PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Revisiting the recombinant history of HIV-1 group M with dynamic network community detection

Abayomi S. Olabode[a,1], Garway T. Ng[a], Kaitlyn E. Wade[a,b], Mikhail Salnikov[c], Heather E. Grant[d], David W. Dick[e], and Art F. Y. Poon[a,b,c,e]

Edited by Malcolm Martin, National Institute of Allergy and Infectious Diseases, Bethesda, MD; received May 19, 2021; accepted March 16, 2022

The prevailing abundance of full-length HIV type 1 (HIV-1) genome sequences provides an opportunity to revisit the standard model of HIV-1 group M (HIV-1/M) diversity that clusters genomes into largely nonrecombinant subtypes, which is not consistent with recent evidence of deep recombinant histories for simian immunodeficiency virus (SIV) and other HIV-1 groups. Here we develop an unsupervised nonparametric clustering approach, which does not rely on predefined nonrecombinant genomes, by adapting a community detection method developed for dynamic social network analysis. We show that this method (dynamic stochastic block model [DSBM]) attains a significantly lower mean error rate in detecting recombinant breakpoints in simulated data (quasibinomial generalized linear model (GLM), $P < 8 \times 10^{-8}$), compared to other reference-free recombination detection programs (genetic algorithm for recombination detection [GARD], recombination detection program 4 [RDP4], and RDP5). When this method was applied to a representative sample of $n = 525$ actual HIV-1 genomes, we determined $k = 29$ as the optimal number of DSBM clusters and used change-point detection to estimate that at least 95% of these genomes are recombinant. Further, we identified both known and undocumented recombination hotspots in the HIV-1 genome and evidence of intersubtype recombination in HIV-1 subtype reference genomes. We propose that clusters generated by DSBM can provide an informative framework for HIV-1 classification.

HIV-1 | recombination | virus evolution | stochastic block model | dynamic network

Significance

Recombination is a major mechanism through which HIV type 1 (HIV-1) maintains genetic diversity and interferes with viral eradication efforts. There is growing evidence demonstrating a recombinant origin of primate lentiviruses including HIV-1 group M (HIV-1/M). Inferring the extent of recombination across the entire HIV-1/M genome is of great importance as it provides deeper insights into the origin, dynamics, and evolution of the global pandemic. Here we propose an alternative method that can reconstruct the extent of genome-wide recombination in HIV-1, uncover reticulate patterns, and serve as a framework for HIV-1 classification. Our method provides an alternative approach for understanding the roles of virus recombination in the early evolutionary history of zoonosis for other emerging viruses.

Understanding the global epidemiology of HIV type 1 (HIV-1) is contingent on our ability to accurately reconstruct the origin of the virus (1). The origin of different lineages of HIV-1 has been traced back to multiple zoonotic transmissions of simian immunodeficiency virus from chimpanzees (SIVcpz) directly or through an intermediate host to human populations in West and Central Africa (2, 3). SIVcpz, on the other hand, originated through multiple cross-species transmissions of SIV from other primates to chimpanzees, with extensive recombination between two or more ancestral variants (4). SIVs in other nonhuman primate species also show evidence of recombinant origins (5), and at least 13 recombinant breakpoints and 14 host-switch events have been identified among SIV lineages overall (6). Furthermore, there is phylogenetic evidence that the ancestor of HIV-1 group N was a mosaic combination of the HIV-1 group M (HIV-1/M) ancestor and an SIV lineage (2). Taken together, these findings imply that recombination has been an important evolutionary force in the deep evolutionary history of primate lentiviruses.

Currently, HIV-1 is classified into four major groups (M, N, O, P). Group M, which is responsible for the global pandemic, is further subdivided into nine "pure" subtypes (A-D, F-H, J, and K) that are defined by substantial genetic divergence (~10 to 25%) and/or bootstrap support for monophyletic clades (7). The first evidence of the existence of recombinant HIV-1 genomes was documented in the late 1980s (8). HIV-1 exhibits a high rate of recombination that is driven by obligate template switching during the reverse transcription stage of its replication cycle (9) and by coinfection at the cellular (10) and host levels (11). Consequently, there are presently close to 100 circulating recombinant forms (CRFs) documented by the Los Alamos National Laboratory (LANL) HIV Sequence Database (https://www.hiv.lanl.gov/), where each CRF comprises a mosaic of two or more pure HIV-1 subtypes that has been sampled from at least three epidemiologically unlinked individuals (7). There are also a large number of unique recombinant forms (URFs) that do not meet the latter criterion (12). Nevertheless, the prevailing view of HIV-1 diversity is that the majority of virus genomes can be classified into one of the subtypes or a relatively small number of CRFs that have reached a high level of global or regional prevalence, such as CRF01_AE (Southeast Asia) (13) or CRF07_BC (China) (14). Thus, intersubtype recombination is considered to be an infrequent event, such that new infections can be routinely classified into one of the subtypes or CRFs by sequencing

Author contributions: A.S.O. and A.F.Y.P. designed research; A.S.O. and A.F.Y.P. performed research; D.W.D. contributed new reagents/analytic tools; A.S.O., G.T.N., K.E.W., M.S., H.E.G., and A.F.Y.P. analyzed data; and A.S.O. and A.F.Y.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: aolabode@uwo.ca.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2108815119/-/DCSupplemental.

Published May 2, 2022.

a specific region of the genome, such as the HIV-1 *pol* gene that encodes the major targets of antiretroviral therapy.

Increasing deployment of whole-genome sequencing technologies for HIV-1 around the world is providing a clearer picture of the prevalence and diversity of recombinant HIV-1 genomes (15). For instance, molecular clock estimates of the origin of the HIV-1 pandemic (16), or the time to the most recent common ancestor (tMRCA), may be inaccurate due to a deep recombinant history, such that a single phylogenetic tree is not an adequate representation of discordant evolutionary histories relating different regions of the virus genome. In previous work (17), we used a sliding-window molecular clock analysis of near full-length HIV-1 genomes to show that different regions of the virus genome yield significantly different estimates of the tMRCA. This result supports the hypothesis that HIV-1/M may have a deep recombinant origin that involved at least two genome fragments with different evolutionary histories. For example, this result could be explained by the introgression of a genome fragment from an unsampled lineage with a more distant common ancestor to the diversity of the present day (18).

This phylogenetic evidence, together with other findings on the deep recombinant history of SIV (6), supports the hypothesis that the evolutionary history of HIV-1/M can be improved if the conventional framework of established pure reference subtypes with limited subsequent recombination is revisited. One of the obstacles to evaluating this hypothesis is that most tools for classifying HIV-1 genomes rely on a reference set of predefined subtypes, such as COMET (context-based modeling for expeditious typing) (19) or SCUEAL (subtype classification using evolutionary algorithms) (20). In other words, these are supervised classification methods. We propose that the best way to test whether the early evolutionary history of HIV-1/M is reticulate is to apply an unsupervised clustering method to the global diversity of virus genomes. For example, GARD (genetic algorithm for recombination detection) (21) is a divisive unsupervised method that attempts to find the optimal partition of the sequence alignment that maximizes the joint likelihood. Fitting GARD to large numbers of near full-length genomes is computationally intensive because phylogenies must be reconstructed for each genome segment defined by every candidate partition. Furthermore, a postprocessing step is required to determine support for a recombination breakpoint from the topological discordance of phylogenies. In addition, there are a number of recombination detection heuristic methods that do not require reference sequences, several of which are implemented in the software package RDP (recombination detection program) (22). However, these methods are limited to pairwise or three-way comparisons among sequences.

This study endeavors to reconstruct the extent of recombination across the entire HIV-1/M genome. Here we describe an unsupervised nonparametric approach to this problem based on adapting a community detection method that was developed for the analysis of dynamic social networks (23). Community detection consists of grouping or partitioning nodes (vertices) of a network graph into the same community based on their relative edge density (24, 25). We note that a connected component in the network, which is often referred to as a "cluster" in the context of genetic epidemiology (26), may comprise multiple communities. Likewise, network communities are often referred to as clusters. To avoid confusion, we explicitly refer to subgraphs in which there are no edges to external nodes as connected components instead of clusters.

