Identification of Reassortant Influenza Viruses At Scale - Algorithm and Applications

Eric J. Ma

Submitted to the Department of Biological Engineering in Partial Fulfillment of the Requirements for the Degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author: Eric J. Ma, Department of Biological Engineering, MIT

Signature

Accepted by: Professor Mark Bathe, Graduate Program Chair and Committee Chair, Department of Biological Engineering, MIT

Signature

Certified by: Professor Jonathan A. Runstadler, Thesis Advisor, Department of Biological Engineering, MIT

Signature

Certified by: Professor Jukka-Pekka Onnela, Thesis Committee Member, Department of Biostatistics, Harvard School of Public Health

---

Signature

Identification of Reassortant Influenza Viruses At Scale - Algorithm and Applications

by

Eric J. Ma

Submitted to the Department of Biological Engineering on on June 2017 in Partial fulfillment of the requirements for the Degree of Doctor of Science in Biological Engineering

Abstract

Reassortment is a reticulate evolutionary process that results in genome shuffling; the most prominent virus known to reassort is the influenza A virus. Methods to identify reassortant influenza viruses do not scale well beyond hundreds of isolates at a time, because they rely on phylogenetic reconstruction, a computationally expensive method. This thus hampers our ability to test systematically whether reassortment is associated with host switching events. In this thesis, I use phylogenetic heuristics to develop a new reassortment detection algorithm capable of finding reassortant viruses in tens of thousands viral isolates. Together with colleagues, we then use the algorithm to test whether reassortment events are over-represented in host switching events and whether reassortment is an alternative 'transmission strategy' for viral persistence.

Thesis Supervisor: Jonathan A. Runstadler

Title: Assistant Professor of Biological Engineering

Contents

1 A Primer on the Influenza A Virus

1.1 The Importance of Studying Influenza Evolution & Ecology

The influenza A virus has inflicted economic damage annually on the order of billions of dollars (1). Being a pathogen with zoonotic origins,[1] it is imperative to study its circulation, evolution and pathogenesis not only in humans, but also in animals (domestic and wild). One major problem of interest pertains to influenza's ability to shuffle its genome with other influenza viruses, and its implication in the ability of the virus to jump between host species. To address this, in this thesis I outline efforts with my colleagues to map out and identify these shuffled viruses at a global scale, and use this systematic, global identification study influenza, reticulate evolution, and ecology.

1.2 Genome Structure & Evolution

The influenza A virus is a negative strand RNA virus, comprised of 8 genomic RNA segments. Its negative strandedness means that it encodes the strand opposite the messenger RNA (mRNA), implying that it needs to first be copied into mRNA before translation can occur. Together, the RNA segments encode its polymerase (PB2, PB1, PA, NP), viral entry and release proteins (HA, NA), a matrix protein (M) and a non-structural protein (NS)

[1] Being of zoonotic origin means that the virus' reservoir is in one or more animal hosts, but "spills over" into humans upon contact. As such, humans are the "spillover host".

(fig. 1).

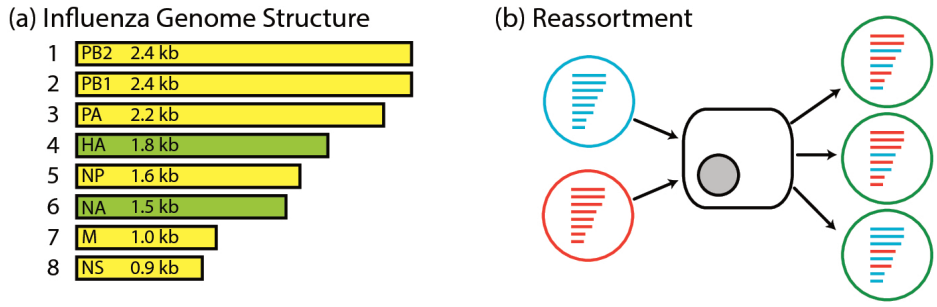

Figure 1: (a) Influenza A virus genome structure. The influenza virus is comprised of 8 RNA segments. (b) Reassortment. Reassortment is the process by which two viruses co-infect the same host cell and produce progeny virus that contain segments from both parental viruses.

Being an RNA virus that carries its own RNA-dependent RNA polymerase, the influenza A virus is prone to copying errors during replication inside a host cell (2). This 'sloppiness' allows the influenza virus to evolve rapidly under neutral conditions, resulting in **evolutionary drift**.

Evolutionary drift coupled with selection contributes to the difference in evolutionary rates that are observed between the external and internal genes. The HA and NA genes are thought to be under immune selection, as they are the external proteins that are targeted by the immune system. The HA and NA proteins, therefore, evolve under dual constraints: they have to continue functioning for cellular entry and release, while also evolving novel epitopes that can successfully evade immune system detection. Evolutionary drift in the HA and NA genes contribute to **antigenic drift**, in which the antigenic characteristics of these two proteins continually evolve under selection. By contrast, the internal proteins do not function under such selective pressures,

and as such are much more highly conserved.

Evolutionary drift is not the only mechanism by which influenza evolves. Its segmented and independently assorting genome allows for **reassortment** as a complementary mode of genomic evolution. Reassortment is thought to be the process resulting from co-infection of two viruses infecting the same host at the same time. If, for example, a red virus and a blue virus were to co-infect the same host cell, the progeny virus would contain any one of 2^8 combinations of red and blue segments (inclusive of the original viruses themselves) (fig. 1). Reassortment, thus, can be viewed as a form of **evolutionary shift**[2] in the genomic structure of the virus.

1.3 Subtype Classification

Influenza A viruses are classically known by their subtypes, e.g. H1N1, H5N1, H3N2. The "H" represents the hemagglutinin subtype, for which there are 16 canonically known ones (H1-H16). The "N" stands for the neuraminidase subtype, for which there are 9 canonically known ones (N1-N9). H17N10 and H18N11, two new subtypes that expand our canonical view of the number of viral subtypes, have been isolated from bats (3).

The hemagglutinin and neuraminidase are proteins expressed on the surface of the viral particle, and as such they are thought to be subject to immune selection and thus evolutionary pressure. This is the best current

[2]Amongst influenza researchers, evolutionary shift almost always refers to the exchange of HA and NA genes to produce viruses with an immunologically novel HA/NA combination. However, in this thesis, evolutionary shift refers more broadly to the exchange of any of the genes resulting in a novel genotype combination.

explanation as to why there is such great diversity in the HA and NA genes, and less diversity in the internal genes (e.g. polymerase genes)

1.4 Phylogenies

The evolutionary history of the influenza virus can be visualized using phylogenies. Phylogenetic trees are a reconstruction of the life history of a virus, and is based on two core concepts in evolutionary biology: common ancestry and descent with modification. There have been three major advances in the history of inference of phylogenies using gene sequence data:

1. Maximum parsimony (non-statistical reconstruction)
2. Maximum likelihood (statistical point estimation of a tree)
3. Bayesian inference (statistical reconstruction of ensemble of trees)

Tree construction is done as follows: given a matrix of **character states** (columns) against **samples** (rows) that are assumed to be independently evolving, we want to find the tree representation of the distance matrix that best reconstructs the evolutionary history of the samples. Prior to the advent of molecular sequence information, the character states that were used were morphological features, such as wings span and bone sizes. With the advent of molecular sequence data, multiple sequence alignments are used as the input data, with the character states being the individual positions[3].

[3]The assumption that character states evolve independently is still used in modern phylogenetic analyses, even though we know that this does not necessarily hold true, such as in the case of co-evolving sites due to epistatic interactions in a protein.

1.4.1 Maximum Parsimony

Maximum parsimony methods for phylogenetic reconstruction follow the logic of "the more similar we look, the closer our common ancestor is". A toy example is shown below. Consider the example where we have the following three samples with 3 binary character states recorded:

Table 1: Toy example of binary character states.

sample	char1	char2	char3
A	1	1	1
B	1	1	0
C	1	0	0

Using the principle of parsimony, we may compute a distance matrix as follows:

Table 2: Distance matrix computed from character states.

sample	A	B	C
A	0	1	2
B	1	0	1
C	2	1	0

Of the three possible trees that can be reconstructed, there are two that fit the data best:

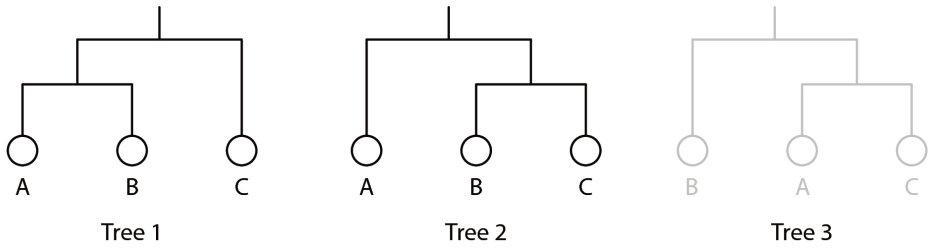

Figure 2: Maximum parsimony-based reconstruction of the character states. The non-parsimonious tree (Tree 3) is greyed out.

1.4.2 Maximum Likelihood

Molecular clock theory essentially states that the number of mutational events observed in a sequence is roughly linearly proportional with time. While in principle, this may seem to suggest that we can use the edit distance (maximum parsimony) to estimate the time of divergence between two sequences, there are problems with this logic.

One of the problems with maximum parsimony methods is that mutational reversions can occur. When a nucleotide changes from A to T, it can continue to mutate to a G or a C, or can revert back to an A. Many generations of replication forward, the edit distance (Hamming or Levenshtein) between the progeny and the original reaches a plateau (fig. 3). When reversions occur, using maximum parsimony to infer evolutionary history masks these reversion events.

Maximum likelihood methods help deal with this problem, by allowing us to calculate the likelihood that a given tree topology fits the sequence data, under an assumed model of sequence evolution that explicitly takes into

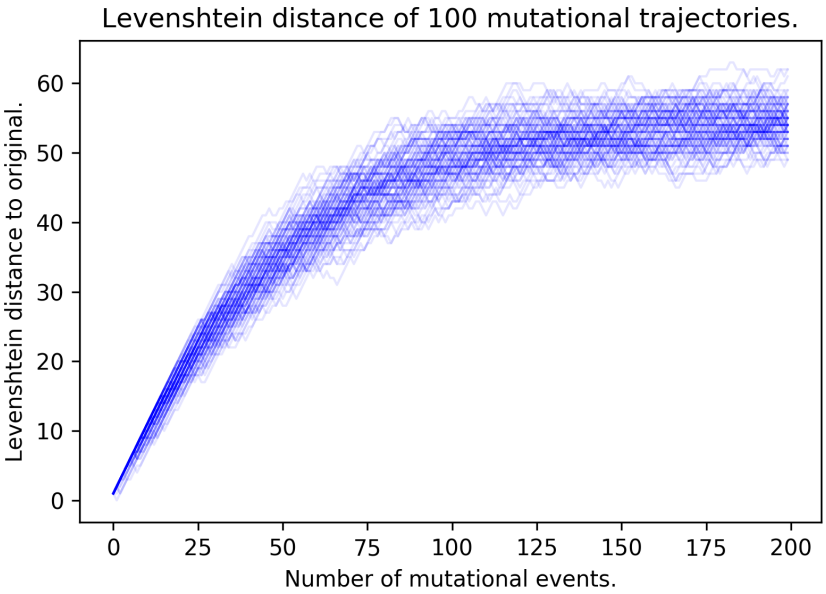

Figure 3: Levenshtein distance of 100 simulated trajectories.

account mutational reversion.

To illustrate how the likelihood calculations are performed, I show a toy example below on how tree likelihoods are computed.

Given the following three samples with the sequence states at a given position j.

Table 3: Toy example of sequence states at a position in a multiple sequence alignment.

Sample	seq_j
1	A
2	A
3	C

We may assume a model of evolution that follows the following sequence mutation (transition) probabilities:

Table 4: Toy example of transition probabilities.

letter	A	T	G	C
A	4/10	2/10	2/10	2/10
T	2/10	4/10	2/10	2/10
G	2/10	2/10	4/10	2/10
C	2/10	2/10	2/10	4/10

Finally, let us consider the following tree topology with two internal node reconstructions, as shown in fig. 4.

We may compute the following log likelihood for the left tree:

$$L_{tree1}(T) = P(A_4 \to A_1) \times P(A_4 \to A_2) \times P(A_5 \to A_4) \times P(A_5 \to C_3)$$

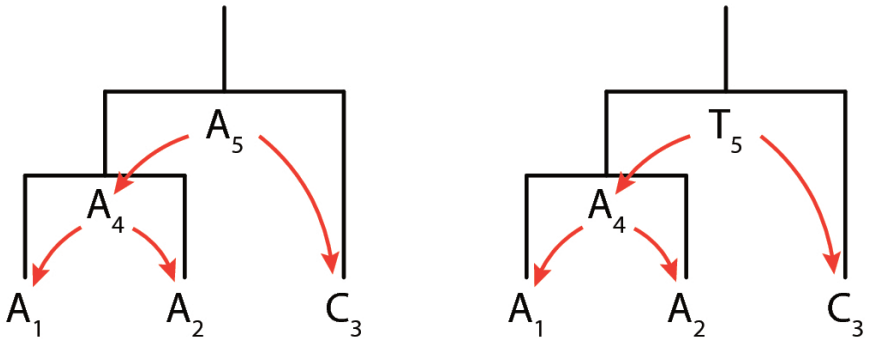

Figure 4: Two trees with internal node reconstructions on which likelihood calculations are performed.

Taking a log transform to prevent underflow in computation yields:

$$log(L_{tree1}(T)) = logP(A_4 \rightarrow A_1) + logP(A_4 \rightarrow A_2) + logP(A_5 \rightarrow A_4) + logP(A_5 \rightarrow C_3)$$

Finally, evaluating the result, we get:

$$log(L_{tree1}(T)) = 3log0.4 + log0.2 = -1.89$$

Doing an analogous computation for the right tree yields a log likelihood score of -2.19. The same computation is performed over all possible nucleotide identities at the internal nodes, given the tree topology, and the log likelihoods are summed up. Tree topologies are compared with respect to their likelihood scores, and the one with the highest score is returned as

the "maximum likelihood" tree.

