PNAS nexus

PNAS Nexus, 2022, **1**, 1–9
https://doi.org/10.1093/pnasnexus/pgac143
Advance access publication date: 5 August 2022
Research Report

Immunity-induced criticality of the genotype network of influenza A (H3N2) hemagglutinin

Blake J. M. Williams[a], C. Brandon Ogbunugafor[a,b,c,d], Benjamin M. Althouse[e,f,g] and Laurent Hébert-Dufresne[a,h,*]

[a]Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405, USA
[b]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA
[c]Santa Fe Institute, Santa Fe, NM 87501, USA
[d]Public Health Modeling Unit, Yale School of Public Health, New Haven, CT 06510, USA
[e]Institute for Disease Modeling, Global Health, Bill & Melinda Gates Foundation, Seattle, WA 98109, USA
[f]Information School, University of Washington, Seattle, WA 98195, USA
[g]Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA
[h]Department of Computer Science, University of Vermont, Burlington VT 05405, USA
*To whom correspondence should be addressed: Email: laurent.hebert-dufresne@uvm.edu

Edited By: Sandro Galea

Abstract

Seasonal influenza kills hundreds of thousands every year, with multiple constantly changing strains in circulation at any given time. A high mutation rate enables the influenza virus to evade recognition by the human immune system, including immunity acquired through past infection and vaccination. Here, we capture the genetic similarity of influenza strains and their evolutionary dynamics with genotype networks. We show that the genotype networks of influenza A (H3N2) hemagglutinin are characterized by heavy-tailed distributions of module sizes and connectivity indicative of critical behavior. We argue that (i) genotype networks are driven by mutation and host immunity to explore a subspace of networks predictable in structure and (ii) genotype networks provide an underlying structure necessary to capture the rich dynamics of multistrain epidemic models. In particular, inclusion of strain-transcending immunity in epidemic models is dependent upon the structure of an underlying genotype network. This interplay is consistent with self-organized criticality where the epidemic dynamics of influenza locates critical regions of its genotype network. We conclude that this interplay between disease dynamics and network structure might be key for future network analysis of pathogen evolution and realistic multistrain epidemic models.

Keywords: genotype network, disease modeling, criticality, complex systems

Significance Statement:

Seasonal influenza threatens global public health, resulting in millions of severe infections every year and a significant economic burden. Vaccination is a key intervention for preventing infections, but vaccine strains must be closely matched to circulating strains to ensure protection. Here, we show that genotype networks provide a map of influenza strains that captures genetic similarity and evolutionary pathways. We argue that genotype networks are necessary for modeling epidemics consisting of multiple strains. In particular, genotype networks enable modeling of diseases in which protection against one strain offers some protection toward other strains. In turn, we find that the dynamics of multistrain epidemics are key to understanding the unique structure of the influenza genotype networks.

Each year, seasonal influenza results in 290,000 to 650,000 deaths globally, 9 million to 36 million cases in the United States alone, and results in significant economic burdens (1–3). Despite widespread vaccination and increased surveillance efforts in recent years, influenza continues to show prominent seasonality in temperate regions and causes a year-round burden in tropical regions (4, 5).

Influenza viruses (INFV) mutate rapidly with antigenic drifts and shifts, leading to the frequent emergence of new strains that are different enough to escape recognition by host immunity (6). As a result, we see frequent epidemics and necessitate yearly updates to vaccine strains based on sequencing data and future projections (7–9). Optimal vaccine strain selection is dependent upon the ability to both forecast prevalent future strains and select a limited number of vaccine strains, such that these strains offer optimal immune protection by leveraging strain-transcending immunity (10, 11). Modern seasonal INFV vaccines induce antibodies for three to four unique strains of INFV, providing direct immunity for these strains and some cross-protective (or strain-transcending) effects toward antigenically similar strains. Similarly, these antibodies are induced in response to a clinical influenza infection (12).

Competing Interests: The authors declare no competing interest.
Received: December 17, 2021. **Accepted:** July 22, 2022

INFV epidemiology has benefited from decades of research using phylogenetics and molecular evolution to carefully interrogate features of INFV evolution (13–15). Exercises in applied evolutionary theory have served as validations for the use of molecular methods toward meaningful predictive evolution (16, 17). These methods, in combination with larger data sets, offer increasingly accurate probabilistic models for INFV evolution. As effective as they have been, these approaches are based on particular population genetic assumptions and limitations. For example, tree-based methods are necessarily acyclic and as such do not fully capture the relatedness of strains.

Phylodynamic approaches have features of neutral networks, defined by genotypes that evolve via drifting through epochal evolution (18, 19). Genotype networks constitute another approach used to study INFV evolution, and are built on different assumptions and constraints than other approaches (20–24). Previous networks have been constructed from the highly antigenic hemagglutinin (HA) protein sequences of INFV (20). The networks revealed features not well represented in phylogenetic trees, such as identical trait evolution in separate lineages (convergent evolution). More importantly, dynamical systems describing the spread of pathogens are often parameterized through genotype networks rather than phylogenetic trees to better capture strain-transcending immunity (25–30). Unfortunately, these previous studies use toy networks as genotype networks are prone to fragmentation in the presence of low sampling rates, reducing the number of observed plausible evolutionary pathways. Sampling has increased dramatically in the last decade, which now allows for a more accurate account of the evolution of INFV genotype networks.

In this study, we utilize a large modern data set of INFV H3N2 sequences (over 28,000) and a genotype network approach to capture the genetic relationship between the 2010 and 2020 INFV H3N2 strains and their evolutionary dynamics. Sequences of the highly antigenic HA protein of INFV A (H3N2) are used to analyze the structure and temporal evolution of the genotype network and its exploration of genotype space. Finally, a multistrain epidemic model is implemented to explore how the density and distribution of edges (or mutation pathways) determine epidemic potential in the context of strain-transcending immunity. We demonstrate the existence of a fundamental structure underlying INFV genotype space, one that captures temporal features of virus evolution and suggests underlying predictability. In doing so, we fortify the relevance of genotype networks as a meaningful approach to the study of virus evolution, one that can complement mathematical and phylodynamic approaches in future efforts to study and predict the dynamics of evolution of INFV and other RNA viruses.

Data and model

Network generation

Protein sequences were obtained for complete INFV A (H3N2) HA samples from the Influenza Research Database (31). Samples acquired from the Influenza Research Database are sourced from databases that include NCBI GenBank and RefSeq. Samples were obtained on 2020 January 16 and restricted to a collection date of 1999 January 4 through 2019 October 1 and collected from human hosts only. A 3-month delay between final sample collection date and data retrieval date was implemented to account for delays in data reporting.

