithms and applications". In: *Bioinformatics and Computational Biology: First International Conference, BICoB 2009, New Orleans, LA, USA, April 8-10, 2009. Proceedings*. Springer, pp. 114–125.

Barrat-Charlaix, Pierre, Timothy G Vaughan, and Richard A Neher (2022). "TreeKnit: Inferring ancestral reassortment graphs of influenza viruses". In: *PLoS Computational Biology* 18.8, e1010394.

Bouvier, Nicole M and Peter Palese (2008). "The biology of influenza viruses". In: *Vaccine* 26, pp. D49–D53.

Briand, Samuel et al. (2020). "A generalized Robinson-Foulds distance for labeled trees". In: *BMC genomics* 21.Suppl 10, p. 779.

Capella-Gutiérrez, Salvador, José M Silla-Martínez, and Toni Gabaldón (2009). "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses". In: *Bioinformatics* 25.15, pp. 1972–1973.

Causey, Douglas and Scott V Edwards (2008). "Ecology of avian influenza virus in birds". In: *The Journal of Infectious Diseases* 197.Supplement_1, S29–S33.

Collienne, Lena and Alex Gavryushkin (2021). "Computing nearest neighbour interchange distances between ranked phylogenetic trees". In: *Journal of Mathematical Biology* 82.1, p. 8.

De Vienne, Damien M (2019). "Tanglegrams are misleading for visual evaluation of tree congruence". In: *Molecular biology and evolution* 36.1, pp. 174–176.

Dugan, Vivien G et al. (2008). “The evolutionary genetics and emergence of avian influenza viruses in wild birds”. In: *PLoS pathogens* 4.5, e1000076.

Frost, Simon DW et al. (2015). “Eight challenges in phylodynamic inference”. In: *Epidemics* 10, pp. 88–92.

Garten, Rebecca J et al. (2009). “Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans”. In: *science* 325.5937, pp. 197–201.

Gong, Xingfei et al. (2021). “Reassortment network of influenza A virus”. In: *Frontiers in Microbiology* 12, p. 793500.

Graziosi, Giulia et al. (2024). “Highly pathogenic avian influenza (HPAI) H5 clade 2.3. 4.4 b virus infection in birds and mammals”. In: *Animals* 14.9, p. 1372.

Gu, Min et al. (2012). *Genome sequence of a natural reassortant H5N2 avian influenza virus from domestic mallard ducks in eastern China*.

Guo, Hongbo et al. (2017). “Highly pathogenic influenza A (H5Nx) viruses with altered H5 receptor-binding specificity”. In: *Emerging infectious diseases* 23.2, p. 220.

Katoh, Kazutaka and Daron M Standley (2013). “MAFFT multiple sequence alignment software version 7: improvements in performance and usability”. In: *Molecular biology and evolution* 30.4, pp. 772–780.

Kelk, Steven and Simone Linz (2020). “New reduction rules for the tree bisection and reconnection distance”. In: *Annals of Combinatorics* 24.3, pp. 475–502.

Kong, Weili et al. (2015). “Novel reassortant influenza viruses between pandemic (H1N1) 2009 and other influenza viruses pose a risk to public health”. In: *Microbial pathogenesis* 89, pp. 62–72.

Lewanski, Alexander L, Michael C Grundler, and Gideon S Bradburd (2024). “The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics”. In: *PLoS Genetics* 20.1, e1011110.

Lewis, Nicola S et al. (2015). “Influenza A virus evolution and spatio-temporal dynamics in Eurasian wild birds: a phylogenetic and phylogeographical study of whole-genome sequence data”. In: *Journal of General Virology* 96.8, pp. 2050–2060.

Li, Chong et al. (2024). “Naturally occurring influenza reassortment in pigs facilitates the emergence of intrahost virus subpopulations with distinct genotypes and replicative fitness”. In: *mBio*, e01924–24.

Lin, Qianying et al. (2024). “The number and pattern of viral genomic reassortments are not necessarily identifiable from segment trees”. In: *Molecular Biology and Evolution* 41.6, msae078.

Lowen, Anice C (2017). “Constraints, drivers, and implications of influenza A virus reassortment”. In: *Annual review of virology* 4.1, pp. 105–121.

Lu, Lu, Samantha J Lycett, and Andrew J Leigh Brown (2014). “Reassortment patterns of avian influenza virus internal segments among different subtypes”. In: *BMC evolutionary biology* 14, pp. 1–15.