Yet, we run into a problem: it is computationally infeasible to compute the likelihood for every single topology! Not only is the tree space large, according to Felsenstein (4):

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

the likelihood over every possible reconstructed ancestral sequence has to be computed as well.

Thus, in practice, tree space is searched iteratively using a greedy algorithm, which is detailed in Felsenstein's book, Inferring Phylogenies (4).

1.4.3 Bayesian Phylogenetic Inference

Bayesian phylogenetic reconstruction methods extend likelihood tree reconstruction methods by allowing us to infer a probability distribution over the tree topology, given the data and an assumed evolutionary model. This yields an ensemble of trees. When paired with phylogeographic inference (5), where geography is modelled as another character state in addition to nucleotide sequence, it is possible to reconstruct an inferred movement of viruses.

As is the case with Bayesian inference in general, the exponential increase in computational power along with advances in tree-space MCMC have been greatly enabling. Bayesian phylogenetic inference has been used successfully

to infer the time of emergence of outbreak viruses such as the Ebola virus (6, 7) and movement swine influenza viruses (8). Nonetheless, while it is the state-of-the-art method, Bayesian phylogenetic tree construction remains computationally expensive; typical real-world runtimes for tree reconstruction, given single core, GPU-enabled compute power, are on the order of weeks for hundreds of taxa and months for thousands of taxa.

1.5 Interpreting Trees

In order to understand how reassortment is detected, we need to first begin with some basic definitions of phylogenetic tree structure.

A bifurcating phylogenetic tree is a directed acyclic graph comprised of leaf nodes (tips), internal nodes, and bifurcating branches at each **internal node** (fig. 5). Branch lengths indicate evolutionary time elapsed from an internal node to another internal node or leaf.

As with any hierarchical clustering method, the leaves can be organized into **clades** (fig. 5), which represent a cluster of isolates on the tree that are closely related. How a clade is defined is often subjective; visual observation is the most common heuristic employed.

A metric of evolutionary distance between any two given isolates is the **patristic distance** (fig. 5) between them. The patristic distance is measured by the sum of branch lengths (in the units that the lengths are defined, or else arbitrary distance) from one isolate to another. As such, isolates that are more evolutionarily related will have a shorter patristic

distance between them.

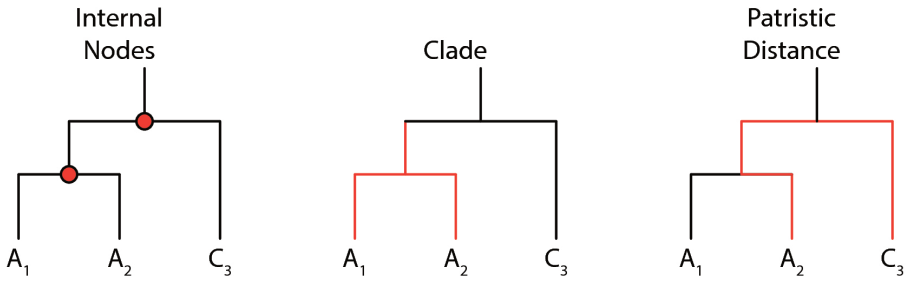

Figure 5: Visual definition of internal nodes, clades, and patristic distances.

1.6 Inferring Reassortment

Now that the basics of phylogenetic tree construction and interpretation have been covered, we will move on to discuss the current methods available for finding reassortant viruses, and their core logic.

1.6.1 Single Virus

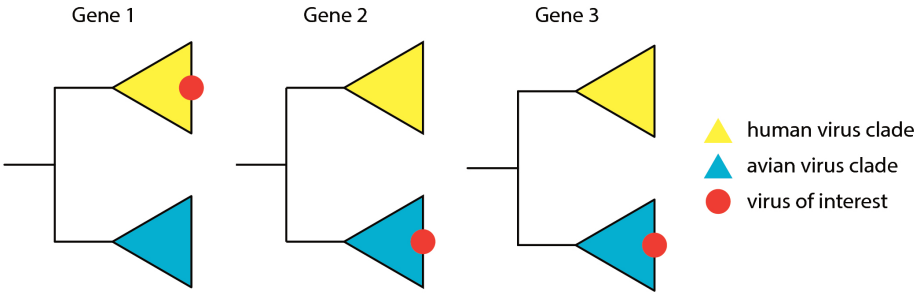

Figure 6: An illustration of how reassortment is inferred for a single virus.

Reassortment is classically inferred on a single virus of interest.

Reassortment can be detected by looking for incongruence in the phylogenetic history of a virus. As a simple example, for the red virus of interest in fig. 6, two of its three genes share closer evolutionary history with avian-isolated viruses, while one gene shares closer evolutionary history with human-isolated viruses. As such, we would infer that this avian virus acquired a human virus' gene through some process of reassortment.

1.6.2 Tree Incongruence

Tree incongruence is another way of identifying reassortant influenza A viruses. Because a bifurcating phylogenetic tree can be defined as a set of splits partitioning the taxa into two sets, "incompatible splits" in the tree can be identified by looking at the partitioned sets and identifying partition sets that have non-null intersections.

Let us look at fig. 7 for an elementary example. Suppose we had two trees with the same set of taxa, $\{t_1, t_2, t_3, t_4\}$. We observe the following splits:

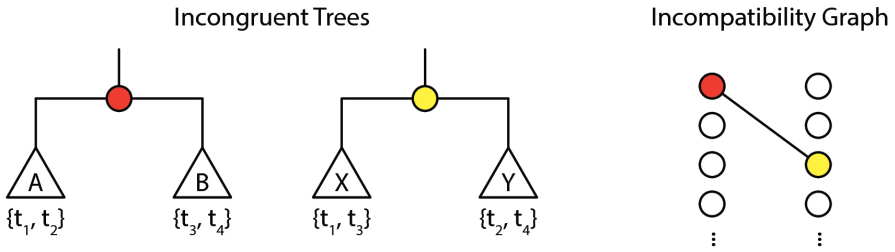

Figure 7: Tree incongruence. Splits are denoted as red or yellow circles on the trees on the left; they are also denoted as nodes in the split incompatibility graph on the right. If two splits are incompatibe, as given by the definition below, then they are joined by an edge in the graph.

- The tree with a red split defines a partition of the four taxa into two splits, $A = \{t_1, t_2\}$ and $B = \{t_3, t_4\}$.
- The tree with a yellow split defines a partition of the four taxa into two splits, $X = \{t_1, t_3\}$ and $B = \{t_2, t_4\}$.

If these two trees are incompatible, then all of the following criteria are true:

- $A \cap X \neq \varnothing$, (i.e. intersection of sets A and X, or set of common items, is not an empty set)
- $A \cap Y \neq \varnothing$,
- $B \cap X \neq \varnothing$, and
- $B \cap Y \neq \varnothing$.

In the case of this pair of trees:

- $A \cap X = \{t_1\} \neq \varnothing$,
- $A \cap Y = \{t_2\} \neq \varnothing$,
- $B \cap X = \{t_3\} \neq \varnothing$, and
- $B \cap Y = \{t_4\} \neq \varnothing$.

Hence, these trees are incompatible, and thus there is evidence that reassortment has happened.

Tree incongruence is a generalization of the logic used to find individual reassortant viruses, and is implemented in the software, GiRaF (9). GiRaF relies on the ensemble of trees returned from Bayesian phylogenetic tree reconstructions; hence, in practice the majority of time spent detecting reassortant viruses is actually spent on tree reconstruction. On the other hand, a nice statistical outcome of sampling tree space is that splits are

only counted if they appear in more than 95% of sampled trees, leading to a natural "95% confidence" for any given reassortment event detected.

1.6.3 3rd Codon Biases

3rd codon sequences are assumed to be under less selective pressure than 1st and 2nd codons in a sequence. If one considers two strains of virus v_a and v_b, and their respective pairs of segments, $s_{i,a}$ and $s_{i,b}$, and $s_{j,a}$ and $s_{j,b}$, we may compute the difference between their segments as follows:

$$d_{i,(a,b)} = EditDistance(s_{i,a}, s_{i,b})$$

$$d_{j,(a,b)} = EditDistance(s_{j,a}, s_{j,b})$$

Plotting the distribution of $d_{i,(a,b)}$ against $d_{j,(a,b)}$ for all pairs of viruses (v_a, v_b) yields fig. 8.

As shown in fig. 8 (adpated from (10)), if no reassortment was present, the hamming distance between the 3rd codons should be correlated under the assumptions that (a) 3rd codons are under neutral selection, and (b) the segments drift at roughly the same rate under neutral conditions. This would result in only blue dots showing up. If reassortment was present, then the hamming distances between two viruses should be non-correlated, and the yellow dots will show up.

This is a computationally simple method, as it only requires the computation of all pairwise edit distances, and as such has the advantage of being scalable

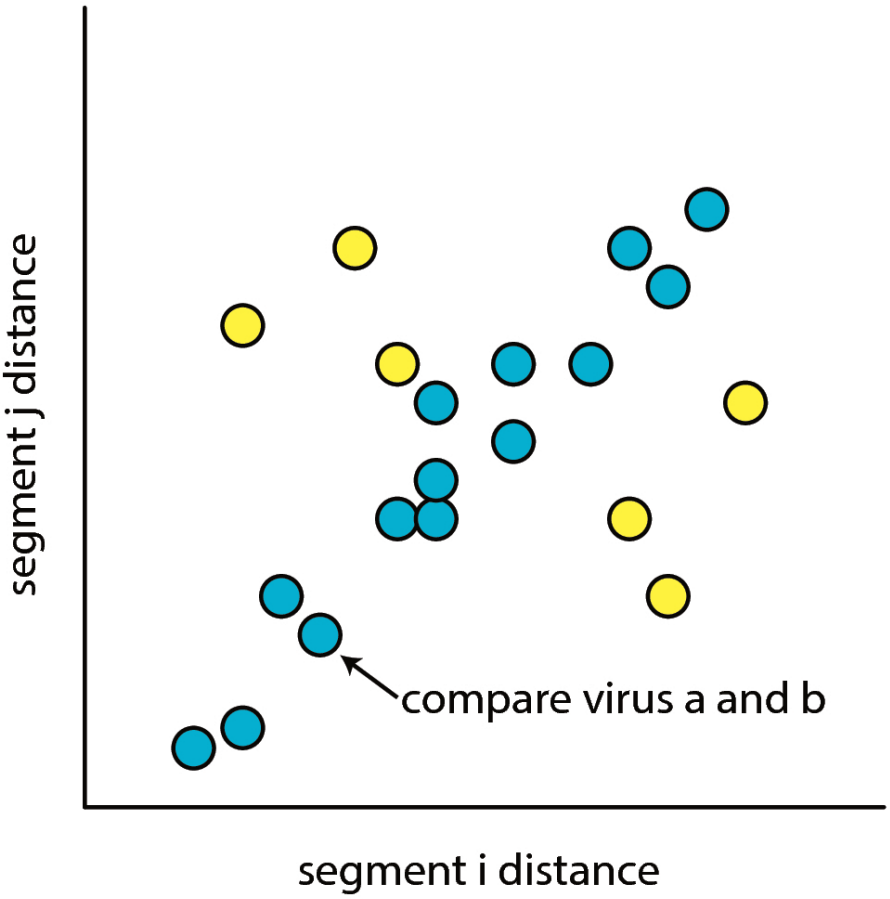

Figure 8: Toy distribution of all pairwise 3rd codon hamming distances between viral isolates. Blue dots: comparisons between viruses that yielded correlated 3rd codon hamming distances. Yellow dots: comparisons between viruses that yielded non-correlated 3rd codon hamming distances.

to large numbers of sequences. However, it ignores the evolutionary history of the virus, making no inference about the source of segments, unless paired with time-stamped data.

1.7 Influenza Biology

1.7.1 Genome Packaging

In the study of the process of reassortment, one cannot escape from the topic of "how viruses are packaged". This is because when two viruses co-infect the same host cell, the resulting mixed pool of genomic segments have to undergo packaging into another live virus. However, the level of abstraction required for understanding this thesis is at the host species level. As such, the details of packaging are not a central and necessary piece of knowledge for understanding influenza reassortment at a global scale. Thus, in lieu of a full description of the current state of knowledge, I have listed the major key points below as follows:

1. There are "packaging signals" located in the coding sequence (imposing a further evolutionary constraint) that determine whether a piece of RNA is selectively packaged into the viral genome. (11–13) (fig. 9)
2. Selective packaging is shown via electron microscopy, where the vast majority of viral particles have a distinct "7+1" arrangement of segments. Only a minority have extra segments. (14)
3. Packaging signals have been exploited to generate influenza viruses

that carry GFP rather than one of the genomic segments, allowing for tracking of viral replication (11). This remains, to date, the strongest evidence in favour of the presence of packaging signals that are part of the coding sequence of each of the 8 genes. This provides the genetic basis for selective packaging, but biochemical mechanisms remain elusive.

