Dynamic Epidemiological Networks: A Data Representation for Modeling and Tracking of SARS-CoV-2 variants

Fiona Senchyna,[1] Rahul Singh[1,2*]

[1] Department of Computer Science, San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132, USA

[2]Center for Discovery and Innovation in Parasitic Diseases, University of California, San Diego

[*]To whom correspondence should be addressed; E-mail: rahul@sfsu.edu.

Keywords: Temporal Networks, Variant Tracking, SARS-CoV-2, Molecular Epidemiology

ABSTRACT

The large-scale real-time sequencing of SARS-CoV-2 genomes has allowed for rapid identification of concerning variants through phylogenetic analysis. However, the nature of phylogenetic reconstruction is typically static, in that the relationships between taxonomic units, once defined, are not subject to alterations. Furthermore, most phylogenetic methods are intrinsically batch-mode in nature, requiring the presence of the entire data set. Finally, the emphasis of phylogenetics is on relating taxonomical units. These characteristics complicate the application of classical phylogenetics methods to represent relationships in molecular data collected from rapidly evolving strains of an etiological agent, such as SARS-CoV-2, since the molecular landscape is updated continuously as samples are collected. In such settings, variant definitions are subject to epistemological constraints and may change as data accumulates.

Furthermore, representing within-variant molecular relationships may be as important as representing between variant relationships. This paper describes a novel data representation framework called dynamic epidemiological networks along with algorithms that underpin its construction to address these issues. The proposed representation is applied to study the molecular development underlying the spread of the COVID-19 pandemic in two countries: Israel and Portugal spanning a two-year period from February 2020 to April 2022. The results demonstrate how this framework could be used to provide a multiscale representation of the data by capturing molecular relationships between samples as well as those between variants, automatically identifying the emergence of high frequency variants (lineages), including variants of concern such as Alpha and Delta, and tracking their growth. Additionally, we show how analyzing the evolution of the dynamic epidemiological network can help identify changes in the viral population that could not be readily inferred from phylogenetic analysis.

1. INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiological agent of Coronavirus disease 2019 (Covid-19), has had a significant impact on the state-of-the-art in molecular epidemiology. From its first recorded outbreak in December 2019, the viral genome of SARS-CoV-2 has been sequenced and made publicly available (Wu *et al*., 2020). As of September 2022, almost 13 million genomes have been sequenced globally (Khare *et al*., 2021), providing a detailed view of the genetic variation accrued by the virus over time. Tools such as Pango (Rambaut *et al*., 2020) and Nextstrain (Hadfield *et al*., 2018) have utilized this data to reconstruct the evolution of the virus through phylogenetic analysis and has provided a nomenclature for emerging, genetically distinct iterations of the virus, also known as "lineages", "clades" or "variants". Phylogenetics has also contributed to the World Health Organization's (WHO) working definitions for variants of interest (VOIs) and concern (VOCs) (Konings *et al*., 2021). However, a classical phylogenetics-based approach may not always be sufficient for specifying lineages or variants of epidemiological importance, particularly for a rapidly evolving disease where information about variants is collected (and stabilizes) over time. For instance, evidence of onward transmission, persistence in a host population, and phenotypic changes are all examples of factors used to define variants by the aforementioned groups, that have been subject to change as the pandemic has progressed (Konings *et al*., 2021; Roemer *et al*., 2022).

A Network is a powerful representation to model evolving phenomena. Among others, networks have been used to model transmissions in viruses, such as the human immunodeficiency virus (HIV) and the Hepatitis C virus (HCV) (Longmire *et al*., 2017; Poon *et al*., 2016). In a network representation, vertices represent viral intra-host sequences and edges connect genetically similar samples. Thus, possible transmission events can be represented in

terms of the network connectivity (Little *et al*., 2014; Longmire *et al*., 2017). Building on this idea, this paper describes a novel data representation called dynamic epidemiological networks (DEN) for modeling and molecular tracking of evolving disease variants along with the algorithms for constructing such a representation. A DEN is defined in a data-driven manner from the molecular constitution of the pathogen samples collected over time in a population. Consequently, the DEN is explicitly parameterized by time. At a particular time-point, in a DEN, vertices represent sampled genomes and edges are defined between samples that are deemed to be genetically close. Furthermore, vertices are grouped into clusters (communities), with each cluster corresponding to a variant. As one proceeds from a particular time point to the next, the DEN represents correspondences between clusters occurring at the successive time points as determined by their genomic contents. If one or more samples at a particular time point are found to significantly differ from existing samples in the DEN, then these new samples are represented by a cluster of vertices that have no prior correspondences. Thus, a DEN can be used for: (1) identification of epidemiologically relevant variants through analysis of the clusters in the network, (2) tracking the evolution of variants, including the emergence of new variants and cessation of old ones by considering the changes in the number and constitution of clusters across time, and (3) obtaining insights about the evolving viral landscape by analyzing the cluster correspondences. Furthermore, by describing within- and between-cluster relationships, DENs support a multiscale data representation. This paper describes the algorithmic underpinnings of DENs and applies this formalism to analyze the spread of COVID-19 in Israel and Portugal based on over 15,000 sequenced viral samples collected from each of the respective populations during February 2020-April 2022. This data spans the initial wave of the pandemic along with the subsequent Alpha and Delta waves. We show how the DEN framework identified

the variants of concern Alpha, Beta, and Delta, along with other region specific, high frequency lineages in these two countries. Furthermore, we compare the results from representing and analyzing the aforementioned data using DEN with those obtained using phylogenetic analysis. It can be seen that unlike in phylogenetic representations, samples have the freedom to "move" between clusters over time in a DEN and thus the variant assignment of a genome is responsive to changes in the overall viral population. Further, through tracking changes in the clusters of the DEN, emergent variants of importance can be found. For instance, the Omicron variant could be noted even when only one Omicron sample had been added to the network, due to its connection to the Alpha cluster.

2. RELATED WORK

The proposed method can be contrasted with other work in the area from two primary perspectives: data representation, and variant identification. From the data representation perspective, technically, the DEN differs from prior work in the area in two crucial ways. The first of these relates to the formulation for determining of network connectivity. Prior techniques for reconstructing HIV and HCV contact networks have employed (static) genetic distance-based thresholds to connect hosts, for example, as in (Campo *et al*., 2016; Wertheim *et al*., 2017). Such an approach is however, unsuitable for variant detection due to differences in the genetic diversity of different variant populations (Weng *et al*., 2022). For example, a distance threshold that successfully clusters one genetically diverse variant population may also place two distinct homogenous variants into one cluster. To avoid this, we create a connected network using relative similarity obtained via *k*-nearest neighbor (*k*-NN) computations. Nearest neighbor-based grouping allows samples to be connected based on genetic similarity relative to all other samples in the dataset rather than the connectivity being defined by a static threshold. Second, networks

representations for studying viral populations (Little *et al*., 2014; Longmire *et al*., 2017; Ramachandran *et al*., 2018; Zarrabi *et al*., 2012) tend to be *static* in that the network structure is typically computed in batch-mode using the entire genomic data. Phylodynamic techniques can model the growth of a viral variant population (Attwood *et al*., 2022). However, such methods are often batch-mode and typically require re-computing if new samples are added. Clearly, the prevalence of a variant in a population varies with time. Consequently, a representation obtained using a static approach may obfuscate temporal information. Moreover, our understanding of the genetic composition of a variant can also change as more samples are collected and sequenced. For instance, the WHO definitions for variants correspond to an initial Pango lineage and the subsequently discovered sub-lineages (World Health Organization (WHO), 2022). The representation proposed by us, on the other hand, is based on creating a dynamic (*i.e*., temporally evolving) network, where the communities representing sample groupings can change over time - providing thereby an evolving perspective on the composition of variants and interactions between them. We conclude our review of the prior research in data representation by noting that modelling time-varying epidemiological dynamics through temporal networks, is as of yet, underutilized compared to static network models (Enright and Kao, 2018). Currently in infectious disease epidemiology, temporal networks have been applied to agent-based contact and transmission networks (Lentz *et al*., 2016; Ruget *et al*., 2021), in epidemiological models (Leitch *et al*., 2019; Nunner *et al*., 2021; Valdano *et al*., 2015), and for determining (local) topological patterns in contact networks (Senchyna and Singh, 2022). However, their use for variant identification and tracking has, to the best of our knowledge, not occurred prior to the current work.

