Bioinformatics, 38(8), 2022, 2144–2152
https://doi.org/10.1093/bioinformatics/btac075
Advance Access Publication Date: 12 February 2022
Original Paper

Phylogenetics

RF-Net 2: fast inference of virus reassortment and hybridization networks

Alexey Markin [1,*], Sanket Wagle[2], Tavis K. Anderson [1] and Oliver Eulenstein[2]

[1]Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA 50010, USA and [2]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on May 5, 2021; revised on January 26, 2022; editorial decision on January 28, 2022; accepted on February 7, 2022

Abstract

Motivation: A phylogenetic network is a powerful model to represent entangled evolutionary histories with both divergent (speciation) and convergent (e.g. hybridization, reassortment, recombination) evolution. The standard approach to inference of hybridization networks is to (i) reconstruct rooted gene trees and (ii) leverage gene tree discordance for network inference. Recently, we introduced a method called *RF-Net* for accurate inference of virus reassortment and hybridization networks from input gene trees in the presence of errors commonly found in phylogenetic trees. While RF-Net demonstrated the ability to accurately infer networks with up to four reticulations from erroneous input gene trees, its application was limited by the number of reticulations it could handle in a reasonable amount of time. This limitation is particularly restrictive in the inference of the evolutionary history of segmented RNA viruses such as influenza A virus (IAV), where reassortment is one of the major mechanisms shaping the evolution of these pathogens.

Results: Here, we expand the functionality of RF-Net that makes it significantly more applicable in practice. Crucially, we introduce a fast extension to RF-Net, called *Fast-RF-Net*, that can handle large numbers of reticulations without sacrificing accuracy. In addition, we develop automatic stopping criteria to select the appropriate number of reticulations heuristically and implement a feature for RF-Net to output error-corrected input gene trees. We then conduct a comprehensive study of the original method and its novel extensions and confirm their efficacy in practice using extensive simulation and empirical IAV evolutionary analyses.

Availability and implementation: *RF-Net 2* is available at https://github.com/flu-crew/rf-net-2.

Contact: alexey.markin@usda.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic analyses have made significant contributions toward our understanding of how genes, genomes and species have evolved. The application of these techniques are far-reaching, affecting conservation biology, agriculture, drug development, epidemiology and pandemic preparedness (Garten *et al.*, 2009; Grenfell *et al.*, 2004; Harris *et al.*, 2013; Jackson, 2004). However, when evolutionary processes such as hybridization, recombination and horizontal gene transfer are involved, a non-treelike evolutionary history emerges. Omitting these processes may lead to incorrect inference (Posada and Crandall, 2002; Woolley *et al.*, 2008) and can confound attempts to apply methods to solve real-world problems, e.g. reconstructing epidemic history and describing virus transmission dynamics (Leitner, 2019). Consequently, more complex evolutionary models in the form of phylogenetic networks have been developed that explicitly model reticulate evolutionary events [see reviews in Posada and Crandall (2001) and Huson and Bryant (2006)].

Phylogenetic networks model evolutionary history by adapting standard tree models (e.g. rooted binary trees) to also include *reticulation* events (Fig. 1, left). Reticulation nodes can capture hybridization and similar convergent processes (Huson *et al.*, 2010). Phylogenetic network inference is a rapidly growing field with recent success in application to important biological questions. For example, Wen *et al.* (2016) developed new reticulation network methods that quantified incomplete lineage sorting and introgression in the malaria vector. Recently, Markin *et al.* (2019) introduced a method that adapted the standard hybridization network problem to address erroneous input trees using the Robinson–Foulds distance. This method had the utility of being able to objectively embed input gene trees into reticulation networks inexactly, but up to a measurable error, and was able to infer phylogenetic networks

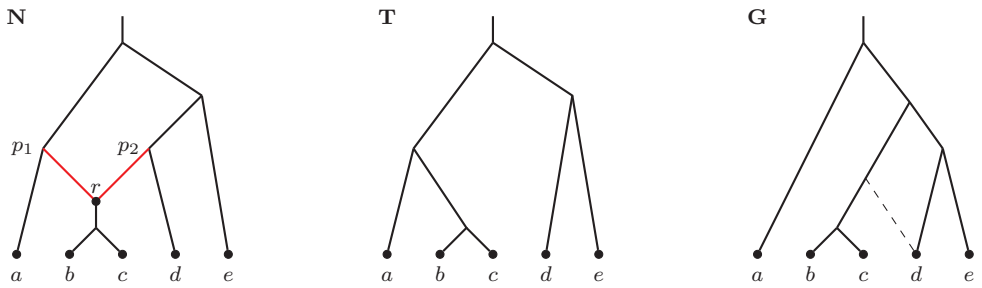

Fig. 1. From left to right: Network N, tree T displayed in N and tree G. N contains one reticulation vertex r with the entering reticulation edges shown in red. T is displayed in N by removing edge (p_2, r). While G is not directly displayed in N, N displays a local modification of G indicated by the dashed edge

accounting for errors in the input gene trees. Unfortunately, while the Markin *et al.* (2019) method, called *RF-Net*, can handle hundreds of taxa, it cannot address networks with large numbers of reticulations (e.g. more than 10); a limitation common to most other network inference methods (Markin *et al.*, 2019; Wen *et al.*, 2016).

The inability to model extensive reticulation in phylogenetic networks is particularly problematic in viruses. A reticulation process, frequently referred to as 'reassortment', has been reported to occur in at least 11 RNA virus families, and is a major driver in the evolution of pathogens such as influenza A virus (IAV), rotavirus A and bluetongue virus (McDonald *et al.*, 2016). In these viruses, reassortment occurs when two single-stranded segmented RNA viruses infect the same cell and exchange complete gene segments (McDonald *et al.*, 2016). In an observational study that tracked gene combinations in H1 swine IAV, and consequently the minimum number of reassortments, Gao *et al.* (2017) detected more than 70 unique genome constellations. Similarly, Rajão *et al.* (2017) identified more than 44 unique genome constellations in swine H3 IAV. Though many of these reassorted viruses appear to be infrequently detected, when viral progeny contain genes derived from more than one parent, it may confer important fitness advantages (Anderson *et al.*, 2021; Crisci *et al.*, 2013). The most notable example of how a reassorted virus can have a far-reaching impact is the 2009 H1N1 influenza pandemic virus. This lineage of IAV contained gene segments with evolutionary origins from bird, human and swine IAV (Smith *et al.*, 2009). Consequently, given the importance of reassortment in virus biology, techniques that identify reassorted genomes and their evolutionary history are critical.

Related work. A major focus in the phylogenetic network literature is on the inference of hybridization networks [cf. a review in Elworth *et al.* (2018)]. The original approach to hybridization network inference was formulated by Baroni *et al.* (2005), which leveraged gene tree discordance. The original Baroni *et al.* work was followed by multiple algorithmic advancements (cf. Albrecht, 2016; Iersel *et al.*, 2014; Whidden *et al.*, 2013). Further, Meng and Kubatko (2009) and Yu *et al.* (2012) introduced the multi-species network coalescent (MSNC) model that combined the hybridization framework of Baroni *et al.* with the classic multi-species coalescent model (Kingman, 1982), which accounts for incomplete lineage sorting in the gene trees. This led to the development of multiple network inference approaches based on MSNC (cf. Solís-Lemus and Ané, 2016; Yu *et al.*, 2012; Yu and Nakhleh, 2015). In addition, Yu *et al.* (2013) proposed a parsimony method, MP-PhyloNet, for the inference of hybridization networks in the presence of incomplete lineage sorting.