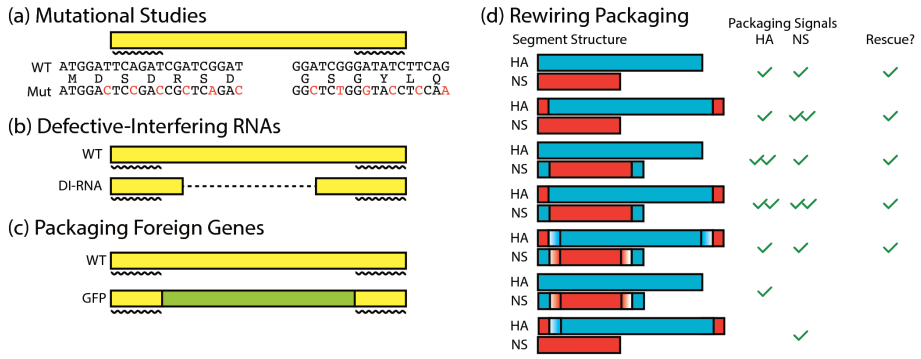

Figure 9: Summary of known results in influenza genome packaging. (a) Mutating the 3rd codon positions in the packaging regions reduces packaging efficiency, thus highlighting their importance. (b) Defective-interfering RNAs harbouring only the packaging signals can interfere with live virion production. (c) Foreign genes, such as GFP, have been packaged into the influenza virus by flanking them with packaging signals. (d) Packaging signals can be swapped between segments, but a packaging signal sequence must be present on each gene in order to rescue live virus.

1.7.2 Host Distribution of Influenza A Virus

In order to understand reassortment in ecological contexts, we need to know the known host range and movement patterns of the influenza A virus.

The influenza A virus has a broad geographic and trophic range. Viral flow

is canonically thought to start in the influenza A virus' reservoir hosts, wild ducks (15). Viruses can occasionally **spillover** from wild birds into domestic animals, such as pigs and chickens. Because of the close proximity of humans to domestic animals, these viruses can also jump from domestic animals into humans, thus leading to pandemics (15).

Though ducks, pigs, chickens and humans are the canonical places we think of flu, influenza is neither solely restricted to these hosts, nor is the flow of virus uni-directional.

Influenza viruses have been isolated in large and small mammals, including cows, horses, dogs, cats (16), seals (17–20), penguins (21) and more. The newest influenza viral subtypes, H17N10 and H18N11, have been isolated from bats (3).

Additionally, cases of **reverse zoonosis**[4] have been reported, where viruses jump back into swine hosts from humans (22). Humans, therefore, are not a "dead-end" host for the virus, as was once thought.

A little detail on influenza biology is necessary to understand reverse zoonosis. Based on studies of the glycans[5] that decorate the surface of cell membranes, human hosts generally have α-(2,3) glycans, while avian hosts generally have α-(2,6) glycans (23, 24). The hemagglutinin attaches to the glycans in the first step of viral infection, thus mediating viral entry.

[4]Zoonotic viruses typically have animal reservoir hosts, and humans are the "spillover" host. Reverse zoonosis occurs when these viruses jump from humans back into animal hosts.

[5]Glycans are branched chains of sugar molecules that have been identified on cell surface membranes. Glycans are also post-translationally added to proteins through glycosylating enzymes.

Pigs have been shown to have both glycans distributed on their cells (25). As such, viruses that are capable of infecting birds can also infect pigs, where they may acquire mutations that allow them to enter human cells; additionally, human viruses are also thus capable of replicating in swine hosts. As such, this knowledge has led to the conclusion that swine hosts may play the role of "mixing vessels" (25), allowing influenza A viruses to reassort in pigs.

1.7.3 Viral and Host Movement

Viruses are obligate parasites, in the sense that they cannot reproduce without the presence of a host to infect. As such, it should also be evident that their mobility is determined by their host's movement range. This has implications for detecting and forecasting reassortment: reassortment between two viruses cannot happen unless their host ranges overlap, or their host geographic ranges overlap.

1.7.4 Evolutionary Consequences of Reassortment

Reassortment can result in novel genotype combinations. An epidemiologically-relevant reassortment is one that occurs between viruses of different subtypes. This is because the HA and NA proteins elicit host immune responses. Thus, novel HA/NA combinations, or the introduction of an antigenically distinct HA or NA of the same subtype, can result in a new virus with the ability to evade immune detection. This would, in turn, help

the virus circumvent existing host (and population) immunity.

In theory, it is also possible for other 'enhancements' to the influenza A virus to be acquired via reassortment. For example, polymorphisms correlated with enhanced polymerase replication capacity in human cells are found in viruses isolated in wild birds, raising the possibility that these mutations can occur naturally and, if reassorted with an immunologically-novel viral subtype, can confer enhanced replication capacity, leading to a much riskier virus.

1.8 Research Questions

Given what we currently know about the ecology of influenza, a key gap in our knowledge is the role that reassortment plays in flu movement and survival. Grounded on this theme, I set out with colleagues to tackle the following questions:

1. Are reticulate evolutionary processes, such as reassortment, important for host switches? If so, can we quantify the importance? Is the principle generalizable?
2. Is reassortment an evolutionary strategy that influenza genes can employ to persist against barriers to transmission?

2 Algorithm

2.1 Description

At a high level, the reassortment detection algorithm works as such. Given a set of sequences, we wish to identify, using the rule of maximal similarity on some given metric, the most likely source of each segment in a virus. Sources, by definition, have to occur prior in time to the virus under consideration. We try to maximize the source similarity score of a virus while minimizing the number of sources needed to explain its existence.

We adapted the SeqTrack algorithm (26) to perform graph construction. Sequences were aligned using Clustal Omega 1.2.1 (27), and the resultant distance matrix was converted into a similarity matrix by taking `1 - distance`. Affinity propagation (28) clustering was performed on each segment's similarity matrix to determine a threshold cutoff similarity value, defined as the minimum (across all clusters for that segment) of minimum in-cluster pairwise identities, below which we deemed it implausible for an evolutionary descent (clonal or reassortment) to have occurred (fig. 17). Because the affinity propagation algorithm does not scale well with sample size, we treated the threshold computation as an estimation problem, and the final threshold was computed as the median threshold of 50 random subsamples of 500 isolates.

We then thresholded each segment's similarity matrix on the basis of its segment's threshold value, summed all eight thresholded similarity matrices, and then for each isolate, we identified the most similar isolate that

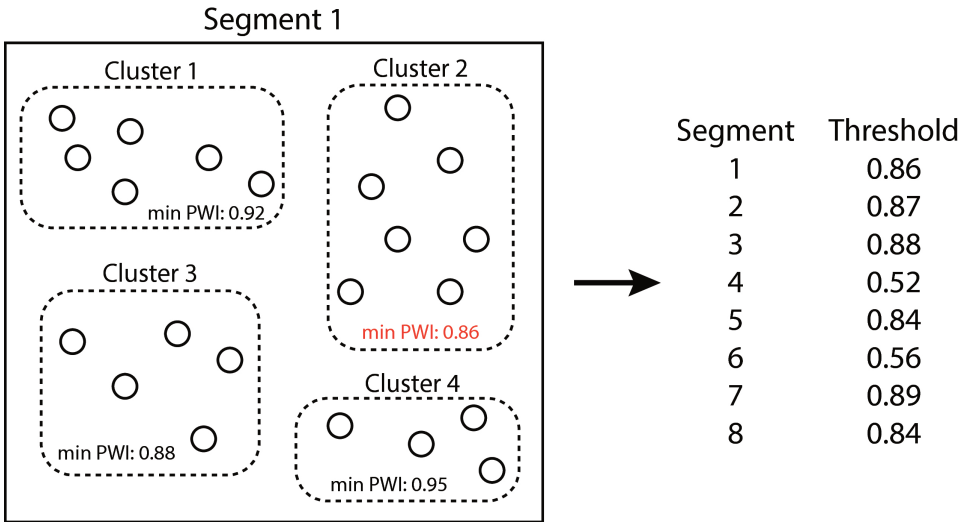

Figure 10: Schematic illustration of for the determination of thresholds. The minimum in-cluster PWI (min PWI) is shown within each cluster's bounding box. The minimum of min PWIs is highlighted in red. Exact threshold values used for the global influenza evolution study, rounded to 2 decimal places, are shown on the right.

occurred before it in time. This yielded the initial "full complement" graph without reassortant viruses. Each edge in this graph has an attached pairwise identity (PWI), which is the sum of PWIs across all eight segments. Within this graph, there are isolates for which no "full complement" of segments could be identified, which are candidate reassortant viruses. In addition, among the isolates for which a full complement of segments could be found from another source, we identified those whose in-edges were weighted at the bottom 10% of all edges present in the graph, which we also identified as candidate reassortant viruses (1,357 of 1,368 such viruses were eventually identified as reassortant; the other 11 were considered to be clonally descended). For these viruses, we performed source pair searches, where we identified sources for a part of the genome from one virus and sources for the complementary part of the genome from another virus. If the summed PWI across the segments for the two viruses was greater than the single-source search, we accepted the source pair as the candidate reassortant.

Pairwise identities were computed using Clustal Omega (version 1.2.1) (27). The algorithm is implemented in the Python programming language (version 3.5); main packages used included NetworkX, `numpy`, `pandas`, and `matplotlib` for visualization. The source code is archived on Zenodo (DOI: 10.5281/zenodo.33421).

2.2 Simulation Studies

To check whether the algorithm was capable of correctly identifying reassortant viruses, simulation studies were conducted. To simplify the problem, we considered the case of a two-segment virus, with each of the two segments having a different nucleotide substitution rate, mirroring the different substitution rates on each of the influenza genome segments. Each simulation run was initialized with anywhere between one and five viruses. At each time step, one virus was chosen at random to replicate (with 0.75 probability) or reassort with another virus (with 0.25 probability). Simulations were run for 50 time steps.

Regardless of replication or reassortment, the progeny virus was subjected to mutations, with the number of mutations in each segment being drawn from a binomial distribution with probability equal to the segment's substitution rate, and the exact positions drawn uniformly across the segment. This process is outlined in fig. 11.

The number of unique starting genotypes and total number of viral isolates being considered was much smaller than the real-world data. Therefore, our graph reconstruction procedure captured the essential parts of the method used in the global analysis, but differed in the details. Here, "full complements" involve only two segments. We did not perform affinity propagation clustering, as we started with completely randomly generated sequences of equal length. Our "null model" graph is where source isolates are chosen uniformly at random from the set of nodes occurring before the

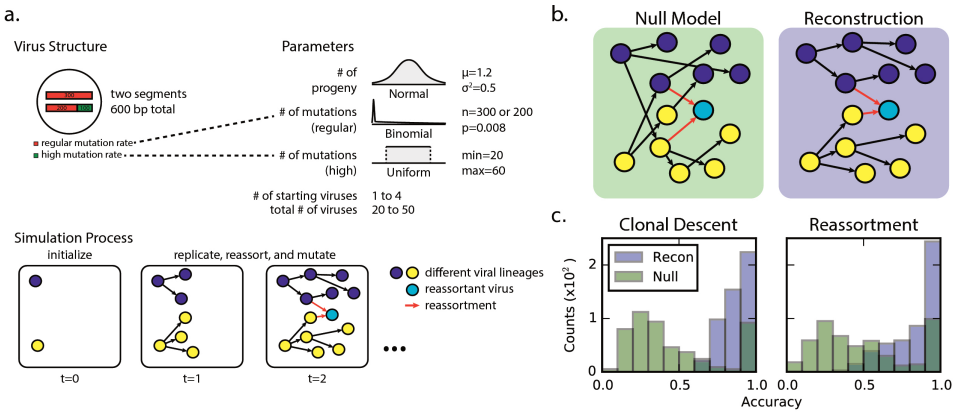

Figure 11: Viral simulation results. (a) Schematic of simulation studies conducted on a model two-segment virus with one segment capable of hypermutating in a short region of it. (b) In the null model, genomic information is ignored and a source virus is picked at random from isolates prior to it in time. Reassortants remain identified as reassortants, but sources are changed. In the proper reconstruction, sources are chosen to minimize genetic distance across two segments. (c) Distribution of proportion of reassortant viruses accurately identified under a proper reconstruction (blue bars) as opposed to a null model (green bars).

sink isolates.

To assess the accuracy of our reconstruction, we defined the path accuracy and reassortant path identification accuracy metrics (fig. 12). Edge accuracy, which is not used for evaluation here, is whether a particular reconstruction transmission between two isolates exists in the simulation. Path accuracy is a generalization of edge accuracy, where a path existing between the source and sink nodes (without considering the direction of edges) in the reconstruction is sufficient for being considered accurate. Reassortant path identification accuracy measures how accurately we identified the reassortant paths, analogous to the regular path accuracy.

Source code for the simulation studies is available on Zenodo (DOI: 10.5281/zenodo.33427).

2.3 Comparative Analysis of Time Complexity

In inferring reassortment, the key step is to identify how viruses are related, given the sequence data. Thus, we compare here the time complexity for inferring the topology of phylogenetic trees vs. the topology of networks, as applied to the detection of reassortant viruses.