A network in which the distribution of edges among nodes remains constant over the observational time period is a static network (27). However, real networks are dynamic as relationships evolve over time, represented by the addition and/or removal of network edges (28, 29). A dynamic network can be represented by a series of static networks that each capture the state of the system at a given point in time. In the context of a dynamic network, communities are a group of nodes that are stably connected by a relatively high density of edges over time (30). The development of methods to detect network communities is an active area of research with a broad domain of application, including biological sciences, for extracting patterns from complex relational data (25).

In this study, we use genetic distances between sequences to generate a network or graph, where each node represents a virus genome, and each edge indicates that the respective genomes have a distance below some threshold. We assume that communities in the resulting graph correspond to phylogenetic clusters such as subtypes. When distances are calculated from different regions of the genome, a node may switch membership from one community to another, which can correspond to a recombination event. This is analogous to an individual in a social network switching affiliations from one community to another. We use stochastic block modeling to detect the community structure of the genetic similarity graph and employ a recently described (31) expectation maximization algorithm to estimate the "migration" rates between communities. Stochastic block models (SBMs) are one of the most widely utilized classes of models for community detection in networks (23). Individuals belong to one of K latent communities, and the probability of an edge between individuals is determined only by community membership; e.g., $P_{ii} > P_{ij}$ for $i, j \in 1, ..., K$ where $i \neq j$. SBMs are designed to uncover hidden structural features in complex networks by clustering nodes based on similar or shared attributes (32), while dynamic stochastic block models (DSBMs) uncover these hidden data structures as a dynamic network changes over discrete time (33). We propose to adapt DSBMs as an unsupervised method to characterize the effect of recombination on the evolutionary history of HIV-1 genomes.

Methods

Data Processing. A total of $n = 3,900$ near full-length (>8,000 nt) HIV-1/M genomes, manually curated from our previous study (17), were used in this study. HIV-1/M has accumulated a considerable amount of genetic diversity that includes numerous sequence insertions and deletions. Consequently, a multiple-sequence alignment generated with conventional methods for a global selection of HIV-1/M genomes tends to contain a large number of gap-rich intervals with spurious alignments of nonhomologous nucleotides (34). To address this issue, we used an automated alignment-free clustering method that we developed in a previous study (17). In summary, we used a k-mer distance (35) to generate a pairwise distance matrix that we converted to a graph comprising a number of connected components. Next, we used a network centrality statistic to select a single representative genome for each component and then generated a consensus sequence from the multiple-sequence alignment of these representatives. Finally, we constructed a reduced multiple-sequence alignment based on the pairwise alignment of sequences against this consensus genome, discarding insertions relative to this reference to filter out regions of relatively low evolutionary homology.

To minimize the computing time of subsequent steps in our analysis, we selected $n = 550$ sequences from this alignment by progressively removing genomes with the shortest genetic distances to other genomes in the dataset. Next, we realigned the remaining sequences using multiple alignment using fast Fourier transform, version 7.271 (36) and then partitioned the alignment into sliding windows of 500 nt at steps of 100 nt, resulting in a total of 82 alignment subsets covering HXB2 nucleotide coordinates 790 to 9,465. (We arrived at these parameter settings after some preliminary tests varying window and step sizes.) Sequences that had deletions spanning more than 20% in one or more subsets were excluded, resulting in a final total of 525 sequence subsets. We used SCUEAL

(20) to classify full-length genomes with respect to the HIV-1/M subtype reference sequences obtained from the LANL HIV Sequence Database.

Community Detection. The aim of this study was to characterize the extent of recombination over a series of networks as snapshots of the dynamic system over time, throughout the evolutionary history of HIV-1/M. We are adapting DSBMs to this problem by drawing an analogy between time and the length of the HIV-1 genome, with recombination breakpoints as discrete events. If recombination is relatively infrequent, then genomes should largely cluster into stable communities (pure subtypes) that diverge over time from a common ancestor. With abundant recombination, however, we would expect sequences to switch frequently between communities. To construct a graph, we used the *dist.dna* function of the R package *ape* (37) to compute the Tamura–Nei (38) (TN93) distances between every pair of sequences within each of the 82 windows extracted from the alignment. We imported the resulting distance matrices into R to generate undirected graphs using the package *igraph* (39). The TN93 distance is the most biologically realistic genetic distance for which a closed-form solution is available, facilitating rapid computation for large alignments (40). It relaxes the assumption of equal transition rates (A-G and C-T), which makes it well suited for comparing HIV-1 sequences that tend to exhibit a significant A-G transition bias due to G-to-A hypermutation (41). By default, the *dna.dist* function treats ambiguous base calls (mixtures) as completely missing values that are ignored when computing distances. Our original HIV-1/M dataset of $n = 525$ sequences contained no mixtures; however, 10 of the 37 sequences in the HIV-1 subtype reference dataset contained a small number of mixtures with a mean of 0.31% (range: 0.01 to 2.13%) mixtures per sequence.

We did not use a fixed distance threshold for all windows because rates of molecular evolution vary substantially from one part of the HIV-1 genome to another. Instead, we applied different percentile thresholds (e.g., the lower quartile distance, 25%) to the observed distribution of TN93 distances for each window and used simulation experiments (*SI Appendix*, Fig. S1) to assess which percentile yielded the most informative graphs. Using a relative percentile threshold means that the actual distance threshold will vary among windows in response to differences in genetic variation. Finally, to reconstruct the distribution of recombination events in the full alignment, we used the R package *dynsbm* (42) to fit a DSBM to the series of graphs. This method jointly estimates the cluster memberships for all nodes in the graph and a transition rate matrix for the movement of nodes to other clusters between graphs. These parameters are estimated by an expectation maximization method (43), which is an iterative algorithm for fitting a model by maximum likelihood when some variables are not observed (i.e., missing or latent). In the case of DSBM, the cluster memberships are latent variables. In brief, the algorithm estimates the latent variables given the current parameters and then updates the parameters by maximum likelihood; the new parameter values are then used to reestimate the latent variables, and so forth. To determine the optimal number of clusters, we used the integrated completed likelihood (ICL) criterion (44), which is commonly used in model-based clustering applications (45, 46).

To predict recombination breakpoints from the DSBM outputs, we used the R package *changepoint* to perform change-point detection on the estimated cluster memberships along the sequence of windows for each genome. Change-point detection methods are useful for detecting abrupt and substantial changes in a time series of observations (47) and have a broad domain of application including finance, climate change, and monitoring medical conditions (47, 48). A change is detected from a shift in a summary statistic, such as the mean or variance, computed for a consecutive series of observations (49). In the context of our work, we draw an analogy between the sliding windows along a sequence alignment and a time series. Hence, change points correspond to estimated recombination breakpoints as determined from shifts in cluster memberships. For each sequence of cluster memberships, we dropped missing values (genome windows that could not be assigned to any cluster) up to a maximum tolerance of 10 missing values, beyond which we did not attempt to predict breakpoints for that genome. We evaluated both the default "at most one change" (AMOC) method (50) and the more recently described pruned exact linear time (PELT) algorithm (51). Both methods require the user to specify a minimum segment length (m), which is the number of consecutive observations for which no change points are tolerated (52). For example, setting $m = 10$ windows means that a new change point cannot be placed on the interval between two existing points if the interval comprises 10 or fewer windows–not even if every window has a different cluster assignment. Reducing m confers more sensitivity at the cost of lower specificity, and vice versa. In both cases, we used the default modified Bayesian information criterion (MBIC) to penalize the addition of change points (53).

Recombination Simulation. Simulating sequence evolution is an important tool for validating new methods in phylogenetic analysis by evaluating their accuracy on data with known parameters. To calibrate our DSBM method, we evaluated its prediction accuracy on simulations under varying parameter settings. In addition, we compared its performance to other recombination detection methods that do not require reference genomes, i.e., GARD and RDP. First, we created a highly simplified recombination scenario using an alignment of 16 reference genomes, 4 for each of the HIV-1/M subtypes A, B, C, and D. We applied one to three breakpoints to produce recombination fragments of roughly equal size, i.e., one breakpoint at position 4,500, two at 3,000 and 6,000, or three at 2,250, 4,500, and 6,750. For each breakpoint, we selected two subtypes at random to recombine, such that all four genomes of one subtype exchanged segments with all four genomes of the other. We then evaluated our ability to reconstruct these simulated recombination breakpoints. This "post hoc" simulation method is similar to the one used to validate the subtyping and recombination detection program COMET (19), which is reference dependent.