Ma, Eric J et al. (2016). “Reticulate evolution is favored in influenza niche switching”. In: *Proceedings of the National Academy of Sciences* 113.19, pp. 5335–5339.

McDonald, Sarah M et al. (2016). “Reassortment in segmented RNA viruses: mechanisms and outcomes”. In: *Nature Reviews Microbiology* 14.7, pp. 448–460.

McHardy, Alice Carolyn and Ben Adams (2009). “The role of genomics in tracking the evolution of influenza A virus”. In: *PLoS pathogens* 5.10, e1000566.

Minh, Bui Quang et al. (2020). “IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era”. In: *Molecular biology and evolution* 37.5, pp. 1530–1534.

Nguyen, Nghia et al. (2022). “Shuffle & untangle: novel untangle methods for solving the tanglegram layout problem”. In: *Bioinformatics Advances* 2.1, vbac014.

Otto, Sarah Perin and Nick H Barton (1997). “The evolution of recombination: removing the limits to natural selection”. In: *Genetics* 147.2, pp. 879–906.

Page, Roderic DM (2003). *Tangled trees: phylogeny, cospeciation, and coevolution*. University of Chicago press.

Pérez-Losada, Marcos et al. (2015). "Recombination in viruses: mechanisms, methods of study, and evolutionary consequences". In: *Infection, Genetics and Evolution* 30, pp. 296–307.

Pinto, Rute M et al. (2021). "Accessory gene products of influenza A virus". In: *Cold Spring Harbor perspectives in medicine* 11.12, a038380.

Pulit-Penaloza, Joanna A et al. (2020). "Characterization of highly pathogenic avian influenza H5Nx viruses in the ferret model". In: *Scientific reports* 10.1, p. 12700.

Rambaut, Andrew et al. (2008). "The genomic and epidemiological dynamics of human influenza A virus". In: *Nature* 453.7195, pp. 615–619.

Rasmussen, David A and Fangfang Guo (2023). "Espalier: Efficient Tree Reconciliation and Ancestral Recombination Graphs Reconstruction Using Maximum Agreement Forests". In: *Systematic Biology* 72.5, pp. 1154–1170.

Revell, Liam J (2024). "phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things)". In: *PeerJ* 12, e16505.

Ryt-Hansen, Pia et al. (2021). "Co-circulation of multiple influenza A reassortants in swine harboring genes from seasonal human and swine influenza viruses". In: *Elife* 10, e60940.

Silva, U Chandimal de et al. (2012). "A comprehensive analysis of reassortment in influenza A virus". In: *Biology open* 1.4, pp. 385–390.

Simon-Loriere, Etienne and Edward C Holmes (2011). "Why do RNA viruses recombine?" In: *Nature Reviews Microbiology* 9.8, pp. 617–626.

Spielman, Stephanie J and Claus O Wilke (2015). "Pyvolve: a flexible Python module for simulating sequences along phylogenies". In: *PloS one* 10.9, e0139047.

Sutton, Troy C (2018). "The pandemic threat of emerging H5 and H7 avian influenza viruses". In: *Viruses* 10.9, p. 461.

Suzuki, Yoshiyuki (2010). “A phylogenetic approach to detecting reassortments in viruses with segmented genomes”. In: *Gene* 464.1-2, pp. 11–16.

Svinti, Victoria, James A Cotton, and James O McInerney (2013). “New approaches for unravelling reassortment pathways”. In: *BMC Evolutionary Biology* 13, pp. 1–14.

Taylor, Kishana Y et al. (2023). “Influenza A virus reassortment is strain dependent”. In: *PLoS pathogens* 19.3, e1011155.

Varsani, Arvind et al. (2018). “Notes on recombination and reassortment in multipartite/segmented viruses”. In: *Current opinion in virology* 33, pp. 156–166.

Vijaykrishna, Dhanasekaran, Reshmi Mukerji, and Gavin JD Smith (2015). “RNA virus reassortment: an evolutionary mechanism for host jumps and immune evasion”. In: *PLoS pathogens* 11.7, e1004902.

Wang, Haiwei et al. (2022). “Recombination in positive-strand RNA viruses”. In: *Frontiers in Microbiology* 13, p. 870759.

Whidden, Christopher, Norbert Zeh, and Robert G Beiko (2014). “Supertrees based on the subtree prune-and-regraft distance”. In: *Systematic biology* 63.4, pp. 566–581.

Wong, Yan et al. (2024). “A general and efficient representation of ancestral recombination graphs”. In: *Genetics* 228.1, iyae100.