2.3.1 Tree Reconstruction Complexity

According to the GiRaF developers, the time required to identify reassortant viruses is dominated by the time to construct the phylogenetic trees for each gene (9). Thus, in calculating the complexity required for phylogeny-based

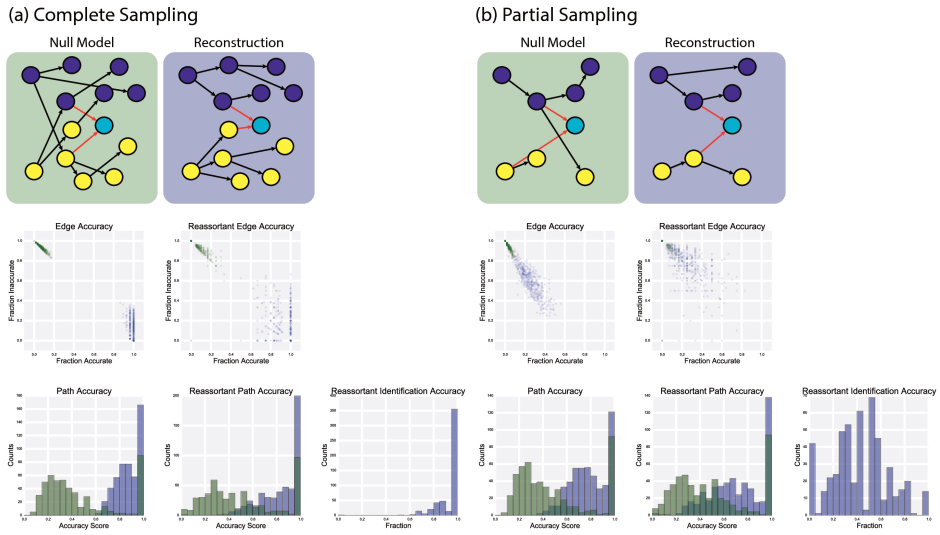

Figure 12: Accuracy scores for simulation studies. Simulations were conducted under (a) complete sampling and (b) incomplete sampling scenarios. Top row: Reconstruction using the algorithm described (blue background) and under a null reconstruction (green background). Middle row: Distribution of edge accuracy metrics (fraction incorrect vs. fraction correct) under null reconstruction (green scatter points) and algorithm reconstruction (blue scatter points). Bottom row: Distribution of path accuracy under a null reconstruction (green) and algorithm reconstruction (blue). The algorithm reconstruction has a consistently higher accuracy in identifying reassortant viruses.

reassortment detection algorithms, tree reconstruction is the major term we need to account for. According to Felsenstein (4), given a set of n labelled sequences, the number of possible rooted, bifurcating trees (which are used for inferring tree inconrguence) is

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

This can be expanded to:

$$\frac{(2n-3)(2n-4)(2n-5)...(n)(n-1)(n-2)(n-3)(n-4)...(1)}{2^{n-2}(n-2)(n-3)(n-4)...(1)}$$

Cancelling the common terms in the numerator and denominator, we get:

$$\frac{(2n-3)(2n-4)(2n-5)...(2n-n)(2n-(n+1))}{2^{n-2}}$$

If we consider only the largest polynomial terms of n, we see that in the numerator, $2n$ is multiplied $k-2$ times, where k is the term subtracted from $2n$. Therefore, the major term (ignoring the smaller polynomials) simplifies to:

$$\frac{(2n)^{n-1}}{2^{n-2}} = \frac{2^{n-1}n^{n-1}}{2^{n-2}} = 2n^{n-1}$$

Under the assumption that a Bayesian reconstruction is only looking for the most optimal tree topologies and is not estimating times of divergence for

internal nodes, then the worst case scenario is that the MCMC sampling algorithm has to search $O(n^n)$ trees in order to find the best topology.

2.3.2 Network Reconstruction Complexity

For each of the major steps in the algorithm developed in this thesis, the time complexity is outlined below:

- Pairwise distance matrix computations is of $O(n^2)$ complexity.
- Finding maximal edges again requires n^2 comparisons to be made.
- In the 2nd search for source pairs, given s segments and n isolates, in the worst case scenario, we have to check all isolates for the source pairs. Thus, we require $\binom{s}{2}$ n^2 comparisons in the worst-case scenario.

Given this analysis, and ignoring the s term (which is the number of segments for a given virus), the worst-case time complexity of the SeqTrack-based algorithm described here should be $O(n^2)$.

2.3.3 Anecdotal Comparisons

In other work not recorded in this thesis, I have helped construct phylogenies for individual virus segments. Anecdotally, it takes on the order of days to sample tree space for tens of isolates, and weeks for hundreds of isolates, using BEAST (29) on the Koch Institute's Sun Grid Engine compute cluster. Part of this is the serial nature of BEAST's MCMC sampling algorithm implementation, though the space complexity that has to be explored is surely the dominating term.

By contrast, even with the manual steps involved, the reassortment finding algorithm takes about 1 week of time to operate on tens of thousands of isolates, and about a day's worth of compute time on thousands of isolates, utilizing roughly the same resources.

3 Applications

3.1 Application 1: Global reticulate evolution study.

This study was conducted with much help from Dr. Nichola J. Hill (MIT Department of Biological Engineering & Division of Comparative Medicine) who gave much technical mentorship on ecology, and two undergraduate students, Kyle Yuan and Justin Zabilansky, both of whom contributed analysis or code to the final paper, which was published in the Proceedings of the National Academy of Sciences (30).

3.1.1 Abstract

Reticulate evolution is thought to accelerate the process of evolution beyond simple genetic drift and selection, helping to rapidly generate novel hybrids with combinations of adaptive traits. However, the long-standing dogma that reticulate evolutionary processes are likewise advantageous for switching ecological niches, as in microbial pathogen host switch events, has not been explicitly tested. We use data from the influenza genome sequencing project and a phylogenetic heuristic approach to show that reassortment, a reticulate evolutionary mechanism, predominates over mutational drift in transmission between different host species. Moreover, as host evolutionary distance increases, reassortment is increasingly favoured. We conclude that the greater the quantitative difference between ecological niches, the greater the importance of reticulate evolutionary processes in overcoming niche barriers.

3.1.2 Significance

Are the processes that result in the exchange of genes between microbes quantitatively advantageous for those microbes when switching between ecological niches? To address this question, we consider the influenza A virus as a model microbe, with its ability to infect multiple host species (ecological niches) and undergo reassortment (exchange genes) with one another. Through our analysis of sequence data from the Influenza Research Database and the Barcode of Life Database, we find that the greater the quantitative difference between influenza hosts, the greater the proportion of reassortment events were found. More broadly, for microbes, we infer that reticulate evolutionary processes should be quantitatively favoured when switching between ecological niches.

3.1.3 Introduction

Reticulate evolutionary processes, such as horizontal gene transfer and genomic reassortment, have been proposed as a major mechanism for microbial evolution (31), aiding in the diversification into new ecological niches (32). In contrast to clonal adaptation through genetic drift over time, reticulate evolutionary processes allow an organism to acquire independently evolved genetic material that can confer new fitness-enhancing traits. Examples include the acquisition of cell surface receptor adaptations (point mutations) in viruses (33) and antibiotic resistance (single genes) (34) and pathogenicity islands (or gene clusters) in bacteria

(35).

Host switching, defined as a pathogen moving from one host species into another, represents a fitness barrier to microbial pathogens. The acquisition of adaptations through reticulate processes either before or after transmission from one species to another may serve to aid successful pathogen host switches by improving fitness and the likelihood of continued transmission (36). In this sense, reticulate evolution may be viewed as an ecological strategy for switching between ecological niches (such as different host species), complementing but also standing in contrast to the clonal adaptation of a microbial pathogen by genetic drift under selection. To test this idea and its importance in host switch events, which are critical for (re)-emerging infectious disease, we provide a quantitative assessment of the relative importance of reticulate processes versus clonal adaptation in aiding the ecological niche switch of a viral pathogen.

Data yielded from influenza genome sequencing projects provide a unique opportunity for quantitatively testing this concept and are suitable for the following reasons. First, the influenza A virus (IAV) has a broad host tropism (15) and is capable of infecting organisms spanning millennia of divergence on the tree of life. With different host-specific restriction factors forming an adaptive barrier, each host species may then be viewed as a unique ecological niche for the virus (37). Second, IAV is capable of and frequently undergoes reassortment, which is a well-documented reticulate evolutionary process (38–41). Reassortment has also been implicated as an adaptive evolutionary mechanism in host switching (42, 43), although

this is most prevalently observed for pandemic viruses of public health interest for which sequences are available (44). Finally, as a result of surveillance efforts during the last 2 decades, whole-genome sequences have been intensively sampled during a long time frame, with corresponding host species metadata, available in an easily accessible and structured format (45). Because reassortant viruses are the product of two or more genetically distinct viruses coinfecting the same host, a more complex process than clonal transmission and adaptation, they are expected to occur less frequently. Hence, the global IAV dataset, which stretches over time and space with large sample numbers, provides the necessary scope to detect reassortant viruses at a scale required to quantitatively assess the relative importance of reticulate events in viral host switching.

3.1.4 Method Validation

We used the phylogenetic heuristic algorithm (described in the Algorithm section) to reconstruct an approximate global phylogeny for all 18,000+ fully-sequenced viruses in the dataset. In this network of viral isolates, clonal descent is mostly structured by host species, with known global patterns of human-to-human (H3N2 & H1N1, and rarer H5N1 & H7N9), chicken-to-chicken (H9N2, H7N9, H5N1) and swine-to-swine (H3N2, H1N1, H1N2) viral circulation captured in the network reconstruction (fig. 13). Edges in the network connected viral isolates with a median genetic similarity of 99.7%, indicating a high degree of genetic similarity captured in the network-based reconstruction (fig. 14). As expected, no clonal descent was identified

between viruses of different subtypes. Moreover, the network recreates the phylogeny of known reassortant viruses, including the 2009 pandemic H1N1 and the recent 2013 H7N9 viruses, further validating the accuracy of our reconstruction (a browser-based `d3.js` visualization is available in Zenodo archive of the Github repository (Materials & Methods)). Small-world simulation studies validated our method as being accurate in detecting reassortment events (fig. 11), while a comparison of edges to a phylogenetic reconstruction on a subset of the data show that our method captures the shorter end of the distribution of patristic distances on a tree, indicating accurate approximation to phylogenetic reconstruction (fig. 15). Hence, our method is capable of detecting reassortment events, which are classically inferred by observing incongruences in phylogenetic tree clustering.

3.1.5 Results

To test whether reassortment or clonal descent was an advantageous strategy when switching hosts, we computed the weighted proportion of reassortant edges (out of all edges) occurring between hosts of the same or different species. When host species were different, reassortant edges were over-represented at 19 percentage points above a null permutation model (permutation test described in Materials & Methods) (fig. 16 (a)), and when host species were the same, reassortant edges were under-represented by 7 percentage points relative to our null model. Thus, reassortment is a strongly favoured strategy when influenza crosses between different host species.

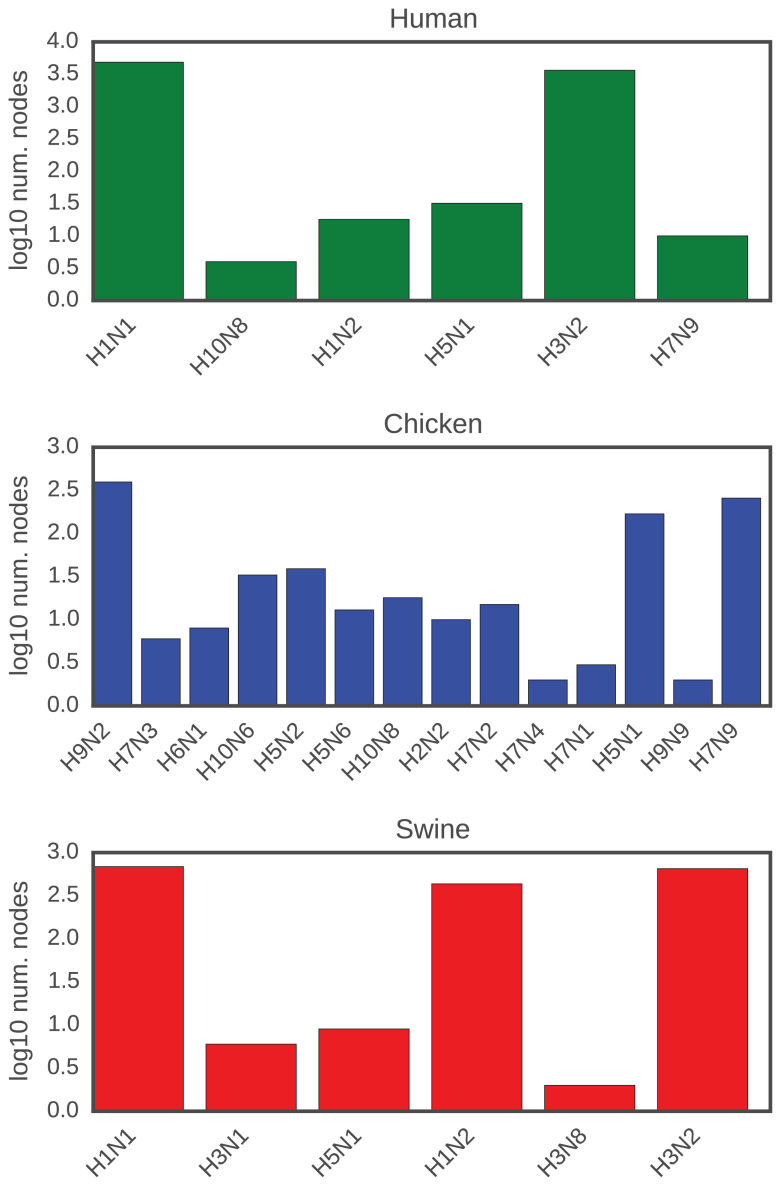

Figure 13: Subtypes involved in clonal descent amongst human, chicken and swine viruses, and their total numbers represented in the dataset.

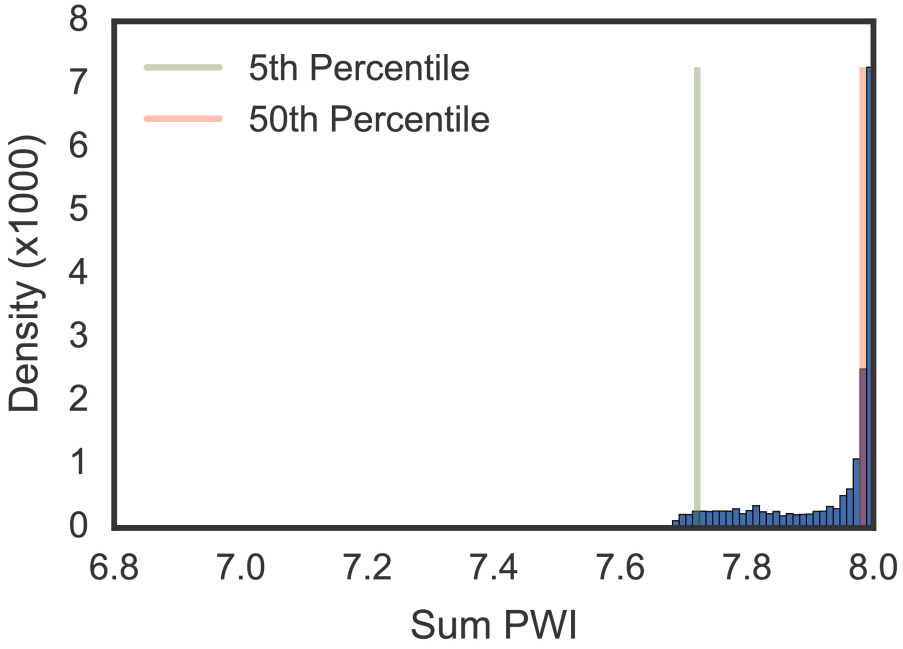

Figure 14: Distribution of pairwise identities across every clonal descent and reassortment event detected. 5th and 50th percentiles of the distribution are shown using a vertical green and red line respectively. Sum PWI: Summed pairwise identity across all 8 segments.

