Topology of viral evolution

Joseph Minhow Chan[a,b], Gunnar Carlsson[c], and Raul Rabadan[a,b,d,1]

[a]Center for Computational Biology and Bioinformatics and Departments of [b]Biomedical Informatics and [d]Systems Biology, Columbia University College of Physicians and Surgeons, New York, NY 10032; and [c]Department of Mathematics, Stanford University, Stanford, CA 94305

Edited* by Arnold J. Levine, Institute for Advanced Study, Princeton, NJ, and approved October 11, 2013 (received for review July 18, 2013)

The tree structure is currently the accepted paradigm to represent evolutionary relationships between organisms, species or other taxa. However, horizontal, or reticulate, genomic exchanges are pervasive in nature and confound characterization of phylogenetic trees. Drawing from algebraic topology, we present a unique evolutionary framework that comprehensively captures both clonal and reticulate evolution. We show that whereas clonal evolution can be summarized as a tree, reticulate evolution exhibits nontrivial topology of dimension greater than zero. Our method effectively characterizes clonal evolution, reassortment, and recombination in RNA viruses. Beyond detecting reticulate evolution, we succinctly recapitulate the history of complex genetic exchanges involving more than two parental strains, such as the triple reassortment of H7N9 avian influenza and the formation of circulating HIV-1 recombinants. In addition, we identify recurrent, large-scale patterns of reticulate evolution, including frequent PB2-PB1-PA-NP cosegregation during avian influenza reassortment. Finally, we bound the rate of reticulate events (i.e., 20 reassortments per year in avian influenza). Our method provides an evolutionary perspective that not only captures reticulate events precluding phylogeny, but also indicates the evolutionary scales where phylogenetic inference could be accurate.

persistent homology | gene flow | topological data analysis

In *On the Origin of the Species* in 1859, Darwin first proposed the phylogenetic tree as a structure to describe the evolution of phenotypic attributes. Since then, the advancement of modern sequencing has spurred development of a number of phylogenetic inference methods (1, 2). The tree structure effectively models vertical or clonal evolution, mediated by random mutations over multiple generations (Fig. 1*A*). Phylogenetic trees, however, cannot capture horizontal, or reticulate, events, which occur when distinct clades merge together to form a new hybrid lineage (Fig. 1*B* and *SI Appendix*, Fig. S1). In nature, horizontal evolution can occur through species hybridization in eukaryotes, lateral gene transfer in bacteria, recombination and reassortment in viruses, viral integration in eukaryotes, and fusion of genomes of symbiotic species (e.g., mitochondria). These horizontal genetic exchanges create incompatibilities that mischaracterize the species tree (3). Doolittle (4) argued that molecular phylogeneticists have failed to identify the "true tree of life" because the history of living organisms cannot be understood as a tree. One may wonder what other mathematical structures beyond phylogeny can capture the richness of evolutionary processes.

Current techniques that detect reticulate events can be divided into phylogenetic and nonphylogenetic methodologies. Phylogenetic methods detect incongruence in the tree structure of different segments (5–8). Nonphylogenetic methods probe for homoplasies (shared character traits independently arising in different lineages by convergent or parallel evolution) or similar inconsistencies in sequence alignment (9–13). Although many of these methods are designed for sensitive detection of viral recombination and bacterial lateral gene transfer, they do not provide comprehensive representation of the evolutionary process.

Perhaps phylogenetic networks exemplify the largest departure from trees, allowing multiple paths between any two leaves. These methods visualize incompatibilities of sequence patterns or tree topologies as reticulation cycles in a network (14–16). Only the subfield of evolutionary networks is amenable to reticulate detection. However, major stumbling blocks abound for such methods. Although phylogenetic network structure is not necessarily unique, all current implementations produce only one network that may represent a suboptimal solution; results may depend on factors as arbitrary as the ordering of samples in the data matrix (16, 17). Moreover, many methods have impractical running times for even small datasets owing to the nondeterministic polynomial-time hard (NP-hard) problem of determining whether a tree exists in an evolutionary network (18). To address these obstacles, ad hoc methods simplify the search space of network structures: k-level, galled, tree-child, and tree-sibling networks. Although some of these methods cease to be NP-hard (19), all prioritize computational tractability over biological modeling (20). For example, galled tree networks minimize the number of inferred recombinations by ensuring that reticulation cycles share no nodes (21). This heuristic is appropriate only for low recombination rates and is not universally applicable.

Here, we propose a comprehensive and fast method of extracting large-scale patterns from genomic data that captures both vertical and horizontal evolutionary events at the same time. The structure we propose is not a tree or a network, but a set of higher-dimensional objects with well-defined topological properties. Using the branch of algebraic topology called persistent homology (throughout this paper, we refer to mathematical homology, not the notion of genetic or structural similarity), we extract robust global features from these high-dimensional complexes. Unlike phylogenetic methods that produce a single, possibly suboptimal, tree or network, persistent homology considers all topologies and their relationships across the entire parameter space of genetic distance. Through analysis of viral and simulated genomic datasets, we show how persistent homology captures fundamental evolutionary aspects not directly inferred

Significance

Evolution is mediated not only by random mutations over a number of generations (vertical evolution), but also through the mixture of genomic material between individuals of different lineages (horizontal evolution). The standard evolutionary representation, the phylogenetic tree, faithfully represents the former but not the latter scenario. Although many elaborations have been developed to address this issue, there is still no agreed-upon method of incorporating both vertical and horizontal evolution. Here, we present an alternative strategy based on algebraic topology to study evolution. This method extends beyond the limits of a tree to capture directly even complex horizontal exchanges between multiple parental strains, as well as uncover broader reticulate patterns, including the segregation of segments during reassortment.

