1 **Title**: Reticulate evolution is favored in microbial niche switching

2 **Authors**: Eric J. Ma, Nichola J. Hill, Justin Zabilansky, Kyle Yuan, Jonathan A. Runstadler

3 **Affiliations**: Department of Biological Engineering, Massachusetts Institute of Technology, 77

4 Massachusetts Ave, Cambridge, MA, 02139

5 **Corresponding Authors:** Jonathan A. Runstadler, jrun@mit.edu; Eric J. Ma,

6 ericmajinglong@gmail.com

7 **Classification:**

8 Major: Biological Sciences

9 Minor: Ecology, Evolution

10 **Keywords:** Ecology, Reticulate Evolution, Influenza,

11

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 favored. We conclude that the greater the quantitative difference between ecological niches, the greater the importance of reticulate evolutionary processes in overcoming niche barriers.

Significance Statement:

Are the processes that result in the exchange of genes between microbes quantitatively advantageous for those microbes when switching between ecological niches? 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, reticulate evolutionary processes are quantitatively favoured when switching between ecological niches.

\body

Manuscript

Reticulate evolutionary processes, such as horizontal gene transfer (HGT) and genomic reassortment, have been proposed as a major mechanism for microbial evolution (*1*), aiding in the diversification into new ecological niches (*2*). 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 (*3*), and antibiotic resistance (single genes) (*4*) and pathogenicity islands (or gene clusters) in bacteria (*5*). Host switching, defined as a pathogen moving from one host species into another, represents a strong fitness barrier to microbial pathogens. The acquisition of adaptations through reticulate processes either prior to 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 (*6*). 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. In order 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 the influenza genome sequencing projects provides a unique opportunity for quantitatively testing this concept, and is suitable for the following reasons. Firstly, the influenza A virus (IAV) has a broad host tropism (*7*), 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 (*8*). Secondly, IAV is capable of and frequently undergoes reassortment, which is a well documented reticulate evolutionary process (*9-12*). Finally, due to surveillance efforts over the past two decades, whole genome sequences have been intensively sampled over a long time frame, with corresponding host species metadata, available in an easily accessible and structured format (*13*). Because reassortant viruses are the product of two or more genetically distinct viruses co-infecting the same host, a more complex process than clonal transmission and adaptation, they are expected to occur less frequently. Hence, the comprehensive 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.

In order to identify reassortment events and the hosts species involved, we adapted a phylogenetic heuristic method (*14*), and mapped out a network of clonal and reassortment descent relationships from a comprehensive set of completely sequenced IAV (18,632 viral genomes) downloaded from the Influenza Research Database (*13*). Briefly, the core logic of the method is as such: for every isolate in the dataset, we look for genomic sources such that the sources found are of maximal similarity across all 8 genomic segments (Materials and Methods). Clonal descent involves single sources, while reassortment descent involves source pairs. Where multiple sources or source pairs correspond to the maximal similarity, all are kept as plausible sources. In the resulting network, nodes are individual viral isolates, and edges are the clonal or reassortment descent relationships.

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 (Supplementary Figure 1). Edges in the network connected viral isolates with a median genetic similarity of 99.7%, indicating high degrees of phylogenetic similarity (Supplementary Figure 2). 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. Small-world simulation studies and phylogenetic comparisons validated our method as being accurate in detecting reassortment events (Supplementary Figures 3 and 4).

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) (Figure 1a), 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 favored strategy when influenza crosses between different host species.

We further sought to explore whether the predominant use of reticulate evolutionary processes in host switch events were correlated with host phylogenetic relatedness. 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. 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 lines in Figures 1b & 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 (*15*)), where shared cellular receptors for viral entry (*16*) 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 highly similar viral subtypes circulating in human populations, which likely obscures the distinction between reassortment and clonal descent. However, we also expect disease control measures to further limit the frequency of co-infection and likelihood of reassortment events. Thus, despite exceptions which are explained by known influenza biology (e.g. human to swine transmissions), reassortment is strongly favored over clonal evolution when crossing between evolutionarily distant hosts.

124 To further explore the relationship between host evolutionary divergence and the predominance
125 of reassortment in transmission events between species, we compared a common phylogenetic
126 measure of species divergence, the cytochrome oxidase I (COI) gene, to the use of reassortment
127 in host switch events. A subset of viral hosts, encompassing a variety of bird and mammal
128 species, have had their cytochrome oxidase I (COI) gene sequenced as part of the barcode of life
129 project (*17*). For the subset of edges in the network for which both the source and sink hosts have
130 a COI gene sequence that fulfilled our criteria for consideration (as described above), we
131 computed the percentage evolutionary distance between the two hosts (Materials and Methods).
132 Applying a similar permutation test and assessment criteria as described for host groups above,
133 we found a trend of increasing over-representation at higher evolutionary distances (Figure 1c).
134 Thus, as host evolutionary distance, or (more broadly, quantitative niche dissimilarity) increases,
135 reticulate evolution becomes increasingly favored for influenza virus niche switch events.