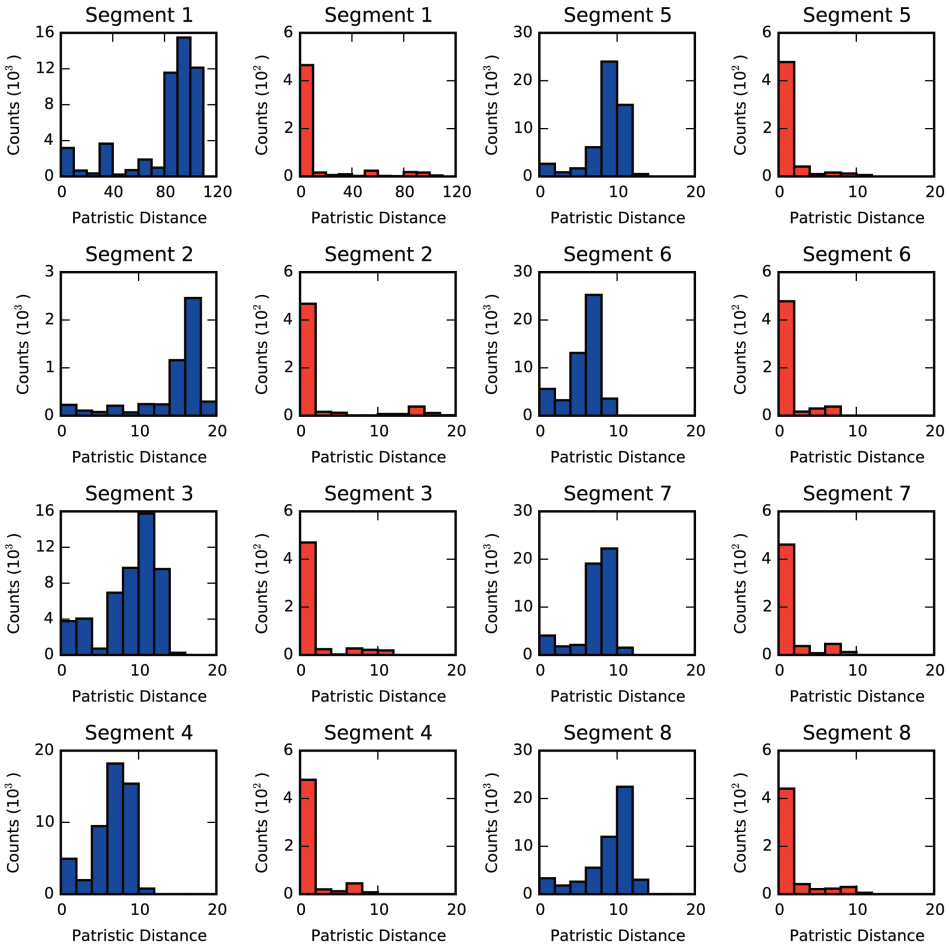

Figure 15: Patristic distance test. All patristic distances (pairwise sum of branch lengths between tree taxa) captured in the network representation vs. individual gene trees of H3N8 isolates from Minto Flats, Alaska. Blue histograms: All pairwise patristic distances represented in the phylogenetic tree. Red histograms: Patristic distances captured by the network reconstruction.

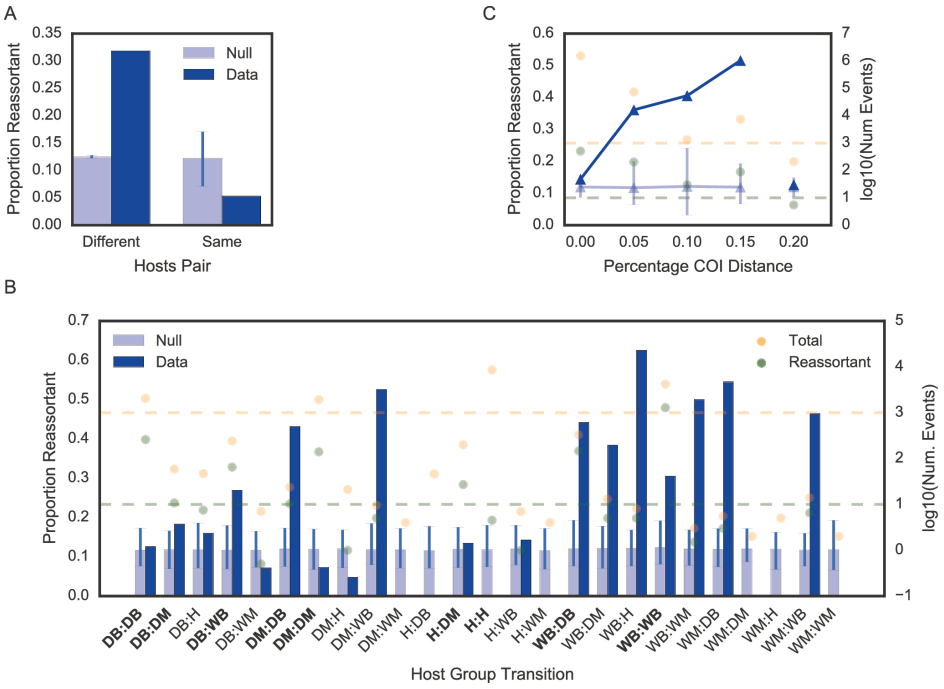

Figure 16: Reassortment is over-represented relative to clonal descent in transmission across host barriers. Proportion of reassortment events when crossing between (a) different or same hosts, (b) different host groups, and (c) hosts of differing evolutionary distance as measured by divergence in the cytochrome oxidase I (COI) gene. Reassortment is over-represented relative to clonal descent in transmission across host barriers. (b) D: Domestic animal, H: Human, W: Wild, B: Bird, M: Mammal. Donor host is labeled first. Bolded x-axis tick labels indicate data for which the weighted sum of all edges exceeded 1000, or the weighted sum of reassortant edges exceeded 10. (c) Pairwise distances between host's cytochrome I oxidase genes are binned in increments of 5%, or 0.05 fractional distance. (a, b, c) Vertical error bars on the null permutation model represent 3 standard deviations from the mean from 100 simulations (a, b), or 95% density intervals from 500 simulations (c). (b, c) Translucent dots indicate the weighted sum of all (clonal and reassortment) descent (yellow) and reassortment (green) events detected in the network under each host group transition. Horizontal yellow and green lines indicates threshold of values of 1000 and 10 respectively.

We further sought to explore whether the predominant use of reticulate evolutionary processes in host switch events were correlated with host phylogenetic relatedness and host ecology. To do this, we first computed the proportion of reassortment when switching between birds, non-human mammals, or humans, which are 3 divergent host groupings with distinct ecological behaviour. (For example, humans are the only known species to employ disease control measures, and affect the ecology of other species (birds and mammals through domestication) at scale.) We further sub-divided avian and mammalian categories into wild and domestic, to assess the impact of anthropological activity on the relative importance of reassortment in host switch interfaces (see Materials and Methods for how AIV was classified as domestic or wild). To ensure that the dataset was sufficient in scope to detect reassortant viruses, we only considered host group transitions with at least 1000 descent events (both clonal and reassortant), or at least 10 reassortment events (dashed yellow and green lines respectively in fig. 16 (b) & (c)). Nonetheless, all data are displayed for completeness.

Here, reassortment is over-represented relative to the null when host groups are different. Only two exceptions occur. The first is between wild birds, where reassortment is over-represented but host groups are not different. In this case, the "wild bird" label encompasses a wide range of host species, and as the natural reservoir for many diverse influenza viral subtypes, we expect to detect reassortment events more frequently between diverse species that may be distantly evolutionarily related.

The second is the human-domestic mammal interface, where reassortment is not over-represented even though the host groups are different. In the case of human to domestic mammal host switches (reverse zoonosis), these are mostly well-documented reverse zoonotic events between human and swine hosts (22)), where shared cellular receptors for viral entry (46) facilitates zoonotic and reverse zoonotic transmission. This may be a case of host convergent evolution inadvertently lowering the adaptive barrier to host switching. Under representation of reassortment at human-to-human transitions is expected because of the limited number of viral subtypes circulating in human populations that undergo serial selective sweeps, resulting in high sequence similarity within the viral pool (47), which likely obscures the distinction between reassortment and clonal descent. However, we also expect antibody-mediated immunity, whether from vaccination or prior exposure, to further limit the frequency of co-infection and likelihood of reassortment events happening amongst humans. Thus, despite the exceptions that may be explained by our current best knowledge of influenza biology (e.g. human to swine transmissions), reassortment is strongly favoured over clonal evolution when crossing between evolutionarily distant hosts.

To further explore the relationship between host evolutionary divergence and the predominance of reassortment in transmission events between species, we compared a common phylogenetic measure of species divergence, the cytochrome oxidase I (COI) gene, to the use of reassortment in host switch events. A subset of viral hosts, encompassing a variety of bird and mammal

species, have had their cytochrome oxidase I (COI) gene sequenced as part of the barcode of life project (48). For the subset of edges in the network for which both the source and sink hosts have a COI gene sequence that fulfilled our criteria for consideration (as described above), we computed the percentage evolutionary distance between the two hosts (Materials and Methods). Applying a similar permutation test and assessment criteria as described for host groups above, we found a trend of increasing over-representation at higher evolutionary distances (fig. 16 (c)). Thus, as host evolutionary distance, or more broadly, as quantitative niche dissimilarity increases, reticulate evolution becomes increasingly favoured for influenza virus niche switch events.

3.1.6 Discussion

In this study, we have quantitatively defined the importance of reticulate evolutionary events in switching ecological niches, using an infectious disease data set with characteristics that are particularly well suited for answering this question. Beyond the viral world, recent reviews have asserted the importance of reticulate evolutionary events as a driver of speciation and niche diversification (49, 50), and recent studies have illustrated heightened fitness effects in hybrid populations (51, 52). However, none have quantitatively tested the importance of reticulate evolutionary strategies in enabling ecological niche switches at a global scale, especially in comparison to clonal adaptation under drift and selection (a task feasible only in fast evolving organisms). Additionally,

no studies to date have examined reticulate evolutionary processes in the context of quantified niche differences, as we have done here by measuring reassortment in the context of host evolutionary distance. Our study provides strong quantitative evidence supporting the hypothesis that reticulate evolutionary processes are advantageous relative to adaptation by drift for pathogen transfer between host species, and therefore more broadly, ecological niche switching.

There are four limitations to this study. Firstly, we recognize that in this study, we have considered only a single pathogen for which abundant genomic data are available, and whose genomic and host tropic characteristics are suitable for this analysis. To specifically answer whether reticulate processes are favoured over clonal transmission for other organisms, using these methods, depends on being able to acquire genome sequences with matched ecological niche metadata.

Secondly, we also note that the global influenza dataset will have unavoidable sampling biases. For example, human isolates predominate in the dataset, and consequently the human-associated subtypes H3N2 and H1N1 also dominate the dataset. Sequences from viral outbreaks will also be over-represented relative to isolates collected through routine surveillance sampling, and will unavoidably lead to a heightened detection of clonal descent in a single host species. In order to deal with this sampling bias, our permutation tests (for the host species and group labels) involve class labels of equal sizes. This allows us to calculate an expected distribution of proportions under ideal assumptions of equal sampling, which in turn

forms the baseline for our conclusions.

Thirdly, our choice to use "host species" as the defined and quantified ecological niche is, in part, borne out of data availability. We naturally expect exceptions to occur if differences between species do not constitute a major barrier, or if barriers are defined by other characteristics of the host. Mallards are one example of such an exception (fig. 17). Amongst mallards, pre-existing immunity (perhaps quantifiable by antibody landscapes (53)) and high subtype diversity may be a strong driving forces for reassortment (54). We note that the necessary data do not currently exist to quantify barriers for other levels of defining and quantifying ecological niches, such as individuals or populations, at a global scale.