From the perspective of variant identification, we note that in phylogenetics, relatedness of viral genomes is determined by computing their hypothetical most-recent common ancestor. The output of phylogenetic analysis is therefore represented as a branched tree where the tips are the sampled genomes, and the internal nodes are the hypothetical ancestors. The lengths of branches (if determined) represent the genetic distance or the time from coalescence of samples. In a phylogenetic representation, clades are sub-trees within the larger tree and include all samples that are the descendant of one common ancestor. Therefore, clades are used to classify a virus into genetically distinct "strains" or "variants" or "lineages". Phylogenetic Assignment of Named Global Outbreak (PANGO) has emerged as the prominent tool for tracking evolution and naming of SARS-CoV-2 lineages (Rambaut *et al*., 2020). Their hierarchical naming system is based on the relative position of sampled genomes in a global SARS-CoV-2 phylogeny. All lineages are descendants of the two founder lineages: A and B. Descendent lineages are assigned a number, and each subsequent generation is separated with a '.' notation. For example, B.1.17 is a descendant of B.1, which is a descendant of B. Given the volume of possible evolutionary changes, lineages are manually curated and only named if they meet certain epidemiological and biological criteria, such as evidence of onward transmission and at least one shared mutation within the lineage (O'Toole *et al*., 2021). Other established phylogenetic-based nomenclature for SARS-CoV-2 include those maintained by Nextstrain (Hadfield *et al*., 2018) and GISAID (Khare *et al*., 2021). Outside the techniques discussed above, methods for determining variant nomenclature have grouped variants by using co-mutating nucleotides and clustering (Melnyk *et al*., 2021; Qin *et al*., 2021). These methods are limited in their ability to model the changing variant population over time. Variant identification and tracking are a dynamic process. Variants emerge and spread in a population for a period of time, and then may secede due to human

imposed factors (*e.g.*, vaccinations and social distancing measures) or the emergence of a new more potent variant. As phylogenetic trees are static and preclude cycles, they are incapable of capturing such dynamics. This leads to the additional reliance on external criteria for identifying variants, as mentioned above. These criteria have undergone several revisions as the pandemic has progressed (Roemer *et al*., 2022; Villabona-Arenas *et al*., 2020), reflecting a distinction between estimating evolution (as is the original purpose of a phylogenetic tree) and tracking of epidemiologically relevant variants.

3. METHODS

3.1. Genomic data collection and preprocessing

SARS-CoV-2 genomes were downloaded from the GISAID database (Khare *et al*., 2021) for two countries, Israel and Portugal. The sequences were filtered for completeness, high coverage, and known dates of collection, (till April 5, 2022). This resulted in 16,929 genomes for the Israel dataset (IDS), and 19,325 genomes for the Portugal dataset (PDS). A reference sequence (originating from the first recorded outbreak in Wuhan, China) was downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/, accession number MN908947). To ensure that all sequences were of the same length, the genomes for each dataset were aligned to the reference sequence with MAFFT (Katoh *et al*., 2019). Subsequently, the first and last 100bp nucleotides in non-coding regions were removed from sequences due to the number of gaps after alignment (the mean proportion of gaps in this region per sequence was 38.9%). If the number of ambiguous nucleotides and/or gaps in a sequence was greater than 1% of the abbreviated sequence length, then corresponding sample was removed from further downstream analysis. Insertions and deletions (indels) were ignored due to lack of clarity between indels and ambiguous nucleotides. After these filtering

steps, 16,526 and 19,097 samples remained in the IDS and PDS sets, respectively. All these sequences had a nucleotide length of 29,703. Further description of these data sets, particularly in terms of the viral lineages present, is provided in Section 4. Next, pairwise genetic distances were calculated as the number of sites where pairs of sequences differ in their nucleotide composition, divided by the total length of the sequence (hamming distance). Gaps in sequences were replaced with the nucleotide present in the same position in the reference sequence so that the hamming distance satisfied the triangle inequality. Israel and Portugal were chosen as sample countries due to their moderate population and geographic size, along with the volume of genomes that have been sequenced and made publicly available from each country. Additionally, the pandemic has been well documented in Israel due to their early mass vaccination program (Goldberg *et al*., 2021; Saban *et al*., 2022), while Portugal gives a snapshot of the SARS-CoV-2 spread in Europe, which was initially one of the hardest hit regions. The complete list of downloaded sequences and associated metadata are available through GISAID (DOI for dataset: 10.55876/gis8.221011cm).

3.2. Data representation and sampling in the DEN

The goal of sampling is to reduce the numerosity of the samples being considered for subsequent analysis without altering the fundamental patterns in the data. Subsequent to sequence alignment and pairwise distance calculation, a $N \times N$ distance matrix D was computed where each sample was characterized by its Hamming distance to all other samples in the dataset. Next, we used classical multidimensional scaling (CMDS) to embed D in a low dimensional space while minimally perturbing the inter-sample distance distribution (Torgerson, 1952). The dimension of the low-dimensional representation space was chosen

to be 10 based on analyzing the eigenvalue distribution underlying CMDS (supplementary Figure 1), except for the first week in the PDS and for the first two weeks in the IDS. In both cases, due to low number of samples, CMDS led to an embedding with fewer than 10 dimensions. The modes of the data distribution were then determined, and the data was sampled using the mean-shift algorithm. Mean shift is a mode finding algorithm used for empirically identifying the maxima of the data density function (Fukunaga and Hostetler, 1975). The only parameter in the method is the bandwidth h, of a weight determining function - called the kernel, making the algorithm particularly useful in problems where the number of clusters is not known *a priori*. In the following, we outline the key steps of the mean shift-based sampling process as used by us.

Let $X = \{x_1, \ldots, x_n\}$, where $x_i \in R^{10}$ be the positional representation of genome samples after CMDS. The mean shift algorithm starts by randomly selecting a sample x_i, and iteratively performs the following steps until all samples are assigned to a cluster.

1) *Weighting of each point:* We employ the flat kernel, $K(x_i - x_j)$, to determine the weight of each point, $x_j \in X$. The point x_j is given a weight of 1 if the distance between x_i and x_j is less than or equal to the bandwidth h, and 0 otherwise.

$$K(x_i - x_j) = \begin{cases} 1 \; if \; \|x_i - x_j\| \leq h \\ 0 \; if \; \|x_i - x_j\| > h \end{cases} \quad (1)$$

2) *Conditional expectation:* The weighted mean of X, $m(x)$, is calculated by conditional expectation determined using kernel regression estimate.

$$m(x) = \frac{\sum_{x_j \in X} K(x_i - x_j) x_j}{\sum_{x_j \in X} K(x_i - x_j)} \quad (2)$$

3) *Kernel shifting:* The kernel is shifted to center around $m(x)$, such that $x_i \rightarrow m(x)$, and steps 1) and 2) are repeated until there is convergence, *i.e.*, $x_i = m(x)$.

4) *Iteration*: The procedure is repeated on another randomly selected point until all points have converged to their local maxima. Points with the same maxima are grouped in the same cluster.

The sampled sequences $S = \{s_1, \ldots, s_n\}$, selected by us are defined to be the sequences whose representation in X are the shortest Euclidean distance to their respective cluster modes. If multiple sequences are closest to the mode, then the sample collected earliest in time is chosen. This step ensures that the selected samples are either the modes or the closest data point(s) to the mode(s). For the viral genomic space, samples centered in locally dense regions represent sequences that are closely related to many other sequences. To capture these mutations with high resolution, a small value for h would result in many sequences sampled, some in sample dense regions. However, the mean shift also identifies sample from low-density regions if there are samples whose distance to all others is greater than h. These regions are of comparatively lesser interest to us since the sequences lying there often contain random low frequency mutations. Therefore, clusters that represent less than a fraction, f, of the dataset at time t are removed from the sampled set. CMDS and mean shift were performed at each time point $t = \{1, \ldots, T\}$ and $S(t)$ represents the sampled sequences at each t. In our studies a weekly temporal resolution was used to analyze the data.

To determine the appropriate values for h and f, the above method was separately applied to each of the IDS and PDS datasets in their entirety. A range of values for $h = \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}, 6 \times 10^{-5}, 7 \times 10^{-5}, 8 \times 10^{-5}, 9 \times 10^{-5}, 1 \times 10^{-4}\}$, and $f = \{0, 0.001, 0.0001\}$ were tested. The values $h = 2 \times 10^{-5}$, and $f = 0.0001$ were empirically chosen based on sample size and inspection of sampling in a two-dimensional space (supplementary table 1). With the chosen value for f, filtering out of low-

density sampled genomes only occurs once 10,000 samples have been collected. An illustration of the sequence sampling is shown in Figure 1 using data from both the IDS and PDS.