Next, we generated more realistic simulations in which recombination events could be distributed throughout the evolutionary history of the "observed" sequences. We generated a multiple-sequence alignment of a larger set ($n = 37$) of HIV-1/M subtype reference sequences from the LANL database, including representatives of subtypes A-D, F-H, J, and K. From this alignment, we used IQ-TREE (version 1.3.11.1) (54) to reconstruct a maximum-likelihood tree relating the subtype reference sequences used in the previous method. We rooted this tree using the sample collection dates by root-to-tip regression using the *rtt* function in the R package *ape* (37). Next, we used BEAST (version 1.10.4) (55) to sample time-scaled phylogenies from the posterior distribution, using the TN93 nucleotide substitution model, with rate distribution modeled by a gamma distribution with four rate categories; an uncorrelated lognormal clock model ($\mu = 1, \sigma = 0.33$); and a skyline tree prior with 10 population sizes. In addition, we set the prior distribution for the time to the most recent common ancestor to a normal distribution with $\mu = 82$ y and $\sigma = 4.1$, based on recent estimates of the origin of HIV-1/M (1). We ran a single-chain sample for 10^8 steps, discarded the first 10^7 steps as burn-in, and generated a maximum clade credibility (MCC) tree from the remainder using TreeAnnotator (version 1.10.4).

We used a Python script to update the MCC tree with recombination events by switching random branches that span a randomly selected time. To generate one, two, or three recombination breakpoints, we selected a random point in time to prune and regraft subtrees on two randomly selected extant branches. We assumed that the leftmost portion of the alignment was always related by the original tree and that the portion of the alignment past the first breakpoint was related by the modified tree. We repeated this process for additional breakpoints by progressively modifying the tree to the immediate left of the breakpoint to relate sequence fragments to the right. In all, we generated 100 simulations of one, two, and three breakpoints for a total of 300 sets of trees. The evolution of a 9,000-nt sequence at the root was simulated along each tree using INDELible (v1.0.3) (56) under a codon substitution model with rate variation modeled by a gamma distribution ($\alpha = 1.5$, $\beta = 3$) discretized into 50 rate categories and a transition/transversion ratio $\kappa = 8.0$ and applied to all tree segments. These settings were taken from a previous study (26) in which the simulation model was calibrated to yield alignments with pairwise TN93 distributions closely resembling the empirical distribution for an alignment of actual HIV-1 sequences. The input tree was rescaled such that the expected number of substitution events per codon was 3.22, which we derived from the total length of the maximum clade credibility tree (643.5 y), the clock rate estimate (0.00167 substitutions per nucleotide per year), and an adjustment of 3 nt per codon.

We used the sequence alignments produced by both simulation methods to evaluate different recombination detection methods, including the DSBM community detection method. We ran the GARD method in HyPhy (version 2.3.11) (21) with the HKY85 nucleotide substitution model and no rate variation. Since this program required a message-passing interface (MPI) parallel computing environment, we ran this analysis with eight threads. In addition, we evaluated

both versions 4 and 5 of the program RDP (22, 57) using the default settings and with the sequence type set to "linear."

To measure the computing times of these different methods, we generated test sets by randomly sampling sequences with replacement from the larger set of 37 reference genomes and added random mutations at 0.1% of positions in each genome on average. Through this process, we generated 10 replicates of 50, 100, and 200 sequences for a total of 30 test sets and then ran each set through DSBM, GARD, RDP4, and RDP5. Since the RDP programs are released only as Windows binary executables, and GARD requires an MPI-enabled environment, it was not feasible to run all tests in the same computing environment. Hence, we focused on characterizing the time complexity of each method (relative change in time with increasing data).

All source code for the DSBM method has been released under a permissive free license at https://github.com/Abayomi-Olabode/dsbm.

Results

Our hypothesis is that community detection with a DSBM (31) can be adapted for the unsupervised detection of recombination events from variation among HIV-1 genomes. We draw an analogy between latent communities in a social network and genetic clusters, such as the clusters that are presently labeled as the HIV-1 subtypes. Hence, a recombination event is analogous to moving from one network community to another. To briefly summarize our approach, we first generate a series of networks by partitioning a multiple alignment of HIV-1 genomes into "sliding windows" of a fixed width and step size and then compute a genetic distance (TN93) between all pairs of subsequences within each window. A network (undirected graph) can be derived from each distance matrix by applying a distance threshold, below which an edge is drawn between the respective nodes (sequences). Finally, we applied the implementation of DSBMs in R (42) to the resulting series of graphs to reconstruct community memberships for all nodes. For brevity, we refer to this method as DSBM.

Simulation Analysis. To evaluate the accuracy of this approach, we first used a simple method to generate alignments of recombinant sequences by combining fragments from one or more HIV-1/M subtype reference genomes at predetermined positions (breakpoints). We also used this simulation experiment to calibrate the threshold for converting TN93 distances to graphs. DSBM performed best at a percentile threshold of 40% with an average accuracy of 88.6%; in comparison, average accuracies were 86.25 and 78.6% at percentile thresholds of 30 and 20%, respectively (*SI Appendix*, Fig. S1). The method was the least accurate for recombination between subtypes B and D, which is consistent with their atypically high similarity (58). Note that the 40% threshold is not a fixed TN93 distance (i.e., 0.04), but rather the distance associated with the lower 40% of the empirical distribution of all TN93 distances for that window. Consequently, every window can have a different threshold distance. We used the percentile as a means of accommodating variation in rates of evolution among different regions of the genome.

Next, we generated more realistic simulations by grafting recombination events into a time-scaled phylogeny reconstructed from $n = 37$ HIV-1 genomes with prior information (1). Breakpoint locations were drawn at random from a uniform distribution over the alignment length. Starting from the left of the alignment, we pruned and regrafted subtrees by selecting two extant branches at random at given time points to produce a new tree for sequence fragments past the next breakpoint. We tested the ability of DSBM to accurately capture the recombination breakpoints in these simulations (Fig. 1*A*). Overall, DSBM was significantly more accurate (measured by error percentage) than the three other methods on these data (quasibinomial generalized linear model (GLM), $t > 5.4$, $P < 7.5 \times 10^{-8}$). Error rates increased significantly with the number of actual recombination breakpoints in the data ($t = 15.6$, $P < 10^{-12}$).

In addition, we evaluated the sensitivity of DSBM to varying the sizes of windows (250, 500, 750, and 1,000 nt) and steps (50, 100, 200, and 300 nt), which determines the number and stability of the graphs generated from the genome alignment. Error percentage was not significantly associated with window size (quasibinomial GLM, $t = -0.33$, $P = 0.74$) or step size ($t = 0.75$, $P = 0.46$), although we noted a slight tendency for error to increase with window size for simulations with one breakpoint and to decrease with window size for three breakpoints (*SI Appendix*, Fig. S2). Furthermore, we expect increasing