A total of 30,175 sequenced samples for HA were obtained. Sequences were further restricted to allow for the precise genetic sequence comparison required for network edge construction. Samples with missing or uncertain residues ($n = 1{,}278$) and sequences with more or less than 566 amino acids ($n = 17$) were removed. The remaining 28,880 samples were condensed into set V of 9,714 unique sequences.

The number of differing amino acids across all sites for sequences v and w, $d_{v,w}$, was found for all pairs of sequences of length $l = 566$:

$$d_{v,w} = \sum_{i=1}^{l} x, \quad \text{where } x = \begin{cases} 1, & \text{if } v_i \neq w_i \\ 0, & \text{if } v_i = w_i \end{cases} \quad v, w \in V.$$

An edge $e_{v,w}$ is formed if $d_{v,w} = 1$. Each edge indicates a plausible, but not definitive, mutation pathway between two viable strains that requires one point mutation, thus no intermediate strains nor multimutation events. The resulting genotype network is defined as $G = (V, E)$, where E is the set of all edges $e_{v,w}$.

Temporal analyses restricted data by year using seasonal trends of the Northern Hemisphere, given its dominance of the data set. Sequences were binned according to a 5-y window, where each year consisted of July 1 through June 30 of the following year. For example, a 5-y window centered on 2010 would contain sequences from 2007 July through 2012 June.

Multistrain epidemic model

Building on previous work (25–30), we assume that the epidemiological dynamics of INFV follow the classic Susceptible–Infectious–Recovered–Susceptible (SIRS) model, and introduce an underlying, data-driven, genotype network that defines potential mutations and allows strain-transcending immunity. An individual infected with strain $i \in [1, N]$ can cause a mutation at a rate μ to a strain $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of first network neighbors of strain i.

All strains spread concurrently in a well-mixed host population. Individuals are susceptible (S) if they possess no previous immunity. Each susceptible individual progresses to infectious state I_i, corresponding to strain i, at a rate βI_i. The basic transmission rate β is held constant for all strains, as we focus on neutral evolution (antigenic drift) as a first approximation.

Infectious individuals in I_i will either (i) recover at rate γ to state R_i and acquire full immunity for strain i and partial immunity to other strains $j \neq i$ or (ii) undergo a mutation to strain j at a rate μ for all strains j in $\mathcal{N}_i$. Recovered individuals in R_i will either (i) lose immunity and progress back to S at rate α or (ii) get infected with strain $j \neq i$ and progress to I_j at a reduced rate β^* due to their partial immunity. Specifically, β^* is an exponentially decaying function of genetic distance between strains i and j,

$$\beta^*_{ij} = \beta \left(1 - e^{-x_{ij}/\Delta}\right),$$

where x_{ij} is the network distance between strains i and j (shortest path between strains i and j in the genotype network, different from $d_{v,w}$ used above) and Δ is the characteristic length of immunity ($0 < \Delta < \infty$) as it transcends specific strains over the genotype network. Note that we make the assumption that an individual's immune response is set by the most recent infection as accounting for a full immune history would result in $N!$ possible immune states.

The model assumes that (i) an individual may be infected by at most one strain at a time, (ii) an individual's immune response is determined by the strain responsible for their last infection, and (iii) transcendence of immunity decays exponentially as a function of the distance between strains. The model was implemented

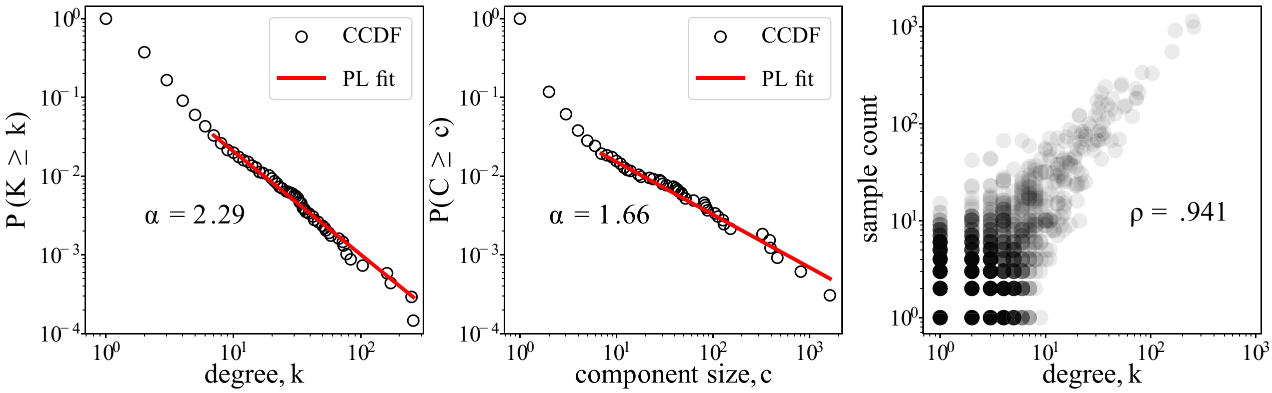

Fig. 1. INFV A (H3N2) HA genotype network degree and component size distribution. (Left) CCDF of degrees. The tail of degree distribution does not significantly differ from a power-law distribution with $\alpha_k = 2.29$ for $k_{\mathrm{min}} = 7$ ($P = 0.11$, $\alpha_{\mathrm{significance}} = 0.05$, 10^3 repetition Kolmogorov–Smirnov test). (Center) CCDF of component sizes. Component size distribution does not significantly differ from a power-law distribution with $\alpha_c = 1.66$ for $c_{\mathrm{min}} = 7$ ($P = 0.59$, $\alpha_{\mathrm{significance}} = 0.05$, 10^3 repetition Kolmogorov–Smirnov test). (Right) Sample count of a sequence versus degree k of corresponding node. Sample count is highly correlated with node degree ($r = 0.941$).

Table 1. Statistics of the entire network G and its giant component GC.

	n	m	$\langle k \rangle$	k_{max}	D	C_{Global}	r
G	9714	7599	1.86	257 (11, 257)	–	0.0096 (0.0005, 0.0112)	−0.13 (0.00, −0.05)
GC	1629	2225	2.73	257 (11, 257)	17 (23.5, 16.5)	0.0010 (0.0004, 0.0112)	−0.20 (−0.03, −0.08)

Number of nodes n and edges m as well as average degree $\langle k \rangle$, maximum degree k_{max}, diameter D, clustering coefficient C_{Global}, and assortativity (degree correlations) coefficient r. Numbers in parentheses correspond to the average values obtained under 100 realizations of two null models, respectively: Erdős–Rényi random graphs parameterized by density only and a configuration model parameterized by the full degree distribution.

with a system of differential equations containing one susceptible state and an infected and recovered state for each strain. The dynamical system describing this model is presented in the "Materials and methods" section, and its dynamics were studied in ref. (30).