Author contributions: J.M.C., and R.R. designed research; J.M.C. performed research; G.C. and R.R. contributed analytic tools; J.M.C. and R.R. analyzed data; and J.M.C., G.C., and R.R. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: rabadan@dbmi.columbia.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1313480110/-/DCSupplemental.

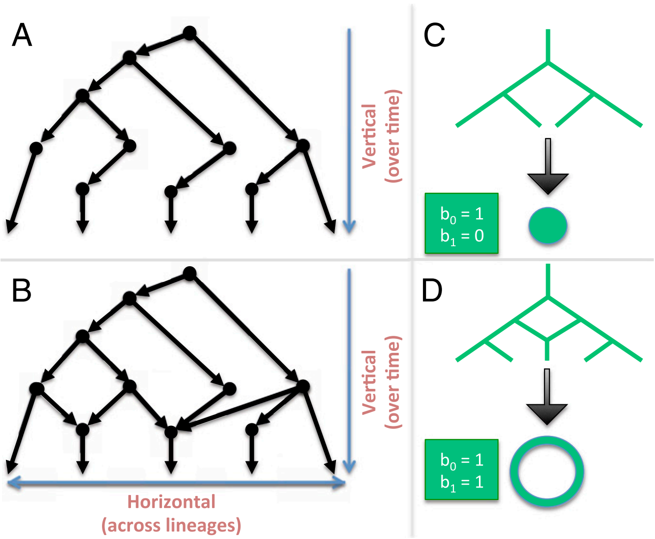

Fig. 1. Linking algebraic topology to evolution. (*A*) A tree depicting vertical evolution. (*B*) A reticulate structure capturing horizontal evolution, as well. (*C*) A tree can be compressed into a point. (*D*) The same cannot be done for a reticulate structure without destroying the hole at the center.

from phylogeny. In addition to representing clonal and reticulate evolution, persistent homology can determine the rate of horizontal genomic events, complex exchanges involving more than two organisms, and statistical patterns of cosegregation (genes more likely to be exchanged as a set). We calibrate our method using viral genomes because they are richly sampled and annotated, with a wide range of reticulate events; however, we foresee broader application to bacteria, eukaryotes, and other datasets.

Results

Persistent Homology in Evolution. We propose a mathematical structure that represents both vertical and horizontal evolutionary events at once. This structure is based on the field of algebraic topology, which characterizes global properties of a geometric object that are invariant to continuous deformation, that is, stretching or bending without tearing or gluing any single part of it. These properties include such notions as connectedness (the number of distinct connected components), as well as the number of holes. We aim to apply these concepts to characterize the topology of evolution. In settings of vertical evolution, a tree can be continuously deformed into a single point or connected component (Fig. 1*C*). The same action cannot be performed for a reticulate structure without destroying the loops within it. The active hypothesis is that the presence of holes results directly from reticulate events (Fig. 1*D*).

We are then interested in computing the number of holes in the evolutionary topological space, and algebraic topology mathematically formalizes these notions. In particular, holes can exist in different dimensions: a loop in one dimension, a void or cavity in two dimensions, and so on. To be precise, we are interested in holes that are "irreducible" cycles: a cycle in dimension k that does not serve as the boundary of a (k + 1)-dimensional object. We can define a topological invariant called the "homology group" H_k as an algebraic structure that encompasses all holes in dimension k, and the "Betti number" b_k is the count of these holes. The special case of the H_0 group addresses how many independent, unconnected components comprise a space.

For our purposes, we can assume that evolution forms some topological space E. Instead of directly observing E, we observe a sample of data points in E, particularly the genomic sequences separated from each other by some genetic distance. The set of these data points and space E do not share the same topology. However, we can estimate the topology of E by defining a function B(x,ε) as the ball centered at data point x with radius of genetic distance ε. We can show that at some value of ε, the union of balls B(x,ε) for all x shares the same topology as E (*SI Appendix*, Fig. S2). The topology of the union of balls is difficult to compute but can be estimated by constructing a corresponding topological space called a simplicial complex (22–24). In short, a simplicial complex is a set of points, lines, triangles, tetrahedra, and higher-dimensional "simplices." Like any topological space, this structure can contain holes of different dimensions (*SI Appendix*, Fig. S3). To build the simplicial complex, one can construct a line if any pair of points is within distance ε of each other, a triangle if any triplet of points are all within ε of each other, and so forth. At some ε, the resulting simplicial complex shares the same topology as the union of balls, and that of E as well (*SI Appendix*, Fig. S2).

However, different scales of ε create different simplicial complexes and reveal different irreducible cycles. A more comprehensive approach would consider all simplicial complexes over the entire parameter space of ε. For irreducible cycle C, we track when C exists over a filtration (subset) of simplicial complexes over a particular interval $[a_C,b_C]$ of genetic distance ε. Here, a_C and b_C are the birth and death of feature C. We then perform persistent homology, which computes the homology groups of dimension k at all scales ε. This process is depicted in a barcode plot, which shows a horizontal bar between $\varepsilon = [a_C,b_C]$ for every independent object C in the homology group. Bars persisting over a large interval of ε are unlikely to derive from noise (24).