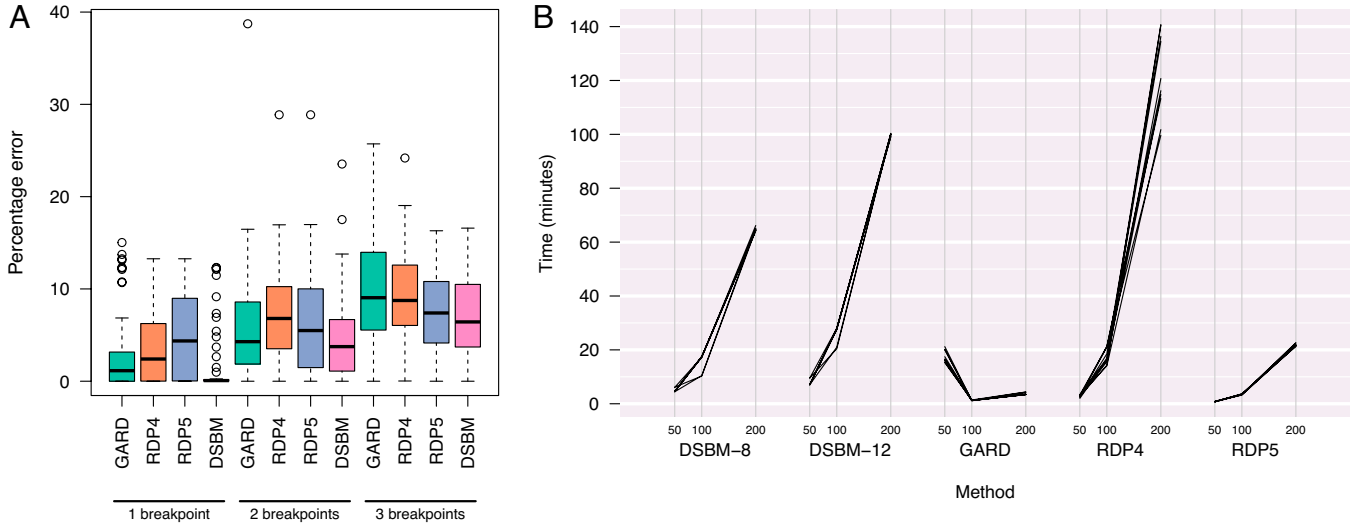

Fig. 1. (*A*) Box-and-whisker plots summarizing the percentage of error of recombination detection methods on sequences simulated from a time-scaled HIV-1/M phylogeny. Percentage of error was calculated by mapping outputs from each program to the same outcome space of cluster assignments by sequence and nucleotide position. (*B*) Slope graphs summarizing the computing times of different recombination detection programs on alignments of 50, 100, and 200 sequences, derived from $n = 37$ HIV-1 subtype reference sequences with random mutations. We evaluated 10 replicate simulations per sample size. DSBM-8 and -12 correspond to dynamic stochastic block model analyses with 8 and 12 clusters, respectively. Note that DSBM computing times do not include the fast calculation of TN93 distances. GARD was run in an MPI environment with eight cores.

window size to constrain our ability to resolve breakpoints that are separated by shorter segments or breakpoints too close to a sequence terminus. Pooling simulation replicates for all numbers of breakpoints, we confirmed that the error rate increased significantly with decreasing length of the shortest segment ($t = -9.46$, $P < 2 \times 10^{-16}$) with no significant effect of window size over the evaluated range (250 to 1,000 nt).

Finally, we measured the runtimes for processing varying numbers of sequences derived from HIV-1 subtype reference genomes with DSBM, GARD, RDP4, and RDP5. Since the purpose of this experiment was to measure computing time, we arbitrarily set the number of DSBM clusters to 8 and 12 rather than optimize the number of clusters. Computing times for DSBM were comparable to those for RDP4, whereas those for GARD and RDP5 were substantially faster (Fig. 1*B*); however, it is difficult to compare the latter two since GARD requires a parallel computing (MPI) environment. Overall, these results indicate that the expected computing times tend to increase faster than linearly with the number of sequences and increase with the number of clusters for DSBM. We also noted an unusual trend in GARD computing times, with faster times obtained for sample sizes above 50 sequences. On examining the output files, we determined that GARD efficiently rejected all models with recombination breakpoints for the larger datasets. In contrast, GARD runs on datasets with 50 sequences were unable to reject the presence of recombination and were burdened with 677.8 candidate models (partitions of the alignment among trees) on average.

Application to HIV-1/M Genomes. Having validated the DSBM method on simulated data in comparison to other unsupervised recombination detection methods, we next applied our method to $n = 525$ full-length HIV-1/M genome sequences to characterize the role of recombination in the evolutionary history of this virus (Fig. 2). These sequences were selected from all available full-length HIV-1 genomes to maximize the representation of global diversity in a reduced dataset. The ICL criterion, used to determine the optimal number of clusters for DSBM, was maximized at 29 clusters (*SI Appendix*, Fig. S4), based on a TN93 distance percentile threshold of 30%. Although our simulation experiments slightly favored 40% over 30%, the memory requirements of the *dynsbm* program for processing graphs generated from this dataset under a 40% threshold exceeded the limits of our compute server.

The results of our analysis are summarized in Fig. 2. Fig. 2*A* displays a heatmap summarizing the transition rate matrix from the DSBM analysis. Some clusters have substantial transition rates (darker shades) between them, such as clusters 3 and 14, 11 and 15, or 21 and 27. This implies a hierarchical structure to clusters akin to the subsubtypes of HIV-1. The frequencies of cluster memberships were consistent across the length of the HIV-1 genome (*SI Appendix*, Fig. S5).

To better understand the associations between these clusters, we mapped the original HIV-1 subtype annotations of the genome sequences to these data. We found that windows from genomes labeled as subtype A1 ($n = 10$) were predominantly assigned to clusters 2, 6, 7, and 13; subtype B ($n = 20$) was associated with clusters 5, 10, and 17; and subtype C ($n = 17$) to clusters 1, 12, 21, and 27 (*SI Appendix*, Fig. S6). Within these major subtypes, cluster assignments tended to become more variable past genome window 50 (*SI Appendix*, Fig. S7), in association with the start of the *env* gene. In addition, subtype D was associated with clusters 11, 15, and 20, while G was associated with clusters 22 and 29. Circulating recombinant form 01AE was strongly associated with cluster 18. Finally, genomes annotated as "complex" intersubtype recombinants did not exhibit a strong association with any cluster.

Fig. 2*B* displays the cluster memberships for all 525 genomes in 82 windows of 500 nt in steps of 100 nt. We mapped cluster memberships to a color gradient such that clusters with similar colors tend to have relatively higher transition rates between them. Under the standard concept of pure HIV-1 subtypes, genomes should tend to maintain the same coloration across their length. In other words, a row representing a non-recombinant genome will maintain a single color across all cells. Indeed, we see an overall stratification of colors in this color map (Fig. 2*B*). We used white to represent windows that the algorithm failed to assigned to any cluster—an outcome that implies the subsequence in that region is too divergent from the other genomes in the dataset, such that its TN93 distance exceeds the threshold in all cases. A total of 489 (1.1%) windows were unassigned; these missing values were distributed across 76 (14.5%) of the genomes.

If we plot the cluster assignments for individual genomes (Fig. 2*C*), we see traces that are consistent with recombination, for instance in Fig. 2 *C2*, *C3*, and *C14*. Other traces are too noisy to be interpreted visually, i.e., Fig. 2 *C10–C13*. To automate the extraction of recombination breakpoints from traces in the presence of random error, we applied two different change-point detection algorithms to the entire set of traces. Seventeen genomes were excluded due to excessive numbers of unassigned windows. The AMOC method classified 496 (97.6%) genomes as recombinant and 12 as nonrecombinant at minimum segment lengths of three and five windows and 494 as recombinant at a minimum of 10 windows.

Next, we used the less restrictive PELT method to estimate the number of breakpoints per genome under varying minimum segment lengths (Fig. 3, *Left*). As expected, the mean number of breakpoints per genome increased with shorter minimum segment lengths. At $m = 5$ windows, for instance, we inferred a mean of 8.3 (interquartile range [IQR] $= 7$ to 10) breakpoints. On the other hand, the distribution of breakpoints across windows was less sensitive to varying the minimum segment length, with significant positive correlations between distributions (Spearman's $\rho > 0.51$, $P < 9.4 \times 10^{-10}$; Fig. 3, *Right*). Local peaks in the frequencies of breakpoints were robust to varying minimum segment length. Consistent with previous work (59), we observed recombination "hotspots" associated with the 5′ and 3′ ends of the *env* gene. We also observed a distinct and robust peak associated with the 3′ end of the *pol* gene and a sharp increase in the frequency of breakpoints within or upstream of *gag*.

Genomes that were classified as recombinant by SCUEAL had significantly higher numbers of breakpoints assigned by DSBM, and this concordance was robust to varying minimum segment lengths (Wilcoxon rank-sum test, $P < 7.8 \times 10^{-13}$; *SI Appendix*, Fig. S9). Finally, we found no significant association between the number of predicted breakpoints and the year of sample collection, which ranged from 1983 to 2016 with a median of 2006, for any of the three minimum segment lengths (Poisson regression, $P < 0.24$; *SI Appendix*, Fig. S8).

Revisiting the HIV-1 Subtype References. In addition to analyzing a large selection of HIV-1/M genome sequences, we applied the DSBM method to evaluate the HIV-1 subtype reference genomes curated by the LANL HIV Sequence Database. These reference genomes have long been used as the "gold standard" against which other genome sequences are evaluated for evidence of recombination. We assessed a subset of the reference genomes covering HIV-1/M subtypes A-D, F-H, J, and K and excluded reference genomes corresponding to the CRFs. For this analysis, the ICL criterion selected $k = 6$ as the optimal number of clusters.