The model itself can run over any genotype network defined as a number of strains $i \in [1, N]$ and a set of neighboring strains $j \in \mathcal{N}_i$ for each strain. In what follows, we therefore couple the model with known generative models of networks that can help explain some key network features found in the genotype data.

Results

INFV A (H3N2) HA genotype network

The INFV A (H3N2) HA genotype network represents 28,880 samples of HA, resulting in 9,714 nodes (unique strains), 7,599 edges (possible point mutations between strains), and 3,262 connected components, of which 384 consist of more than one node. With 29.6% of nodes of degree $k = 0$ and 44.0% of $k = 1$, the network features a skewed degree distribution, stretching up to a maximum degree of $k = 256$. The tail of the complementary cumulative distribution function (CCDF) of degree, $P(K \geq k)$, exhibits power-law behavior: $P(K \geq k) \ \propto k^{-\alpha_k}$ with an estimated scale exponent $\alpha_k = 2.29$, Fig. 1 (left panel). This is in agreement with the heavy-tailed degree distribution found by Wagner in the largest connected component of a smaller data set from 2002 to 2007 (20). In growing networks, this degree distribution points to generative models with approximately linear preferential attachment underlying the dynamics of the observed genotype network (32–34). Linear attachment is a critical mechanism such that a growing network produces power-law degree distributions, at a transition between exponential distributions under sublinear attachment and condensation to a star graph under super-linear attachment (35, 36).

The distribution of component sizes of the genotype network is similarly skewed. The tail of the CCDF of component sizes $P(C \geq c)$ follows a power-law distribution, where $P(C \geq c) \propto c^{-\alpha_c}$ with scale exponent $\alpha_c = 1.66$, Fig. 1 (center panel). This scaling is also be suggestive of another critical process in the formation of the genotype network, as this distribution of component sizes with scale exponent $\alpha_c = 1.5$ is a well-known result for the critical point of percolation processes and random graphs (37).

The degree of a node and the number of times its corresponding sequence was sampled are highly correlated, Fig. 1 (right panel). Structurally important nodes of high degree (hubs) are therefore robust to reduced sampling, given that the duplicate sample count of a strain may be a proxy for its population prevalence. The network also contains numerous cycles amidst its heterogeneous tree-like structure. Its 500 triangles indicate mutations at the same site between three sequences, while sparse squares indicate potential convergent evolution (20). These structures are clearly displayed in genotype networks, while phylogenetic tree construction do not include convergent evolution structurally. The tree-like topology of the network prevents longer cycles from forming. Further network summary statistics are shown in Table 1 for the entire network G and the giant component GC. The triangles are captured by global clustering C_{global}, which is equivalent to the proportion of triplets (three connected nodes) that form a closed triangle. Despite the biological relevance of triangles (20), we find that they are neither overrepresented nor underpresented when compared to random networks with a fixed degree distribution; as shown in Table 1.

The degree assortativity r represents the correlation between the degree of a node and that of its neighbors (Table 1). A negative value for both the entire network G and the giant component GC indicate that high degree nodes tend to attach to low degree nodes. In fact, these negative degree correlations are the only

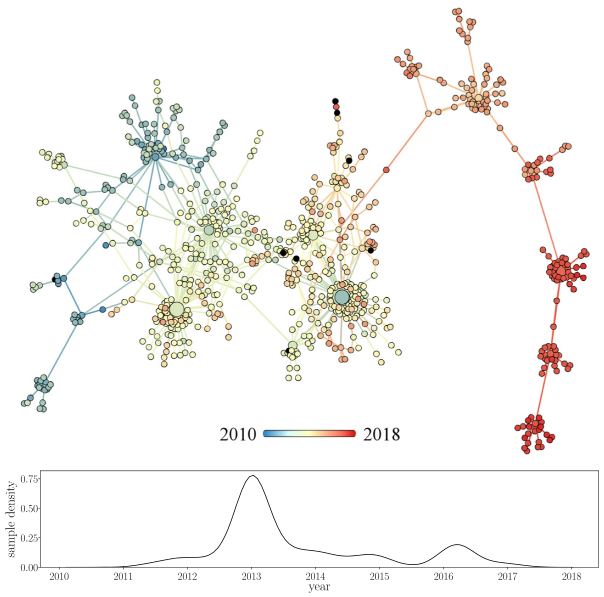

Fig. 2. Sample dates among strains of second largest network component. (Top) Nodes colored by first sample date (eight nodes with lacking sample dates colored black), with a larger radius corresponding to more samples (max sample count 337). (Bottom) Sample date distribution across all dated samples of strains within the above network.

feature that appears statistically significant when compared to random networks with fixed-degree distributions; again, this is consistent with growing random networks under positive attachment kernel (32, 34–36).

Network topology in time

The genotype network grows in time as new strains emerge and are sampled. For example, the growth of the second largest component is shown in Fig. 2, with each node colored by the first sample date for each strain. This component is large enough to span several years while remaining small enough to qualitatively observe network growth in time. The blue-shifted nodes represent the earliest observed strains among those belonging to this component, the first of which was sampled in late 2010. The majority of unique strains were sampled from 2012 to 2015, including multiple high-degree strains and their neighbors. The most recent strains from this network component are red-shifted, clearly depicting the tree-like growth process.

Numerous hubs are seen throughout the network, with the largest hubs existing around the 2012 and 2013 flu season that contributed numerous strains to this component (Fig. 2, bottom). Seasonality is reflected in the sample date distribution of this component, with multiple peaks around the start of the calendar year during flu season.

Features of the genotype network remain fairly stable in time, even in the presence of a constantly increasing sampling rate. Genotype networks were constructed using samples within a 5-y window, sweeping across the entire sample set. These temporally restricted genotype networks display the structure of the network local in time—an important consideration given that strains emerge and fall out of circulation. These networks display the increased availability of sequenced samples with each successive year, with notable increases in sampling since 2008 (Fig. 3, left panel). The number of nodes and edges has grown steadily in the past two decades across both the entire network of the 5-y windows and its giant component.