136 We have quantitatively defined the importance of reticulate evolutionary events in switching
137 ecological niches, using an infectious disease data set with characteristics that are particularly
138 well suited for answering this question. Beyond the viral world, recent reviews have asserted the
139 importance of reticulate evolutionary events as a driver of speciation and niche diversification
140 (*18*, *19*), and recent studies have illustrated heightened fitness effects in hybrid populations (*20*,
141 *21*). However, none have quantitatively tested the importance of reticulate evolutionary
142 strategies in enabling ecological niche switches at a global scale, especially in comparison to
143 clonal adaptation under drift and selection, a task feasible only in fast evolving organisms. To
144 date, no studies have examined reticulate evolutionary processes in the context of quantified
145 niche differences, as we have done here by measuring reassortment in the context of host
146 evolutionary distance. Our study provides strong quantitative evidence supporting the hypothesis

147 that reticulate evolutionary processes are advantageous relative to adaptation by drift for
148 pathogen transfer between host species, and therefore more broadly, ecological niche switching.

149 We recognize that in this study, we have considered only a single pathogen for which abundant
150 genomic data are available, and whose genomic and host tropic characteristics are suitable for
151 this analysis. To specifically answer whether reticulate processes are favored over clonal
152 transmission for other organisms, using these methods, depends on being able to acquire genome
153 sequences with matched ecological niche metadata. We also note that the global influenza
154 dataset will have unavoidable sampling biases. For example, human isolates predominate in the
155 dataset, and consequently the human-associated subtypes H3N2 and H1N1 also dominate the
156 dataset. Sequences from viral outbreaks will also be over-represented relative to isolates
157 collected through routine surveillance sampling, and will unavoidably lead to a heightened
158 detection of clonal descent in a single host species. In order to deal with this sampling bias, our
159 permutation tests (for the host species and group labels) involve class labels of equal sizes. This
160 allows us to calculate an expected distribution of proportions under ideal assumptions of equal
161 sampling, which in turn forms the baseline for our conclusions.

162 In summary, using data available from a model zoonotic viral pathogen, we have shown that a
163 reticulate evolutionary strategies are important in enabling pathogen host switches. For the
164 influenza virus, reticulate evolution predominates when crossing between hosts. More broadly,
165 the greater the quantitative difference between ecological niches, the greater the importance of
166 reticulate evolutionary processes in enabling niche switches. While the quantitative importance
167 of reticulate evolution may differ for different organisms evolving in different niches, we expect
168 that further sequencing efforts from across broad domains of microbial life, and a further
169 characterization and definition of their ecological niches, will elucidate whether this principle

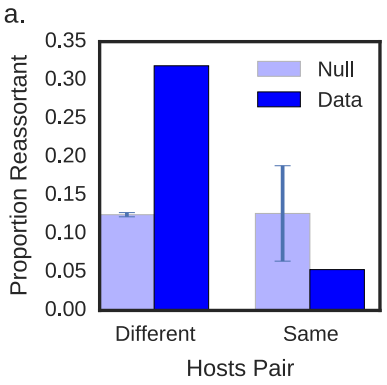
170 holds more broadly. Beyond its relevance to evolutionary ecology, reticulate evolution also has
171 public health consequences. Reassortant influenza viruses have been implicated in all past
172 human pandemic strains (*22*-*25*), and the ancestry of HIV-1 involved a hybrid SIV (*26*). Hence,
173 knowing how reticulate events shape disease emergence may help the ecology and evolution of
174 infectious disease become a more predictive science, leading to insight important to disease
175 prevention and mitigation (*27*).