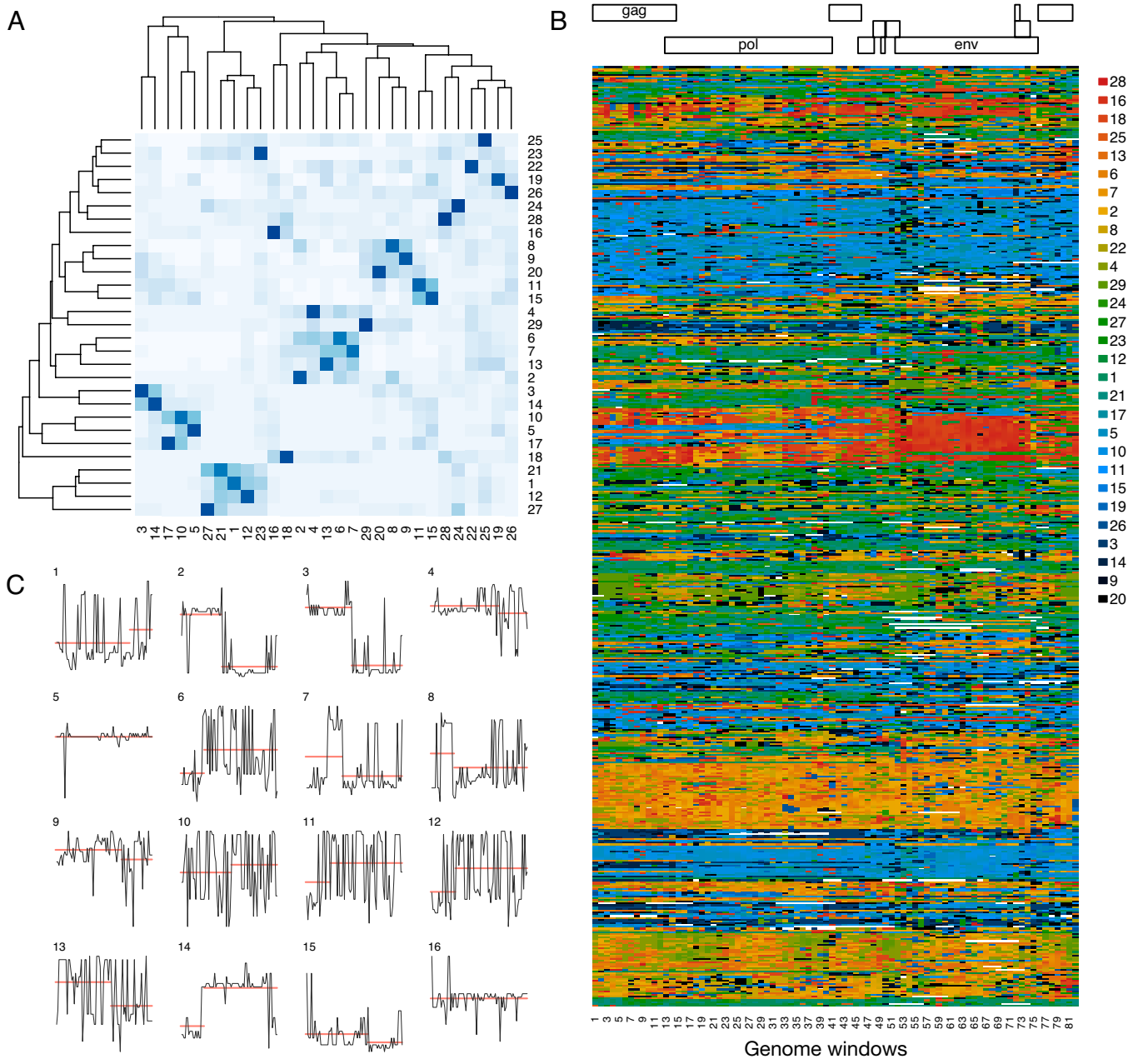

Fig. 2. Summary of results from DSBM analysis of $n = 525$ HIV-1/M genomes. (*A*) A heatmap depicting the matrix of transition rates among 29 clusters. A transition occurs when an individual switches cluster memberships from one genomic window to the next. Cells are shaded in proportion to rates, and rows and columns of the matrix were reordered by hierarchical clustering (represented by dendrograms above and to the left of the heatmap) to bring together clusters linked by higher transition rates. (*B*) A color map displaying the cluster assignments for $n = 525$ genomes, partitioned into $n = 82$ windows of 500 nt in steps of 100 nt. Clusters linked by higher transition rates were mapped to similar colors along a gradient. A white cell indicates that DSBM failed to assign the corresponding window to a cluster. A color-blind accessible version is provided in *SI Appendix*, Fig. S3. (*C*) Step charts depicting the cluster assignments for an arbitrary sample of 16 genomes across 82 windows, where we used the hierarchical clustering permutation order from *A* to minimize the vertical distance between clusters related by higher transition rates. Line segments (red) indicate the location of change points, as determined by a conservative AMOC method with a minimum segment length of 20.

This is clearly fewer than the recognized number of HIV-1 subtypes. We observed that subtypes B and D were assigned to the same cluster. Furthermore, genomic windows from subtypes F and K tended to be assigned to the same cluster and subtypes G and J to a third cluster. These cluster assignments are consistent with the placement of the respective subtypes in a phylogeny of HIV-1/M subtypes and CRFs (60).

Employing the same postprocessing method as in our previous analysis, we detected evidence of recombination breakpoints in 18 (48.6%) of 37 reference genomes when the minimum segment length was set to 5. The PELT method of change-point detection identified one, two, and four breakpoints in 2, 13, and 3 of the remaining genomes, respectively (Fig. 4). These breakpoints tended to be concentrated in genomes classified by SCUEAL into subtypes F (subsubtypes F1 and F2), J, and K. We also observed consistent patterns of recombination in the subtype G genomes spanning windows 42 to 51 and 67 to 78. These putative breakpoints, which are consistent with previous work (18), were not picked up by the PELT method because the affected clusters were annotated as being similar, even though we did not observe substantial clustering in transition rates between the six clusters in this analysis. These results highlight an important limitation

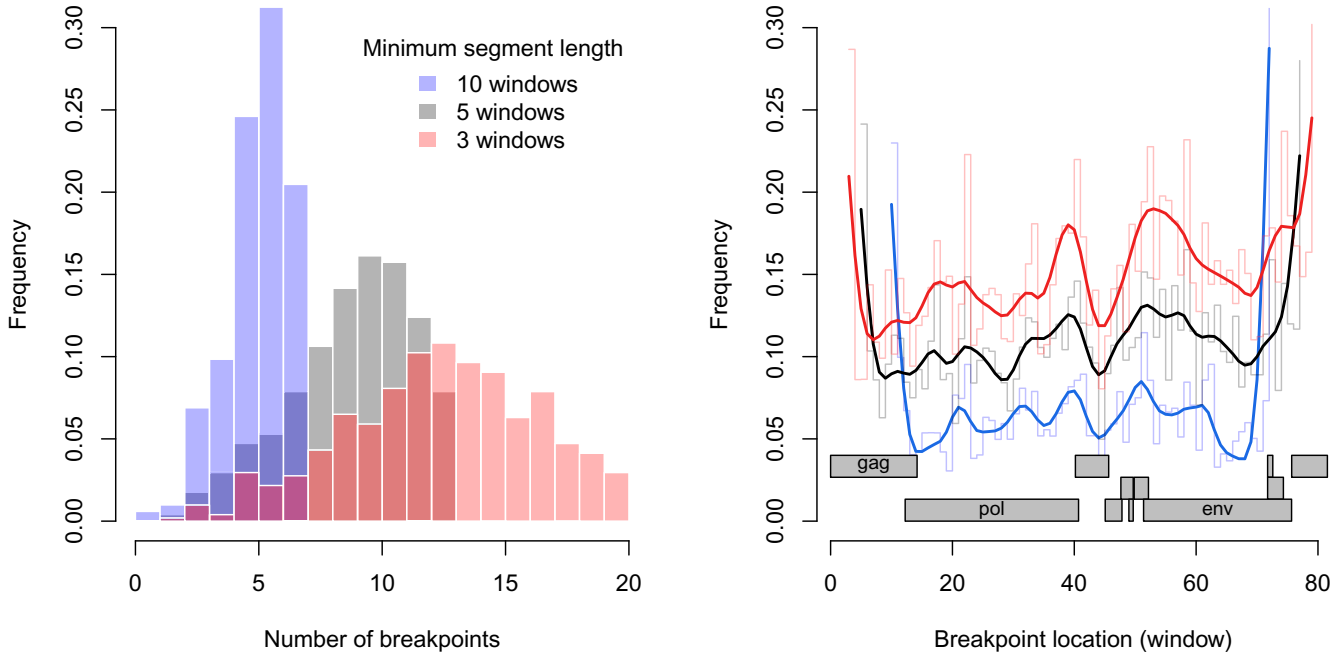

Fig. 3. Distribution of DSBM predicted recombination breakpoints in HIV-1/M genomes. (*Left*) Histogram of the number of breakpoints per genome, where breakpoints were extracted from DSBM cluster assignments by the PELT change detection method. We varied the minimum segment length to 3, 5, and 10 windows (see color key). (*Right*) Smoothed splines summarizing the distribution of predicted breakpoints among sliding windows of the HIV-1 genome alignment. The raw frequency distributions are drawn as step charts in lighter colors. A diagram of the HIV-1 reading frames, mapped from the HXB2 reference coordinates to our alignment, is displayed at the base.