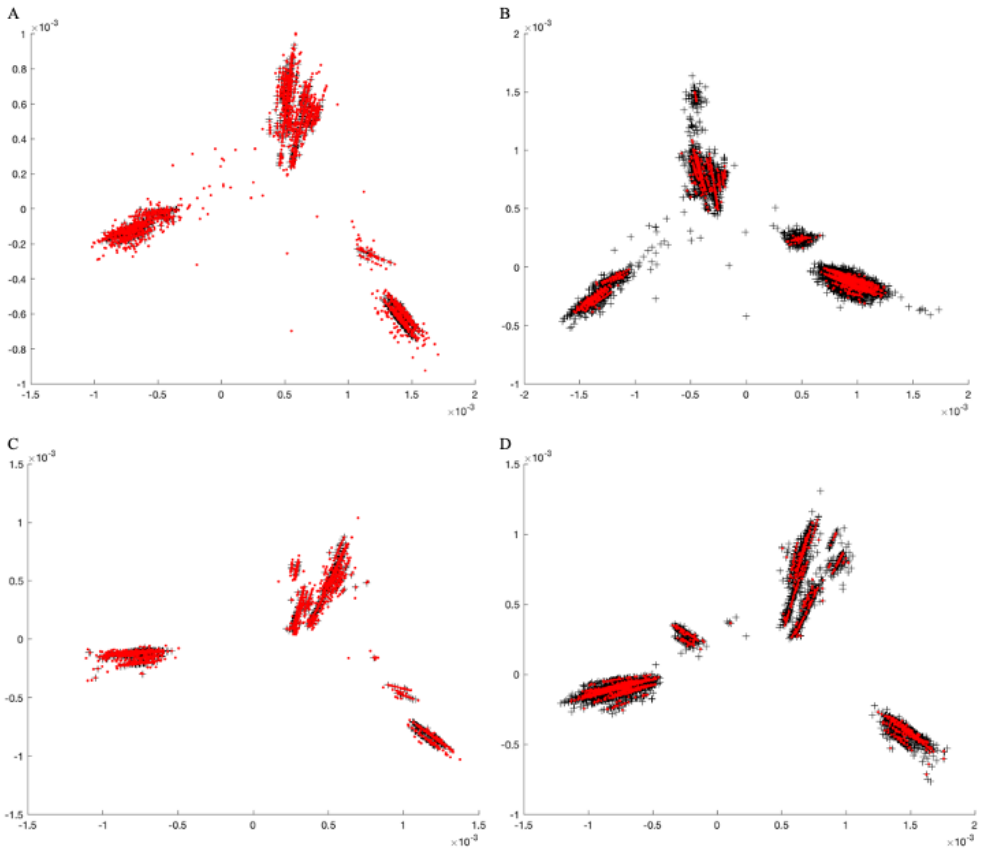

Figure 1. Illustration of the data sampling process using a two-dimensional projection of the data. The two largest eigenvectors from CMDS are used as the X- and Y-axes respectively. The mean-shift procedure is subsequently employed to select the representative samples. Sampled genomes are represented in the figures as red dots and unsampled genomes are shown as black crosses. The top two figures are of IDS on A) July 11th, 2021 (n=9,095), and B) March 27th, 2022 (n=16,526). The bottom two figures are PDS on C) June 20th, 2021 (n=8,422), and D) April 3rd, 2022. (n=19,097). Filtering of low-density clusters occurs in B) and D) but not in A) and C) as the number of samples in the two former figures is greater than 10,000.

3.3. Modeling the evolution of variants

The DEN can be thought of as a "network of networks" consisting of two-levels at which the data and relationships within it are represented (figure 2).

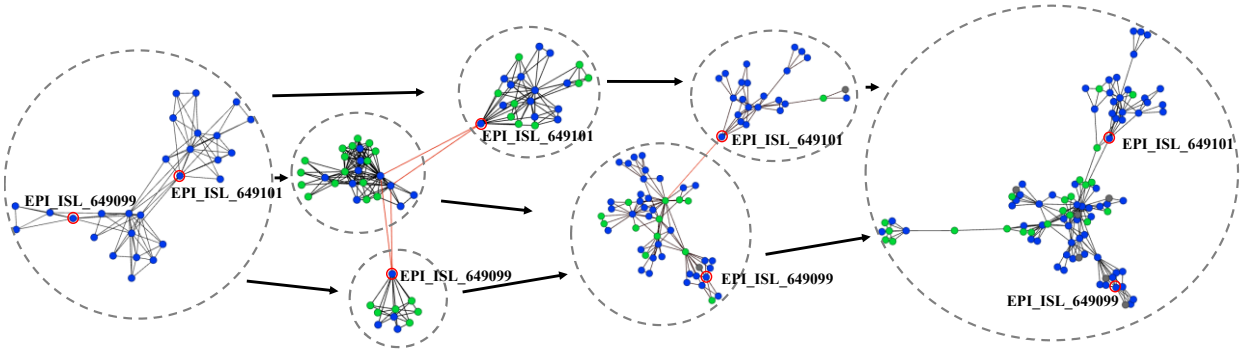

Figure 2. Visualization of network formation, graph cuts and clustering, and vertex correspondence over time. Each network represents a week of samples from the IDS collected during the consecutive weeks of March 8th - March 29th, 2020. The blue vertices are the persisting vertices (*i.e.*, they were present in the previous week). The green are the new vertices (*i.e.*, they were not present in the previous week), and the gray are the non-persisting vertices (*i.e.*, they will not be present in the following week). Red edges denote those edges that were cut when partitioning the graph into clusters. These clusters are outlined with the gray dashed circles, which are also the vertices in the DEN. Black arrows are the directed edges in the DEN. For contextualization across time, two persisting samples are labeled: EPI_ISL_649099 and EPI_ISL_649101. These two samples were in two separate clusters on March 15th (2nd network in the figure).

At the first level, for each time-point t, the undirected network $G(t) = (S(t), E(t))$ captures the similarity of samples. In $G(t)$, the set of vertices, $S(t)$, represents the samples and $E(t)$

represents the set of edges that connect samples whose genomes are deemed to be similar. In a network, an important notion is that of a community: a cluster of (connected) vertices that are more similar to each other than to other vertices. For our data, samples of the same variant form a community, and at each time-point $t \in [1, T]$, the samples $s \in S(t)$ are partitioned into a set of disjoint clusters, $C(t) = \{c_{1(t)}, \ldots, c_{n(t)}\}$. Tracking cluster dynamics in temporal networks is a challenging problem (Alotaibi and Rhouma, 2022; Cazabet and Rossetti, 2019); with the passage of time, these clusters can grow, shrink, split, merge, appear, and disappear. To computationally track the evolution of variants in a data-driven manner, a correspondence between clusters occurring at successive points in time is established and constitutes the second level of the network representation. The correspondence between clusters, is represented as a mapping between two sets of clusters at consecutive points in time: $C(t)$ and $C(t+1)$ with an edge connecting the pair of clusters ($c_{k(t)}$, $c_{l(t+1)}$), where $c_{k(t)} \in C(t)$ and $c_{l(t+1)} \in C(t+1)$ if a correspondence can be established between them. Let κ denote the set of all such correspondences between clusters at consecutive time points $C(t)$ and $C(t+1)$, $t \in [1, T]$. Then the dynamic epidemiological network DEN for the entire set of sampled sequences S collected over time $t \in [1, T]$ is defined as given in Eq. (3). The set of correspondences in the DEN are represented as edges between the corresponding clusters:

$$\mathrm{DEN}(S) = G(t) \cup \kappa \tag{3}$$

In the following we identify and discuss four issues that are important for modeling the dynamics how variants evolve in our proposed framework.

3.3.1 Temporal sample selection. As CMDS and mean shift operations are performed independently at every time-point t, the sets $S(t)$ and $S(t+1)$ may differ due both to the addition of new samples as well as the different eigenvectors defining the representation space. Consider two time-points t and $t+1$; compared to $S(t)$, new samples may be added to $S(t+1)$ as additional modes are found, and samples may be removed when modes cease to exist or if the corresponding cluster is small (of size less than f). Additionally, a sample in $S(t+1)$ may replace another present in $S(t)$ due to a shift in the mode. It's necessary to formally specify these changes and account for them prior to downstream analysis.

Consider the sample sets $S(t)$ and $S(t+1)$ and a sample $s_t \in S(t)$ such that $|s_t - s_{t+1}| > h$ for all $s_{t+1} \in S(t+1)$. Then the vertex, $v_{t,}$ corresponding to s_t is deemed a *non-persisting vertex* and is present in $V(t)$ but not in $V(t+1)$. From a biological perspective, this implies that the genotype corresponding to vertex v_t did not propagate to time z. If there exists a $s_{t+1} \in S(t+1)$ such that $|s_t - s_{t+1}| > h$ for all $s_{t+1} \in S(t+1)$, then, the vertex v_{t+1} representing $s_{t+1}$is called a *new vertex* since it is present in $V(t+1)$ but not $V(t)$. Biologically, such vertex represents a genotype not observed earlier. Finally, a *persisting vertex*, v_{t+1}, present in both $V(t)$ and $V(t+1)$ exists if there is a $s_t \in S(t)$ and $s_{t+1} \in S(t+1)$ such that $|s_t - s_{t+1}| \leq h$. In this case, the persisting vertex, v_{t+1}, represents the sample s_t rather than s_{t+1}. It should be noted that samples $s_{t+1} \in S(t+1)$ that satisfy the constraint $|s_t - s_{t+1}| \leq h$ with $s_t \in S(t)$ but do not constitute the closest neighbors at time-points t and $t+1$ represent non-persisting (new) vertices. For example, if there are two samples, $s_{t+1(i)}$ and $s_{t+1(j)}$, whose distance from the sample $s_t \leq h$, but $|s_t - s_{t+1(i)}| > |s_t - s_{t+1(j)}|$, then the persisting vertex will be represented by s_t, and $s_{t+1(i)}$ will represent a new vertex at time $t+1$. The exception to this analysis being when $t = 1$, as there is no *prior time-point and* therefore $S_t = V_t$ at the first time-point.

3.3.2. Network formation. We seek to define the connectivity of $G(t)$ in a manner that captures the neighborhood (proximity) relationships amongst viral variants. The approach described in this paper draws upon work in defining representation spaces for evolving phenotypic responses of parasites causing the disease schistosomiasis (Singh *et al*., 2018). Specifically, we employ parameterized neighborhood graphs, where vertices correspond to samples and the neighborhood relationships are represented by connecting the corresponding vertices to indicate relative genetic similarity. In particular, we use the family of *k*-nearest neighbor graph (NNG), with the parameter *k* characterizing each network as described below.