Finally, we do not specifically identify whether reassortment occurs prior to or after host switching, but only identify host transitions across which reassortment is implicated. A reassortment event may occur within a host species, during transfer between two host species, or after the transfer; reassortment's association with host switching will depend on when the reassortant virus is detected, and consequent clonal expansion of the reassortant strain will be identified as "clonally descended". Our method does not identify when the reassortment event happens, and this is both a limitation of our method and of IAV surveillance being less dense than necessary to distinguish between these two scenarios. Without better prior knowledge on whether reassortment happens prior to or after host switching, our method assumes that the detected reassortment events are the best possible representation of ground truth. It is with this limitation

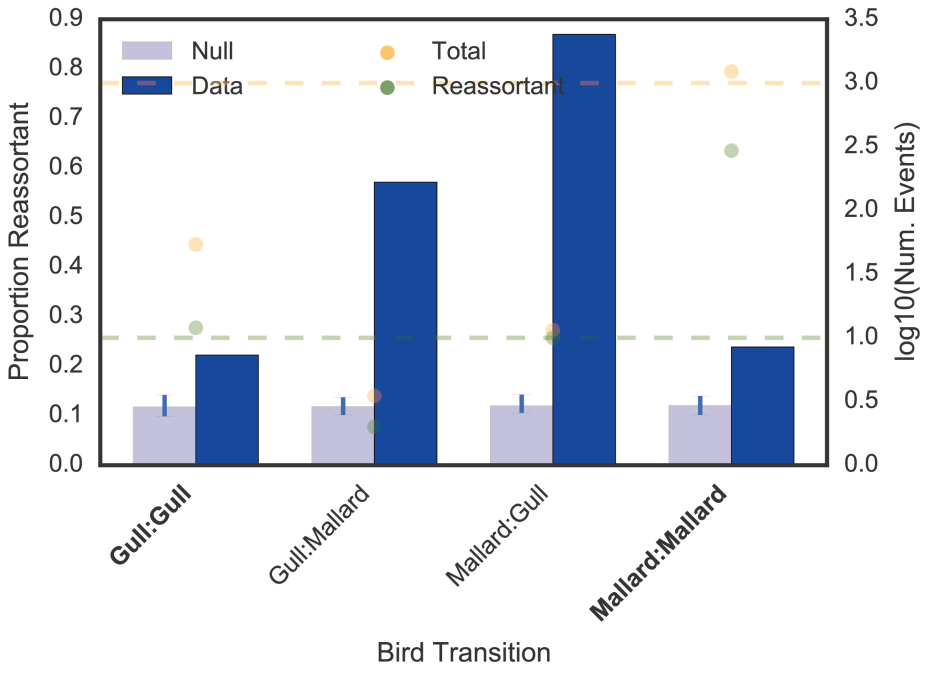

Figure 17: Analysis of reassortment amongst gulls and mallards. Vertical error bars on the null permutation model represent 99% density intervals from the mean from 100 simulations. Translucent dots indicate the weighted sum of all (clonal and reassortment) descent (yellow) and reassortment (green) events detected in the network under each host group transition. Horizontal yellow and green lines indicates threshold of values of 1000 and 10 respectively.

in mind that we identify associations of reassortment events with host switches, or more broadly across ecological niches. Whether reticulate evolution is causal for ecological niche switching will require further study.

In summary, using data available from a model zoonotic viral pathogen, we have shown that reticulate evolutionary processes are important in enabling pathogen host switches. For the influenza virus, reticulate evolution predominates when crossing between hosts. More broadly, the greater the quantitative difference between ecological niches, the greater the importance of reticulate evolutionary processes in enabling niche switches. While the quantitative importance of reticulate evolution may differ for different organisms evolving in different niches, we expect that further sequencing efforts from across broad domains of microbial life, and a further characterization and definition of their ecological niches, will elucidate whether this principle holds more broadly. Beyond its relevance to evolutionary ecology, reticulate evolution also has public health consequences. Reassortant influenza viruses have been implicated in all past human pandemic strains for which we have sequence data (55–58), and the ancestry of HIV-1 involved a hybrid SIV (59). Hence, knowing how reticulate events shape disease emergence may help the ecology and evolution of infectious disease become a more predictive science, leading to insight important to disease prevention and mitigation (60).

3.1.7 Methods

We describe here the methods specifically used for this application.

Permutation tests. Null models were constructed by permuting node labels; equal class size permutation was performed for host species (for fig. 16 (a)) and host group (for fig. 16 (b)). For example, if there were 9 'human', 4 'swine' and 2 'chicken' nodes (15 total), labels would be randomly shuffled amongst the nodes such that each label were equally represented (5 each).

Host Group Labelling. Host species were manually classified into the "human", "domestic animal" and "wild animal" groups, based on the country of isolation. For example, "ducks" would be considered a "wild animal" in North America, while it would be considered a "domestic animal" in East Asian countries. Ambiguous host species, while remaining in the dataset, were excluded from the analysis.

Host Evolutionary Distance. Host species' scientific names were sourced from the Tree of Life database (www.tolweb.org). Only host species with unambiguous scientific names recorded were considered. Cytochrome oxidase I genes were sourced from the Barcode of Life Database (www.boldsystems.org) on 31 October 2015. Sequences had to be at least 600 n.t. long to be considered, and only positions with fewer than 3 gap characters were concatenated into the final trimmed alignment. Evolutionary distance was computed from the trimmed alignment as the proportion of mismatched nucleotides. Further details are available in the Jupyter notebooks.

Phylogenetic Reconstruction and Patristic Distance Comparison. Phylogenetic reconstruction was done for a subset of H3N8 viruses isolated from Minto Flats, AK, between 2009 and 2010 as part of a separate study.

Briefly, each segment of the viral genomes were individually aligned using Clustal Omega (27), and their genealogies reconstructed using BEAST 1.8.0 (29). A minimum of 3 MCMC runs that converged on a single optimal tree were chosen to compute the maximum clade credibility (MCC) tree. Burn-in ranged from 10 to 39 million steps out of 40, with median 24 million steps. Patristic distances were calculated using the DendroPy package (61). In the graph reconstruction on the Minto Flats study, we extracted the edges and nodes involving only the H3N8 isolates, and computed the tree patristic distances between isolate pairs linked by an edge in the graph.

Edge Weighting and Proportion Reassortment Calculations. The proportion of reassortment events was calculated by first weighting each incoming edge to every virus. The weighting procedure is described here: If the virus is detected to be plausibly clonally descended from n other viruses, as determined by maximal similarity, it is given a weight of $\frac{1}{n}$. If the virus is detected to be a reassortant, then edges are weighted by the the fraction of times that it is involved in a max similarity source pair. (fig. 18) For example, if a given node A has a plausible source of segments in B, C, D, with (B and C) and (B and D) being plausible sources, then the edge B-A would be given a weight of 0.5, and the edge C-A and D-A would be given a weight of 0.25 each. (fig. 18) The proportion of reassortment edges was then calculated by taking the sum of weights across all reassortant edges (for a particular transition, e.g. between or within host species), divided by the sum of weights across all reassortant and clonal edges (for the same particular transition). With this weighting scheme, multiple plausible sources that lead to the same

virus are not double-counted.

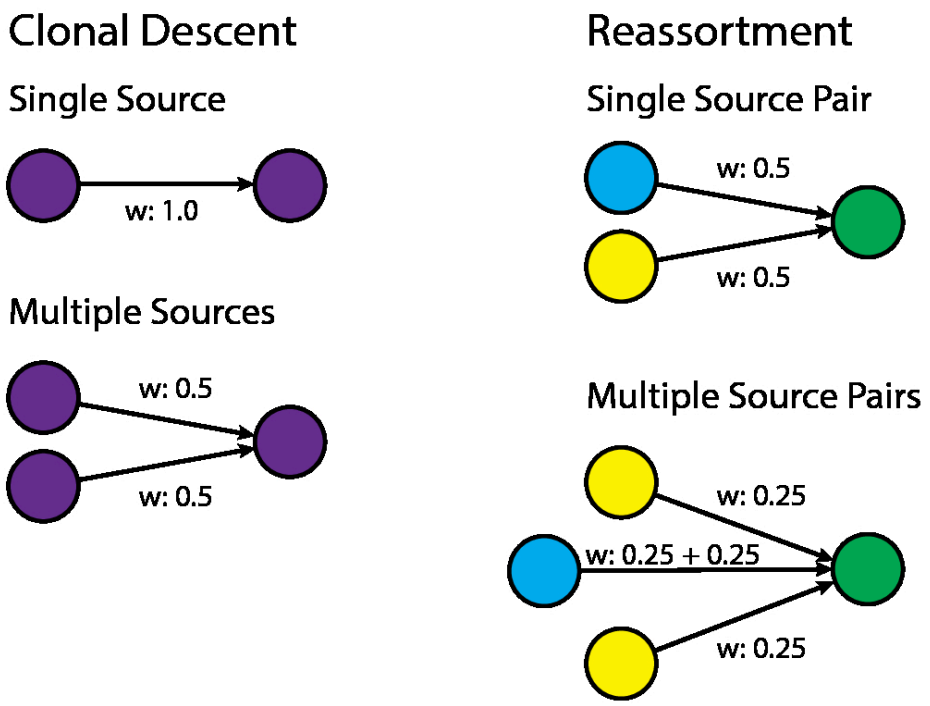

Figure 18: Illustration of how our network reconstruction method deals with multiple plausible sources, for the "Clonal Descent" and "Reassortment" scenarios. Weightings, rather than summed pairwise identities, are shown on the edges. Colours represent different viral lineages.

3.2 Application 2: Viral persistence.

The bulk of this study, including the writing, was led by Dr. Nichola J. Hill and published in Ecology Letters (62); as such, apart from the abstract and two of the figures, I provide a synopsis of the work rather than the original text. Additionally, much effort was dedicated to sample collection in Fairbanks, AK, led by Brandt Meixell, without which this study would

not have been possible, and I would like to acknowledge the effort put in by the sample collection team in enabling this study.

3.2.1 Abstract

Influenza A Viruses (IAV) in nature must overcome shifting transmission barriers caused by the mobility of their primary host, migratory wild birds, that change throughout the annual cycle. Using a phylogenetic network of viral sequences from North American wild birds (2008–2011) we demonstrate a shift from intraspecific to interspecific transmission that along with reassortment, allows IAV to achieve viral flow across successive seasons from summer to winter. Our study sup- ports amplification of IAV during summer breeding seeded by overwintering virus persisting locally and virus introduced from a wide range of latitudes. As birds migrate from breeding sites to lower latitudes, they become involved in transmission networks with greater connectivity to other bird species, with interspecies transmission of reassortant viruses peaking during the winter. We propose that switching transmission dynamics may be a critical strategy for pathogens that infect mobile hosts inhabiting regions with strong seasonality.

3.2.2 Research Question

The influenza A virus has a broad host range. Canonically, it is thought that the virus' reservoir host[6] are wild birds, namely *Anseriformes* and *Charadriiformes* (15). Being migratory birds, they do not form stable and geographically-restricted populations; rather, their migration patterns are seasonal, and their movement depends heavily on environmental cues. Given the influenza A virus' dependence on its viral hosts, the fluctuating host environment gives rise to the question of whether host seasonality changes the virus' transmission strategy, here defined as being clonal descent or through reassortment. For example, does reassortment predominate during the breeding state of wild birds, or does it predominate during the stages after the birds have left their breeding grounds?

What do our time-stamped sequence data and source-sink reconstruction (from genomic sequence) of influenza transmission tell us about reassortment and its role in viral persistence?

3.2.3 Research Methods and Findings

Working with Dr. Nichola Hill, we integrated densely-sampled influenza A virus sequence data from the Minto Flats State Game Refuge (AK) with host collection metadata to answer this question. Data were collected over 4 years from 2008-2012, with 14004 wild birds sampled yielding 545 influenza genomes. To interrogate migration patterns of the virus, we

[6]The term "reservoir host" is typically defined as a host population in which a virus can persistently circulate in without the host incurring a (large degree of) fitness cost.

also included 1242 fully sequenced influenza viruses sampled across North America between 2008 and 2012, yielding 1787 viruses that were used for network reconstruction.

A modification to the algorithm was done to make the virus epidemiologically-relevant for this study; because the date of sampling may not necessarily correspond to the date of infection, we modified the network reconstruction algorithm to permit sources to occur after sinks by up to 6 days, which is roughly the length of time that an infected bird sheds viruses. The resultant graph had two types of edges, "full complement" (whole genome) transmissions, and "reassortment" transmissions, similar to our global study.

Through analysis of the time-span of edges, we found that full complement edges mostly spanned short time scales, dominating short chain transmissions, while reassortment edges spanned annual seasons. As shown in fig. 19 (a), full complement edges (top panel) generally started and ended in viruses isolated from the same years. On the other hand, while the majority of reassortment edges (bottom panel) were also constrained to the same season, there was a greater proportion that spanned longer time frames, as shown by the cluster of off-diagonal scatter points that begin at an early source date and end at a sink date separated by hundreds of days. One additional novel finding, which was previously difficult to quantify, was the presence of inter-annual persistence of viruses, mediated primarily through reassortment.

Additionally, we observed that reassortment edges spanned greater

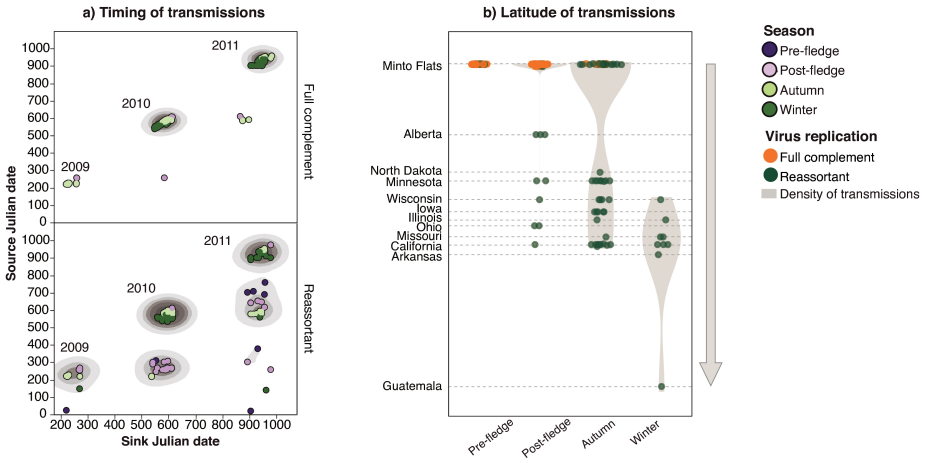

Figure 19: The distribution of transmissions at Minto Flats, Alaska according to (a) time and (b) latitude. Temporal analysis (a) depicts the density of transmissions among ducks at Minto Flats (grey concentric circles) during 2009–2011. The majority of full complement transmissions (top panel) were sourced locally from Minto Flats and involved ducks with source and sink dates that overlapped (i.e. transmission within the same breeding season). However, reassortant viruses at Minto Flats (lower panel) were also seeded from virus shed by ducks up to two years earlier, originating primarily in autumn and winter from prior annual cycles. Data from 2008 is included but not depicted as this year was a source, but not sink of interannual transmissions. Spatial analysis (b) indicates the latitudinal span of locations that were sourced by virus originating at Minto Flats. The majority of viruses introduced into lower latitudes were detected as reassortants.

geographic scales (fig. 20 (a) and (b)) than full complement edges. As shown in (fig. 19 (a)), viruses that occurred later in the calendar year, as birds migrated to lower latitudes, were detected as reassortant viruses rather than the product of short-chain transmission. Additionally, when visualizing the co-variation between source and sink longitude (fig. 20 (a)) and latitude (fig. 20 (b)), full complement edges tended to co-vary with one another, beginning and ending in the same longitude or latitude. On the other hand, reassortment edges had a greater spread, as indicated by the 90% ellipse.