of change-point detection based on changes in the mean. For comparison, we ran GARD and RDP on this same alignment. GARD detected two breakpoints associated with windows 32 and 43, respectively, but it was not possible to associate these breakpoints with specific sequences. Neither RDP4 nor RDP5 predicted any significant breakpoints from these data.

Discussion

Recombination is a major contributing factor to the extensive genetic variability observed in RNA viruses (61). Since many comparative methods assume that sequences are related through a single, nonrecombinant evolutionary history, screening for recombination is an important step for the analysis of these viruses (21). Furthermore, ongoing advancements in next-generation sequencing are driving the proliferation of large HIV-1/M genome databases (62, 63), providing more opportunities to reconstruct the role of recombination in the evolutionary history of this virus. Here, we have adapted a community detection method from network science (DSBMs) (23, 31) to detect the residual evidence of past recombination events without requiring a reference set of nonrecombinant genomes. Hence, DSBM is an unsupervised method for detecting recombination.

Comparison to Other Unsupervised Methods. Compared to the commonly used unsupervised methods for recombination detection [GARD, RDP4 (57), and RDP5 (22)], DSBM is more sensitive to detecting recombination in our simulation experiments. In addition, DSBM offers unique advantages by striking a balance between the level of detail offered by RDP and the statistical power of GARD. RDP employs a number of efficient, nonparametric heuristics for detecting recombination by the direct comparison of extant sequences. For example, the BOOTSCAN (64) component of RDP calculates genetic distances in sliding windows between every pair of sequences from the input alignment. DSBM can be viewed as a generalization of RDP, in that it uses the same distance-based approach to directly compare all sequences in the alignment. It uses these distances to generate a series of undirected graphs, to which it fits a hidden Markov model in which the probability of an edge connecting two individuals is determined by their unobservable (latent) and dynamic cluster memberships. Consequently, the placement of a breakpoint in a sequence by DSBM is informed by the entire dataset, which confers greater power than pairwise comparisons at the cost of increased computational complexity.

Similarly, GARD analyzes the joint distribution of all sequences, using a genetic algorithm to explore the model space defined by the placement of hypothetical breakpoints in the alignment. The likelihood of a given model is calculated by fitting maximum-likelihood phylogenies to every interval between breakpoints (21). Hence, GARD is the most biologically realistic of the unsupervised methods, since it is explicitly modeling the impact of recombination on evolutionary histories. The minimum length of an interval must increase with the number of sequences, since a requisite number of phylogenetically informative substitutions are needed to reconstruct larger trees. This requirement sets an upper limit to the number of breakpoints for a given length of alignment. Since GARD operates on phylogenies, it is less sensitive to more recent recombination events that affect only a small number of sequences. It does not readily scale with increasing amounts of recombination because every breakpoint requires fitting an additional tree to fewer data. In contrast, adding breakpoints does not increase the number of model parameters in either RDP or DSBM. Additionally, GARD does not report which sequences are affected by each recombination event. Like RDP, DSBM infers recombination events at the level of individual sequences because it makes direct comparisons between extant sequences. Therefore, we characterize DSBM as a compromise between the efficient heuristics in RDP and the model-based approach of GARD.

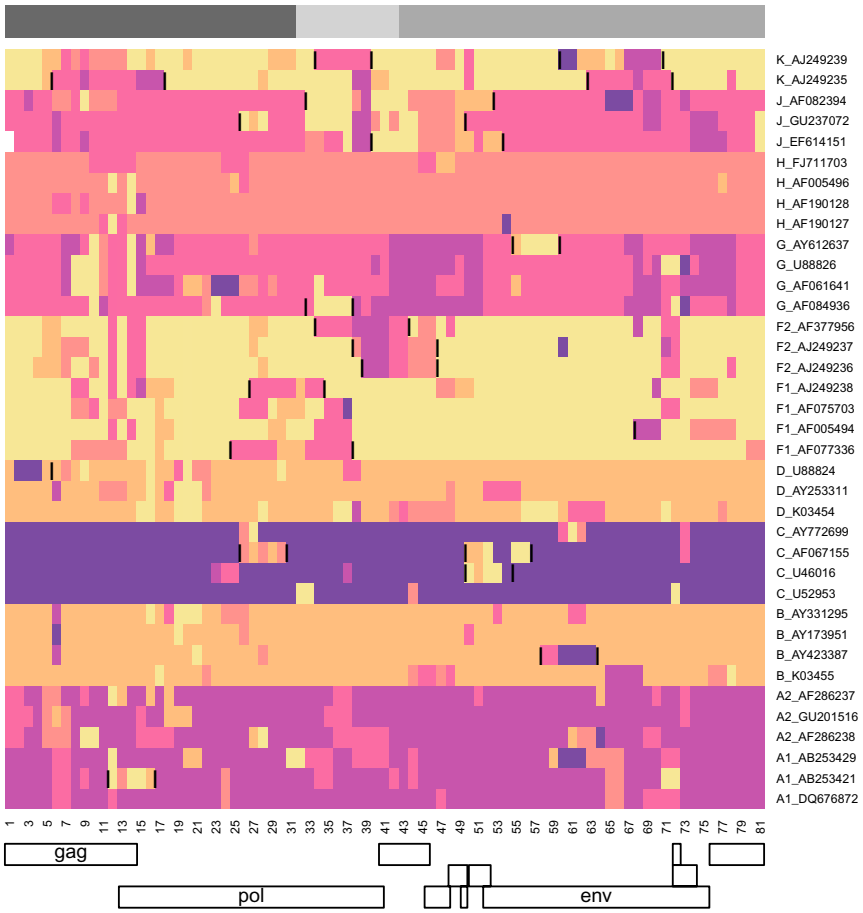

Fig. 4. Detection of recombination in curated HIV-1 subtype reference genomes by community detection with a DSBM. Predicted breakpoints (using the PELT change-point detection method with a minimum segment length of five windows) are marked directly on the heatmap with vertical line segments. We used a different color palette to emphasize the different number of clusters between this analysis ($k = 6$) and our previous analysis of $n = 525$ HIV-1/M genomes ($k = 25$). The bar at the top represents the location of the breakpoints predicted by GARD (the colors we used solely to depict the location and sizes of the recombinant fragments).

Limitations. DSBM is time consuming to compute and slower than the other methods, in part due to the quadratic time complexity of parameterizing the matrix of transition rates between clusters (communities). Like GARD, the computing time can be ameliorated by running DSBM in a parallel computing environment. Our extension of DSBM involves a number of tuning parameters—namely, a genetic distance threshold, the window size, and the window step size. Our simulation experiments indicated that results from DSBM are not sensitive to varying the window and step size parameters. However, choosing an appropriate distance threshold, which determines how sequence variation maps to network topologies, was critical to our analysis (*SI Appendix*, Fig. S1). Furthermore, we used change-point detection to map cluster assignments from DSBM to recombination breakpoints, which introduced additional tuning parameters such as the minimum segment length. We evaluated the effect of varying minimum segment lengths on the distribution of recombination breakpoints, for instance. Although the overall number of breakpoints varied with this parameter, the density of breakpoints was qualitatively consistent (Fig. 3). Even so, it will be important to report all parameter settings when applying DSBM to other data, as well as evaluating the robustness of results to varying settings. Finally, another limitation of DSBM is that it assumes the rates of transitions between clusters are "time" homogeneous, i.e., that recombination rates between specific lineages are constant through the genome. Since DSBMs are a recent innovation in network science, it may eventually become possible to fit models with transition rates that vary over the genome length; however, the current model already estimates a large number of parameters from the data.