For a fixed t, consider $G(t)$ comprising n samples $\{s_1, s_2, \ldots s_n\}$. Let φ_k denote the NNG for scale parameter k. φ_k is defined as shown in Eq. (4), where for each sample s_i, $d_m(s_i)$ denotes the m_{th} closest sample in terms of their genetic distance. That is, $d_1(s_i)$ denotes the closest genotype to s_i, $d_2(s_i)$ denotes the second closest genotype to s_i and so on.

$$\varphi_k = (V, E): V = \{s, \ldots, s_n\} \wedge (s_i, s_j) \in E, \qquad if\ s = d_k(s_i) \tag{4}$$

The complete connectivity structure for $G(t)$ at a time point t can therefore be obtained by considering all possible *k*-nearest neighbor graphs as shown in Eq. (5). The reader may note that such a representation would be multi-scale and capture the entirety of the neighborhood information present between the samples. In practice, we iterate up to the nearest neighbor value n, at which the number of connected components in $G(t)$ equals 1.

$$\varphi = \varphi_1 \cup \varphi_2 \cup \ldots \cup \varphi_n \tag{5}$$

We note that the nearest neighbor relationship may be computed either directly using the genetic distances or via Euclidean distances between the corresponding points in the low-dimensional CMDS embedding. To empirically assess and best preserve consistency in k across time, we evaluated the network formation using both these options. Comparing the

former to the latter, the median for k increased from 21 (interquartile range, IQR, 4-71) to 31 (IQR, 8-302) in IDS, and from 50 (IQR, 11-74.75) to 99 (IQR, 84.25-214) in PDS. The substantially higher variation associated with the lower dimensional embedding led to the use of the original distance matrix for network formation.

3.3.3 Cluster identification. We use Laplacian spectral partitioning for detecting communities. This method does not assume prior knowledge of the number of communities in the data. For each graph, $G(t)$, community detection via spectral clustering is performed by recursively splitting the graph through Laplacian eigendecomposition. The graph Laplacian $L(G)$ for the graph G, is calculated as follows:

$$L(G) = D(G) - A(G) \tag{6}$$

In Eq. (6), $D(G)$ is the diagonal degree matrix of $G(t)$ and $A(G)$ is the adjacency matrix. The eigendecomposition of $L(G)$ clusters vertices into two subgraphs g_1 and g_2, with the approximate minimum cut (Fiedler, 1973). By taking the eigenvector corresponding to the second smallest eigenvalue, the vertices are split such that those with an eigenvector value greater than a partition value, in this case 0, are placed in g_1 and the remaining vertices are placed in g_2. Each of the subgraphs g_1 and g_2 can be recursively cut (subdivided) following the same procedure until the resultant subgraphs contain only one vertex. However, this would be uninformative for community detection. Therefore, to stop the recursive division process at the point at which communities are well clustered together, certain cuts can be rejected based on their quality (*i.e.*, the proportion of edges removed). This can be assessed by the normalized cut value, *nc*:

$$nc = \frac{IPC(g_1,\ g_2)}{vol(g_1)} + \frac{IPC(g_1,\ g_2)}{vol(g_2)} \quad (7)$$

In Eq. (7), the inter-partition connectivity (IPC) of g_1 and g_2 is the number of edges needed to be removed to partition the graph and $vol(g_x)$ denotes the number of edges connected to the vertices in g_x before the cut. The value of *nc* increases as the proportion of edges needing to be removed increases. Therefore, a cut threshold *r*, is used to reject cuts with large values of *nc*. The network formation and community detection procedure can be summarized in terms of the following four steps:

1) The network, $G(t)$, is formed (Algorithm 1).
2) Spectral clustering splits $G(t)$ into g_1 and g_2.
3) If the *nc* of $G(t) > r$ the cut is rejected, and no communities within $G(t)$ are found.
4) If the normalized cut of $G(t) \leq r$, the graphs g_1 and g_2 are each recursively split using spectral partitioning until no further cuts meet the acceptability threshold *r*.

The subgraphs found with this procedure are the clusters, $C(t)$, within G(t). Several values were tested for *r* (0.001, 0.005, 0.01, 0.05, 0.1, 0.5), and $r = 0.1$ was chosen for its better overall performance (Supplementary table 2 and 3). A list of all parameters and values used for network formation are given in table 1.

Table 1. Parameters for data sampling, network formation, and community detection

Parameter (*symbol, if used*)	Section Applied	Value
Number of eigenvectors	Dimensionality of mean shift sampling	10
Bandwidth (*h*)	Mean shift sampling and node selection	2×10^{-5}
Filter (*f*)	Mean shift sampling	1×10^{-4}
Cut threshold (*r*)	Community detection	0.1

3.3.4. Determining correspondences between clusters over time. For each pair of clusters, $c_{k(t)}$ and $c_{l(t+1)}$, where $c_{k(t)} \in C(t)$ and $c_{l(t+1)} \in C(t+1)$, if there are sampled genomes in common, then the two clusters are defined to have a correspondence. Such a cluster correspondence is represented by a directed edge between the two clusters in the DEN.

3.3.5. Analyzing the network evolution process. With the above methodology, the DEN can characterize the evolution of the disease in terms of the appearance, disappearance, and merging/splitting of viral genotype clusters over time. If only a single viral lineage were to persist over time, then the DEN would consist of a series of vertices having a one-to-one relationship across time. In other words, every vertex in the network would have an indegree and outdegree no greater than 1 with the corresponding clusters putatively increasing in size over time. However, during real-world outbreaks as a virus spreads, mutations are accumulated. In some cases, certain mutations translate to an evolutionary advantage, resulting in increased spread and thus an increase in the observed frequency of this mutation(s). Eventually, accumulated changes can lead to a new variant. In our directed network, this phenomenon is represented by a "split" in a cluster, where a vertex at time t has an outdegree greater than one connecting it to multiple vertices at time t+1. It can also be the case that a variant fails to grow, or that it does not become sufficiently distinct from the rest of the viral population. In such a case, the cluster containing this variant at time t may "merge" into another cluster at time t+1. A merging of clusters is represented as a vertex at time t+1 with an in-degree greater than one. Clearly, the difference between $S(t)$ and $S(t+1)$

due to sampling may lead to splitting and merging events. For instance, the nearest neighbors of a sample can change with the removal and/or addition of samples in the network. Moreover, sampling can also indirectly cause splitting and merging events. This occurs when a change in sampling leads to a change the value of k in the k-nearest neighbors network formation. The readers may recall that the value used for k is the minimum number of nearest neighbors a sample is required to form an edge with, in order to create a connected graph. In the viral networks considered by us, it is often the case that groups of samples share the same nearest neighbors, and so as the group increases in size, k must be increased in order to connect a group to the remainder of the samples in the network. This change in k can cause groups to merge and split, even if there is no change in sampling in the group itself. This situation has parallels in phylogenetic analysis and more generally in clustering: a group of samples that are a distinct clade at time t may appear as a small sub-clade at time t+1 with the addition of new samples that form a much larger subsuming clade. Further analysis of the approach using the entire IDS and PDS are presented in Section 4.

4. EXPERIMENTS AND RESULTS

4.1. Description of datasets

4.1.1. IDS. Samples were collected for 97 non-consecutive weeks between February 23rd, 2020, and March 27th, 2020. There were 227 distinct lineages present in the dataset. The majority (52.75%, 8,717/16,526) were the B.1.617.2 and descendant lineages (Delta), with descendant lineages comprising 95.98% (8,367/8,717) of the B.1.617.2 group. Lineage B.1.1.7 (Alpha) followed B.1.617.2 by prevalence (31.54%, 5,213/16,526). B.1.1.7 descendant lineages were less than 1% (49/5,213) of the B.1.1.7 group. The WHO classified variants, Beta (lineage B.1.351 and descendants) and Omicron (B.1.1.529 and descendants)

were found to be approximately 0.8% (131/16,526) and 0.8% (132/16,526) of the dataset, respectively. Other lineages of notable frequency, including their respective descendants, were B.1.1.50 (5.97%, 987/16,526), B.1.362 (2.27%, 375/16,526), and B.1.1.294 (0.5%, 90/16,526). Of the remaining lineages, B.1.1 was 0.5% (86/16,256), B.1 was 1.8% (293/16,526), and B was less than 0.01% (6/16,526). Lastly, 0.08% (13/16,526) of samples were within the A lineage group, and the remaining samples comprised of 77 B sub-lineages not outlined here (3.5%, 574/16,526). The relative frequency of these lineages over time can be seen in figure 3A. As many lineages were low frequency, sub-lineages were grouped with their ancestral lineages unless otherwise stated. In the initial year of the pandemic, B.1 was the most frequent lineage, followed by B.1.1, B, and A. By December 2020, lineage B.1.1.7, B.1.1.50, and B.1.362 began to emerge and grow until approximately September 2021, when B.1.617.2 became dominant.