Recently, RF-Net (Markin *et al.*, 2019) extended the Baroni *et al.* hybridization framework to account for errors in the input gene trees (Steel and Rodrigo, 2008). Accounting for errors is crucial in many biological systems, where evolutionary inference is error-prone, e.g. (Hahn, 2007; Rasmussen and Kellis, 2011) and the classic incomplete lineage sorting model (the multi-species coalescent) is not entirely applicable, e.g. RNA viruses. To accommodate potential errors in gene trees, Markin *et al.* (2019) introduced the concept of *embeddings* and *embedding cost*. Given a gene tree G and a network N, an embedding of G in N is a tree G' that is *displayed* in N, which is most similar to G (i.e. G' can be obtained from N by removing some reticulation edges). If N is the *true* phylogenetic network, then we can expect the embedding G' to be the error-corrected version of G. The embedding cost is then the Robinson–Foulds distance between G and its embedding. RF-Net uses the sum of embedding costs over all gene trees as its objective function to perform local search on the space of phylogenetic networks. The local search is driven by the SubNet Prune and Regraft (SNPR) edit operation introduced by Bordewich *et al.* (2017). Note that SNPR was proven to 'connect' the space of phylogenetic networks (Bordewich *et al.*, 2017; Janssen *et al.*, 2018).

Our contribution. In this work, we expand the functionality of RF-Net (Markin *et al.*, 2019) to advance its practical applicability. We develop the *Fast-RF-Net* approach that is not bound to a limited number of reticulations, which resolves a major drawback in the base RF-Net method. The core idea behind Fast-RF-Net is that, as RF-Net adds a new reticulation to a candidate phylogenetic network, different genes are likely to 'favor' one reticulation path to another based on the respective gene tree topology. As the search gradually progresses, we hypothesize that these preferences are unlikely to change, as each added reticulation improves the fitness of the network. This implies that we do not have to recompute gene tree embeddings 'from scratch' for every new candidate network, but we can re-use the previous embeddings.

Second, we address a common problem in phylogenetic network inference and develop an approach that can determine the appropriate number of reticulations for a given dataset (Cai and Ané, 2021; Hejase *et al.*, 2018). In Section 3.2, we propose an automated stopping criterion for RF-Net that does not require a user-specified number of reticulations. In addition, RF-Net can plot the dynamic of the fitness function (embedding cost) as it changes with the number of reticulations, which provides practitioners a visualization to aid in preventing over-fitting and selecting the best inferred network (Hejase *et al.*, 2018).

We then conduct a comprehensive study of RF-Net and its novel components on (i) a small validation IAV dataset with clear patterns of reassortment (Section 4); (ii) simulated data (Section 5); and (iii) a large swine H3 IAV dataset with 429 whole genome strains and highly incongruent gene evolution (at least 9 unique gene constellations)—see Section 6. Crucially, the simulation study demonstrated that both RF-Net and Fast-RF-Net were robust to gene tree estimation errors caused by short gene sequences with low phylogenetic signal.

2 Background

In this section, we introduce the preliminaries and the core optimization problem.

2.1 Preliminaries

Phylogenetic networks. A *(phylogenetic) network* is a directed acyclic graph (DAG) without parallel edges. It has a designated root vertex of in-degree zero and out-degree one and all other vertices are either of in-degree one and out-degree two (*tree vertices*), in-degree two and out-degree one (*reticulation vertices*) or in-degree one and

out-degree zero (*leaves*). An example network with one reticulation vertex r is depicted in Figure 1 (left).

Let N be a network, then its vertices, edges, root and leaves are denoted by $V(N)$, $E(N)$, $\rho(N)$ and $\mathsf{L}(N)$, respectively. The edges in $E(N)$ are distinguished by the edges that are entering (i) reticulation vertices (*reticulation edges*) and (ii) tree vertices or leaves (*tree edges*). A *tree-path* in N is a directed path that consists only of tree edges.

A vertex $v \in V(N)$ is a *descendant* of $w \in V(N)$ when there is a directed path from w to v (we consider each vertex to be a descendant of itself). A *(hardwired) cluster* of vertex v, C_v, is the set of leaves that are descendants of v. Finally, $\mathcal{C}(N)$ is the set of all clusters in N.

Phylogenetic trees. A *(phylogenetic) tree* T is a network with no reticulation vertices. Given a set $L \subseteq \mathsf{L}(T)$, $T|_L$ denotes a restriction of T to the leaf-set L. That is, $T|_L$ is obtained from the minimal subtree T' of T, which connects all leaves in L and the root, by suppressing all out-degree one nodes except for the root.

Displayed trees. Tree T is *displayed* in network N (with the same leaf set) if one can remove exactly one reticulation edge from each reticulation node, remove all potentially appearing non-labeled vertices with out-degree zero, and obtain a subdivision of T. Figure 1 demonstrates an example of tree T (middle) displayed in network N (left).

The set of all trees displayed by N is denoted by $\mathcal{P}_N$.

Tree-child networks. A network is called *tree-child* if each non-leaf vertex has at least one outgoing tree edge (i.e. a child that is a tree-vertex). Observe that each vertex in a tree-child network must have a tree-path going to some leaf.

Robinson–Foulds (RF) distance. The Robinson–Foulds distance between two trees is defined as the symmetric difference of their cluster representations (Robinson and Foulds, 1981). In particular, $RF(G,T) := |(\mathcal{C}(G) \setminus \mathcal{C}(T)) \cup (\mathcal{C}(T) \setminus \mathcal{C}(G))|$.

When one of the trees has an incomplete set of taxa, we apply the standard *minus-method* approach (Cotton and Wilkinson, 2007). That is, if $\mathsf{L}(G) \subset \mathsf{L}(T)$, we have $RF(G,T) := RF(G, T|_{\mathsf{L}(G)})$.

2.2 RF-networks

We now briefly re-iterate the definition of RF-Networks and the respective optimization problem from (Markin *et al.*, 2019). We first introduce our core 'fitness' function for phylogenetic networks and then formulate the respective optimization problem.

Embeddings. To score the 'fitness' of network N against tree G, we need to consider the trees displayed in N. In particular, we define the *embedding cost* as follows (recall that $\mathcal{P}_N$ denotes the set of all trees displayed in N)

$$\delta(G,N) := \min_{T \in \mathcal{P}_N} RF(G,T).$$

A tree T displayed in N with $RF(G,T) = \delta(G,N)$ is called an *embedding* of G in N. Note that the leaf-set of G should be a subset of the leaf-set of N.