Scaling of both degree distribution and component size distribution tails remain fairly constant in time. The scale exponent for degree averaged 2.34 across these networks, varying from $2.10 < \alpha_k < 3.06$. We find over a decade of consistency near its mean (about 2.2 or 2.5) even as the network grew several times larger, Fig. 3 (center panel). Similarly, the power-law exponent for component size averaged 1.63 and varied within $1.44 < \alpha_c < 1.73$, demonstrating consistency in time and a comparable independence from sample rate as the network grew, Fig. 3 (center panel). Here $c_{\min}$ and $k_{\min}$ were fixed at 7, enabling a direct comparison with the entire network.

Local cycles continue to remain prevalent in the network through time. The global clustering coefficient varied within $6.78 \times 10^{-3} < C_{\text{global}} < 1.27 \times 10^{-2}$, showing greater variability than scaling factors, Fig. 3 (right panel). Similarly, degree assortativity varied within $-0.365 < r < -0.124$, demonstrating variability but preserving the disassortative structure of the network. The above features demonstrate that in the presence of variable sequence sampling rates, genotype networks possess fairly consistent topological features that are highly predictable from recent years.

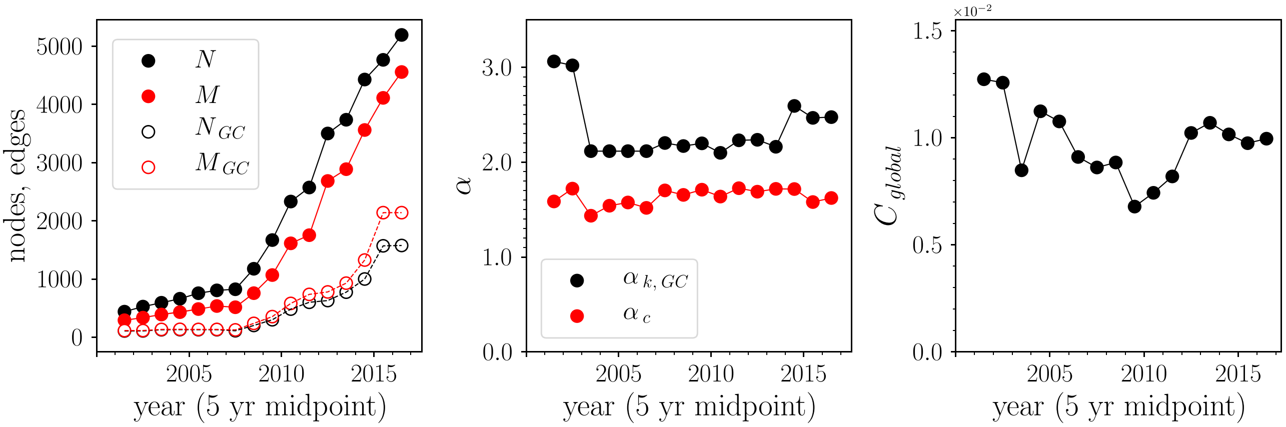

Fig. 3. Network statistics in time. INFV A (H3N2) HA genotype networks generated using samples within a sweeping 5-y window from 1999 July through 2019 June, shown at midpoint. (Left) Number of nodes and edges for entire network and giant component. (Center) Power-law scale exponents α_c and $\alpha_{k,\text{GC}}$ obtained by fitting the tail of the distribution above $c_{\min} = k_{\min} = 7$ following Fig. 1. (Right) Global clustering coefficient C_{global} over time.

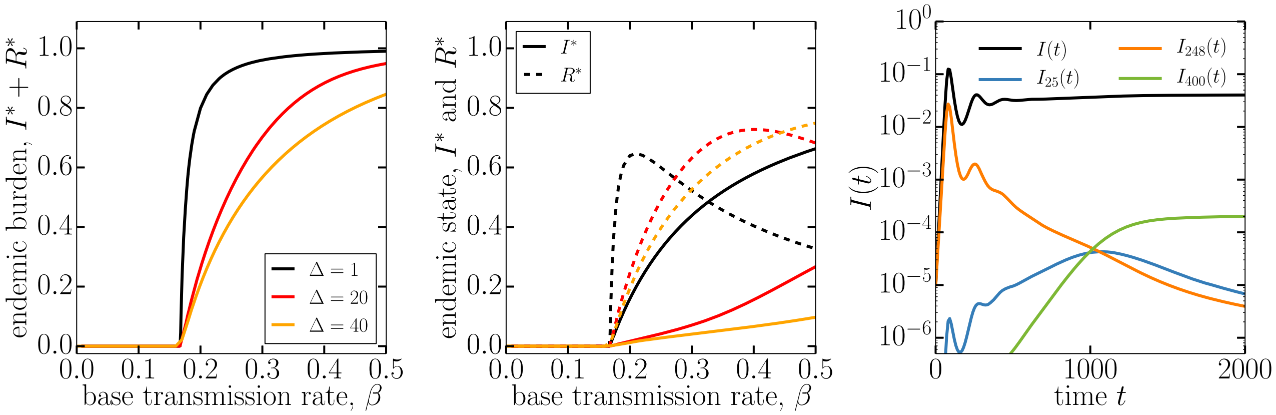

Fig. 4. Epidemics on a large genotype network. We run multistrain epidemic dynamics in a well-mixed population using the network presented in Fig. 2 as an underlying genotype network. (Left) Endemic burden, defined as the steady-state sum of infectious and recently recovered individuals for different values of strain-transcending immunity Δ. (Center) We now separate the density of currently infectious individuals and recently recovered individuals. This highlight the existence of two important epidemic transitions: An epidemic threshold (equivalent to $R_0 = 1$) where a nontrivial endemic state emerges and a threshold of immune invasion, which occurs at higher transmission rates. This second transition is noticeable as an inflection point in I^* (or a maximum in $dI^*/d\beta$), visible around $\beta = 0.35$ for $\Delta = 20$. (Right) Example of a time series to show how we arrive at a steady-state value. We plot the overall prevalence $I(t)$ as well as some example strains. All endemic results are integrated up to time $t = 25{,}000$.

Multistrain epidemics with underlying genotype networks

Before running our dynamical system for a multistrain epidemics on empirical and synthetic genotype networks, it is useful to clarify what structure is represented by these networks. In theory, there is a true, fixed, *full* genotype network, which represents all possible sequences of INFV H3N2 regardless of viability and fitness. In practice however, only a subset of these sequences actually emerge and are viable, leaving us with a subgraph corresponding to the *realized* genotype network. To make their way into our dataset, this network is further sampled by the sequencing process, leaving us with a subgraph for the *observed* genotype network.