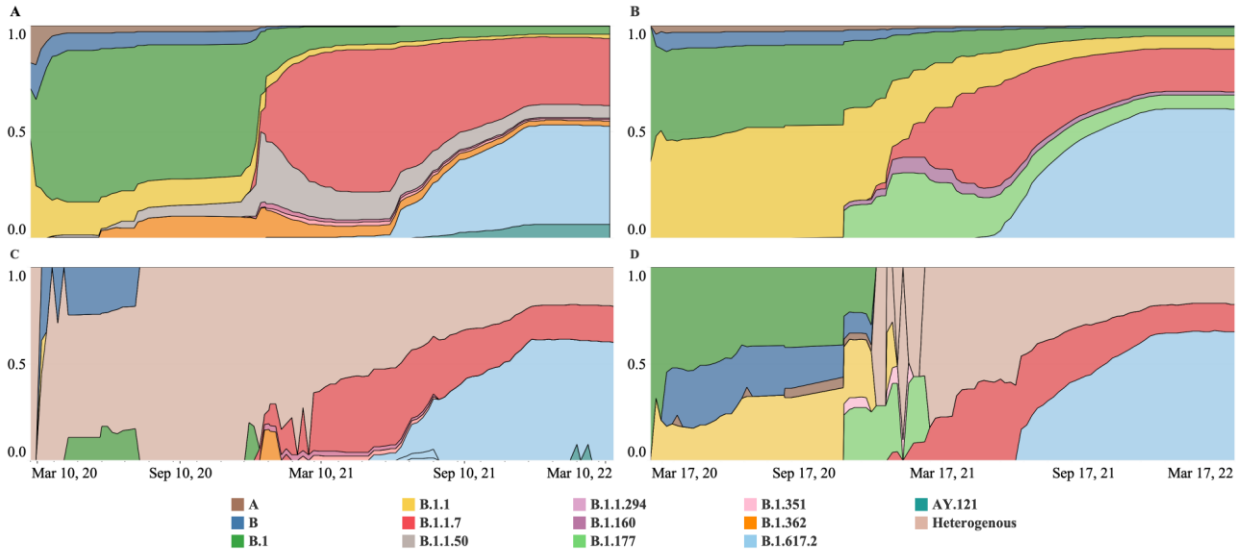

Figure 3. Changes in the proportions of each lineage group in the dataset and cluster sizes in the DEN, over the period of analysis. Actual proportions of the entire datasets are shown in (A) IDS and (B) PDS. The cluster sizes in the DEN over the entire tracked period are shown in

(C) IDS and (D) PDS. The color-coding of each lineage group in (A) and (B) and the corresponding cluster in (C) and (D) is shown in the figure legend. Clusters are numbered in the order of their emergence and the color of a cluster corresponds to the majority lineage of that cluster. Clusters that are not predominantly one lineage group are called "heterogenous". Changing cluster sizes in (C) and (D) approximately follow the growth of the lineage groups in (A) and (B), respectively. A major exception being B.1.1.50 in IDS, which fails to becomes a cluster (A and C). The size of the heterogeneous cluster is also approximately equivalent to the combined size of those lineages that are in decline (*e.g.*, lineages A, B, B.1).

4.1.2. PDS. The first sample collected was on March 1st, 2020, after which samples were collected for 99 non-consecutive weeks until April 1st, 2022. The dataset contained 254 distinct lineages. Lineage B.1.617.2 and descendants comprised 60.6% (11,574/19,097); 98.12% (11,357/11,574) of which were descendant lineages. B.1.1.7 and descendants were 20% (3,825/19,097) of the dataset. The majority of this group was the B.1.1.7 lineage itself (99.4%, 3,802/3,825), with only 0.6% (23/3,825) being descendants of the B.1.1.7 lineage. The lineages B.1.177 and B.1.160 along with their descendants were 6.53% (1,247/19,097) and 1.6% (307/19,097) of the dataset, respectively. Other WHO defined variants Omicron (B.1.1.529 and descendants), Gamma (P.1 and descendants), and Beta (B.1.351 and descendants) only reached 0.43% (83/19,097), 0.81% (154/19,097), and 0.41% (78/19,097) of the collected samples by the end of the tracked period, respectively. Of the ancestral lineages, the B lineage comprised 0.2% (24/19,097) of the dataset, while B.1 comprised 1.7% (321/19,097) and B.1.1 comprised 3.3% (635/19,097). In addition to the B lineages already mentioned, there were 77 more lineages which, when combined, comprised 4.2% (810/19,097) of the dataset. The remaining samples were within the A lineage (0.3%,

55/19,097). B.1.1 was found to be the dominant lineage in the first months of the pandemic (Figure 3B). Subsequently, B.1.1.7, B.1.177, and B.1.160 grew in frequency from late 2020 onward, until the emergence and dominance of the B.1.617.2 strain beginning in the latter half of 2021. Like IDS, sub-lineages were grouped with their ancestral lineage unless stated otherwise in the figure caption.

4.2. Assessment of clustering by graph cuts

4.2.1. Cluster Correspondences. The DEN for each dataset can be seen in figures 4 (IDS) and 5 (PDS). Vertices are represented as pie charts which shows the Pango lineage-composition of each cluster. The numbers overlayed on the vertices identify the clusters they represent in the network. Obviously, a cluster propagates in time if it has a one-to-one correspondence with another cluster across time. However, clusters may also merge and/or split. Additionally, there are instances where a merge is immediately followed by a split, such as on February 7th, 2021, in figure 4. In such cases, determining the similarity of clusters in terms of their sample composition before and after a split/merge may be of interest. Such similarity is difficult to determine by considering the edges of the directed network alone. Therefore, to find cluster correspondence in the absence of a one-to-one relationship, we determine the Jaccard index between all previously identified clusters and the merged or split cluster. The Jaccard index measures the similarity between two sets, C_x and C_y, as the size of their intersection divided by the size of their union:

$$J(C_x, C_y) = (|C_x \cap C_y| / |C_x \cup C_y|) \quad (8)$$

The range of the Jaccard index, $J(C_x, C_y)$, is a number between 0 and 1, where 0 implies the absence of the sets being compared overlapping while 1 implies a complete overlap between the two sets. To assess if there was a correspondence between a cluster, *A*, at time *t* and a

previously identified cluster, the Jaccard index is calculated between the set of persisting samples in cluster *A* and the set of persisting samples in all previously identified clusters. Cluster *A* is said to corresponded with a previous cluster if it (a) has a maximum overlap with the cluster in question (as given by the Jaccard index) among all other candidate clusters, and (b) the value of the Jacquard index exceeds the (empirically set) threshold of 0.6 (*i.e.*, the majority of persisting samples moved into cluster *A*). This resulted in 15 cluster correspondences in IDS and 11 in PDS.

4.2.2. Cluster correlation with Pango lineages. Many clusters were found to approximately align with Pango lineage-groups. In IDS, cluster 2 contained a majority of B lineage samples, cluster 3 contained samples of lineage B.1.1, cluster 5 and 6 were mostly lineage B.1, cluster 7 and 10 comprised lineage B.1.1.7, and clusters 12, 13, 14, and 15 aligned with lineage B.1.617.2, with cluster 15 comprising completely of the descendant lineage, AY.121 (alias of B.1.617.2.121). Cluster 1 and cluster 4 were less specific; cluster 4 in particular occurred due to the merging of other clusters (figure 4). Similarly, in PDS, cluster 1 corresponded to the lineage B.1.1, cluster 2 to lineage B.1, cluster 3 to lineage B, cluster 4 to lineage A.2, cluster 5 to lineage B.1.160, cluster 6 to lineage B.1.177, cluster 9 to lineage B.1.1.7 and cluster 11 to lineage B.1.617.2 respectively. Clusters 7, 8, and 10 were a combination of multiple merged clusters (Figure 5). The cluster formation underlying the DEN can be assessed in terms of precision, recall and accuracy. To explain the logic underlying such an assessment, let a lineage group simply imply a lineage and all those descendants that are not a majority lineage in another cluster at the same time-point. Then, for each cluster $c_{x(t)} \in C(t)$, $s(l) \in c_{x(t)}$ are the samples in $c_{x(t)}$ corresponding to the samples of the majority lineage group of cluster

$c_{x(t)}$ (true positives), $s(l') \in c_{x(t)}$ are the samples in $c_{x(t)}$ corresponding to the samples that are not the majority lineage group of cluster $c_{x(t)}$ (false positives), $s(l) \in c_{x(t)}'$ are the samples not in $c_{x(t)}$ whose lineage is the majority lineage group of $c_{x(t)}$ (false negatives) and $s(l') \in c_{x(t)}'$ are the samples not $c_{x(t)}$ and whose lineage are not the majority lineage of $c_{x(t)}$ (true negatives). The measures of cluster assessment can now be defined as follows:

$$\text{Precision } (P) = \frac{\sum s(l) \in c_{x(t)}}{\sum s \in c_{x(t)}} \tag{9}$$

$$\text{Recall } (R) = \frac{\sum s(l) \in c_{x(t)}}{\sum s(l) \in c_{x(t)} + \sum s(l) \in c_{x(t)'}} \tag{10}$$

$$\text{Accuracy } (A) = \frac{\sum s(l) \in c_{x(t)} + \sum s(l') \in c_{x(t)'}}{\sum s \in G(t)} \tag{11}$$

The average results over all time points and clusters of the same lineage groups can be found in table 2 and table 3. There was a low average precision for clustering the B lineage in both datasets (0.74 in IDS and 0.78 in PDS) as it was clustered with the A lineage. Lineage B.1 clustering had a high precision in both datasets (0.96 in IDS and 0.98 in PDS) but a low recall (0.63) in IDS due to its split between the two clusters 4 and 5 and subsequently 5 and 6. Clustering B.1.1 lineage (IDS and PDS) and A.2 lineage (PDS) were at least 0.97 in all measures. The average precision, recall, and accuracy for clustering B.1.1.7 was above 0.99 in both datasets. These measures remained high for B.1.1.7 in IDS as even though there were two B.1.1.7 clusters present in the DEN at the same time (cluster 7 and 10), cluster 10 was very small, comparatively, and only lasted one week. However, the multiple clusters (clusters 12, 13, and 14) over multiple weeks did have an impact on the recall (0.78) and accuracy (0.94) for clustering B.1.617.2 in IDS, while these measures were all 1 for B.1.617.2 in PDS. Precision, recall, and accuracy were all 1 for lineage AY.121 in IDS, and all measures were at least 0.96 for the region-specific clusters of B.1.362 (cluster 9) and B.1.351 (cluster 11) in

IDS, and B.1.160 (cluster 5) and B.1.177 (cluster 6) in PDS. Clustering of B.1.1.294 in IDS had a slightly lower precision (0.91), recall (0.84), and accuracy (0.95) as a portion of B.1.1.294 samples were in cluster 4.