As an example, consider Figure 1. T in that example is displayed in N and therefore $\delta(T,N) = 0$. At the same time, tree G is not displayed in the network, while a small modification of G indicated using the dashed edge is displayed; let us denote this modified tree as G'. It is then not difficult to see that $\delta(G,N) = RF(G,G') = 2$.

Then, for a set of input trees $\mathcal{G}$ and a network N (with $\mathsf{L}(G) \subseteq \mathsf{L}(N)$ for all $G \in \mathcal{G}$) the *total embedding cost* is the sum of individual embedding costs. That is, $\delta(\mathcal{G},N) := \sum_{G \in \mathcal{G}} \delta(G,N)$.

Markin *et al.* (2019) demonstrated that computing $\delta(G,N)$ is NP-hard even when N is tree-child.

Optimization problem. We now formulate our core (NP-Hard) optimization problem that we approach with RF-Net (Markin *et al.*, 2019).

Problem 1. RF-Network
Input: Set of input trees $\mathcal{G}$ and maximum number of reticulations r;
Output: Find network N with at most r reticulations minimizing the total embedding cost $\delta(\mathcal{G},N)$. N should contain all leaves (taxa) from the input trees

3 Materials and methods

Markin *et al.*, (2019) proposed a 'layered' approach to RF-Network inference. In particular, we call a set of all networks with exactly k reticulations a *kth layer*, denoted by $\mathcal{N}_k$. Then the optimization search starts with the 0th layer (i.e. tree layer) and then incrementally explores each higher layer until it reaches the user-specified upper bound on the number of reticulations, r. More precisely, the method first finds a locally optimal supertree N^0 for the input gene trees. Note that when N is a tree, the embedding cost is the Robinson–Foulds distance. We use the highly efficient RF Supertree method by Bansal *et al.* (2010) for this step.

Next, RF-Net repeats the following procedure until it reaches the layer with r reticulations (or the total embedding cost is reduced to 0). It starts with $k = 0$ and we denote a starting network for layer k by N^k.

i. Add a reticulation to N^k in a way that minimizes the total embedding cost. Let N^{k+1} denote the resulting network. If there are multiple optimal ways to add a reticulation, RF-Net randomly chooses p optimal N^{k+1} networks and performs step (ii) for each of them sequentially. Here, p is a user-specified parameter.
ii. Explore layer $\mathcal{N}_{k+1}$ using SNPR edit operations (Bordewich *et al.*, 2017), starting with an N^{k+1} network. At each local search iteration, if there are multiple optimal networks in the SNPR neighborhood, RF-Net proceeds by choosing one of them uniformly at random.

Note that RF-Net implements two major algorithmic advancements as described in Markin *et al.* (2019). These algorithms (i) make the computation of the total embedding cost for each network significantly faster, and (ii) accelerate SNPR neighborhood traversals for each local search iteration. Finally, RF-Net can constrain the network search space to tree-child networks only. As in practice, many reticulate histories can be expected to be tree-child (Cardona *et al.*, 2009; Markin *et al.*, 2019), this setting can noticeably reduce the search space and the runtime of RF-Net. That is, on each local search iteration, RF-Net will only compute embedding costs for the tree-child subset of all neighboring networks.

3.1 Fast-RF-Net: scaling to highly entangled datasets

Despite the effective optimizations proposed in Markin *et al.* (2019), the computational difficulty of calculating the RF embedding cost prevents RF-Net from inferring networks with more than 10–11 reticulations in a reasonable time (for larger datasets with 300–500 taxa). To overcome this in practice stringent limitation, we now describe a novel fast extension to RF-Net that can scale to an arbitrary number of reticulations.

Let $n_1, n_2, \ldots, n_k$ denote the reticulation vertices in network N^k in the kth layer. For each input tree G, RF-Net computes an embedding in N^k. That is, for each reticulation vertex n_i, it chooses one of its parent-edges for the embedding. We can then represent an embedding for G as $(n_1/p_1, n_2/p_2, \ldots, n_k/p_k)$, where $p_i \in \{1,2\}$ represents whether the 'first' or the 'second' parent of n_i is chosen (or left/right).

Then during the traversal of the SNPR neighborhood of N^k, to evaluate a network N', Fast-RF-Net uses the same $(n_1/p_1, n_2/p_2, \ldots, n_k/p_k)$ embedding for G instead of recomputing the optimal embedding from scratch. That way, the approximate embedding cost for N' can be computed in $O(|\mathsf{L}(N')| \cdot |\mathcal{G}|)$ time instead of $O(2^k |\mathsf{L}(N')| \cdot |\mathcal{G}|)$ time required by RF-Net.

Similarly, when adding a new reticulation vertex n_{k+1}, Fast-RF-Net maintains the previous parent assignments $(n_1/p_1, n_2/p_2, \ldots, n_k/p_k)$ for G and then chooses the best parent assignment for n_{k+1} between $n_{k+1}/1$ and $n_{k+1}/2$ by computing the respective RF distances.

In summary, Fast-RF-Net eliminates the exponential factor 2^k from the RF-Net's time complexity as a trade-off for potentially considering sub-optimal embeddings. We validated this approach in the simulation study (see Section 5) and showed that Fast-RF-Net generally matches the accuracy of the original RF-Net method.

3.2 Determining the number of reticulations

It is generally hard (or even infeasible) to estimate the exact number of hybridization/reassortment events that took place in the past for a particular dataset. Therefore, to assist practitioners with model selection and determining the proper number of reticulations, we implemented an automated heuristic and a visual tool as described below. The visual tool plots total embedding costs; i.e. costs decrease as the number of reticulations grows. This plot can identify where RF-Net stops picking on the main reassortment signal, and instead starts 'over-fitting' based on the topological error in gene trees (noise).

Automated stopping criteria. Following the similar idea to our graphical visualization of decrease in the embedding cost, we propose a stopping criterion to determine the appropriate number of reticulations heuristically.

Let C_0 denote the initial embedding cost (the total RF distance) for N^0. We then introduce a (percentage) parameter $0 \leq t < 100$ and require a difference between the embedding costs for two neighboring layers to be at least $\frac{t}{100} \cdot C_0$.

More formally, let F_k denote the final network for a kth layer. Then $\delta(\mathcal{G}, F_{k-1}) - \delta(\mathcal{G}, F_k) \geq \frac{t}{100} \cdot C_0$ for all layers that RF-Net reports in the output. Once the improvement on the next layer is lower than the threshold, the search terminates. From now, on we will refer to t as a *reticulation threshold*.

3.3 Error-corrected gene trees

Computation of the embedding costs allows RF-Net to account for errors in the input gene trees. Consequently, in addition to a reassortment/hybridization network, RF-Net can produce the respective embeddings of the gene trees into the network. That is, RF-Net can output the error-corrected gene trees, which we empirically confirm in the simulation study (see Section 5).

When the embedding of an input tree into the computed network is not unique, RF-Net outputs a strict consensus of all embeddings for such input tree. On the other hand, Fast-RF-Net maintains a single embedding structure throughout the entire search and, therefore, outputs that particular embedding.