Our aim, then, is to apply persistent homology to the study of evolution. We consider a set of genomes and calculate the genetic distance between each pair of sequences. Using the pairwise distance matrix, we calculate the homology groups across all genetic distances ε in different dimensions. We can refine our original hypothesis now and assert that zero-dimensional topology provides information about vertical evolution. At a particular scale ε, for example, b_0 represents the number of different strains or subclades. However, one-dimensional topology provides information about horizontal evolution, because reticulate structures contain loops (Fig. 2). We hypothesize that even higher-dimensional homology groups $H_{i\geq 2}$ result from multiple horizontal exchanges or complex reticulate events involving multiple parental strains. The "generator," the set of sequences that represents a particular irreducible cycle, can describe such complex genomic mixtures, as we will see in simulations and real data (HIV and avian influenza). In sum, we propose to connect the principles of algebraic topology and evolution and provide a dictionary that translates vocabulary in both fields (Table 1).

Topological Obstruction to Phylogeny. We can mathematically formalize the role of phylogeny within the framework of persistent homology. By definition, trees cannot contain holes; therefore, higher-dimensional irreducible cycles in a simplicial complex constructed from genomic data preclude phylogenetic construction at genetic distance ε. This intuition is proven for holes in dimension one and higher (*SI Appendix*, *Theorem 2.1*). We define an additive tree as a phylogeny where the distance between two leaves is the sum of the branch lengths connecting them. If there exists a nonzero Betti number of dimension greater than zero, then no additive tree exists that appropriately represents the genomic data. A related concept is the set of genetic distances I where nonzero topology vanishes. I reflects the evolutionary scales at which there is no evidence of reticulate exchange that can confound tree construction (Fig. 2*B*).

We define the topological obstruction to phylogeny (TOP) to be the L-∞ norm, or maximum, of B. If TOP is nonzero, then no additive tree exists that can appropriately represent the data. Another important concept is stability: the amount of fluctuation in the results owing to statistical noise, sequencing errors, or incomplete sampling. We show that TOP is a stable measure that is bounded by the Gromov–Hausdorff distance to the additive tree (*SI Appendix*, *Theorem 2.2*). Because additive structures have vanishing higher-dimensional homology, small deviations from additivity generate only small bars in the barcode.

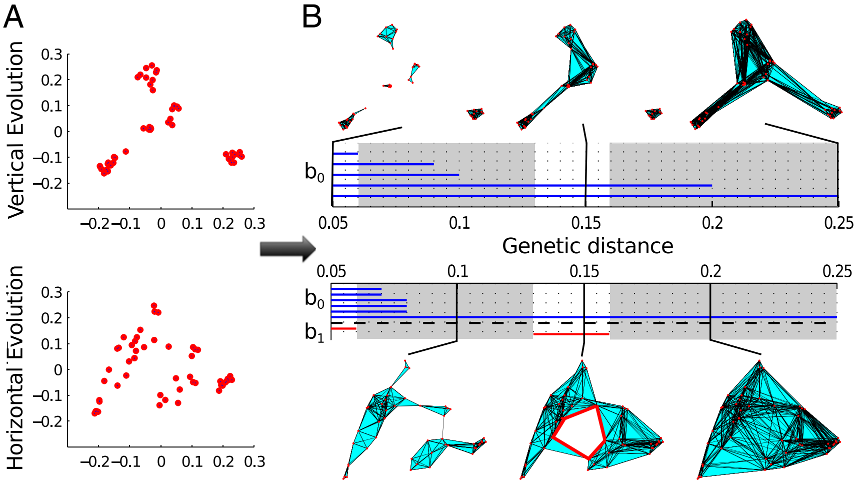

Fig. 2. Persistent homology characterizes topological features of vertical and horizontal evolution. Evolution was simulated with and without reassortment (*SI Appendix*, Supplementary Text). (*A*) A metric space of pairwise genetic distances d(i,j) can be calculated for a given population of genomic sequences $g_1,\ldots, g_n$. We visualize these data points using principal coordinate analysis (PCoA) (*SI Appendix*, Supplementary Text). (*B*) In the construction of simplicial complexes, two genomes are considered related (joined by a line) if their genetic distance is smaller than ε. Three genomes within ε of each other form a triangle, and so on (*SI Appendix*, Supplementary Text). From there, we calculate the homology groups at different genetic scales. In the barcode, each bar in different dimensions represents a topological feature of a filtration of simplicial complexes persisting over an interval of ε. A one-dimensional cycle (red highlight) exists at ε = [0.13, 0.16 Hamming distance] and corresponds to a reticulate event. The evolutionary scales *l* where b_1 = 0 are highlighted in gray.

Topological Estimates of Recombination/Reassortment Rates. Most approaches to estimating recombination rates are based on observed variance in site differences between pairs of haplotypes (25) or maximum likelihood estimators (26). Typically, these estimators assume constant population size, panmictic populations, and constant rates. Persistent homology, however, provides a lower bound for these rates by considering independent irreducible cycles for all ε in a period. The irreducible cycle rate (ICR) then is defined as the average number of the one-dimensional irreducible cycles per unit of time, ICR = (total number of one-dimensional bars for all ε)/(time frame). The numerator of this quantity is the L-0 norm (number of higher-dimensional bars). We normalized ICR based on the time interval between the earliest 5% and most recent 5% of the sequence dataset. Simulations show that ICR is proportional to and provides a lower bound for recombination/reassortment rate (*SI Appendix*, Fig. S5 *C* and *D* and *G* and *H*). In this way, sequence data add a temporal dimension to our methodology, a unique aspect not shared by other applications of algebraic topology.