4.2.3. Normalized cuts. The normalized cut values obtained during the initial cluster identification can be seen in figure 6A. The median values are 4×10^{-3} (interquartile range, IQR, $3 \times 10^{-4} - 2 \times 10^{-2}$) and 5×10^{-3} (IQR, $8 \times 10^{-5} - 3 \times 10^{-2}$) in IDS and PDS, respectively. As the vast majority of these cuts are well below the 0.1 threshold, this indicates that the clusters are not an artefact of the normalized cut threshold parameter but represent fundamental patterns in the data. To further assess the quality of the clustering, the normalized cut values were investigated for each of the cluster correspondences, separately, in both datasets (figures 6B and C). Given that the normalized cuts were initially obtained recursively (*i.e.,* they may have been calculated over a subnetwork rather than the entire network), equation (5) was repeated for each cluster with g_1 as the cluster of interest and g_2 as the remainder of the network. Figures 6B and 6C show that the clusters with the highest median normalized cut value either 1) had a short temporal duration in the DEN or 2) split and merged with another cluster several times. In IDS, clusters 2, 3, 5, 6, 10, and 15 had the highest median normalized cuts, those being 0.05 (IQR, 0.04-0.05), 0.08, 0.03 (IQR, 0.03-0.03), 0.07 (IQR, 0.06-0.08), 0.05, and 0.09 (0.08-0.09), respectively. Clusters 2 and 15 split and merged from clusters 4 and 12, respectively, on at least two occasions. While clusters 3, 4, 6, and 10 were only present in the network for at most two weeks. Similarly, cluster 4 in PDS split and merged from cluster 3 on three occasions and had the highest median normalized cut 0.06 (IQR, 0.06 – 0.06), equal only with cluster 8 (0.06, IQR, 0.06 –

0.06). Cluster 8 was present in the directed network for only 2 weeks (January 3rd – 10th, 2021), the shortest duration of any cluster in PDS.

Conversely, those cluster correspondences present in the directed network for the longest duration had some of the lowest median normalized cuts in the dataset. Specifically, the clusters that were large in size and specific for one variant. This included cluster 7, the B.1.1.7 majority cluster, and cluster 12, the B.1.617.2 majority cluster, in IDS, which had the median normalized cuts 9×10^{-5} (IQR, 5×10^{-5} - 5×10^{-3}), and 9×10^{-4} (IQR, 3×10^{-4} - 2×10^{-3}), respectively. In PDS, also, it was the clusters corresponding to the lineage groups B.1.1.7 (cluster 9) and B.1.617.2 (cluster 11) that had the lowest median cut vale of 4×10^{-5} (IQR, 3×10^{-5} - 6×10^{-3}) and 4×10^{-3} (IQR, 2×10^{-3} - 8×10^{-3}), respectively, along with cluster 10, which had a median normalized cut of 4×10^{-3} (IQR, 5×10^{-5} - 7×10^{-3}).

Table 2. Precision, recall, accuracy, and proportion of active period covered by clusters for a specific lineage group in the IDS. The column "Cluster week(s)" lists the specific weeks a given cluster was found to be present in the network. The column titled "Proportion collected at splitting" represents the proportion of total samples of a lineage collected before it became a cluster. The proportion of samples of a lineage collected while the cluster was present in the network is given by the difference between the corresponding proportion collected at splitting values and proportion collected at merging values.

Lineage	Clusters	Precision	Recall	Accuracy	Cluster week(s)	Proportion collected at splitting (n/N)	Proportion collected at merging (n/N)
B	2	0.74	1.00	0.93	Mar 15th-22nd, 2020	0.33 (8/24)	0.75 (18/24)
					Apr 5th, 2020	0.88 (21/24)	1.00 (24/24)
					Apr 19th-Jul 20th, 2020	1.00 (24/24)	1.00 (24/24)
B.1	5	0.96	0.63	0.68	Mar 15th-22nd, 2020	0.03 (20/612)	0.14 (88/612)
					Apr 5th 2020	0.24 (148/612)	0.27 (166/612)
					Apr 19th-Jul 20th, 2020	0.32 (196/612)	0.37 (228/612)
					Dec 7th-13th, 2020	0.41 (252/612)	0.4 (246/612)
B.1.1	3	1.00	1.00	1.00	Mar 15th, 2020	0.01 (12/1350)	0.02 (23/1350)
B.1.1.7	7 & 10	1.00	0.99	1.00	Dec 20th, 2020-Jan 31st, 2021	0 (1/5213)	0.34 (1749/5213)
					Feb 14th, 2021	0.42 (2168/5213)	0.45 (2364/5213)
					Feb 28th, 2021-Mar 27th, 2022	0.54 (2825/5213)	1.00 (5213/5213)
B.1.1.294	8	0.91	0.84	0.95	Dec 27th, 2020-Aug 1st, 2021	0.06 (5/90)	1.00 (90/90)
B.1.362	9	1.00	1.00	1.00	Dec 27th, 2020-Jan 10th, 2021	0.13 (50/375)	0.76 (285/375)
B.1.351	11	1.00	1.00	1.00	Feb 14th, 2021	0.69 (91/131)	360.71 (93/131)
					Feb 28th-Aug 1st, 2021	0.8 (105/131)	1.00 (131/131)
B.1.617.2	12, 13, 14, & 15	1.00	0.78	0.94	May 16th, 2021-Mar 27th, 2022	0.01 (63/8717)	1.00 (8717/8717)
AY.121	15	1.00	1.00	1.00	Feb 1st, 2022	1.00 (1055/1055)	1.00 (1055/1055)
					Feb 15th, 2022	1.00 (1055/1055)	1.00 (1055/1055)

Table 3. Precision, recall, accuracy, and proportion of active period covered by clusters for a specific lineage group in PDS. Description of column headers is analogous to that in Table 2.

Lineage	Clusters	Precision	Recall	Accuracy	Cluster week(s)	Proportion collected at splitting (n/N)	Proportion collected at merging (n/N)
B.1	1	0.98	0.98	0.98	Mar 8th-Dec 14th, 2020	0.01 (9/1096)	0.54 (597/1096)
B.1.1	2	0.97	1.00	0.99	Mar 1st-Dec 14th, 2020	0 (0/0)	0.64 (779/1212)
B	3	0.78	1.00	0.94	Mar 22nd-Dec 14th, 2020	0.36 (32/88)	1.00 (88/88)
A.2	4	1.00	1.00	1.00	Apr 5th, 2020	0.73 (22/30)	0.77 (23/30)
					Jul 5th, 2020	0.87 (26/30)	0.87 (26/30)
					Aug 23rd-Dec 14th, 2020	0.87 (26/30)	0.87 (26/30)
B.1.160	5	1.00	1.00	1.00	Nov 8th-Dec 7th, 2020	0 (0/307)	0.12 (36/307)
					Jan 3rd-Jan 31st, 2021	0.21 (64/307)	0.73 (225/307)
B.1.177	6	0.96	1.00	0.97	Nov 8th, 2020-Jan 17th, 2021	0 (6/1247)	0.71 (886/1247)
					Jan 31st-Feb 21st, 2021	0.75 (937/1247)	0.93 (1155/1247)
B.1.1.7	9	1.00	1.00	1.00	Jan 10th-Jan 17th, 2021	0.03 (98/3825)	0.05 (201/3825)
					Jan 31st, 2021-Apr 3rd, 2022	0.06 (246/3825)	1.00 (3825/3825)
B.1.617.2	10	1.00	1.00	1.00	Jun 27th, 2021-Apr 3rd, 2022	0.09 (1088/11574)	1.00 (11574/11574)