4 Validation of RF-Net with swine IAV data

To evaluate the efficacy of RF-Net at inferring phylogenetic networks and accurately recreating reassortment events, we apply it to a small dataset of swine IAVs that include strains with evidence for reassortment based upon gene tree incongruence. We collated hemagglutinin (HA) and neuraminidase (NA) gene sequences for 22 swine IAV strains collected in the US as part of a national surveillance program (Anderson *et al.*, 2013). These data included 21 strains from the H1N2 subtype within the delta-1a HA genetic clade paired with a N2-1998 lineage or N2-2002 lineage NA gene. We included a single H1N1 strain that had the same delta-1a HA genetic clade, but was paired with a Classical swine lineage N1 NA gene. We aligned each gene dataset in mafft v7.453 (Katoh *et al.*, 2005), and inferred maximum likelihood gene trees for each alignment using FastTree v2.1.11 (Price *et al.*, 2010). We then rooted the inferred trees with TreeTime v0.7.5 (Sagulenko *et al.*, 2018).

The NA gene tree, Figure 2, demonstrates three different clades represented by the branch colors (red, green and blue) reflecting the two subtypes and three lineages. The HA gene tree reflects a single lineage, but when the branches are colored by NA gene, distinct reassortment patterns are evident. While the H1N1 strain (blue) is nested within the HA gene tree, it is a distinct lineage in the NA tree. In the NA gene tree, we note two large monophyletic clades reflecting the two distinct N2 lineages, N2-1998 versus N2-2002: in the HA tree, the N2-1998 green strains are nested within N2-2002 red strains. This topological gene tree incongruence supports the proposition that at least two reassortment events have occurred during the evolution of these swine IAV strains (Boni *et al.*, 2010).

We subsequently conducted an RF-Net analysis to infer a network with two reassortments with the rooted HA and NA gene trees as input (Fig. 3). RF-Net correctly inferred the reassortment event at the H1N1 strain, and was also able to infer the reassortment event demonstrated by the N2-1998 green strains, with a dashed line indicating the acquisition of this novel N2-1998 gene. Hence, we observe that RF-Net is able to recreate expected reassortment events with a high degree of resilience to small errors or inconsistencies in the input trees (e.g. derived from sequence error, alignment error or lack of phylogenetic signal in highly similar IAV gene sequences).

5 Simulation study

We evaluate RF-Net and its novel extensions in a simulation study using the credible model of IAV evolution from Müller *et al.* (2020). In this study, we deliberately use erroneous gene trees to evaluate RF-Net's ability to perform error-correction.

5.1 Simulation setup

Recently, Müller *et al.* (2020) introduced a statistical model of IAV evolution by augmenting the classical coalescent model (Kingman, 1982) with reassortment. We apply their model, referred to as *coalescent with reassortment (CoalRe)*, to simulate reticulation networks and respective gene trees following the process described in Müller *et al.* (2020). We also simulated sequence alignments given the true gene trees, estimated (erroneous) gene trees from the alignments, and executed RF-Net on the estimated gene trees. We evaluated RF-Net and its extensions against MP-PhyloNet (Yu *et al.*, 2013) and CoalRe. As we evaluate CoalRe on the same evolutionary model that it optimizes, it serves as an optimal baseline in this study.

Simulating networks and gene trees. The CoalRe model introduced a parameter referred to as the *rate of reassortment* (per lineage per year). That is, in CoalRe, each lineage (going backward in time) has a fixed probability, ρ, of becoming a product of reassortment. In addition, CoalRe uses effective population size as a standard parameter in coalescent models.

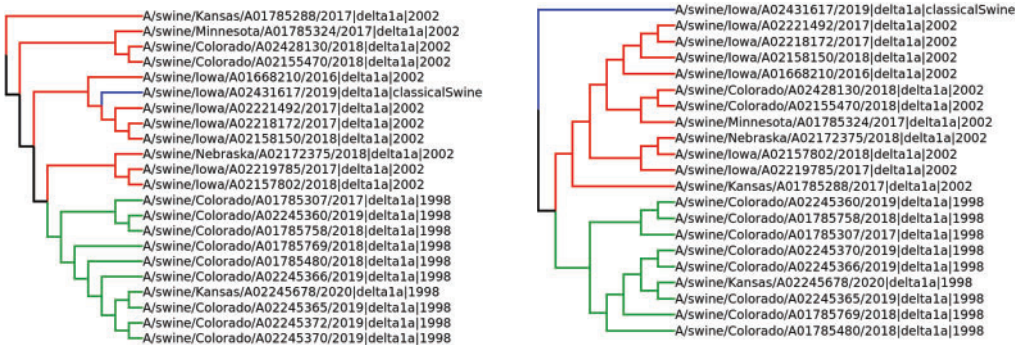

Fig. 2. Evolutionary history of swine IAV H1 delta-1a HA genes (left) with paired N1 or N2 genes (right). IAV strains are annotated by strain name, HA genetic clade and NA genetic clade. The NA genetic clades are highlighted by blue, red and green colors to represent N1-Classical swine, N2-2002 and N2-1998 clades, respectively

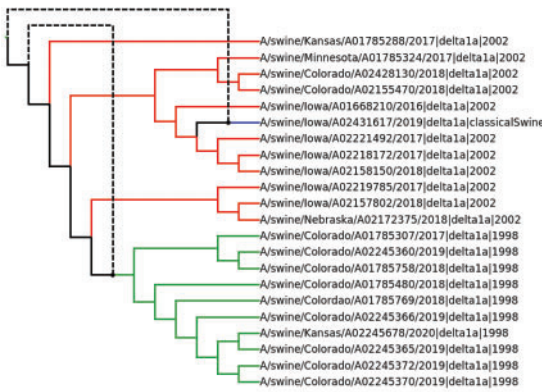

Fig. 3. RF-Net network demonstrating two inferred reassortment events represented by dashed lines. RF-Net correctly inferred the novel HA and NA gene pairing of the H1N1 strain, and detected a reassortment within the H1 delta-1A lineage gene where reassortment resulted in a novel gene pairing of H1 delta-1a HA with a N2-1998 gene versus the dominant N2-2002 gene

We then use the following parameters in CoalRe simulations: (i) 50 and 75 for the number of taxa; (ii) a normal distribution with mean 0.05 and standard deviation 0.01 for the rate of reassortment; (iii) a normal distribution with mean 5 and standard deviation 0.1 for the population size; (iv) a uniform distribution over the $[0, 10]$ interval as the sampling time (in years) for the taxa. Finally, we set the number of genes to 8, matching the number of genes in IAV. That is, each reassortment network is accompanied by eight gene trees displayed in it. Note that we choose the mean value of five for the population size, as it corresponds to the population size estimates for human IAVs by Müller *et al.* (2020). The chosen reassortment rate resulted in networks with the number of reassortments between 0 and 12. The average number of reassortments was 4.4 for instances with 50 taxa and 5 for instances with 75 taxa.

In addition, we simulated larger networks with 200 taxa and a higher mean reassortment rate of 0.1. The resulting networks had between 8 and 30 reticulations (18.6 on average). We generated 25 instances under each model condition (75 instances overall).