World Health Organization (1980). “A revision of the system of nomenclature for influenza viruses: a WHO memorandum”. In: *Bulletin of the World Health Organization* 58.4, pp. 585–591.

Wu, Yufeng (2009). “A practical method for exact computation of subtree prune and regraft distance”. In: *Bioinformatics* 25.2, pp. 190–196.

Supplementary Figures

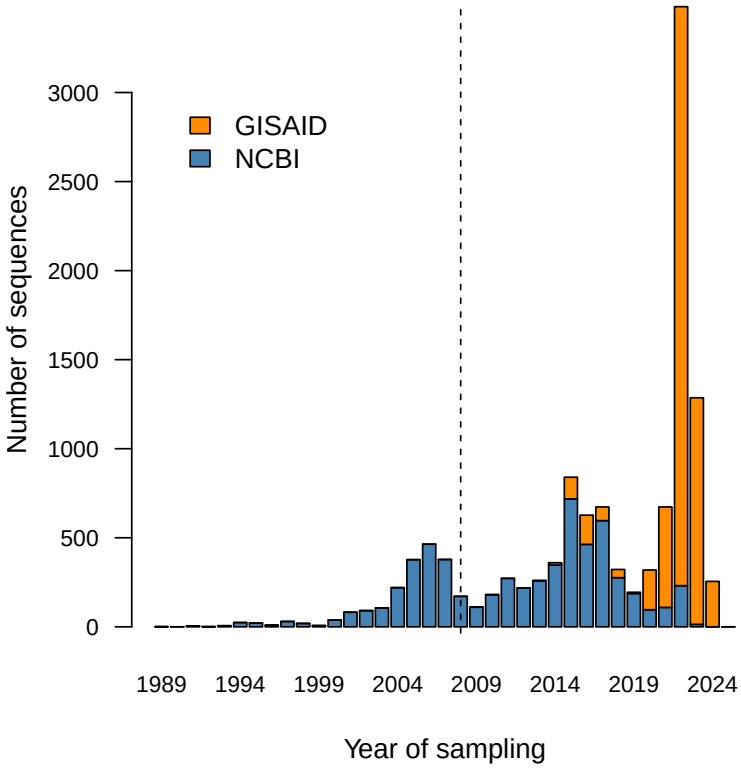

Figure S1: Distributions of sample collection dates for hemagglutinin (HA) sequences from the NCBI Genbank and GISAID databases.

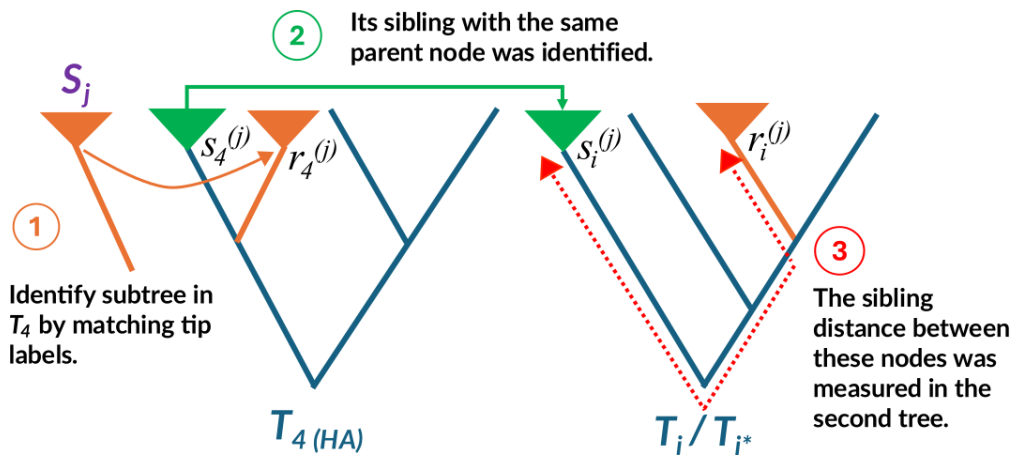

Figure S2: Diagram of the process for calculating the sibling distance of a subtree-prune-regraft (SPR) operation. For each subtree (S_j) in the maximum agreement forest, we identify the corresponding subtree in the segment 4 (HA) tree (T_4) by locating the internal node that is ancestral to the smallest set of tip labels matching all labels in S_j. Next, we locate the sibling internal node in T_4 ($s_4^{(j)}$) that shares an immediate common ancestor with $r_4^{(j)}$. We use the same label matching method to find the analogues of both $r_4^{(j)}$ and $s_4^{(j)}$ in a second tree relating either real (T_i) or simulated (T_{i*}) sequences of the i-th segment (where $i \neq 4$). Finally, we extract the total path length between these nodes ($r_i^{(j)}$ and $s_i^{(j)}$) as the sibling distance of the SPR.