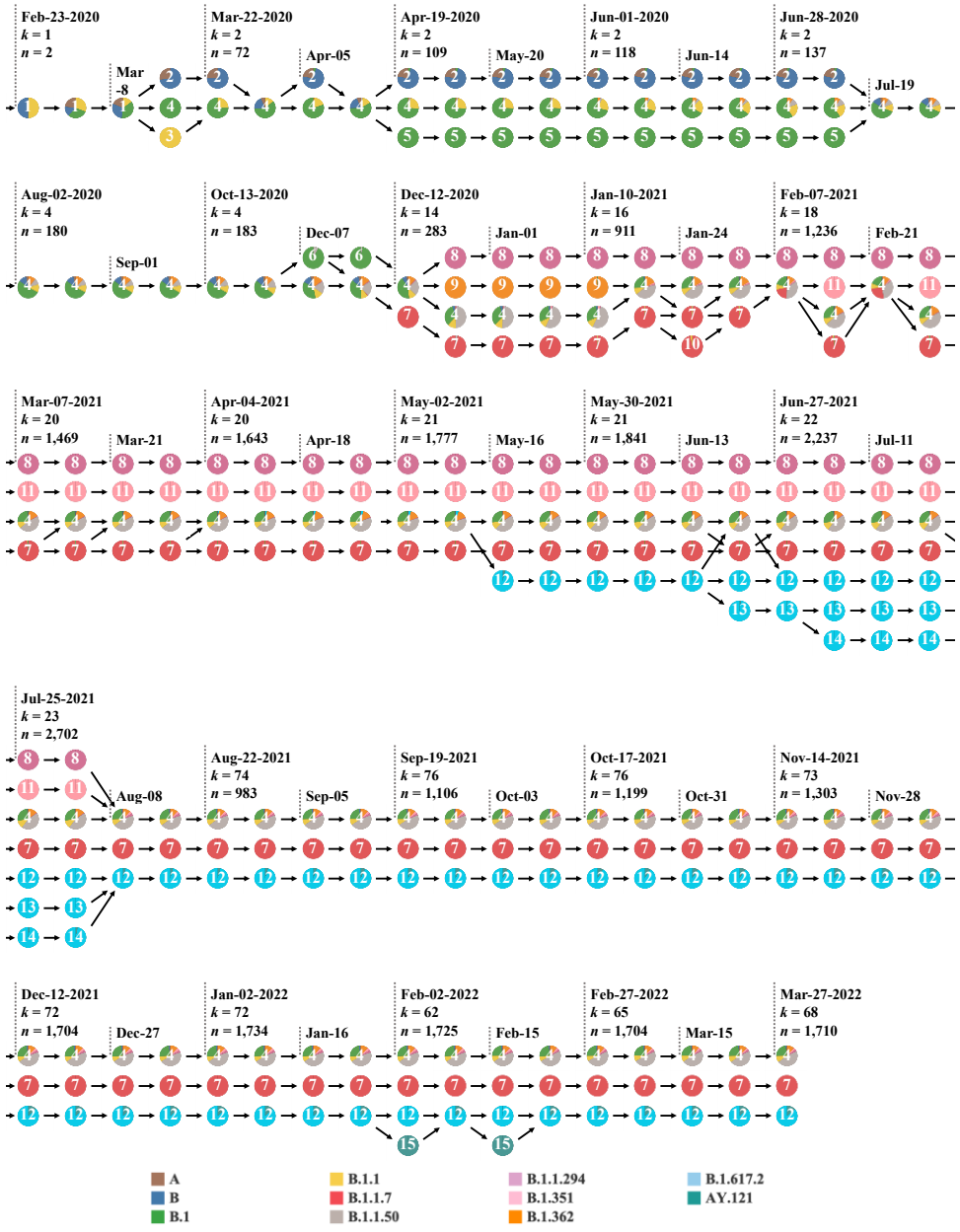

Figure 4. The DEN for the IDS and its evolution over time. The sample collection date is shown every two weeks, with the value of *k* and number of sampled genomes additionally shown for every other label. Each vertex in the network represents a uniquely numbered cluster. Vertices in the network are shown as pie charts, depicting the Pango lineage composition of the

samples in the corresponding cluster. These lineages are color coded as shown at the bottom of the figure.

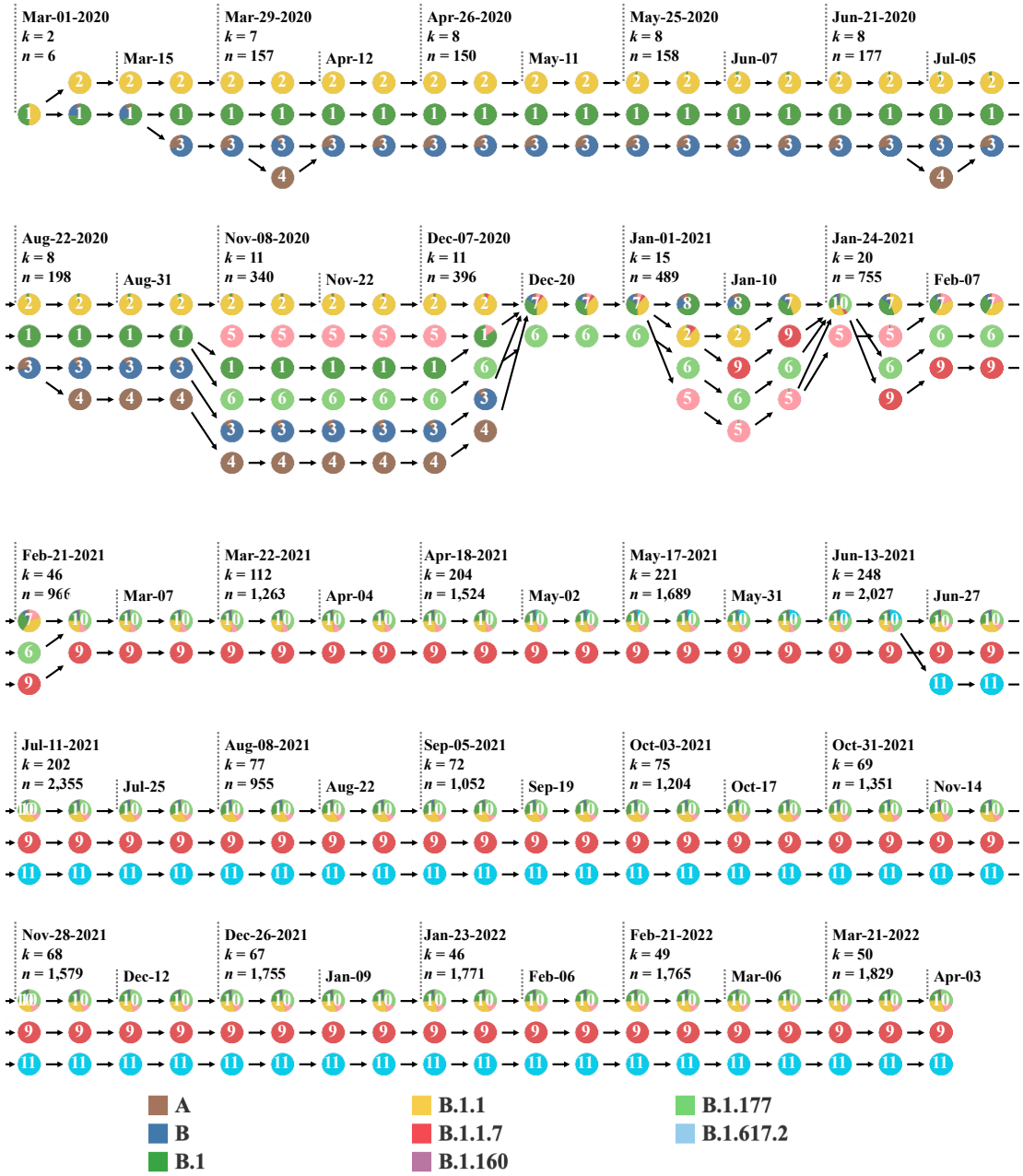

Figure 5. The DEN for the PDS and its evolution over time. The sample collection date is shown every two weeks, with the value of *k* and number of sampled genomes additionally shown for every other label. Each vertex in the network represents a uniquely numbered cluster.

Vertices in the network are shown as pie charts, depicting the Pango lineage composition of the samples in the corresponding cluster. These lineages are color coded as shown at the bottom of the figure.

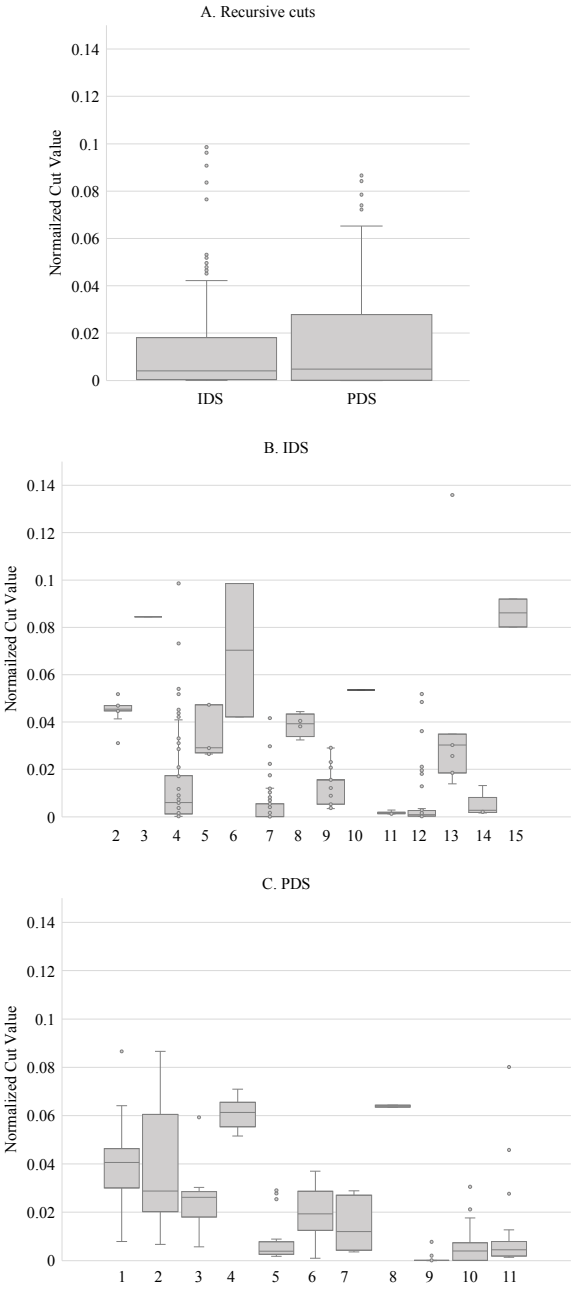

Figure 6. Box and whisker plots of normalized cuts from (A) recursive cutting of the network during initial cluster identification and for each cluster correspondence in (B) IDS

and (C) PDS. Cluster correspondence number is given on the x-axis and normalized cut values are on the y-axis. The plots demonstrate the observation that the majority of normalized cut values are well below the normalized cut threshold of 0.1 and therefore the clusters represent fundamental patterns in the data rather than being artefacts of the clustering algorithm.