Sequence alignments and estimated gene trees. To model gene tree estimation error, we generated nucleotide sequence alignments of length 1000 along the edges of the true gene trees. The substitution rates were sampled from a normal distribution with mean 0.005 and standard deviation 0.001 independently for each gene tree. The mean substitution rate of 0.005 aligns with the estimates in Zeller *et al.* (2020) for swine influenza HA genes. The substitutions were driven by the Generalized Time-Reversible (GTR) model (Tavaré *et al.*, 1986) with 4 discrete Gamma categories (Yang, 1994). The GTR+Gamma model parameters were sampled from realistic distributions estimated from empirical IAV data (see Supplementary Section S1 for more detail). With the simulated sequence alignments, we inferred gene trees using RAxML v8.2.12 (Stamatakis, 2014) implementing a GTR+Gamma substitution model. Each inferred gene tree was rooted using TreeTime v0.8.4 (Sagulenko *et al.*, 2018). Importantly, each process in this analysis—though standard in phylogenetics (Dereeper *et al.*, 2008; Kapli *et al.*, 2020)—unintentionally introduces a substantial number of errors to the estimated gene trees. These errors were due to the limited number of sites in the alignment and uncertainty at the rooting step.

All simulations were carried out in BEAST 2 (Bouckaert *et al.*, 2014) with the CoalRe package.

Methods. First, we evaluated Fast-RF-Net and the automated stopping criteria extensions against the base RF-Net method, MP-PhyloNet (Yu *et al.*, 2013) and CoalRe (Müller *et al.*, 2020) on smaller instances with 50 and 75 taxa.

We ran six variations of RF-Net on each of the simulated instances using the estimated (erroneous) gene trees as input. Specifically, we conducted runs with:

(i) RF-Net with the true number of reassortments taken from the true network;

(ii) RF-Net with an automated stopping criterion—5% reticulation threshold (see Section 3.2);

(iii) RF-Net with 3% reticulation threshold;

(iv–vi) Fast-RF-Net with the same three stopping criteria.

For convenience, we denote the above six variations of the method as *RF-Net, t5, t3, Fast-RF-Net, Fast-t5, Fast-t3*, respectively. Here, t5 and t3 refer to the employed 5% and 3% reticulation thresholds, respectively.

MP-PhyloNet is a popular method for inference of hybridization networks. It was set to use the estimated trees and the true number of reassortments as input. To obtain better results we increased the number of samples per iteration performed by MP-PhyloNet to 500 from the default 100. *CoalRe* was executed in BEAST with true priors on the reassortment rate [i.e. $\mathcal{N}(0.05, 0.01^2)$ for smaller instances and $\mathcal{N}(0.1, 0.01^2)$ for large instances with $n = 200$ taxa] and GTR with 4-category Gamma model for substitutions. The MCMC chain length was set to 1 000 000.

Second, we executed the two most scalable methods, Fast-RF-Net and CoalRe, on the large instances with 200 taxa and 19 reticulations on average.

All methods were given at most 2 h to complete on a single instance for smaller datasets with 50 and 75 taxa. For instances with 200 taxa, we restricted the runtime to at most 7 h, which was sufficient for Fast-RF-Net and CoalRe to complete. This study was conducted on the USDA-ARS SCINet Ceres high-performance computing cluster https://scinet.usda.gov. Each method was given a single 2.4 GHz core with 8 GB of RAM.

5.2 Simulation results

Network reconstruction accuracy. We evaluate the methods' accuracy in terms of similarity between the reconstructed and true simulated networks under the generalized Robinson–Foulds (gRF) measurement (Cardona *et al.*, 2009). In particular, we compute the similarity between two networks by calculating the number of *hardwired* clusters they have in common and then normalizing that value by the total number of hardwired clusters in the networks. The similarity value of one then represents the maximum accuracy.

Figure 4 shows the accuracy (similarity to the true network) of RF-Net, CoalRe and MP-PhyloNet under all model conditions. We observe that CoalRe was generally the most accurate method with a slight edge over RF-Net and its variations on smaller datasets with up to 75 taxa. This outcome was expected, as the simulations were carried out under the model of reassortment implemented in CoalRe. However, we note that Fast-RF-Net nearly matched the accuracy of CoalRe on the large instances with 200 taxa and up to 30 reticulations (Fast-RF-Net and CoalRe had an average accuracy of 87.6% and 88.3%, respectively). Fast-RF-Net had strictly higher accuracy than CoalRe on 10/25 instances with 200 taxa.

Further, we observed that Fast-RF-Net accuracy generally matched the base RF-Net accuracy on all instances where both methods were tested. Overall, all tested extensions of RF-Net performed very similarly to the base RF-Net method, which supports their applicability.

MP-PhyloNet was not able to complete on all instances within 2 h: it failed on 6/25 instances with 50 taxa and 8/25 instances with 75 taxa. For those instances, where MP-PhyloNet completed, it produced networks with at most 5 reticulations on instances with 50 taxa and 4 reticulations on instances with 75 taxa. On average, MP-PhyloNet performed worse than RF-Net and CoalRe and had the broadest variation in accuracy.

Evaluating RF-Net error-correction. We evaluated the embeddings (see Section 3.3) of the input trees into an inferred network. Our main goal was to see whether the RF-Net embeddings were closer to the true simulated gene trees than the erroneous estimated gene trees. That is, we wanted to verify that RF-Net performs error-correction on the input trees.

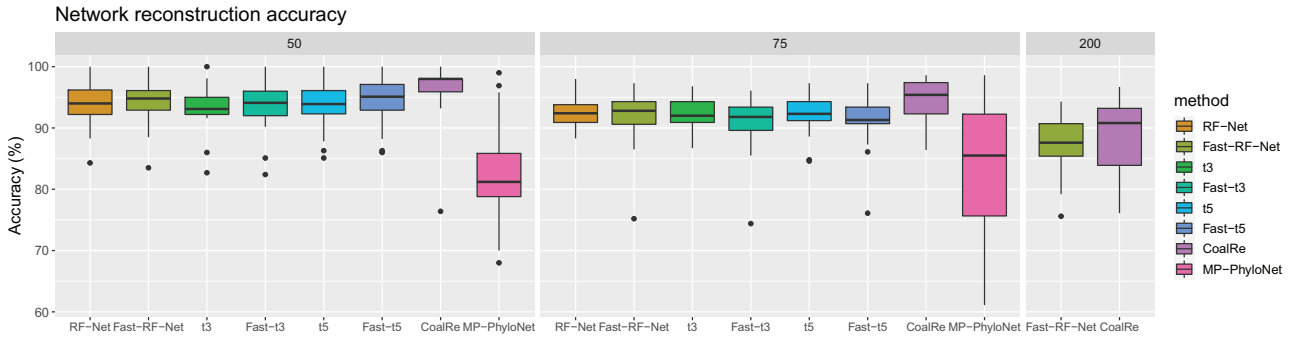

Fig. 4. The precision of network inference methods, in terms of the similarity of the reconstructed networks with the respective true simulated networks, for 50 taxa (left), 75 taxa (middle) and 200 taxa (right)