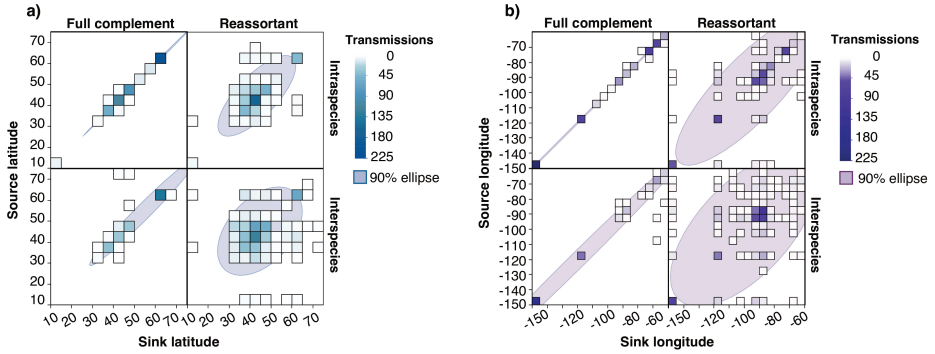

Figure 20: Spatial pattern of viral transmissions between wild birds in North America, 2008–2011 according to (a) latitude and (b) longitude. Heatmaps indicate the number of transmission events highlighted by colour (blue: latitude, purple: longitude) and the shaded areas (90% ellipse) where transmissions are concentrated. Full complement transmissions are localized both latitudinally and longitudinally, indicated by the source and sink locations overlapping. In contrast, transmissions involving reassortment are more spatially diffuse. The combination of reassortant and interspecies transmission resulted in greatest dispersal both latitudinally and longitudinally.

Taken together, the data support a model where reassortment and expansion of host species is a way for the influenza A virus to persist in wild birds.

This may be because direct transmission becomes progressively harder as the seasonal cycle progresses, due to transmission barriers that may arise.[7]

3.3 Caveats

A caveat common to this study and the global influenza study before is that sampling efforts are ad-hoc (often in response to a new outbreak or directed towards testing some hypothesis), non-uniform over space and time (owing to the movement of viral hosts and logistical issues), and yield varying amounts of sequence data (owing to differences in prevalence over time). This potentially could bias the inference of reassortment events. To tackle this in the first application, we tested the sensitivity of the inferences to biased vs. equal sampling in our null model; the conclusions did not change. In the second application, we progressively sub-sampled an increasing maximum number of sequences for a host species, and investigated how it affected the metrics quantified. No trends were observed, indicating that sampling bias did not affect our conclusions. Nonetheless, while resampling and permutation tests can help us measure how sampling biases may change the conclusion, the gold standard (and politically/financially tougher) method would be to re-structure sampling efforts to be constant over space and time.

We have used permutation tests (global study) and chi-square tests (viral migration study)[8] to test whether the what we observed was likely to

[7] By transmission barriers, we refer to biotic factors such as population immunity, and abiotic factors such as an unfavourable climate for environmental persistence.

[8] Intuitively, I believe that the permutation test can be simplified to a chi-square test

happen by a null model of random chance. However, this nonetheless raises the question, "What is the most appropriate null model?" This is a difficult philosophical question, as with network models there potentially many different "nulls" to compare against,[9] and I have no good answer to this.

itself, but at the time of paper writing I was not sufficiently confident in concluding this, and hence I proceeded with the permutation test.

[9] As an example, rather than shuffling the node labels in a graph, one may instead choose to rewire the connectivity of the graph randomly instead.

4 Remaining Challenges & Future Work

4.1 Scientific

4.1.1 Viral Packaging and Reassortment

An open question that has not been addressed is whether packaging signals can bias the segments that co-assort with one another. In exploratory work conducted on the side, I observed that the polymerase segments had a tendency to co-assort with one another (fig. 21), concurring with observations made before (63). One hypothesis is that the polymerase segments tend to co-assort because they have to work together; there is negative selective pressure against polymerase segments from different viruses, because they may not confer high fitness (e.g. fast replication rate). On the other hand, another hypothesis may be the compatibility of packaging signals from two viruses. The relative contributions of the two to packaging compatibility, and hence co-assortment frequencies, are unclear, and it remains an open question to be investigated.

4.1.2 Quantification of Ecological Niche Differences

Ecological niches are defined as being the set of activities that help an organism survive and reproduce, and is influenced by abiotic (e.g. temperature, sunlight) and biotic factors (presence or absence of predators). Biotic factors are traditionally described in qualitative or categorical terms: predation (X eats Y) and symbiosis (A helps B which

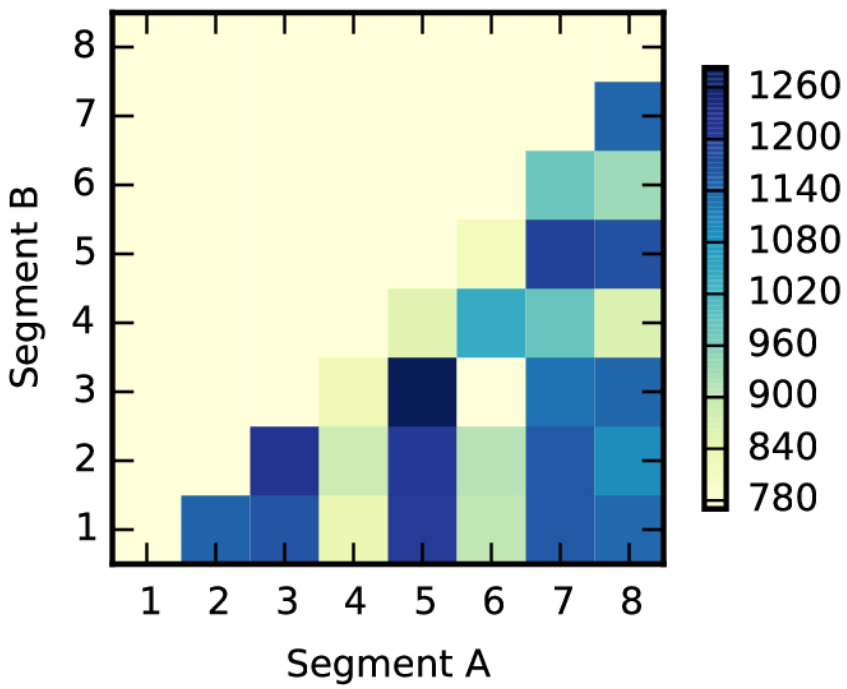

Figure 21: Coassortment counts between influenza genome segments as observed in the global analysis of reassortment.

also helps A) being the two most common categories, in addition to parasitism, neutral interactions, and more. As a global population, the influenza A virus lives essentially in a parasitic relationship with its hosts (as far as we can observe), and one idea put forth in the global reticulate evolutionary study was the idea of quantitative niche differences, as quantified by host evolutionary distance.

I recognize that cytochrome oxidase I sequence is a crude approximation of virus host differences; mechanistically speaking, it is more likely that quantitative differences in the immune system repertoire between each hosts would play a greater functional role in viral fitness. As an example, one might opt to sequence the B-cell receptor (BCR) repertoire of two viral hosts, and pass their sequences through a dimensionality reduction algorithm (e.g. multi-dimensional scaling, variational autoencoders), and use that lower-dimension representation as a way of quantifying differences in the immune system of a variety of hosts.

4.1.3 Observation of Viral Subtypes

With 16 canonical hemagglutinin and 9 neuraminidase subtypes identified, there are theoretically 144 possible influenza subtypes that can form. Of them, we have observed (in the sequence dataset) about only $\frac{2}{3}$ of them (fig. 22). One may wonder, then, why we have not observed all 144 of them yet? Can we forecast which subtypes (or reassortant viruses) could be detected next?

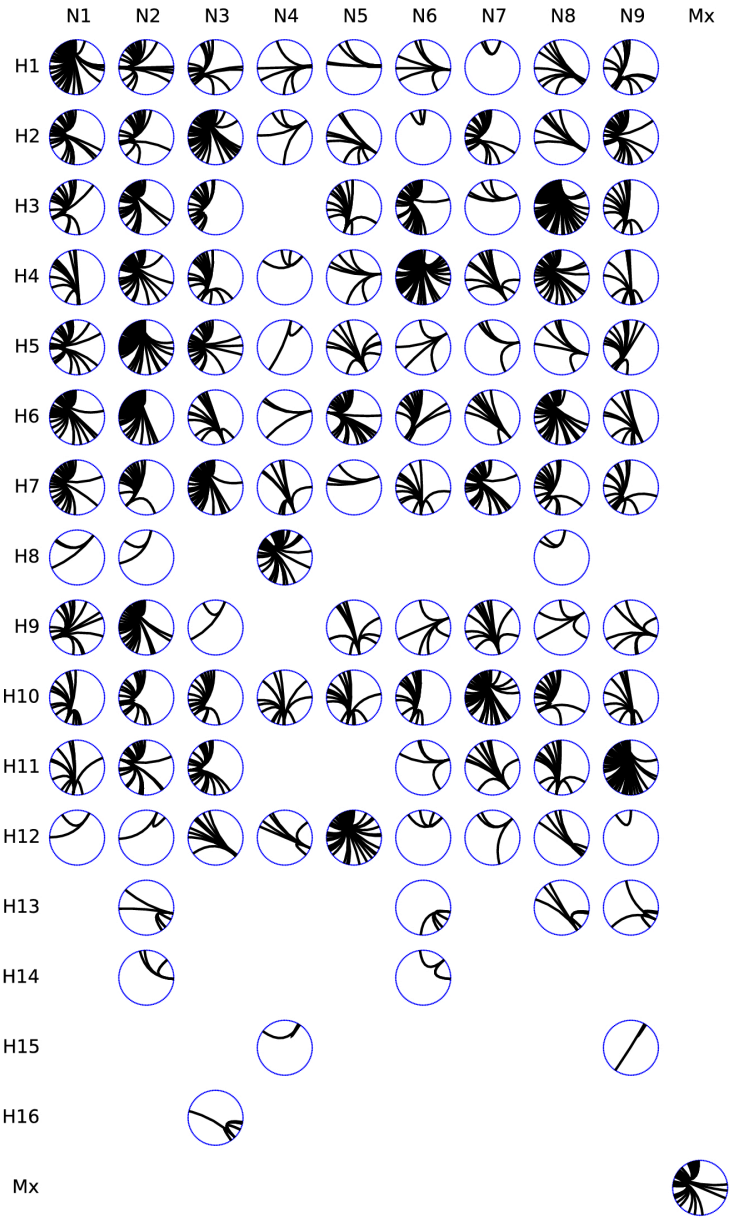

Figure 22: Circos panel depicting the connectivity of a particular HA & NA subtype combination with other subtypes. Within each circos plot, subtypes are ordered from the 12 o'clock position in increasing connectivity, starting with the lowest-ranked at 12 o'clock and increasing clockwise. The highest-connected subtype, H3N8, is found just before the 12 o'clock position. Mx: "mixed subtype".

A naive statistical model of new subtype emergence would assume that reassortment between subtypes happens only by chance, and that us not having observed all subtypes would merely be a function of time. This would assume that identity of new subtype emergence is essentially unpredictable.

Another more sophisticated statistical model of new subtype emergence would take greater advantage of the observation that certain subtypes of virus are more heavily associated with particular viral hosts. Because viruses have a very strong reliance on hosts for movement, it would make sense that where hosts overlap geographically, and even interact at a "host ecological niche" level, there would be a greater probability of viral subtypes associated with those host pairs to reassort, and hence produce new viral subtypes.

4.1.4 Denser Sampling

As things stand right now, we do not detect whether a reassortment event happens prior to or after the host switch event. This caveat was addressed in Application 1, where we argued that it did not matter whether reassortment was helpful prior to or after host switching, as the fitness gains in the new host could be acquired at either step. Nonetheless, it may be of scientific interest to decide between these two hypotheses, in which case denser sampling is required to be able to identify whether a reassortment event happened prior to or after a host switch event. Additionally, single viral particle sequencing from a viral host would be a very enabling technology here, and if paired with denser sampling, would open the doors to quantifying the relative viral load in a mixed infection,

including reassortant viruses that emerge from the coinfection.

4.1.5 Homologous Reassortment

Detection of reassortment between genotypically similar (but non-identical) viruses, which we might want to call "homologous reassortment", remains an oft-cited challenge for reassortment detection. Reassortment frequency is high and fairly unbiased under neutral fitness conditions (64) (i.e. there is no change in viral protein sequence, but only nucleotide sequence). However, this knowledge also raises the question about the utility of knowing how much homologous reassortment happens - if there are minimal (or no) fitness differences, then is homologous reassortment consequential for the evolution of a virus? My own answer, based on intuition, is that homologous reassortment likely has little impact on the evolution of a virus, and that the surveillance community would be better concerned with heterologous reassortment.

4.1.6 Probabilistic Identification of Reassortant Viruses

The algorithm, as it stands right now, only provides a deterministic identification of the parental sources, based on the evolutionary distance computed by Clustal Omega (27). I define this as "deterministic" because only the viruses that (when taken together) give the highest summed PWI, will be chosen as parental viruses.