Detection of Simulated Reticulate Events. To evaluate sensitivity and specificity of persistent homology to capture complex evolutionary processes, we simulated four scenarios: clonal evolution, population admixture, reassortment, and homologous recombination. Each simulation represents a population of constant size evolving over generations under a Wright–Fisher model with substitution rate μ, reassortment/recombination rate *r*, number of reassorting segments *S*, and subsamples with or without ancestral sequences (*SI Appendix*, Supplementary Text). Simulations show that (*i*) nontrivial homology appears when *r* is nonzero, (*ii*) one-dimensional ICR increases proportionally to *r*, and (*iii*) multiple reassortment/recombinant events can produce 2D topology. A precise account of simulations, comparison with other methods to identify recombination, and discussion of results can be found in *SI Appendix*, Supplementary Text and Figs. S4–S6.

Vertical Evolution in Influenza. The evolution of influenza, a segmented single-stranded RNA orthomyxovirus, is punctuated by frequent reassortment. To characterize influenza A evolution, we applied persistent homology to four influenza datasets from several hosts (avian, swine, and human), each numbering as many as 1,000 genomic sequences (*Dataset S1*, Table S1). When applied to a single viral segment unaffected by reassortment, higher-dimensional homology groups vanish, suggesting that no significant reticulate events have taken place (Fig. 3 *A* and *B*). Alignments of single segments are therefore more suitable for phylogenetic analysis.

In settings of vertical evolution, we can transform a filtration of simplicial complexes of dimension zero into an equivalent distance-based dendrogram. Fig. 4*A* represents the zero-dimensional topology of the hemagglutinin (HA) segment of avian influenza viruses. The zero-dimensional generators (*Dataset S1*, Table S2) at higher genetic distances ε indicate the major clusters, coinciding with the major antigenic subtypes (H1–H16). From the bar sizes of the barcode (Fig. 4*A*), we can create a dendrogram (Fig. 4*B*) that recapitulates classic phylogenetic analyses (27, 28), depicted in Fig. 4*C*. Importantly, our method deduces the two major HA groups and, in particular, tight clusters of H3 and H4; H7, H10, and H15; H8, H9, and H12; and H13 and H16.

Reassortment in Influenza Evolution. Owing to reassortment, each individual influenza segment can carry its own unique evolutionary history. Consequently, persistent homology of concatenated segments that are adjoined into a single sequence demonstrates evidence of reassortment that precludes phylogeny. For avian influenza, individual segments of HA (Fig. 3*A*) and neuraminidase

Table 1. Dictionary between persistent homology and evolutionary concepts

Persistent homology	Viral evolution
Filtration value ε	Genetic distance (evolutionary scale)
Zero-dimensional Betti number at filtration value ε	Number of clusters at scale ε
Generators of Zero-dimensional Betti number homology	A representative element of the cluster
Hierarchical relationship among generators of Zero-dimensional Betti number homology	Hierarchical clustering
1D Betti number	Number of reticulate events (recombination and reassortment)
Generators of 1D homology	Reticulate events
Generators of 2D homology	Complex horizontal genomic exchange
Nonzero high-dimensional homology (topological obstruction to phylogeny)	No phylogenetic representation
No. of higher-dimensional generators over time (irreducible cycle rate)	Lower bound on rate of reticulate events