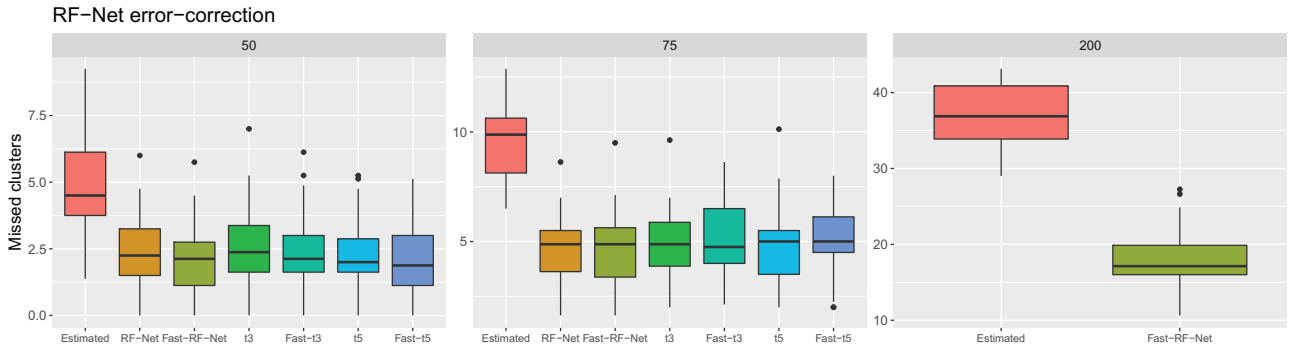

Fig. 5. Number of missed clusters (false negatives) in the estimated gene trees and in the embeddings of the gene trees computed by variants of RF-Net

The standard approach to measure error in trees is to count the number of incorrect clusters (false positives) and the number of missing clusters (false negatives). In our case, the true simulated gene trees are always fully resolved (binary), whereas the embeddings cannot be fully resolved. In these circumstances, the number of false negatives is always at least as large as false positives. Therefore, we only considered the false negatives (as the worst case).

Figure 5 summarizes the performance of RF-Net in terms of error-correction, where errors are measured as the number of missing clusters. First of all, we observed that all variants of RF-Net effectively performed error-correction by significantly reducing the number of errors in the estimated gene trees. In fact, all variants reduced the number of errors in the estimated gene trees by more than 45% on average for all instances. For example, RF-Net with a reticulation threshold of 5% (t5) on average eliminated 59.4% of errors for 50 taxon instances and 52.0% of errors for 75 taxon instances. Finally, similarly to Figure 4, we did not observe significant differences among the RF-Net variations in terms of error-correction: Fast-RF-Net and RF-Net coupled with automated stopping criteria matched the performance of the base RF-Net method.

5.3 Discussion

The simulations showed that Fast-RF-Net matched the accuracy of the base RF-Net method and was close to the accuracy of CoalRe, which implements the underlying reassortment model we used for simulations. Note that the CoalRe reassortment model makes several simplifying assumptions that might not hold in practice. For example, CoalRe assumes a constant population size of the influenza viruses and unconstrained interactions among them. In contrast, it is known that, e.g. swine IAV is not homogeneously geographically distributed—different countries (or even individual swine farms) often have unique influenza strains and influenza dynamics (Anderson *et al.*, 2021; Diaz *et al.*, 2017). This gives RF-Net a potential advantage, as it is not constrained to a particular model of evolution and can be more flexibly applied, e.g. to hybridization problems and complex reassortment problems.

RF-Net and Fast-RF-Net generally outperformed MP-PhyloNet in this simulation study. We observed that MP-PhyloNet typically outputs networks with a smaller number of reticulations than specified in the input, and it was able to compute networks with at most 5 reticulations on instances with 50 taxa. This shows that MP-PhyloNet is unlikely to be applicable in studies with large numbers of taxa and many hybridization/reassortment events. Other popular methods that infer hybridization networks [SNaQ (Solís-Lemus and Ané, 2016), ML-PhyloNet (Yu *et al.*, 2014), MPL-PhyloNet (Yu and Nakhleh, 2015) and SpeciesNetwork (Zhang *et al.*, 2018)] were previously shown to be only applicable to datasets with less than 30 taxa and a handful of reticulations (Hejase and Liu, 2016; Zhang *et al.*, 2018). In contrast, we show that Fast-RF-Net can infer accurate networks and perform error-correction on instances with 200 taxa and up to 30 reticulations. Fast-RF-Net runtime scales linearly with the number of gene trees, which makes it potentially applicable for large hybridization studies spanning hundreds or even thousands of genes. As a reference point, Fast-RF-Net took at most four minutes (two on average) to complete on any dataset with 75 taxa and 8 genes. Further, assuming that the number of iterations that Fast-RF-Net performs on each layer is constant, its runtime also scales linearly with the number of reticulations.

6 Empirical study

The observed success of many pathogens likely depends on the coordinated function of combinations of genes. In IAV this is exemplified by the interplay between the two surface proteins—the hemagglutinin (HA) and neuramindase (NA). These genes function in a coordinated effort to replicate within and transmit between hosts: balance in NA and HA cellular interactions are necessary for IAV infection to result in successful host-to-host transmission (Das *et al.*, 2011; Mitnaul *et al.*, 2000; Neverov *et al.*, 2014). In addition to these links, there is the potential for synergism with the remaining six genes in IAV. Consequently, understanding the evolutionary dynamics of IAV across all genes will inform and improve control of the

virus in swine, and help minimize the risk of swine-origin IAV being transmitted to humans.

In this empirical example, we focus on an H3.2010.1 swine IAV lineage derived from a human-seasonal H3 introduction to swine (Anderson *et al.*, 2021; Rajão *et al.*, 2015). Generally, human IAV infection in swine results in low replication and rare pig-to-pig transmission, but some human-origin IAV become endemic, and this is typically associated with genetic differences from the precursor strain (Powell *et al.*, 2021; Rajão *et al.*, 2018), or reassortment with endemic host-adapted viruses. The H3.2010.1 lineage follows this paradigm, and assessing reassortment in this lineage allows us to generate testable hypotheses on how expansion in the genomic diversity of IAV in swine populations may facilitate IAV capable of spillover from swine to humans.

Data collection. To study the evolution of the H3.2010.1 virus lineage, we downloaded all swine H3N2 complete genomes ($n = 1563$) from the Influenza Research Database (Zhang *et al.*, 2017). The genomes were separated into the eight constituent genes, and the genetic clade and evolutionary lineage was inferred using the octoFLU pipeline (Chang *et al.*, 2019). Genomes that contained a H3.2010.1 HA gene were extracted, and strains were annotated by NA and evolutionary lineage for the six internal genes. Strains that reflected single outbreak events at agricultural fairs (Bowman *et al.*, 2017) were removed from analyses resulting in 429 H3 whole genomes. An outgroup of a dead-end human-to-swine H3 spillover was included (A/swine/Guatemala/CIP049-IP040078/2010). To generate the required input trees for RF-Net, the eight genes were aligned using mafft v7.453 (Katoh *et al.*, 2005), and maximum likelihood phylogenetic trees were inferred following automatic model selection using IQ-TREE (Nguyen *et al.*, 2015). Rooting was inferred using TreeTime (Sagulenko *et al.*, 2018) and gene tree topology was consistent with the IAV gene classifications (Chang *et al.*, 2019); the only gene that did not have an apparent rooting was NS, as it did not have sufficient phylogenetic signal due to high sequence similarity across the relatively short gene. RF-Net relies upon tree topology; consequently, the NS gene was excluded from network inference.