Because our dynamical system is represented by a set of continuous and deterministic equations, the model itself blurs the line between the realized and the observed genotype network. This assumption is meaningful since sampling of complex networks can often alter their structure in nontrivial ways (38). However, sequences in our dataset are sampled an average of 2.97 times each and structurally important nodes are generally sampled proportionally to their degree as shown in Fig. 1 (right). Furthermore, Fig. 3 as already shown that key features are relatively fixed in time even when the size of the temporal samples vary by an order of magnitude.

To illustrate the output of our multistrain model, we this directly run the equations on one of the largest components in our dataset in Fig. 4. These results reproduce some of the main results known from the study of toy genotype networks (27–30, 39). First, increasing the depth of strain-transcending immunity in the genotype network does not alter the epidemic threshold of the system but does lower endemic burden and change its composition between currently infectious individuals [$I(t)$] and recovered individuals [$R(t)$, recently infectious]. Second, we find in Fig. 4 (center) that depending on the interplay between the depth of immunity (Δ) and its waning rate (α) there can exist a regime of localization where the fraction of infectious individuals in the endemic state (I^*) grows very slowly with the transmission rate (β) before a second transition [previously theorized as an immune invasion threshold (30)]. Finally, in Fig. 4 (right) we show a representative time series that illustrate the rich strain-specific dynamics that emerge even in a deterministic model, with heterogeneous time of emergence and cyclical dynamics that eventually settle at an endemic state (28).

Epidemics with random genotype networks

To investigate how the observed genotype network structure may be influenced by the spread of disease and learned host immunity, we ran our multistrain SIRS model with varied sytnthetic genotype networks. The incorporation of a genetic strain structure allows for both mutation between neighboring strains and cross-protective immune effects, defined as a function of network distance. Generative networks models then allow us to better study how network features affect the overall disease prevalence. Our hypothesis being that the evolutionary dynamics of INFV A (H3N2) would localize observed strains preferentially in regions of its full genotype network that have a structure consistent with high disease prevalence; the local network structure itself acting as an selection pressure on the realized and observed genotype networks.

The connectivity or edge density of a genotype network may influence its endemic infection capacity, as suggested by cross-protective immune effects and the observed criticality within the genotype network structure. Here, the effects of connectivity were investigated with the implementation of the multistrain model on fully random networks, namely $G(n, P)$ Erdős–Rényi random graphs (40), with a given number of nodes n and edge probability P controlling connectivity for an average degree of $\langle k \rangle = P(N - 1)$. We measure the endemic disease burden $I^* + R^*$, summed over all strains, once the epidemic dynamics has reached an equilibrium (i.e. after a long period of transient dynamics). This disease burden was observed across varying edge densities and levels of immunity transcendence in Fig. 5 (top left) to determine the relationship between connectivity and endemic infections for a genotype network of a given size.

Nontrivial dynamics are revealed by the multistrain epidemic model with an underlying genotype network structure of Erdős–Rényi random networks. Endemic disease burdens are lowered in random genotype networks in the presence of high connectivity and nonzero transcending-immunity parameter Δ, producing cross-protective immune effects that outweigh the increase in mutations. On the opposite end, extremely low connectivity also

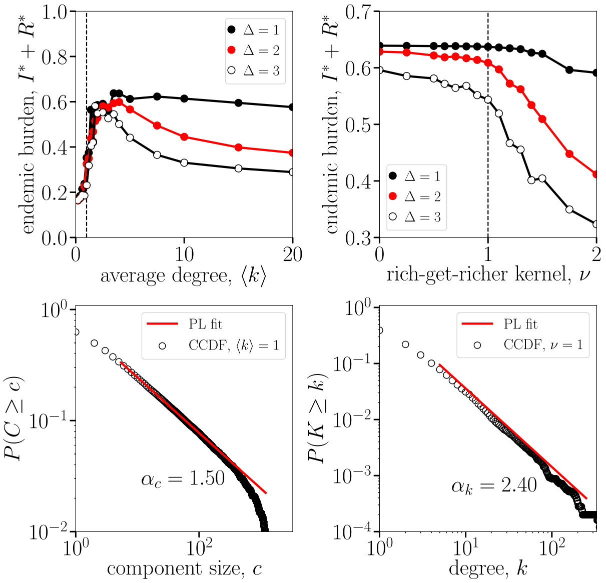

Fig. 5. Endemic disease burden as a function of the connectivity and heterogeneity of the underlying genotype network. (Top left) We look at the disease burden $I^* + R^*$ in an SIRS model with an underlying, random, genotype network. The network is specified as an Erdős–Rényi random graph with varying average degree. (Top right) The network is now generated by a nonlinear preferential attachment scheme with a fixed density (corresponding to the empirical INFV genotype network) and varying attachment kernel ν. In this scheme, $\nu = 0$ corresponds to uniform attachment, $\nu = 1$ to scale-free networks, and $\nu = 2$ to star-like networks. Other parameters: Network size $n = 250$, mutation rate $\mu = 1/50$, transmission rate $\beta = 1/2$, recovery rate $\gamma = 1/6$, and immune loss rate $\alpha = 1/100$. (Bottom left) Component size of random nodes found in the networks with highest endemic burden in the top left panel, i.e. $\langle k \rangle = 1$. (Bottom right) Degree distribution of nodes found in the networks with strongest preferential attachment before endemic burden decreases due to condensation in the genotype network, i.e. $\nu = 1$.

lowers disease burden through increased network fragmentation, resulting in numerous components that restrict mutation pathways between all strains. Together these dynamics produce an optimal connectivity that maximizes disease burden. While slightly affected by parameters, we find that the optimal average degree is observed increasingly close to $\langle k \rangle = 1$ as the pervasiveness of immunity Δ increases. This density is a critical point of the network structure where a giant component emerges. Around this critical transition, we find a power-law distribution of component sizes with exponent 1.5. This distribution shown in Fig. 5 (bottom left) is similar to that empirically observed in Fig. 1 (center).

This critical component size distribution is not found at an average degree $\langle k \rangle = 1$ in the INFV genotype network since its structure is far from that of Erdős–Rényi random networks. Most notably, the degree distribution of the real network is not homogeneous: The power-law degree distribution shown in Fig. 1 is radically different from the Poisson degree distributions of Erdős–Rényi networks. As previously stated, the observed degree distributions and negative correlations are both consistent with preferential attachment models. To explore degree heterogeneity, we therefore turn to a nonlinear preferential attachment model where networks are grown according to a rich-get-richer process where new strains are a mutation of existing strains chosen randomly but proportionally to their current degree to some power ν, controlling the network heterogeneity (35, 36). In Fig. 5 (top right), we find that the strongest rich-get-richer effect that a genotype network can support before decreases in disease burden is a linear attachment effect, reminiscent of the relationship observed in Fig. 1 (right). Under this linear preferential attachment, we find a power-law degree distribution with scale exponent 2.43, close to the exponent of 2.3 observed in the INFV genotype network in Fig. 1.