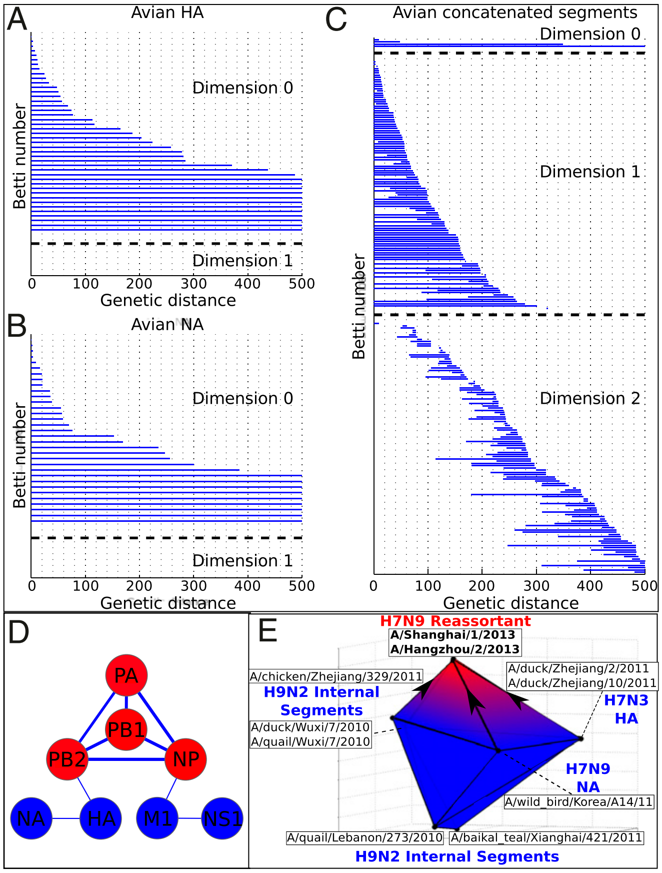

Fig. 3. Persistent homology of reassortment in avian influenza. Analysis of (*A*) HA and (*B*) NA reveal no significant one-dimensional topological structure. (*C*) Concatenated segments reveal rich 1D and 2D topology, indicating reassortment. For specific parameters, see *SI Appendix*, Supplementary Text. (*D*) Network representing the reassortment pattern of avian influenza deduced from high-dimensional topology. Line width is determined by the probability that two segments reassort together. Node color ranges from blue to red, correlating with the sum of connected line weights for a given node. For specific parameters, see *SI Appendix*, Supplementary Text. (*E*) b_2 polytope representing the triple reassortment of H7N9 avian influenza. Concatenated genomic sequences forming the polytope were transformed into 3D space using PCoA (*SI Appendix*, Supplementary Text). Two-dimensional barcoding was performed using Vietoris–Rips complex and a maximum scale ε of 4,000 nucleotides.

(NA) (Fig. 3*B*) independently produce only zero-dimensional homology groups. However, concatenating these segments reveals a complex high-dimensional topology (Fig. 3*C* and *Dataset S1*, Tables S3 and S4). These results confirm that persistent homology can detect pervasive reassortment in influenza.

To illustrate how higher-dimensional topology captures reassortments, we analyzed 1,000 human H3N2 genomes and identified three generators of one-dimensional homology when joining the PB2 and HA segments. As an example, the [G3] generator with the longest bar (*Dataset S1*, Table S5) is represented by an oriented one-dimensional irreducible cycle, implying at least one reassortment involving PB2 and HA of the isolates or their ancestors. The number of sequences in the generator serves as an upper bound on the number of candidate reassortants. Simple observation of the resulting sequence alignment reveals two divergent allelic patterns between informative sites in PB2 and HA, as reflected in incongruent trees (*SI Appendix*, Fig. S8 *A* and *B*) and reticulate cycles of the phylogenetic network (*SI Appendix*, Fig. S8*C*).

Our analysis of the concatenated H1N1pdm genome identified two nontrivial cycles, nominating candidate H1N1pdm reassortments in humans (*Dataset S1*, Table S6). Given the greater homogeneity of H1N1pdm sequences due to increased sampling, the number of informative sites among [G2] sequences was too small to perform maximum likelihood phylogenetic analysis. We therefore visually inspected informative sites, which suggested potential reassortment of two viral strains each contributing [PB2, M1, NS1] and [PB1, PA, HA] (*SI Appendix*, Fig. S9*A*). Phylogenetic network analysis supports these incompatibilities (*SI Appendix*, Fig. S9*B*).

One-dimensional ICR provides a lower-bound estimate of reassortment rate (*SI Appendix*, Fig. S10*B*). We calculate ICR < 1 event per year for classic H1N1 swine and H3N2 human influenza, supported by previous phylogenetic estimates (29, 30). In contrast, we calculate a high rate of 22.16 reassortments per year for avian influenza A (*Dataset S1*, Table S16). This difference could be explained by the high diversity and frequent coinfection of avian viruses (31) and correlates with the high proportion of avian reassortants reported in previous studies (32).

Nonrandom Reassortment Pattern in Avian Influenza. Although previous phylogenetic studies confirmed a high reassortment rate in avian influenza, none has identified a clear pattern of gene segment association (32). To determine whether any segments cosegregate more than expected by chance, we applied persistent homology to avian influenza. We first considered all pairs of concatenated segments and estimated the number of reassortments by b_1. We then ascertained the significance of observing a number of reassortments between each pair of segments given the total estimate of reassortments in the concatenated genome (*SI Appendix*, Supplementary Text). Analysis of avian influenza reveals a statistically significant configuration of four cosegregating segments: polymerase basic 2 (PB2), polymerase basic 1 (PB1), polymerase acidic (PA), and nucleoprotein (NP) (Fig. 3*D*). Interestingly, this pattern mimics previous in vitro results that suggest that effective protein–protein interaction between the polymerase complex and the NP protein constrain reassortment (31).

Recapitulating the H7N9 Avian Influenza Triple Reassortment. In April 2013, an outbreak of H7N9 avian influenza in humans and poultry began in the Jiangsu province of China and spread to kill 32 out of 131 positive cases as of May 8, 2013 (33). By constructing a series of trees per gene and observing conflicting structure by eye, Gao et al. (34) determined that the novel virus was a triple reassortment of an H7N3 A/duck/Zhejiang/12/2011-like lineage, an H7N9 A/wild bird/Korea/A14/2011-like lineage, and an H9N2 A/brambling/Beijing/16/2012-like lineage donating HA, NA, and internal segments, respectively. Persistent homology of concatenated H7N9, H9N2, and H7N3 avian genomes identified a 2D irreducible cycle representing the H7N9 triple reassortment. This 2D cavity is enclosed by a six-sided b_2 polytope formed by joining two tetrahedra at the top and bottom (Fig. 3*E*). The top tetrahedron is formed at the apex by H7N9 reassortants and at the base by members of the three parental lineages. This finding recapitulates the triple reassortment in a succinct, visually interpretable manner.