It should be possible to generalize the application of DSBM from HIV-1 genomes to those other viruses. Like HIV-1 and other retroviruses (65), many other viruses are characterized by frequent recombination, substantial genetic variation (66), and rate heterogeneity among sites (67). For instance, many viruses in the family Picornaviridae have an abundance of genetic diversity and frequent recombination over short and long evolutionary time scales (68). Although we propose DSBM as a general-purpose method for detecting recombination in virus genomes, it would likely be necessary to recalibrate the cutoff parameters for generating networks to apply this method to other viruses. Some aspects may be more robust to changing evolutionary contexts. For instance, we used the empirical distribution of TN93 distances to normalize cutoffs across sliding windows. By not retuning the cutoffs for such viruses, normalization would rescale the TN93 distances but the underlying distribution within that range may be different. The end result will be a dataset skewed toward having a few but very large clusters or pockets of small clusters with few memberships and many "singletons," i.e., isolated nodes.

Recombination and HIV-1. An interesting outcome of our analysis is the putative recombination hotspots in association with the *gag* and *nef* gene sequences, which are adjacent to or overlap

the 5′ and 3′ long terminal repeats (LTR), respectively (Fig. 3). Based on the composition of the CRF genomes documented by the LANL HIV Sequence Database, breakpoints are often found in these regions. For example, CRF01_AE contains a breakpoint at the 5′ end of the 3′ LTR sequence (HXB2 nucleotide coordinate 9,086), and CRF02_AG contains one at the 3′ end of the 5′ LTR (coordinate 789). Overall, we observed that 57.1% ($n = 77$) and 45.6% ($n = 92$) of the documented CRFs contain breakpoints upstream of HXB2 coordinates 1,496 and 8,757, respectively (adjusting for CRFs without sequence coverage in these regions). Furthermore, recombination in these regions of the HIV-1 genome has been observed as non-CRFs (69, 70). In a comparative analysis of HIV-1 *gag*, Minin et al. (71) detected a recombination hotspot in association with an instability element in the region encoding the capsid protein. However, this hotspot was not reproduced in subsequent work by Archer et al. (59), who reported hotspots associated with both ends of the *env* gene. Resolving these differences will require the reconciliation of methods and datasets or, more effectively, experimental validation with an in vitro system (72).

Our analysis revealed that there was no significant difference in the number of detected recombination breakpoints over time, when older and contemporary sequences were compared. This result is consistent with frequent and ongoing recombination throughout the evolutionary history of HIV-1/M. Rather than continuing to accumulate novel recombination breakpoints over time, the lack of a significant trend suggests that the number of breakpoints has already approached an upper limit by the time of our earliest samples. This would be consistent with the existence of recombination hotspots, such that the probability of breakpoints is not uniformly distributed across the genome. However, we should be careful not to overinterpret this result; for instance, reconstructing recombination events by the direct comparison of sequences may be increasingly uncertain for less frequent early genomes where the "parental" genomes are less likely to have been sampled.

Previous studies have proposed reclassifying some of the defined HIV-1 subtypes and circulating recombinant forms. In some cases, this was motivated by the availability of whole genome sequencing data (73, 74). For example, a recent analysis of the major HIV-1 genes *gag*, *pol*, and *env* culminated in a proposal to further partition of subtypes A and D into subsubtypes and to merge subtypes B and D into a single subtype (58). A nonrecombinant HIV-1/M subtype L was proposed ∼30 y after the genomes were first sampled in 1983 and 1990 in what is now the Democratic Republic of the Congo (75). Furthermore, there has been some debate surrounding the status of HIV-1/M subtype G. For instance, Abecasis et al. (18) and Lemey et al. (76) have both found evidence suggesting that subtype G is a recombinant offspring of three distinct parental genomes that are also categorized as pure subtypes or circulating recombinant forms, including subtypes A, J, and CRF02_AG. Other groups (77) have countered with evidence that CRF02_AG is an offspring of subtypes A and G. Nevertheless, the current HIV-1 nomenclature system classifies subtype G as a pure subtype. The most recent set of subtype reference genomes curated by the LANL HIV Sequence Database continues to include subtype G, for example.

Our analysis is consistent with the hypothesis that many of the HIV-1/M subtype reference genomes are actually recombinant (17, 18). These findings support the idea that the current HIV-1 nomenclature (7), which has served as an important framework for our understanding of HIV-1 diversity and evolution, should be revisited in light of recent genomic evidence. For example, our DSBM analysis of a large sample of HIV-1/M genomes identified $k = 29$ as the optimal number of clusters. In this context, a cluster is roughly analogous to a nonrecombinant subtype. However, what we recognize as a subtype may also comprise multiple clusters (*SI Appendix*, Fig. S7). A key challenge to developing a revision proposal is that nonrecombinant genomes seem to be rare under our current working definition. None of the $n = 525$ in our largest analysis had windows uniformly assigned to a single cluster along its entire length. Given that this dataset was designed to summarize the global diversity of HIV-1, we are pessimistic that expanding the scope of our analysis will yield a substantial number of genomes that are each representative of a single cluster. However, it may be feasible to utilize cluster assignments from a DSBM analysis to reconstruct consensus genomes as a mosaic of observed genomes, which could play the same role as a full-length representative genome.

Data Availability. Previously published data were used for this work (https://github.com/Abayomi-Olabode/dsbm).

ACKNOWLEDGMENTS. This work was supported in part by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-131), by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant 05516-2018 RGPIN), and by the Canadian Institutes of Health Research (Grants PJT-155990, PJT-156178, FRN-130609, and BOP-149562).

Author affiliations: [a]Department of Pathology & Laboratory Medicine, Western University, London, ON, N6A 5C1 Canada; [b]Department of Computer Science, Western University, London, ON, N6A 5B7 Canada; [c]Department of Microbiology & Immunology, Western University, London, ON, N6A 3K7 Canada; [d]Institute of Evolutionary Biology, University of Edinburgh, EH9 3JT Edinburgh, United Kingdom; and [e]Department of Applied Mathematics, Western University, London, ON, N6A 5B7 Canada

1. N. R. Faria *et al.*, The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61 (2014).
2. F. Gao *et al.*, Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**, 436–441 (1999).
3. B. F. Keele *et al.*, Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**, 523–526 (2006).
4. E. Bailes *et al.*, Hybrid origin of SIV in chimpanzees. *Science* **300**, 1713–1713 (2003).
5. M. J. Jin *et al.*, Mosaic genome structure of simian immunodeficiency virus from West African green monkeys. *EMBO J.* **13**, 2935–2947 (1994).
6. S. M. Bell, T. Bedford, Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission. *PLoS Pathog.* **13**, e1006466 (2017).
7. D. L. Robertson *et al.*, HIV-1 nomenclature proposal. *Science* **288**, 55–56 (2000).
8. W. H. Li, M. Tanimura, P. M. Sharp, Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5**, 313–330 (1988).
9. D. Cromer, A. J. Grimm, T. E. Schlub, J. Mak, M. P. Davenport, Estimating the in-vivo HIV template switching and recombination rate. *AIDS* **30**, 185–192 (2016).
10. R. A. Neher, T. Leitner, Recombination rate and selection strength in HIV intra-patient evolution. *PLOS Comput. Biol.* **6**, e1000660 (2010).
11. A. D. Redd, T. C. Quinn, A. A. Tobian, Frequency and implications of HIV superinfection. *Lancet Infect. Dis.* **13**, 622–628 (2013).
12. M. A. Rodgers *et al.*, ARCHITECT HIV combo Ag/Ab and realtime HIV-1 assays detect diverse HIV strains in clinical specimens. *AIDS Res. Hum. Retroviruses* **34**, 314–318 (2018).
13. K. Angelis *et al.*, Global dispersal pattern of HIV type 1 subtype CRF01_AE: A genetic trace of human mobility related to heterosexual sexual activities centralized in Southeast Asia. *J. Infect. Dis.* **211**, 1735–1744 (2015).
14. M. Zhang *et al.*, The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* **7**, 1–15 (2010).
15. H. E. Grant *et al.*, Pervasive and non-random recombination in near full-length HIV genomes from Uganda. *Virus Evol.* **6**, veaa004 (2020).
16. N. R. Faria *et al.*, Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. *PLoS Pathog.* **15**, e1007976 (2019).
17. A. S. Olabode *et al.*, Evidence for a recombinant origin of HIV-1 Group M from genomic variation. *Virus Evol.* **5**, vey039 (2019).
18. A. B. Abecasis *et al.*, Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype G is a circulating recombinant form. *J. Virol.* **81**, 8543–8551 (2007).
19. D. Struck, G. Lawyer, A. M. Ternes, J. C. Schmit, D. P. Bercoff, COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res.* **42**, e144 (2014).
20. S. L. Kosakovsky Pond *et al.*, An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLOS Comput. Biol.* **5**, e1000581 (2009).