Results and Discussion. The expanded functionality of RF-Net enabled us to infer the evolutionary history of a virus that is shaped by clonal and non-clonal processes. Given the frequency of reassortment in IAV, methods that do not consider reticulation processes may result in error if there is a reliance on single-gene inference. In analyzing our H3 swine IAV data, we were able to track the evolution of a major H3 lineage in swine as it reassorted that coincided with the genetic lineage becoming a major component of swine IAV diversity (Zeller *et al.*, 2018).

Our analysis recapitulates the three major reassortment events as this IAV lineage evolved (Powell *et al.*, 2021; Rajão *et al.*, 2015). The initial detection (A/swine/Missouri/A01476459/2012) contained a human seasonal H3 hemagglutinin (HA), N2 neuraminidase (NA) and internal genes from the 2009 pandemic H1N1 (H1N1pdm09). The second reassortment event occurred as the human N2-NA was replaced by a classical swine N1-NA. The third major reassortment event exchanged the N1-NA to an N2-NA derived from endemic swine 2002 N2 genes, the inclusion of the Matrix (M) gene from H1N1pdm09, with the remaining internal genes associated with the endemic swine triple reassortant internal gene (TRIG) constellation (Powell *et al.*, 2021). Given the known minimum number of reassortment events in this IAV lineage, our method adequately recreates its evolutionary history. The phylogenetic networks computed by RF-Net, with over 400 strains, are at https://github.com/flu-crew/rf-net-2 and may be interactively explored with IcyTree (Vaughan, 2017).

A significant feature of the expanded RF-Net tool is the exploration of networks with different numbers of reticulations r and the inclusion of an automatic stopping criterion (t). We explored networks with r ranging from 0 to 20 and determined whether biologically plausible reassorted strains were detected. In doing so, we noted additional reassortment events, with some occurring in contemporary swine strains (e.g. A/swine/Illinois/A02218757/2017 and A/swine/Pennsylvania/A02218184/2017). These reassorted viruses have been associated with recent human infections [see Bowman *et al.* (2017); Anderson *et al.* (2021); Duwell *et al.* (2018)]. We also noted an almost linear decrease in the embedding cost as we increased the number of reticulations present in the network. This suggests that the empirical data best fit in networks with at least 20 reticulations (see Supplementary Fig. S1 in Supplementary Material). This is supported by empirical single-gene phylogenetic studies that demonstrated considerable amounts of reassortment; Rajão *et al.*, 2017 detected at least 40 reassorted genomes in a different swine H3 IAV lineage. Our data also reveal what appear to be a number of intralineage reassortment events, where highly similar IAV coinfect cells and exchange genes, resulting in novel progeny. The intralineage reassortment process would be difficult to detect with standard methods such as tanglegrams that track topological incongruence (Zeller *et al.*, 2020) as the evolutionary gene lineages would appear to be the same despite being acquired from different parents. The consequence of such high levels of intralineage reassortment is unknown, but we hypothesize that this process contributes to the expansion in the genetic diversity of IAV in swine in the US (Neverov *et al.*, 2015; Zeller *et al.*, 2020).

7 Conclusion

We extended RF-Net with a suite of methods to advance its applicability in practice. The simulation study validated our hypothesis that the reticulation paths in the embeddings of gene trees do not change significantly as the search progresses, which allowed us to scale RF-Net to large numbers of reticulations. In addition, the simulations showcased the applicability of our proposed automated stopping criterion for the inference of the proper number of reticulate events. With the additional feature for the output of gene tree embeddings, we evaluated and confirmed the ability of RF-Net to successfully error-correct input trees. Finally, we were able to analyze the evolution of a major H3 swine IAV lineage and observe patterns of both inter- and intra-lineage reassortments. In IAV, reassortment increases the genetic and antigenic diversity of IAV within the swine 'mixing vessel' host (Scholtissek, 1990); and it increases the likelihood that a virus with zoonotic potential may emerge via a sampling effect. Though our analyses were post-hoc, RF-Net could be developed to run as an online algorithm, where surveillance data are screened for reassorted viruses, and novel viruses could be assessed to preemptively determine pandemic potential.

Acknowledgement

The authors thank Zebulun Arendsee for providing the validation dataset on IAV in swine. The authors also thank the anonymous reviewers whose comments and suggestions significantly helped to improve this manuscript.

Funding

This work was supported in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services [Contract No. 75N93021C00015]; National Science Foundation [1617626]; the Department of Defense [PREEMPT, HR00112020034]; USDA-ARS [0500-00093-001-00-D and 5030-32000-231-000-D]; and ORISE [DE-AC05-06OR23100]. The funding sources had no role in study design, data collection and interpretation, or the decision to submit the work for publication. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Albrecht,B. (2016) Computing hybridization networks using agreement forests. PhD Thesis, Ludwig-Maximilians-Universität München, Munich, DE.

Anderson,T.K. *et al.* (2013) Population dynamics of cocirculating swine influenza A viruses in the United States from 2009 to 2012. *Influenza Other Respir. Viruses*, 7, 42–51.