Topology of HIV, HEPC1, Dengue, and Other Viruses. The retrovirus HIV is notorious for high diversity mediated by not only a high mutation rate, but also frequent homologous recombination, leading to antiretroviral resistance (35) and immune evasion (36). These factors cloud studies that classically rely on phylogenetics, such as estimates of the origin of the HIV pandemic (37), underscoring the need for nonphylogenetic approaches. We apply our methodology to the independent and concatenated alignments of HIV-1 gag, pol, and env, the three largest genes of the genome. Like influenza, the concatenated alignment reproduces one-dimensional topology, indicative of nonclonal evolution. However, individual gene alignments also reveal one-dimensional homology groups, suggesting that recombination breakpoints exist within as well as between individual genes (Fig. 5 *A*–*D* and *Dataset S1*, Tables S8–S10).

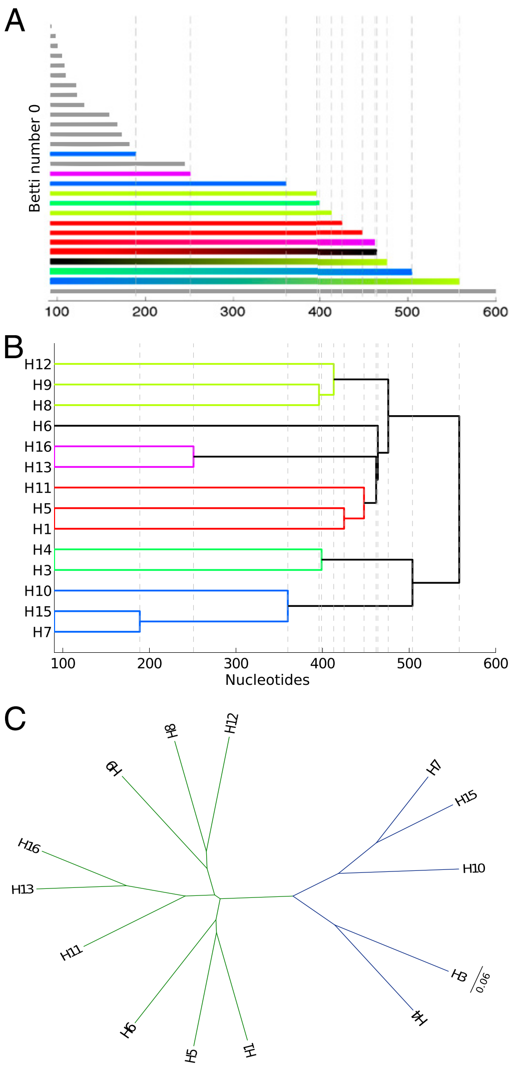

Fig. 4. Reconstructing phylogeny from persistent homology of avian influenza HA. (*A*) Barcode plot in dimension 0 of all avian HA subtypes. Each bar represents a connected simplex of sequences given a Hamming distance of ε. When a bar ends at a given ε, it merges with another simplex. Gray bars indicate that two simplices of the same HA subtype merge together at a given ε. Solid color bars indicate that two simplices of different HA subtypes but same major clade merge together. Interpolated color bars indicate that two simplices of different major clades merge together. Colors correspond to known major clades of HA. For specific parameters, see *SI Appendix*, Supplementary Text. (*B*) Phylogeny of avian HA reconstructed from the barcode plot in *A*. Major clades are color-coded. (*C*) Neighbor-joining tree of avian HA (*SI Appendix*, Supplementary Text).

In addition, persistent homology of the concatenated HIV gag, pol, and env produced 2D topology derived from complex recombination events (*Dataset S1*, Table S12). One example is the [G4] generator depicted by the b_2 polytope in Fig. 5*E*. This 2D cavity identifies CRF13_cpx and CRF0209, two complex circulating recombinant forms (CRFs) that result from recombination between viruses of different subtypes. CRF13_cpx recombinant derives from subtypes A, G, J, and U, and CRF01_AE. CRF0209 is a recombinant between CRF02_AG and CRF09_cpx (subtypes A, G, and U). The b_2 polytope includes other viral subtypes A, B, C, D, and F, as well.

Flaviviruses are positive single-stranded RNA viruses, whose ability to perform homologous recombination through RNA polymerase template switching has been debated. Sporadic recombinants have been detected for flaviviruses such as hepatitis C (38, 39), dengue virus (40, 41), and West Nile virus (42). However, some of these reports, as in dengue, have been shown to be the product of sequencing error (43), and it is generally agreed that if recombination occurs, it is rare. To assay the extent of flavivirus recombination, we applied our methodology to the polyproteins of hepatitis C subtype 1, dengue subtypes 1–4, and West Nile virus. We found high-dimensional topology for hepatitis C virus (*Dataset S1*, Table S13), although at lower TOP and ICR than in HIV. However, we found little to no high-dimensional structure for dengue (*SI Appendix*, Fig. S11) and no evidence of recombination for West Nile virus, suggesting that recombination rarely occurs in these viruses.