The results of these two experiments are consistent with our hypothesis. Namely, regions of a full genotype network that are at critical points in terms of density and rich-get-richer processes lead to higher disease burden such that there is a selection pressure for the realized genotype network to localize around these regions. Importantly, these experiments do not test actual mechanism for the growth of the realized and observed genotype networks. While growing networks with a power-law degree distribution can imply a preferential attachment statistic (33, 34), other mechanisms can produce similar networks (41). We can, however, venture two hypotheses for how a rich-get-richer comes into play in the observed genotype network. It can emerge from either (i) mutation patterns, as strains with more neighbors are more likely to re-emerge and re-explore their neighborhood or (ii) prevalence patterns, as the individual strain fitness and reproduction rate can be estimated structurally from strain degree.

Discussion

In this study, we utilize a large data set and a genotype network to examine INFV evolution from 1999 to 2020. In doing so, we reveal features suggestive of a fundamental structure underlying INFV genotypic space, and by extension, virus evolution. The INFV genotype networks explore a subspace of all networks that is predictable in structure as they grow in time. Features such as scale-free degree distributions and component size distributions, both related to underlying critical phenomena (42), remained present and consistent in networks generated using temporal subsets of strain samples.

It may be hypothesized that selecting vaccine strains near to hubs can provide a set of candidate strains for vaccine selection. It may also be hypothesized that well-selected strains would not be near observed hubs, were they to effectively neutralize spread of strains near to it in the network space. Strains selected for vaccines are recommended in part by the antigenic similarity between candidate strains and strains circulating during the targeted INFV season. Interestingly, we see the A/Texas/50/2012(H3N2)-like virus (WHO vaccine recommendation from 2014 to 2015) and A/Victoria/361/2011(H3N2)-like virus (2012 to 2014) as edgeless nodes, respectively, two and four mutations from their nearest strains in the network. The 2010–2012 H3N2 recommendation, A/Perth/16/2009(H3N2)-like virus, has four neighboring strains, including a hub two strains away. Future work will include a thorough analysis of the relationship between vaccine strains and temporal network structure. This relationship introduces further dynamical interplay as vaccines often include strains already spreading successfully but then, in turn, affect the epidemic dynamics and therefore limit further spread and mutations around these strains.

Given the numerous mutations possible, it may not be realistic to use genotype networks to predict new strains with meaningful accuracy. However, it may be possible to predict their genetic relationship to strains existing in the network structure. Any such efforts would effectively create a map of the genotype space currently occupied by INFV, and suggestive of where in that space it may evolve (see Fig. 6). Here, we find that INFV evolutionary dynamics never returns to regions of the network left more than a year or two in the past, in line with descriptions based on

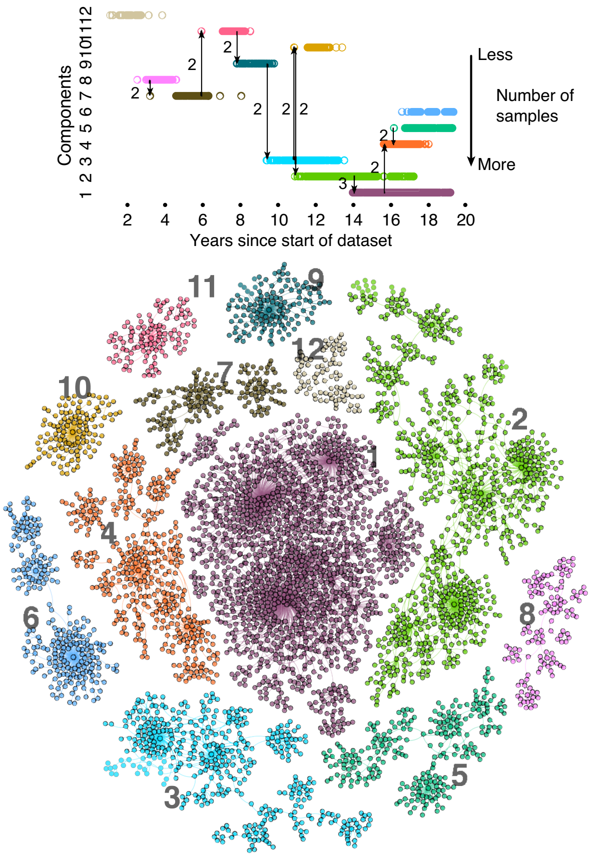

Fig. 6. Higher-order mutations help explain the global structure of the genotype network. We show the 12 largest network components. In the top panel, we show how double and triple mutations help explain almost all jumps across components. We also note that components are never rediscovered after more than a few months without new strains emerging therein. Altogether, this analysis shows that the sampling of strains might be better than originally expected, but also that higher-order networks structures (paths of multiple mutations) might eventually help us better understand the global patterns of INFV immune evasion.

travelling wave models (43) but adding a descriptive layer for the growth and local structure of the observed network. Assuming the genetic distance is proportional to antigenic distance (22, 44, 45), this is a consequential development with regards to our understanding of cross-protective immune effects and vaccination strain selection. That is, the outlined approaches may offer perspective on which specific genotypes of a given INFV strain might offer the best cross-protective immunity. Future work may include development of models further able to predict and refine the space of plausible future strains.

The predictable statistics of the genotype network topology indicates that the INFV A (H3N2) HA genotype network is influenced by strain-transcending immunity. This is consistent with the dynamics of a multistrain epidemic model. As the pervasiveness of learned immunity increases, the peak endemic burden expected in our multistrain model shifts closer to, and becomes narrower around, a critical network density corresponding to the emergence of a giant component and a power-law distribution of component sizes. In the future, knowing what mechanisms help shape the genotype network could allow network inference frameworks to identify critical new strains as they emerge (46, 47).

The strong positive relationship between degree and sample count implies preferential attachment based on degree; however, node age implements a consequential maximum age at which a node may acquire new neighbors. This corresponds to the point at which the strain is not widely circulating or extinct in the host population. Furthermore, the multistrain model indicates that strain-transcending immunity drives this strain extinction process as cross-protective effects increase population immunity toward strains in time. As shown by the model, any stronger preferential attachment mechanisms would also decrease the expected epidemic burden.