21. S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, S. D. Frost, GARD: A genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).
22. D. P. Martin *et al*., Rdp5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2021).
23. E. Abbe, Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18**, 6446–6531 (2017).
24. B. Yang, D. Liu, J. Liu, "Discovering Communities from Social Networks: Methodologies and Applications" in *Handbook of Social Network Technologies and Applications*, B. Furht, Ed. (Springer, 2010), pp. 331–346.
25. B. S. Khan, M. A. Niazi, Network community detection: A review and visual survey. *arXiv* [Preprint] (2017). https://arxiv.org/abs/1708.00977. Accessed 1 April 2021.
26. A. F. Poon, Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evol.* **2**, vew031 (2016).
27. T. Aynaud, J. L. Guillaume, "Static community detection algorithms for evolving networks" in *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (IEEE, New York, NY), pp. 513–519 (2010).
28. R. Cazabet, G. Rossetti, F. Amblard, "Dynamic community detection" in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj, J. Rokne, Eds. (Springer, New York, NY, 2017), pp. 404–414.
29. G. Chen, Y. Wang, J. Wei, A new multiobjective evolutionary algorithm for community detection in dynamic complex networks. *Math. Probl. Eng.* **2013**, pp. 1–7 (2013).
30. N. K. Ahmed, F. Berchmans, J. Neville, R. Kompella, "Time-based sampling of social network activity graphs" in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, U. Brefeld, L. Getoor, S. A. Macskassy, Eds. (Association for Computing Machinery, New York, NY, 2010), pp. 1–9.
31. C. Matias, V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. B* **79**, 1119–1141 (2017).
32. M. Qiao, J. Yu, W. Bian, Q. Li, D. Tao, Adapting stochastic block models to power-law degree distributions. *IEEE Trans. Cybern.* **49**, 626–637 (2019).
33. K. Xu, "Stochastic block transition models for dynamic networks" in *Artificial Intelligence and Statistics*, G. Lebanon S. V. N. Vishwanathan, Eds. (Proceedings of Machine Learning Research, 2015), pp. 1079–1087.
34. A. Löytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635 (2008).
35. A. Zielezinski *et al*., Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, 1–18 (2019).
36. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
37. E. Paradis, J. Claude, K. Strimmer, Ape: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
38. K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
39. G. Csardi *et al*., The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).
40. J. O. Wertheim *et al*., Social and genetic networks of HIV-1 transmission in New York City. *PLoS Pathog.* **13**, e1006000 (2017).
41. R. A. Russell, M. D. Moore, W. S. Hu, V. K. Pathak, APOBEC3G induces a hypermutation gradient: Purifying selection at multiple steps during HIV-1 replication results in levels of G-to-A mutations that are high in DNA, intermediate in cellular viral RNA, and low in virion RNA. *Retrovirology* **6**, 1–15 (2009).
42. V. Miele, C. Matias, Revealing the hidden structure of dynamic ecological networks. *R. Soc. Open Sci.* **4**, 170251 (2017).
43. C. Matias, S. Robin, Modeling heterogeneity in random graphs through latent space models: A selective review. *ESAIM Proc. Surv.* **47**, 55–74 (2014).
44. C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 719–725 (2000).
45. J. P. Baudry *et al*., Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electron. J. Stat.* **9**, 1041–1077 (2015).
46. M. Bertoletti, N. Friel, R. Rastelli, Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron* **73**, 177–199 (2015).
47. S. Aminikhanghahi, D. J. Cook, A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **51**, 339–367 (2017).
48. G. J. van den Burg, C. K. Williams, An evaluation of change point detection algorithms. *arXiv* [Preprint] (2020). https://arxiv.org/abs/2003.06222. Accessed 4 January 2021.
49. S. Sharma, D. A . Swayne, C. Obimbo, Trend analysis and change point techniques: A survey. *Energy Ecol. Environ.* **1**, 123–130 (2016).
50. D. V. Hinkley, Inference about the change-point in a sequence of random variables. *Biometrika* **57**, 1–17 (1970).
51. R. Killick, P. Fearnhead, I. A. Eckley, Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107**, 1590–1598 (2012).
52. R. Killick, I. Eckley, Changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **58**, 1–19 (2014).
53. N. R. Zhang, D. O. Siegmund, A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22–32 (2007).
54. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
55. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
56. W. Fletcher, Z. Yang, INDELible: A flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26**, 1879–1888 (2009).
57. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).
58. N. Désiré *et al*., Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. *Retrovirology* **15**, 1–7 (2018).
59. J. Archer *et al*., Identifying the important HIV-1 recombination breakpoints. *PLOS Comput. Biol.* **4**, e1000178 (2008).
60. D. M. Tebit, E. J. Arts, Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect. Dis.* **11**, 45–56 (2011).
61. M. Worobey, E. C. Holmes, Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80**, 2535–2543 (1999).
62. C. Kuiken, B. Korber, R. W. Shafer, HIV sequence databases. *AIDS Rev.* **5**, 52–61 (2003).
63. D. Pillay *et al*.; PANGEA-HIV Consortium, PANGEA-HIV: Phylogenetics for generalised epidemics in Africa. *Lancet Infect. Dis.* **15**, 259–261 (2015).
64. D. P. Martin, D. Posada, K. A. Crandall, C. Williamson, A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* **21**, 98–102 (2005).
65. A. Onafuwa-Nuga, A. Telesnitsky, The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol. Mol. Biol. Rev.* **73**, 451–480 (2009).
66. J. J. Bujarski, "Recombination of viruses" in *Encyclopedia of Virology*, A. Granoff, R. G. Webster, Eds. (Academic Press, New York, NY, 1999), pp. 1446–1454.
67. E. Domingo *et al*., The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: Biological relevance–A review. *Gene* **40**, 1–8 (1985).
68. A. N. Lukashev, Recombination among picornaviruses. *Rev. Med. Virol.* **20**, 327–337 (2010).
69. J. T. Blackard *et al*., Transmission of human immunodeficiency type 1 viruses with intersubtype recombinant long terminal repeat sequences. *Virology* **254**, 220–225 (1999).
70. U. Neogi, V. Sood, N. Goel, A. Wanchu, A. C. Banerjea, Novel HIV-1 long terminal repeat (LTR) sequences of subtype B and mosaic intersubtype B/C recombinants in North India. *Arch. Virol.* **153**, 1961–1966 (2008).
71. V. N. Minin, K. S. Dorman, F. Fang, M. A. Suchard, Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes. *Genetics* **175**, 1773–1785 (2007).
72. H. A. Baird *et al*., Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. *Nucleic Acids Res.* **34**, 5203–5216 (2006).
73. G. Yebra *et al*.; ICONIC Consortium, A high HIV-1 strain variability in London, UK, revealed by full-genome analysis: Results from the ICONIC project. *PLoS One* **13**, e0192081 (2018).
74. M. Rubio-Garrido *et al*., Current and historic HIV-1 molecular epidemiology in paediatric and adult population from Kinshasa in the Democratic Republic of Congo. *Sci. Rep.* **10**, 1–13 (2020).
75. J. Yamaguchi *et al*., Brief report: Complete genome sequence of CG-0018A-01 establishes HIV-1 subtype I. *J. Acquired Immune Deficiency Syndromes (1999)* **83**, 319 (2020).
76. P. Lemey, M. Lott, D. P. Martin, V. Moulton, Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. *BMC Bioinformatics* **10**, 1–18 (2009).
77. I. Bulla *et al*., HIV classification using the coalescent theory. *Bioinformatics* **26**, 1409–1415 (2010).