Anderson,T.K. *et al.* (2021) Swine influenza A viruses and the tangled relationship with humans. *Cold Spring Harbor Perspect. Med.*, **11**, a038737.
Bansal,M.S. *et al.* (2010) Robinson-Foulds supertrees. *Algorithms Mol. Biol.*, **5**, 18.
Baroni,M. *et al.* (2005) A framework for representing reticulate evolution. *Ann. Comb.*, **8**, 391–408.
Boni,M.F. *et al.* (2010) Guidelines for identifying homologous recombination events in influenza A virus. *PLoS One*, **5**, e10434.
Bordewich,M. *et al.* (2017) Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks. *J. Theor. Biol.*, **423**, 1–12.
Bouckaert,R. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **10**, e1003537.
Bowman,A.S. *et al.* (2017) Influenza A (H3N2) virus in swine at agricultural fairs and transmission to humans, Michigan and Ohio, USA, 2016. *Emerg. Infect. Dis.*, **23**, 1551–1555.
Cai,R. and Ané,C. (2021) Assessing the fit of the multi-species network coalescent to multi-locus data. *Bioinformatics*, **37**, 634–641.
Cardona,G. *et al.* (2009) Metrics for phylogenetic networks I: generalizations of the Robinson-Foulds metric. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **6**, 46–61.
Chang,J. *et al.* (2019) octoFLU: automated classification for the evolutionary origin of influenza A virus gene sequences detected in US Swine. *Microbiol. Resour. Announc.*, **8**, e00673-19.
Cotton,J.A. and Wilkinson,M. (2007) Majority-rule supertrees. *Syst. Biol.*, **56**, 445–452.
Crisci,E. *et al.* (2013) Influenza virus in pigs. *Mol. Immunol.*, **55**, 200–211.
Das,S.R. *et al.* (2011) Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. *Proc. Natl. Acad. Sci. USA*, **108**, E1417–E1422.
Dereeper,A. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.
Diaz,A. *et al.* (2017) Complete genome sequencing of influenza A viruses within swine farrow-to-wean farms reveals the emergence, persistence, and subsidence of diverse viral genotypes. *J. Virol.*, **91**, e00745-17.
Duwell,M.M. *et al.* (2018) Influenza A (H3N2) variant virus outbreak at three fairs—Maryland, 2017. *Morbidity Mortality Weekly Rep.*, **67**, 1169–1173.
Elworth,R.L. *et al.* (2018) Advances in Computational Methods for Phylogenetic Networks in the Presence of Hybridization. In: Warnow, T. (ed.) *Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret*. Springer International Publishing, Cham, pp. 317–360. https://doi.org/10.1007/978-3-030-10837-3_13.
Gao,S. *et al.* (2017) The genomic evolution of H1 influenza A viruses from swine detected in the united states between 2009 and 2016. *J. Gen. Virol.*, **98**, 2001–2010.
Garten,R.J. *et al.* (2009) Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. *Science*, **325**, 197–201.
Grenfell,B.T. *et al.* (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**, 327–332.
Hahn,M.W. (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.*, **8**, R141–R149.
Harris,S.R. *et al.* (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.*, **13**, 130–136.
Hejase,H.A. and Liu,K.J. (2016) A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, **17**, 422.
Hejase,H.A. *et al.* (2018) Fastnet: fast and accurate statistical inference of phylogenetic networks using large-scale genomic sequence data. In: *RECOMB International Conference on Comparative Genomics*. Springer, Sherbrooke, QC, CA, pp. 242–259.
Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–267.
Huson,D.H. *et al.* (2010) *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, Cambridge, UK.
Iersel,L.v. *et al.* (2014) A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees. *BMC Bioinformatics*, **15**, 127.
Jackson,A.P. (2004) A reconciliation analysis of host switching in plant-fungal symbioses. *Evolution*, **58**, 1909–1923.
Janssen,R. *et al.* (2018) Exploring the tiers of rooted phylogenetic network space using tail moves. *Bull. Math. Biol.*, **80**, 2177–2208.
Kapli,P. *et al.* (2020) Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.*, **21**, 428–444.
Katoh,K. *et al.* (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
Kingman,J.F.C. (1982) The coalescent. *Stochastic Process. Appl.*, **13**, 235–248.
Leitner,T. (2019) Phylogenetics in HIV transmission: taking within-host diversity into account. *Curr. Opin. HIV AIDS*, **14**, 181–187.
Markin,A. *et al.* (2019) Robinson-Foulds reticulation networks. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Niagara Falls, NY, USA, pp. 77–86.
McDonald,S.M. *et al.* (2016) Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat. Rev. Microbiol.*, **14**, 448–460.
Meng,C. and Kubatko,L.S. (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.*, **75**, 35–45.
Mitnaul,L.J. *et al.* (2000) Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *J. Virol.*, **74**, 6015–6020.
Müller,N.F. *et al.* (2020) Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proc. Natl. Acad. Sci. USA*, **117**, 17104–17111.
Neverov,A.D. *et al.* (2014) Intrasubtype reassortments cause adaptive amino acid replacements in H3N2 influenza genes. *PLoS Genet.*, **10**, e1004037.
Neverov,A.D. *et al.* (2015) Coordinated evolution of influenza A surface proteins. *PLoS Genet.*, **11**, e1005404.
Nguyen,L.-T. *et al.* (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
Posada,D. and Crandall,K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.*, **16**, 37–45.
Posada,D. and Crandall,K.A. (2002) The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.*, **54**, 396–402.
Powell,J.D. *et al.* (2021) Characterization of contemporary 2010.1 H3N2 swine influenza A viruses circulating in United States pigs. *Virology*, **553**, 94–101.
Price,M.N. *et al.* (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
Rajão,D.S. *et al.* (2015) Novel reassortant human-like H3N2 and H3N1 influenza A viruses detected in pigs are virulent and antigenically distinct from swine viruses endemic to the United States. *J. Virol.*, **89**, 11213–11222.
Rajão,D.S. *et al.* (2017) Reassortment between swine H3N2 and 2009 pandemic H1N1 in the United States resulted in influenza A viruses with diverse genetic constellations with variable virulence in pigs. *J. Virol.*, **91**, e01763-16.
Rajão,D.S. *et al.* (2018) Antigenic and genetic evolution of contemporary swine H1 influenza viruses in the United States. *Virology*, **518**, 45–54.
Rasmussen,M.D. and Kellis,M. (2011) A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.*, **28**, 273–290.
Robinson,D. and Foulds,L. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
Sagulenko,P. *et al.* (2018) Treetime: maximum-likelihood phylodynamic analysis. *Virus Evol.*, **4**, vex042.
Scholtissek,C. (1990) Pigs as 'mixing vessels' for the creation of new pandemic influenza A viruses. *Med. Principles Pract.*, **2**, 65–71.
Smith,G.J. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122–1125.
Solís-Lemus,C. and Ané,C. (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.*, **12**, e1005896.
Stamatakis,A. (2014) Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
Steel,M. and Rodrigo,A. (2008) Maximum likelihood supertrees. *Syst. Biol.*, **57**, 243–250.
Tavaré,S. *et al.* (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.*, **17**, 57–86.
Vaughan,T.G. (2017) IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, **33**, 2392–2394.
Wen,D. *et al.* (2016) Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol. Ecol.*, **25**, 2361–2372.
Whidden,C. *et al.* (2013) Fixed-parameter algorithms for maximum agreement forests. *SIAM J. Comput.*, **42**, 1431–1466.
Woolley,S.M. *et al.* (2008) A comparison of phylogenetic network methods using computer simulation. *PLoS One*, **3**, e1913.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yu,Y. and Nakhleh,L. (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, **16**, 1–10.

Yu,Y. *et al.* (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.*, **8**, e1002660.

Yu,Y. *et al.* (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Syst. Biol.*, **62**, 738–751.

Yu,Y. *et al.* (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. USA*, **111**, 16448–16453.

Zeller,M.A. *et al.* (2018) ISU FLUture: a veterinary diagnostic laboratory web-based platform to monitor the temporal genetic patterns of Influenza A virus in swine. *BMC Bioinformatics*, **19**, 397.

Zeller,M.A. *et al.* (2020) Spatial and Temporal Coevolution of N2 Neuraminidase and H1 and H3 Hemagglutinin Genes of Influenza A Virus in United States Swine. *Virus Evolution*, 2021. https://doi.org/10.1093/ve/veab090.

Zhang,C. *et al.* (2018) Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, **35**, 504–517.

Zhang,Y. *et al.* (2017) Influenza research database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res.*, **45**, D466–D474.