A logical next step would be to extend the algorithm to identify not merely

the parental sources that are of highest PWI, but to also assign a likelihood score to the parental source, given the evolutionary distance. The simplest thing that could be attempted is to use the summed PWI as a probability metric itself, as follows:

$$\frac{\sum_{k=1}^{S} p_k}{S}$$

where p is the PWI for any given segment, and S is the number of segments. To retrieve a probability score for any given PWI, one would then apply a 'softmax' normalization across all valid parental combinations, thus normalizing the probability scores to sum to 1. No doubt this still requires the assumption of the algorithm being a phylogenetic heuristic, rather than an attempt at ground truth reconstruction.

4.2 Engineering

4.2.1 Deployment

Algorithms developed in the academic world often require "just some engineering" to be made ready for deployment to the real-world. It remains on my personal wish-list to have turned this reassortment detection algorithm into a standalone software package, similar to BEAST (29) or GiRaF (9). It is also my desire to have a continually-updated monitoring system, similar to NextFlu (65), to continually detect newly-sequenced reassortant viruses as they show up. These goals were hampered by my

personal lack of software engineering training, and as such I never got around to doing a proper refactoring of the code into reusable modules. Thankfully, at the moment, the Influenza Research Database team, led by Prof. Richard Schuermann at the JCVI, is considering incorporating the reassortment detection code as part of their data platform, and I would like to thank them for their interest.

4.2.2 Automation

The code, as it stands right now, was designed for execution on a Sun Grid Engine (SGE) compute cluster. This design enabled manual parallelism wherever the code was embarrassingly parallel, in a map-reduce paradigm. For example, one key step is the creation of a multiple sequence alignment for each influenza A virus segment. Because the the alignment of one segment is not dependent on the alignment of another segment, they could be aligned in parallel, with the alignment of longer segments taking longer than the alignment of shorter segments. However, a few steps after that, there is a "reduction" step that is dependent on having all 8 evolutionary distance matrices computed fully, and this was one example of a step that was not automated because of (1) a lack of expertise in parallel computation and (2) the nature of the SGE scheduling system not being accessible from an external API.

With the development of Python-based software schedulers (e.g. Dask (66)) enabling automatic execution of complex, arbitrary computation graphs, a rework of the code could be performed to make it executable with a single

command from the command line. Dask has the added advantage of being able to scale from single cores to cloud infrastructure, though at the moment SGE clusters are not supported.

5 Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Jonathan Runstadler, for providing guidance and mentorship. I entered both worlds of infectious disease and computational research from scratch, and Jon provided the support and environment that enabled me to grow in both fields.

Secondly, I would like to thank my committee for their support and mentorship. Profs. Mark Bathe (MIT) and Jukka-Pekka Onnela (Harvard School of Public Health) both provided me with opportunities to interact with their research groups. These opportunities were valuable to my learning during tenure as a graduate student.

Thirdly, I would like to thank my wife, Dr. Nan Li, for her unwavering support and companionship. Her intelligence is a constant source of inspiration, and her patience has been a source of comfort for me.

Fourthly, I would like to acknowledge colleagues with whom I have worked closely on research both described here and not described here. Dr. Nichola Hill has been a wonderful collaborator and has taught me many ecology concepts, a topic I was not deeply trained in. Dr. Islam Hussein has been a close collaborator on experimental work; alongside Mia Lieberman (DCM), I have had much fun helping them with statistical analysis problems. Drs. David Duvenaud and Matthew Johnson, now at U Toronto and Google Brain respectively, gave me a wonderful introduction to the world of deep learning when I camped myself up at their office in Harvard.

Fifthly, I would like to acknowledge the undergraduate students that I have had a privilege to work with: Andrea Nickerson, Justin Zabilansky, Kyle Yuan, Ellie Laukitis, and Vivian Zhong. You are an amazing bunch, and have gone on to do wonderful things. I wish you all the best for your careers - continue doing amazing stuff!

I would also like to acknowledge the financial support and provision of resources by the Department of Biological Engineering, the BioMicroCenter, and the Broad Institute of Harvard and MIT. The research funds provided and access to compute resources have been instrumental in conducting my research, including research projects done outside of this thesis.

Finally, all praise be to the Lord Jesus Christ, who has graciously provided all that I needed throughout my time in graduate school. My research has helped me refine a much more nuanced view on the relationship between the divine and our physical world, while simultaneously leaving me in awe at the complexity of nature. I have also learned the meaning of worship through excellence in our work. While we continue the fight against infectious disease, I look forward to that day, when "there will be no more death or mourning or crying or pain, for the old order of things has passed away" (Revelation 21:4), and infectious diseases, which have given me a topic for work, will finally be done with and be no more.

References

1. Molinari N-AM, et al. (2007) The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* 25(27):5086–5096.

2. Holmes EC (2003) Error thresholds and the constraints to RNA virus evolution. *Trends in microbiology* 11(12):543–546.

3. Wu Y, Wu Y, Tefsen B, Shi Y, Gao GF (2014) Bat-derived influenza-like viruses H17N10 and H18N11. *Trends in microbiology* 22(4):183–191.

4. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer).

5. Lemey P, Rambaut A, Welch JJ, Suchard MA (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution* 27(8):1877–1885.

6. Gire SK, et al. (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (New York, NY)* 345(6202):1369–1372.

7. Park DJ, et al. (2015) Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* 161(7):1516–1526.

8. Nelson MI, et al. (2015) Global migration of influenza A viruses in swine. *Nature communications* 6:6696.

9. Nagarajan N, Kingsford C (2011) GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic acids*

research 39(6):e34–e34.

10. Rabadan R, Levine AJ, Krasnitz M (2008) Non-random reassortment in human influenza A viruses. *Influenza and Other Respiratory Viruses* 2(1):9–22.

11. Goto H, Muramoto Y, Noda T, Kawaoka Y (2013) The genome-packaging signal of the influenza A virus genome comprises a genome incorporation signal and a genome-bundling signal. *Journal of virology* 87(21):11316–11322.

12. Hutchinson EC, Wise HM, Kudryavtseva K, Curran MD, Digard P (2009) Characterisation of influenza A viruses with mutations in segment 5 packaging signals. *Vaccine* 27(45):6270–6275.

13. Gog JR, et al. (2007) Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic acids research* 35(6):1897–1907.

14. Gerber M, Isel C, Moules V, Marquet R (2014) Selective packaging of the influenza A genome and consequences for genetic reassortment. *Trends in microbiology* 22(8):446–455.

15. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. *Microbiological reviews* 56(1):152–179.

16. Kuiken T, et al. (2004) Avian H5N1 influenza in cats. *Science (New*

York, NY) 306(5694):241.

17. Hussein ITM, et al. (2016) New England harbor seal H3N8 influenza virus retains avian-like receptor specificity. *Scientific Reports* 6:21428.

18. Anthony SJ, et al. (2012) Emergence of fatal avian influenza in New England harbor seals. *mBio* 3(4):e00166–12.

19. Lang G, Gagnon A, Geraci JR (1981) Isolation of an influenza A virus from seals. *Archives of virology* 68(3-4):189–195.

20. Fereidouni S, Munoz O, Von Dobschuetz S, De Nardi M (2014) Influenza Virus Infection of Marine Mammals. *EcoHealth* 13(1):1–10.

21. Wallensten A, et al. (2006) Mounting evidence for the presence of influenza A virus in the avifauna of the Antarctic region. *Antarctic …* 18(03):353–356.

22. Nelson MI, Vincent AL (2015) Reverse zoonosis of influenza to swine: new perspectives on the human-animal interface. *Trends in microbiology* 23(3):142–153.

23. Gagneux P, et al. (2003) Human-specific regulation of alpha 2-6-linked sialic acids. *Journal of Biological Chemistry* 278(48):48245–48250.

24. Matrosovich M, Zhou N, Kawaoka Y, Webster R (1999) The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *Journal of virology* 73(2):1146–1155.

25. Ito T, et al. (1998) Molecular basis for the generation in pigs of influenza

A viruses with pandemic potential. *Journal of virology* 72(9):7367–7373.

26. Jombart T, Eggo RM, Dodd PJ, Balloux F (2011) Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 106(2):383–390.

27. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7(1):539–539.

28. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science (New York, NY)* 315(5814):972–976.

29. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* 29(8):1969–1973.

30. Ma EJ, Hill NJ, Zabilansky J, Yuan K, Runstadler JA (2016) Reticulate evolution is favored in influenza niche switching. *Proceedings of the National Academy of Sciences of the United States of America* 113(19):201522921–5339.

31. Hernández-López A, et al. (2013) To tree or not to tree? Genome-wide quantification of recombination and reticulate evolution during the diversification of strict intracellular bacteria. *Genome biology and evolution* 5(12):2305–2317.

32. Peris D, et al. (2014) Population structure and reticulate evolution of Saccharomyces eubayanus and its lager-brewing hybrids. *Molecular Ecology*

23(8):2031–2045.

33. Garten RJ, et al. (2009) Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. *Science (New York, NY)* 325(5937):197–201.

34. Smillie CS, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–244.

35. Antonenka U, Nölting C, Heesemann J, Rakin A (2005) Horizontal transfer of Yersinia high-pathogenicity island by the conjugative RP4 attB target-presenting shuttle plasmid. *Molecular microbiology* 57(3):727–734.

36. Remold SK, Rambaut A, Turner PE (2008) Evolutionary Genomics of Host Adaptation in Vesicular Stomatitis Virus. *Molecular biology and evolution* 25(6):1138–1147.

37. Duggal NK, Emerman M (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. *Nature reviews Immunology* 12(10):687–695.

38. Li C, et al. (2010) Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. *Proceedings of the National Academy of Sciences* 107(10):4687–4692.

39. Mehle A, Dugan VG, Taubenberger JK, Doudna JA (2012) Reassortment and Mutation of the Avian Influenza Virus Polymerase PA Subunit Overcome Species Barriers. *Journal of virology* 86(3):1750–

1757.

40. Lam TTY, et al. (2011) Reassortment Events among Swine Influenza A Viruses in China: Implications for the Origin of the 2009 Influenza Pandemic. *Journal of virology* 85(19):10279–10285.

41. Tao H, Steel J, Lowen AC (2014) Intra-host dynamics of influenza virus reassortment. *Journal of virology* 88(13):JVI.00715–14–7492.

42. Ince WL, Gueye-Mbaye A, Bennink JR, Yewdell JW (2013) Reassortment complements spontaneous mutation in influenza A virus NP and M1 genes to accelerate adaptation to a new host. - PubMed - NCBI. *Journal of virology* 87(8):4330–4338.

43. Steel J, Lowen AC (2014) Influenza A Virus Reassortment. *Influenza Pathogenesis and Control-Volume I* (Springer International Publishing, Cham), pp 377–401.

44. Furuse Y, Suzuki A, Oshitani H (2010) Reassortment between swine influenza A viruses increased their adaptation to humans in pandemic H1N1/09. *Infection, Genetics and Evolution* 10(4):569–574.

45. Squires RB, et al. (2012) Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses* 6(6):404–416.

46. Ma W, Kahn RE, Richt JA (2008) The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *Journal of molecular and genetic medicine : an international journal of biomedical research*

3(1):158–166.

47. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics* 192(2):671–682.

48. Ratnasingham S, Herbert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular ecology notes* 7(3):355–364.

49. Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD (2015) The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Molecular Ecology* 24(9):2095–2111.

50. Molofsky J, Keller SR, Lavergne S, Kaproth MA, Eppinga MB (2014) Human-aided admixture may fuel ecosystem transformation during biological invasions: theoretical and experimental evidence. *Ecology and evolution* 4(7):899–910.

51. Verhoeven KJF, Macel M, Wolfe LM, Biere A (2011) Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proceedings of the Royal Society B: Biological Sciences* 278(1702):2–8.

52. Keller SR, Fields PD, Berardi AE, Taylor DR (2014) Recent admixture generates heterozygosityfitness correlations during the range expansion of an invading species. *Journal of Evolutionary Biology* 27(3):616–627.

53. Fonville JM, et al. (2014) Antibody landscapes after influenza virus

infection or vaccination. *Science (New York, NY)* 346(6212):996–1000.

54. Wille M, et al. (2013) Frequency and patterns of reassortment in natural influenza A virus infection in a reservoir host. *Virology* 443(1):150–160.

55. Wu A, et al. (2013) Sequential Reassortments Underlie Diverse Influenza H7N9 Genotypes in China. *Cell host & microbe* 14(4):446–452.

56. Pu J, et al. (2014) Evolution of the H9N2 influenza genotype that facilitated the genesis of the novel H7N9 virus. *Proceedings of the National Academy of Sciences of the United States of America* 112(2):201422456–553.

57. Lam TT-Y, et al. (2013) The genesis and source of the H7N9 influenza viruses causing human infections in China. *Nature* 502(7470):241–244.

58. Gao R, et al. (2013) Human infection with a novel avian-origin influenza A (H7N9) virus. *The New England journal of medicine* 368(20):1888–1897.

59. Bailes E, et al. (2003) Hybrid origin of SIV in chimpanzees. *Science (New York, NY)* 300(5626):1713–1713.

60. Wasik BR, Turner PE (2013) On the biological success of viruses. *Annual review of microbiology* 67(1):519–541.

61. Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics (Oxford, England)* 26(12):1569–1571.

62. Hill NJ, et al. (2016) Transmission of influenza reflects seasonality of wild birds across the annual cycle. *Ecology letters* 19(8):915–925.

63. Chan JM, Carlsson G, Rabadán R (2013) Topology of viral evolution.

pnasorg 110(46):18566–18571.

64. Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC (2013) Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathogens* 9(6):e1003421.

65. Neher RA, Bedford T (2015) nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics (Oxford, England)* 31(21):3546–3548.

66. Team DD *Dask: Library for dynamic task scheduling.*

List of Tables

List of Figures