This study introduces the use of network growth processes that could be used in parallel to other methods used to study pathogen evolution. These include phylodynamics (43), genealogical trees (48), antigenic cartography (23, 24), and other network approaches. With regard to phylodynamics, our approach requires few of the population genetic (and other) assumptions that are embedded in phylodynamic approaches. Moreover, our results offer improvements over existing network models through the additional insights: the identification of critical properties in INFV genotype networks and the offering of mechanisms for its underlying structure. Our observations are consistent with our hypothesis that the observed INFV genotype network explores a subspace predictable in structure, influenced by the effects of strain-transcending immunity. A more realistic network growth process [involving, for example, convergent evolution, correlated mutations, and epistatic effects (49)] would be necessary to better fit the observed genotype network structure. Likewise, this observed structure is also impacted by the imperfect sampling of INFV strains. Future efforts may utilize more densely sampled populations.

In summary, we stress that increased genomic surveillance of multistrain pathogens will allow for similar analyses of other diseases with variable antigenic properties. As the evolutionary forces acting on multistrain pathogens differ, we may expect differing network structures from pathogen to pathogen. For instance, HIV has unique pressures from lifetime infection and pathogen evolution, highly active antiretroviral therapy used in its management, as well as bottleneck transmission events and selection biases (50)—all mechanisms that could lead to unique network features. Rapidly changing pathogenicity and virulence in emergent viruses, such as SARS-CoV-2, could yield dynamic network features. As the COVID-19 pandemic has generated data at an unprecedented pace and level of granularity, it may offer the opportunity for an analogous comparison (51).

More broadly, our findings support the importance of multiple methods—utilizing both existing phylodynamic approaches and network and graph theoretical methods—toward a comprehensive picture of virus evolutionary dynamics. The use of multiple methods can be complementary, as standard canon from evolutionary theory and methods from complex systems can each offer useful information about pathogen evolution.

In the future, we might be able to characterize the underlying physics of RNA virus infection networks that can be used to predict long-term patterns, toward improved public health interventions: vaccine strain selection, analysis of evolutionary trajectories, and refinement of the understanding of cross-protective immunity.

Materials and methods

Statistical methods

Distribution tails were fitted with power laws using the “poweRlaw” package (52, 53). For the full network generated from all years of the data set, we fit power-law distribution tails for observed values (degree, component size), where tail implies the distribution of observed values greater than some minimum. Minimum values (k_{min}, c_{min}) of 5 or greater were considered, and the best goodness of fit was observed at $k_{\mathrm{min}} = c_{\mathrm{min}} = 7$ for both degree and component size. These minimum values were then constrained to 7 for networks consisting of 5 y of data in our temporal analysis.

Multistrain epidemic model

The dynamics of the model described in the main text and ref. (30) can be tracked with the following set of ordinary differential equations:

$$\frac{dS}{dt} = -\beta \sum_{i=1}^{N} \frac{SI_i}{N} + \alpha \sum_{i=1}^{N} R_i,$$

$$\frac{dI_i}{dt} = \beta \frac{SI_i}{N} - \gamma I_i + \mu \sum_{j=1}^{N} A_{i,j}(I_j - I_i) + \sum_{j=1}^{N} \beta^*_{ij} \frac{I_i R_j}{N},$$

$$\frac{dR_i}{dt} = \gamma I_i - \alpha R_i - \sum_{j=1}^{N} \beta^*_{ij} \frac{I_j R_i}{N},$$

with $\beta^*_{ij} = \beta\left(1 - e^{-x_{ij}/\Delta}\right)$.

Experiment on Erdős–Rényi networks

In the left column of Fig. 5, we present the endemic disease burden $\sum_i I_i + R_i$ (i.e. recent infections) of our multistrain epidemic model on Erdős–Rényi networks (40). The endemic state is defined as the fixed point where all derivatives of the system are equal to zero. Erdős–Rényi networks are obtained by generating a set of $n = 250$ nodes and connecting each possible pair of nodes with probability $P = \langle k \rangle/(N - 1)$ such that the expected degree of all nodes (number of first network neighbors) is set by $\langle k \rangle$.

Experiment on preferential attachment networks

In the right column of Fig. 5, we present the endemic disease burden $\sum_i I_i + R_i$ of our multistrain epidemic model on networks grown through preferential attachment (35, 36). These networks are obtained by starting from a pair of connected nodes and growing the network until we reach a network of size $n = 250$ nodes.

The networks are grown through the following discrete stochastic process. At each time step, we either connect an existing pair of nodes with probability P or connect a new node to an existing node with complementary probability $1 - P$. The probability P sets the expected density of the network, since after t time steps we expect t edges and $(1 - P)t$ nodes for an average degree $\langle k \rangle = 2/(1 - P)$. In our experiment, P is chosen to fix the average degree to that observed in the giant component of our empirical data, i.e. $\langle k \rangle = 2.73$.

At every time step, we therefore need to pick either two existing nodes (probability P) or one existing node (complementary probability $1 - P$). These existing nodes are chosen proportionally to their degree k proportionally to the kernel k^ν. Meaning a given node i of degree k_i will be chosen with probability $k_i^\nu / \sum_j k_j^\nu$. A kernel with $\nu = 0$ corresponds to uniform attachment, whereas the linear kernel $\nu = 1$ corresponds to the much studied linear attachment model of Barabási and Albert (32).

Funding

B.J.M.W. and L.H.-D. acknowledge support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award.

Acknowledgment

The authors would like to thank Samuel V. Scarpino for helpful discussions.

Authors' contributions

B.J.M.W. compiled the data and produced all statistical analyses; B.J.M.W. and L.H.-D. developed the model; and C.B.O., B.M.A., and L.H.-D. supervised the study. All authors wrote the manuscript.

Data availability

The full network data associated with this manuscript, as well as software for the integration of the mathematical model using the data, are available at https://github.com/LaurentHebert/infAH3N2-genotype-criticality.

References

1. Putri WCWS, Muscatello DJ, Stockwell MS, Newall AT. 2018. Economic burden of seasonal influenza in the United States. Vaccine. 36:3960–3966.
2. Molinari NAM, *et al.* 2007. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 25:5086–5096.
3. Rolfes MA, *et al.* 2018. Annual estimates of the burden of seasonal influenza in the United States: a tool for strengthening influenza surveillance and preparedness. Influenza Other Respir Viruses. 12:132–137.
4. Iuliano AD, *et al.* 2018. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 391:1285–1300.
5. Nair H, *et al.* 2011. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. Lancet. 378:1917–1930.
6. Guan Y, *et al.* 2010. The emergence of pandemic influenza viruses. Protein Cell. 1:9–13.
7. Barr IG, *et al.* 2010. Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009–2010 Northern Hemisphere season. Vaccine. 28:1156–1167.
8. Klimov AI, *et al.* 2012. WHO recommendations for the viruses to be used in the 2012 Southern Hemisphere influenza vaccine: epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from February to September 2011. Vaccine. 30:6461–6471.
9. Barr IG, *et al.* 2014. WHO recommendations for the viruses used in the 2013–2014 Northern Hemisphere influenza vaccine: epidemiology, antigenic and genetic characteristics of influenza A(H1N1)pdm09, A(H3N2) and B influenza viruses collected from October 2012 to January 2013. Vaccine. 32:4713–4725.
10. Carrat F, Flahault A. 2007. Influenza vaccine: the challenge of antigenic drift. Vaccine. 25:6852–6862.