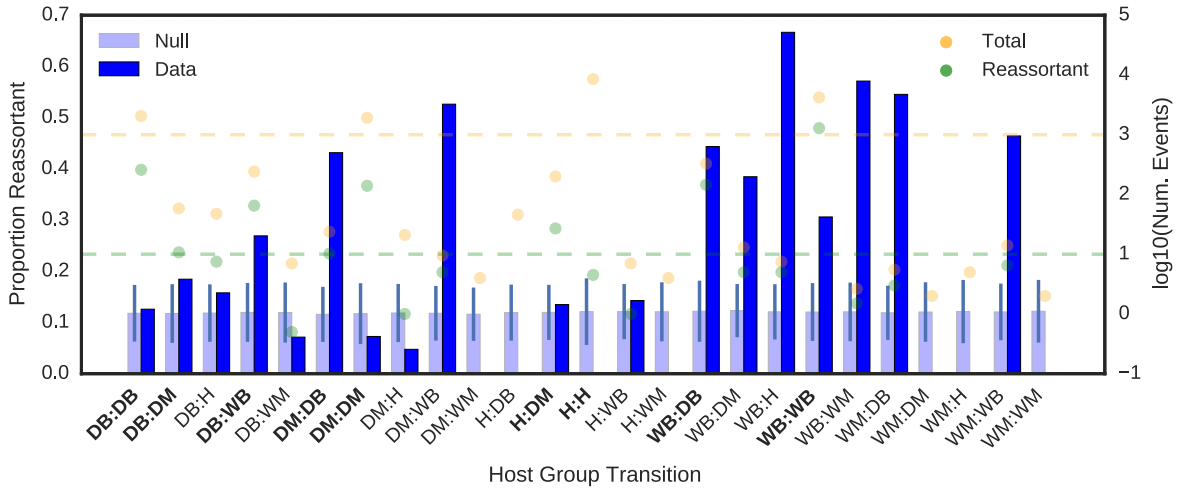
176 **Acknowledgments**

177 The authors would like to thank Dr. Nan Li for technical assistance on graph computation during
178 the earlier stages of this work. We would like to acknowledge funding from the NIH/NIAID
179 CEIRS Program (Contract no.: HHSN272014000008C) and the MIT Department of Biological
180 Engineering. The authors would also like to thank William R. Hesse of the MIT BE
181 Communications Lab for assistance in reviewing the manuscript.

182

Figures





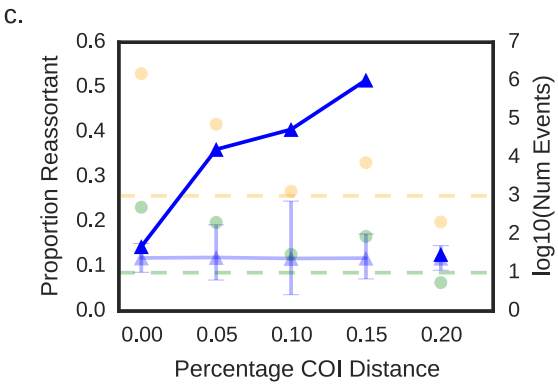


Figure 1. **Reassortment is over-represented relative to clonal descent in transmission across host barriers.** Proportion of reassortment events when crossing between (a) different or same hosts, (b) different host groups, and (c) hosts of differing evolutionary distance as measured by

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. (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 clonal 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.

Materials and Methods

Sequence data source and preprocessing. Sequence and metadata was downloaded from the Influenza Research Database on 2 September 2015. Search parameters on the IRD were as such: Segment/Nucleotide data, Virus Type A, All Segments, All Hosts, All Geographic Groupings, Complete Genomes Only, Include pH1N1 sequences, Date Range 1980-Present. Advanced options included: Laboratory Strains excluded. Segment FASTA files were downloaded, with a Custom Format where only Accession Numbers were included.

Source code and Data. Digital Object Identifiers (DOIs) hosted on zenodo.org can be found for the following source code.

- Source code for graph construction: 10.5281/zenodo.33421.
- Source code for analysis, derivative data (network node and edge lists, computed threshold values) and figure construction as part of a series of Jupyter notebooks: 10.5281/zenodo.33422.
- Source code for the patristic distance tests: 10.5281/zenodo.33425.
- Source code and Jupyter notebooks for simulation studies: 10.5281/zenodo.33427

Reassortant virus identification. We adapted the SeqTrack algorithm (*14*) to perform graph construction. Sequences were aligned using Clustal Omega 1.2.1 (*28*), and the resultant distance matrix was converted into a similarity matrix by taking $1 - distance$. Affinity Propagation (*29*) clustering was done on each segment's similarity matrix to determine a threshold cut-off similarity value, defined as the minimum of minimum in-cluster pairwise identities. Because the Affinity Propagation algorithm does not scale well with sample size, we treated the threshold

221 computation as an estimation problem, computed as the median threshold of 50 random sub-
222 samples of 500 isolates.

223 We then thresholded each segment's similarity matrix based on its segment's threshold value,
224 summed all 8 thresholded similarity matrices, and then for each isolate, we identified the most
225 similar isolate that occurred prior to it in time. This yielded the initial "full complement" graph
226 without reassortant viruses. Each edge in this graph has an attached PWI, which is the sum of
227 PWIs across all 8 segments. Within this graph, there are isolates for which no "full complement"
228 of segments could be identified, which are candidate reassortant viruses. Additionally, amongst
229 the isolates for which a full complement of segments could be found from another source, we
230 identified those whose in-edges were weighted at the bottom 10% of all edges present in the
231 graph, which we also conservatively identified as candidate reassortant viruses. For these
232 viruses, we performed source pair searches, where we identified sources for a part of the genome
233 from one virus and sources for the complementary part of the genome from another virus. If the
234 summed PWI across the segments for the two viruses was greater than the single-source search,
235 we accepted the source pair as the candidate reassortant.