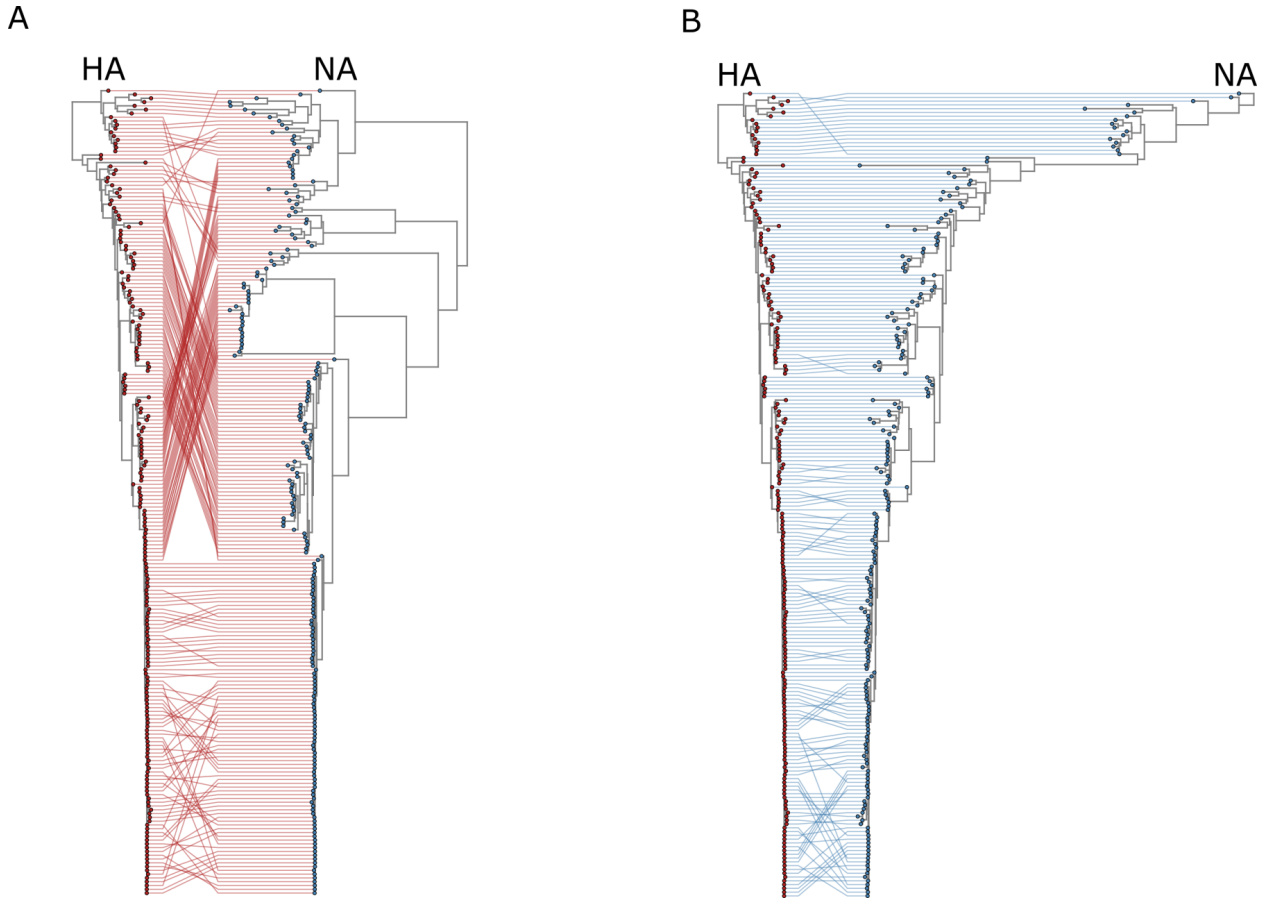

Figure S3: Tanglegram visualizing the correspondence between phylogenetic trees of the HA and NA segments of influenza viruses, generated with the Python module *baltic* (https://github.com/evogytis/baltic). (A) Tanglegram based on real data, showing the relationship and potential reassortment events between matched strains for each segment. (B) Tanglegram comparison of the HA tree real with the simulated scaled NA. This panel shows potential reassortment events that arise in the simulated data.