11. Hensley SE. 2014. Challenges of selecting seasonal influenza vaccine strains for humans with diverse pre-exposure histories. Curr Opin Virol. 8:85–89.
12. Peeters B, *et al.* 2017. Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses: consequences for vaccine strain selection. Virology. 503: 83–93.
13. Taubenberger JK, Kash JC. 2010. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe. 7:440–451.
14. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev. 56:152–179.
15. Nelson MI, Holmes EC. 2007. The evolution of epidemic influenza. Nat Rev Genet. 8:196–205.
16. Lässig M, Mustonen V, Walczak AM. 2017. Predicting evolution. Nat Ecol Evol. 1:1–9.
17. Morris DH, *et al.* 2018. Predictive modeling of influenza shows the promise of applied evolutionary biology. Trends Microbiol. 26:102–118.
18. Koelle K, Cobey S, Grenfell B, Pascual M. 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science. 314:1898–1903.
19. Van Nimwegen E. 2006. Influenza escapes immunity along neutral networks. Science. 314:1884–1886.
20. Wagner A. 2014. A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin. Proc R Soc B: Biol Sci. 281:20132763.
21. Łuksza M, Lässig M. 2014. A predictive fitness model for influenza. Nature. 507:57–61.
22. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. 2016. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci USA. 113:E1701–E1709.
23. Smith DJ, *et al.* 2004. Mapping the antigenic and genetic evolution of influenza virus. Science. 305:371–376.
24. Fonville JM, *et al.* 2014. Antibody landscapes after influenza virus infection or vaccination. Science. 346:996–1000.
25. Andreasen V, Lin J, Levin SA. 1997. The dynamics of cocirculating influenza strains conferring partial cross-immunity. J Math Biol. 35:825–842.
26. Gog JR, Grenfell BT. 2002. Dynamics and selection of many-strain pathogens. Proc Natl Acad Sci. 99:17209–17214.
27. Kamo M, Sasaki A. 2002. The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. Physica D: Nonlinear Phenom. 165:228–241.
28. Minayev P, Ferguson N. 2009. Improving the realism of deterministic multi-strain models: implications for modelling influenza A. J R Soc Interface. 6:509–518.
29. Kucharski JGAJ, Andreasen V. 2016. Capturing the dynamics of pathogens with many strains. J Math Biol. 72(1–2): 1–24.
30. Williams BJM, St-Onge G, Hébert-Dufresne L. 2021. Localization, epidemic transitions, and unpredictability of multistrain epidemics with an underlying genotype network. PLoS Comput Biol. 17:e1008606.
31. Zhang Y, *et al.* 2017. Influenza research database: an integrated bioinformatics resource for influenza virus research. Nucleic Acids Res. 45:D466–D474.
32. Barabási AL, Albert R. 1999. Emergence of scaling in random networks. Science. 286:509.
33. Eriksen KA, Hörnquist M. 2001. Scale-free growing networks imply linear preferential attachment. Phys Rev E. 65:017102.
34. Hébert-Dufresne L, Allard A, Young JG, Dubé LJ. 2016. Constrained growth of complex scale-independent systems. Phys Rev E. 93:032304.
35. Krapivsky PL, Rodgers GJ, Redner S. 2001. Degree distributions of growing networks. Phys Rev Lett. 86:5401.
36. Krapivsky PL, Redner S. 2001. Organization of growing random networks. Phys Rev E. 63:066123.
37. Cohen R, Havlin S. 2009. Complex media and percolation theory. In: Sahimi M, Hunt AG, editors. Encyclopedia of complexity and systems science series. New York (NY): Springer. p. 419–431.
38. Stumpf MP, Wiuf C, May RM. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci. 102:4221–4224.
39. Ferguson N, Andreasen V. 2002. The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In: Castillo-Chavez C, Blower S, van den Driessche P, Kirschner D, Yakubu AA, editors. Mathematical approaches for emerging and reemerging infectious diseases: models, methods, and theory. New York (NY): Springer. p. 157–169.
40. Erdős P, Rényi A, *et al.* 1960. On the evolution of random graphs. Publ Math Inst Hung Acad Sci. 5:17–60.
41. Fox Keller E. 2005. Revisiting "scale-free" networks. BioEssays. 27:1060–1068.
42. Dorogovtsev SN, Goltsev AV, Mendes JF. 2008. Critical phenomena in complex networks. Rev Modern Phys. 80:1275.
43. Yan L, Neher RA, Shraiman BI. 2019. Phylodynamic theory of persistence, extinction and speciation of rapidly adapting pathogens. eLife. 8:e44205.
44. Peeters B, *et al.* 2017. Genetic versus antigenic differences among highly pathogenic H5N1 avian influenza A viruses: consequences for vaccine strain selection. Virology. 503:83–93.
45. Bedford T, *et al.* 2014. Integrating influenza antigenic dynamics with molecular evolution. eLife. 3:e01914.
46. Young JG, *et al.* 2019. Phase transition in the recoverability of network history. Phys Rev X. 9:041056.
47. Cantwell GT, St-Onge G, Young JG. 2021. Inference, model selection, and the combinatorics of growing trees. Phys Rev Lett. 126:038301.
48. Neher RA, Russell CA, Shraiman BI. 2014. Predicting evolution from the shape of genealogical trees. eLife. 3:e03568.
49. Tria F, Pompei S, Loreto V. 2013. Dynamically correlated mutations drive human influenza a evolution. Sci Rep. 3:1–7.
50. Carlson JM, *et al.* 2014. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science. 345:1254031.
51. Yin C. 2020. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics. 112:3588–3596.
52. Gillespie CS. 2015. Fitting heavy tailed distributions: the poweRlaw package. J Stat Soft. 64:1–16.
53. Clauset A, Shalizi CR, Newman ME. 2009. Power-law distributions in empirical data. SIAM Rev. 51:661–703.