236 **Proportion Reassortment Calculations.** The proportion of reassortment events was calculated
237 by first weighting each incoming edge to every virus. The weighting procedure is described here:
238 If the virus is detected to be plausibly descended from n other viruses, as determined by maximal
239 similarity, it is given a weight of $\frac{1}{n}$. If the virus is detected to be a reassortant, then edges are
240 weighted by the the fraction of times that it is involved in a max similarity source pair. For
241 example, if a given node A has a plausible source of segments in B, C, D, with (B and C) and (B
242 and D) being plausible sources, then the edge B-A would be given a weight of 0.5, and the edge
243 C-A and D-A would be given a weight of 0.25 each.

Permutation Tests. Null models were constructed by permuting node labels; equal class size permutation was performed for host species (for Figure 1a) and host group (for Figure 1b). 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.

Simulation Studies. In our simulations, we sought to model the process of obtaining sequences as simply as possible. Therefore, we used a kinetic Monte Carlo-style process to simulate the generation of new viral sequences under the processes of replication, mutation and reassortment. Briefly, we simulated simple two-segment model viruses, with each of the two segments having a different substitution rate. Each simulation run was initialized with anywhere between 1 and 5 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 is 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. Specific implementation is viewable in the IPython HTML notebooks and Python class definitions available on Github, available at the URL provided in Code and Data Deposition. This simulation process thus gives rise to a fully known ground-truth graph, which all reconstructions can be compared against.

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 employed 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 prior to the sink isolates.

In order to assess the accuracy of our reconstruction, we defined the path accuracy, and reassortant path identification accuracy metrics. 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.

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 aligned using Clustal Omega (*28*), and the genealogies reconstructed using BEAST 1.8.0 (*30*). 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 (*31*). 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.

References

1. 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 Biol Evol* 5(12):2305–2317.

2. Peris D, et al. (2014) Population structure and reticulate evolution of Saccharomyces eubayanus and its lager-brewing hybrids. *Molecular Ecology* 23(8):2031–2045.

3. Garten RJ, et al. (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* 325(5937):197–201.

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

5. 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. *Mol Microbiol* 57(3):727–734.

6. Remold SK, Rambaut A, Turner PE (2008) Evolutionary genomics of host adaptation in vesicular stomatitis virus. *Mol Biol Evol* 25(6):1138–1147.

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

8. Duggal NK, Emerman M (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat Rev Immunol* 12(10):687–695.

9. Li C, et al. (2010) Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. *Proc Natl Acad Sci USA* 107(10):4687–4692.

10. Mehle A, Dugan VG, Taubenberger JK, Doudna JA (2012) Reassortment and mutation of the avian influenza virus polymerase PA subunit overcome species barriers. *J Virol* 86(3):1750–1757.

11. Lam TT-Y, et al. (2011) Reassortment Events among Swine Influenza A Viruses in China: Implications for the Origin of the 2009 Influenza Pandemic. *J Virol* 85(19):10279–10285.

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

13. 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.

14. Jombart T, Eggo RM, Dodd PJ, Balloux F (2011) Reconstructing disease outbreaks from

genetic data: a graph approach. *Heredity* 106(2):383–390.

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

16. Ma W, Kahn RE, Richt JA (2008) The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *J Mol Genet Med* 3(1):158–166.

17. Ratnasingham S, Herbert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes* 7(3):355–364.

18. 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.

19. 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. *Ecol Evol* 4(7):899–910.

20. Verhoeven KJF, Macel M, Wolfe LM, Biere A (2011) Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proc R Soc B* 278(1702):2–8.

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

22. Wu A, et al. (2013) Sequential Reassortments Underlie Diverse Influenza H7N9 Genotypes in China. *Cell Host Microbe* 14(4):446–452.

23. Pu J, et al. (2014) Evolution of the H9N2 influenza genotype that facilitated the genesis of the novel H7N9 virus. *Proc Natl Acad Sci USA*:201422456.

24. 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.

25. Gao R, et al. (2013) Human infection with a novel avian-origin influenza A (H7N9) virus. *N Engl J Med* 368(20):1888–1897.

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

27. Wasik BR, Turner PE (2013) On the Biological Success of Viruses. *http://dxdoiorg/101146/annurev-micro-090110-102833* 67(1):519–541.

28. Sievers F, Dineen D, Wilm A, Higgins DG (2013) Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics* 29(8):989–995.

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

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

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