Recombination is considered even less frequent in negative-sense RNA viruses such as Newcastle virus, with only scattered studies reporting positive findings (44). Interestingly, persistent homology of Newcastle virus confirmed a low ICR but a nonvanishing TOP (*Dataset S1*, Table S14). *SI Appendix*, Fig. S10 and Tables S15 and S16 summarize TOP and ICR for all viruses

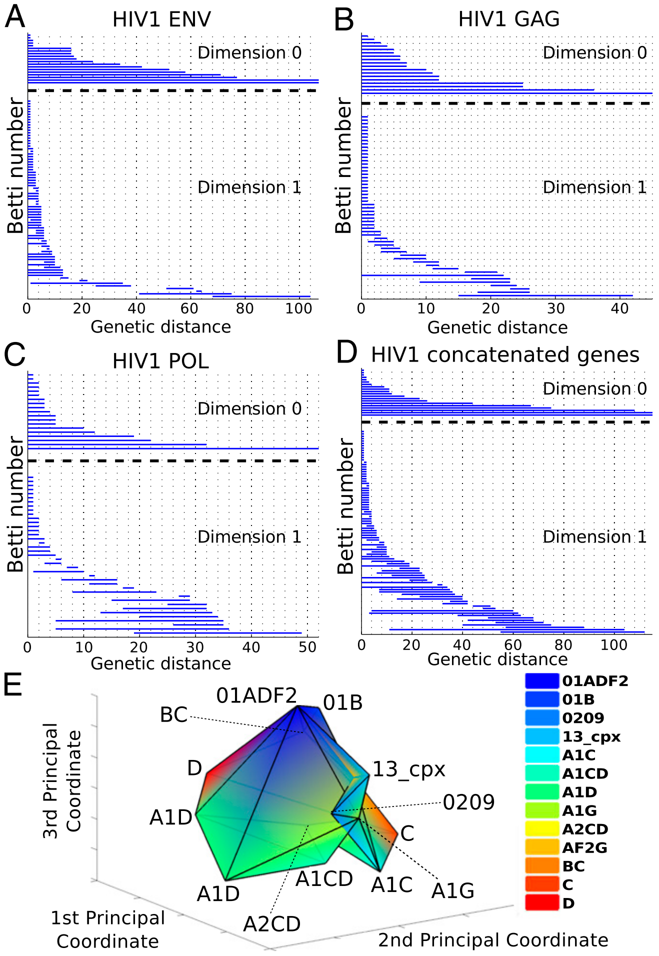

Fig. 5. Barcoding plots of HIV-1 reveal evidence of recombination in (*A*) env, (*B*), gag, (*C*) pol, and (*D*) the concatenated sequences of all three genes. One-dimensional topology present for alignments of individual genes as well as the concatenated sequences suggests recombination. (*E*) b_2 polytope representing a complex recombination event with multiple parental strains. Sequences of the [G4] generator of concatenated HIV-1 gag, pol, and env were transformed into 3D space using PCoA (*SI Appendix*, Supplementary Text) of the Nei–Tamura pairwise genetic distances. Two-simplices from the [G4] generator defined a polytope whose cavity represents a complex recombination. Each vertex of the polytope corresponds to a sequence that is color-coded by HIV-1 subtype. For specific parameters, see *SI Appendix*, Supplementary Text.

in this paper. Additional results are detailed in *SI Appendix*, Supplementary Text.

Discussion

Persistent homology can quickly determine robust topological properties of evolution from big genomic datasets. Barcodes generated by persistent homology represent the structure of these properties at all evolutionary scales. The unveiled structure is equivalent to a tree in cases where only vertical genetic exchange takes place, as can be seen in single-segment analysis of influenza A, dengue, West Nile virus, and rabies. However, persistent homology can also capture horizontal events such as recombination and reassortment. Generators of nontrivial homology identify specific reticulate events, and the normalized count provides a lower bound for the recombination/reassortment rate. Using this strategy, we estimated these rates in several influenza strains, HIV, flaviviruses, rabies, and Newcastle virus. Moreover, we used higher-dimensional topology to uncover complex evolutionary patterns, such as cosegregating segments during influenza reassortment. As a guide, we provide a dictionary that links evolutionary concepts to principles of algebraic topology (Table 1).

Persistent homology proposes a departure from the tree paradigm of evolution. Where many phylogenetic methods produce a single, possibly suboptimal, tree or network, persistent homology analyzes the invariant topological characteristics of all simplicial complexes across the entire parameter space of genetic distance. Moreover, our methodology exhibits stability to small fluctuations in input data.

In this paper, we propose a phenomenological approach to study evolution by capturing topological features of the relationships between genomes without imposing preconceived structure. These relationships could be phylogenetic or not, between organisms of the same or different species. By uncovering global, topological features that cannot be represented by phylogeny, persistent homology analyzes viral genomic datasets to (*i*) represent general evolutionary processes that may include reticulate events, (*ii*) estimate reticulate rates, and (*iii*) extract complex patterns in these processes. In the future, we foresee further application of persistent homology to other nonviral taxa and to addressing the species problem from genomic data.

Methods

Here, we harness persistent homology to characterize viral evolution. To this end, we apply our method to both real and simulated viral datasets. Collection or generation of datasets, mathematical background and implementation of persistent homology, comparison with other methods detecting recombination, and other tests used in the paper are detailed in *SI Appendix*, Supplementary Text.

ACKNOWLEDGMENTS. We thank Dr. Stephen Goff, Dr. David Hirsh, Dr. Siddhartha Mukherjee, Suzanne Christen, Kevin Emmett, Ola Jacunski, Oliver Eliott, and Dr. Francesco Abate for their comments. R.R. was supported by Grant U54 CA121852-05, Defense Threat Reduction Agency Grant W81XWH-13-2-0029, and the Stewart Foundation. G.C. was supported by National Science Foundation Grants DMS 0905823 and DMS 0964242, Air Force Office of Scientific Research Grants FA9550-09-1-0643 and FA9550-09-1-0531, and National Institutes of Health Grant I-U54-ca149145-01.

1. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer Associates, Sunderland, MA).
2. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161(3):1307–1320.
3. Nei M (1986) Stochastic errors in DNA evolution and molecular phylogeny. *Prog Clin Biol Res* 218:133–147.
4. Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284(5423): 2124–2129.
5. Posada D, Crandall KA, Holmes EC (2002) Recombination in evolutionary genomics. *Annu Rev Genet* 36:75–97.
6. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* 21(1):98–102.
7. McGuire G, Wright F (2000) TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* 16(2):130–134.
8. Minin VN, Dorman KS, Fang F, Suchard MA (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* 21(13):3034–3042.
9. Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? *Proc Natl Acad Sci USA* 90(10):4384–4388.
10. Maynard Smith J, Smith NH (1998) Detecting recombination from gene trees. *Mol Biol Evol* 15(5):590–599.
11. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681.
12. Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* 12(4):291–295.
13. Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* 34(2):126–129.
14. Fitch WM (1997) Networks and viral evolution. *J Mol Evol* 44(Suppl 1):S65–S75.
15. Holland BR, Huber KT, Moulton V, Lockhart PJ (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol Biol Evol* 21(7):1459–1461.
16. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254–267.
17. Morrison DA (2011) *Introduction to Phylogenetic Networks* (RJR Productions, Uppsala), p 216.
18. Kanj IA, Nakhleh L, Than C, Xia G (2008) Seeing the trees and their branches in the network is hard. *Theor Comput Sci* 401(1-3):153–164.
19. van Iersel L, Kelk S, Rupp R, Huson D (2010) Phylogenetic networks do not need to be complex: Using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26(12):i124–i131.
20. Kelk S (2011) Phylogenetic networks: Concepts, algorithms and applications. *Syst Biol* 61(1):174–175.
21. Gusfield D, Eddhu S, Langley C (2004) Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J Bioinform Comput Biol* 2(1):173–213.
22. Edelsbrunner H, Letscher D, Zomorodian A (2000) Topological persistence and simplification. *Proceedings 41st Annual Symposium on Foundations of Computer Science, 2000* (IEEE, Washington, DC), pp 454–463.
23. Zomorodian A, Carlsson G (2005) Computing persistent homology. *Discrete Comput Geom* 33(2):249–274.
24. Collins A, Zomorodian A, Carlsson G, Guibas LJ (2004) A barcode shape descriptor for curve point cloud data. *Comput Graph-Uk* 28(6):881–894.
25. Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genet Res* 50(3):245–250.
26. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160(3):1231–1241.
27. Fouchier RA, et al. (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J Virol* 79(5):2814–2822.
28. Hifumi E, Fujimoto N, Ishida K, Kawawaki H, Uda T (2010) Characteristic features of InfA-15 monoclonal antibody recognizing H1, H3, and H5 subtypes of hemagglutinin of influenza virus A type. *J Biosci Bioeng* 109(6):598–608.
29. Lycett SJ, et al.; Combating Swine Influenza Initiative-COSI Consortium (2012) Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *J Gen Virol* 93(Pt 11):2326–2336.
30. Holmes EC, et al. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol* 3(9):e300.
31. Lubeck MD, Palese P, Schulman JL (1979) Nonrandom association of parental genes in influenza A virus recombinants. *Virology* 95(1):269–274.
32. Dugan VG, et al. (2008) The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog* 4(5):e1000076.
33. Holmes D (2013) The world waits for H7N9 to yield up its secrets. *Lancet Infect Dis* 13(6):477–478.
34. Gao R, et al. (2013) Human infection with a novel avian-origin influenza A (H7N9) virus. *N Engl J Med* 368(20):1888–1897.
35. Nora T, et al. (2007) Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J Virol* 81(14):7620–7628.
36. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM (2004) Dynamics of HIV-1 recombination in its natural target cells. *Proc Natl Acad Sci USA* 101(12):4204–4209.
37. Rambaut A, Posada D, Crandall KA, Holmes EC (2004) The causes and consequences of HIV evolution. *Nat Rev Genet* 5(1):52–61.
38. Colina R, et al. (2004) Evidence of intratypic recombination in natural populations of hepatitis C virus. *J Gen Virol* 85(Pt 1):31–37.
39. González-Candelas F, López-Labrador FX, Bracho MA (2011) Recombination in hepatitis C virus. *Viruses* 3(10):2006–2024.
40. Worobey M, Rambaut A, Holmes EC (1999) Widespread intra-serotype recombination in natural populations of dengue virus. *Proc Natl Acad Sci USA* 96(13):7352–7357.
41. Tolou HJG, et al. (2001) Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome sequences. *J Gen Virol* 82(Pt 6): 1283–1290.
42. Pickett BE, Lefkowitz EJ (2009) Recombination in West Nile Virus: Minimal contribution to genomic diversity. *Virol J* 6(1):165–171.
43. Rico-Hesse R (2003) Microevolution and virulence of dengue viruses. *Adv Virus Res* 59: 315–341.
44. Chare ER, Gould EA, Holmes EC (2003) Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *J Gen Virol* 84(Pt 10): 2691–2703.