4.3. Case studies capturing the dynamics of viral evolution

The period tracked for both the IDS and the PDS (figures 4 and 5) can be conceptually grouped into three broad stages that characterize the pandemic: 1) the first wave caused by the spread of the ancestral lineages, 2) the emergence of B.1.1.7 (Alpha) and other, region-specific lineages, and 3) cessation of previous dominant lineages and rapid spread of lineage B.1.617.2 (Delta). In the following we examine the network evolution induced by the data in light of the aforementioned conceptual groupings:

4.3.1. March 2020 – November 2020. Samples split into clusters corresponding to the ancestral lineages (lineages A, B, B.1, and B.1.1):

A. IDS. Tracking began on February 23rd, 2020. For the first three weeks there was only one cluster: (cluster 1). On the fourth week, cluster 1 split into cluster 2 (lineages A and B), cluster 3 (lineage B.1.1), and cluster 4 (lineage B.1). In the two subsequent weeks, clusters 2 and cluster 3 merged with cluster 4. By April 19th, 2020, cluster 2 and cluster 5 (also a majority B.1 lineage) split from cluster 4. The median normalized cut value for clusters 2 and 5 was relatively high (figure 6B). Although, the median normalized cut value for cluster 4 was low (6×10^{-3}), for its first 12 weeks (until its merging with clusters 2 and 5 on July 19th, 2020), its normalized cut value was higher (>0.04) than in any subsequent week. The splitting and merging of these clusters along with their high normalized cut values, indicate that there was substantial connectivity between them during this period; sufficient accumulation of mutations had not occurred to support stable clusters.

B. PDS. Clusters were found to mainly split along the lineages B (cluster 3), B.1 (cluster 1), and B.1.1 (cluster 2). Intermittently, a proportion of lineage A also split from the B cluster,

leading to the cluster 4. Figure 6B shows the relatively higher normalized cut value of cluster 4, indicating its instability as a cluster.

4.3.2. November 2020 – May 2021. Beginning November 2020, a B.1.1.7 cluster and several other region-specific clusters emerged:

A. IDS. Between December 20th and 27th, 2020, clusters 7 (lineage B.1.1.7), 8 (lineage B.1.1.294), and 9 (lineage B.1.362) split from cluster 4. Splitting occurred less than 1% (45/5,213), 6% (5/90) and 13% (50/375) of samples of their respective lineages had been collected (table 2). Additionally, 87.3% (4,553/5,213), 87.8% (79/90), and 62.7% (235/375) were collected in this setting (table 2), and the growth of cluster 7 in figure 3C corresponded with the growth of the B.1.1.7 lineage in figure 3A. Indeed, cluster 7 reaches its peak size on the same week that B.1.1.7 becomes the most frequent lineage in the dataset (May 16th, 2021). This is important as tracking of lineages is most relevant when the lineages are "active" in the population (*i.e.*, when samples of a given lineage are being collected). The exception to this is cluster 11 (lineage B.1.351), which split from cluster 4 on February 21st, 2021, when 69.5% (91/131) of the lineage samples had already been collected.

Initially, cluster 7 merged and split with cluster 4 twice (January 17th, 2021, and February 7th, 2021). Relatedly, all normalized cut values above the 75th percentile of cluster 7 occurred during the first 15 weeks of its existence as a separate cluster (figure 6B). Its straightforward to conclude that the connectivity between clusters 7 and 4 fluctuated across time; particularly it decreased when more B.1.1.7 samples were added. An opposite observation can be made regarding cluster 10 - another B.1.1.7 cluster that split from cluster 7 on January 24th, 2021. During the following week, cluster 10 was found to have merged with cluster 7 and did not

split-out again. This phenomenon was due to the fact that additional B.1.1.7 samples increased the connectivity between these two clusters. Generally, the majority of samples within each lineage that correspond to a specific cluster were collected while the cluster persisted in the DEN.

B. PDS. Cluster 5 (lineage B.1.160), seen in DEN beginning November 8th, 2020, is the only example of a cluster with an indegree of 0 when time $t > 1$, *i.e.*, no cluster 5 samples had propagated from the previous week. This is understandable as the first B.1.160 samples (n = 35) were collected on November 8th, 2020. Coalescing of the B.1.160 samples in a cluster of their own demonstrates the ability of the network evolution model to correctly represent completely new data that does not conform to the existing organization of the information. On the same week, cluster 6 (lineage B.1.177), split from cluster 1. Then, on December 20th, 2020, all persisting clusters except for cluster 6 merged into cluster 7. These clusters already had higher normalized cut values (figure 6B), and their proportion of the overall network decreased with the growth of cluster 6 (figure 3D). Beginning December 14th, 2020, lineage B.1.1.7 became its own cluster in two stages. First, cluster 2, the majority B.1.1 cluster, and cluster 8 split from cluster 7. Then, cluster 9 (lineage B.1.1.7) split from cluster 2 when only 4% (170/3,825) of B.1.1.7 samples had been collected (table 3). Subsequently, cluster 2 merged back with cluster 8. Like IDS, cluster 9 merged with the heterogenous cluster 7 in the initial weeks of its persistence. As cluster 9 continued to grow and encompassed a larger proportion of the network, clusters 5 and 6 merged with cluster 7. For lineage collections, 62.9% (193/307) of B.1.160, 73.2% (913/1,247) of B.1.177, and 96.3% (3,683/3,825) of B.1.1.7 samples were collected while each lineage was as a persisting cluster in DEN. Again,

the week B.1.1.7 was the most frequent in the dataset (May 9th, 2021) aligned with the week cluster 9 was the largest in size (figure 3B and D).

4.3.3. May 2021 – April 2022. Smaller, region-specific clusters merged together and B.1.617.2 (Delta) became a cluster:

A. IDS. Cluster 12 (lineage B.1.617.2) split from cluster 4 on May 16th, 2021, when less than 1% (69/8,717) of B.1.617.2 had been collected. In the weeks following, it split again into two further clusters: clusters 13 and 14. Both of these clusters were very small, with an average of 72.6 samples in cluster 13 and 30.2 samples in cluster 14 compared to 417 samples in cluster 12. Clusters 13 and 14 did not grow (figure 3C), but instead merged back with cluster 12 as more B.1.617.2 samples were added. On February 1st and 15th, 2022, cluster 15 split twice from cluster 12. Cluster 15 comprised only of the sub-lineage AY.121 (alias of B.1.617.2.121). Although this splitting indicates a substantial within connectivity of this sub-lineage compared to its connectivity to other sub-lineages, its normalized cut value was still quite high (figure 6B), and so it merged back with cluster 12. Additionally, 100% of samples of lineage AY.121 had already been collected before the split (table 2), further enforcing that this cluster was not sustainable as there was no further growth of the corresponding lineage.

B. PDS. Samples of lineage B.1.617.2 split from cluster 10 on June 27th, 2021, when 13% (1,521/11,574) of B.1.617.2 samples had been collected. After this split, the cluster continued to grow in size to become the dominant cluster (figure 3D), and there were no further splitting or merging events in the remaining of the tracked period.

With few exceptions, lineage groups that became clusters were also the most frequent lineages in the dataset. The exception in PDS was lineage B.1.91 (*n*=207), which reached a frequency of 0.18 on March 22nd, 2020, but never became a cluster. The lineage remained highly connected with other B lineages. On average, a B.1.91 sample had 8.2 neighbors of the same lineage and 5.2 neighbors of another lineage.

Lineage B.1.1.50, the third largest lineage group in IDS (*n*=987) also never became a cluster. Again, this was due to its high connectivity to samples outside its own lineage. For example, on the week of December 27th, 2020, when B.1.1.50 reached its peak of 33% of the dataset (*n*=225), the smaller lineages B.1.362, B.1.1.294, and B.1.1.7 were all clusters. The latter three lineages had an average of 0.4, 0.05, and 0.3, neighbors outside of their lineage, respectively, while B.1.1.50 samples had an average of 3.1 neighbors outside their lineage. This was also seen with the B.1.177 (*n*=167) lineage in IDS which reached a frequency of 3% on January 17th, 2021, higher than the 2% reached by B.1.1.294 (*n*=90). During its peak in the dataset, B.1.177 samples had on average 3.63 neighbors outside their lineage compared to the average of 0.06 in B.1.1.294.

4.4. Comparison with Phylogenetic trees

A DEN models ongoing interactions between viral samples through time and is updated as new data is gathered. Phylogenetic trees on the other hand need to be recomputed when new genomes are collected. In phylodynamics, which involves the prediction of epidemic spread through phylogenetic analysis (Volz *et al*., 2013), the aforementioned limitation persists: when new information is added, predictions or classifications are made solely based on the recomputed tree, with no explicit incorporation of information from the previous tree. By definition, a variant is simply a mutational change in the genome. However, in the

epidemiological and public health context, genomes are grouped into a variant by both shared genotypic and phenotypic traits. Therefore, characterization of variants using mutations is a dynamic process, putatively changing over time as more genomes are collected and a better understanding of the molecular basis of the variant is developed. By modeling variant clustering as dynamic entities, several aspects of variant identification and tracking can be approached from a new perspective – as we show in the following for the problems of variant classification and determination of relationships between variants.

4.4.1. Dynamic clustering of samples. Ideally, emerging variants would form a distinct subgroup, genetically separate from all other samples in the population. In reality, due to homoplasy, reverse mutations, and lack of complete data, there may exist ambiguity in the variant designation of some genomes. For example, a genome can be placed on an isolated branch that is equidistant to multiple clades. The variant designation of this genome is susceptible to change depending on the addition of genomes sampled with the passage of time. The phylogenetic approach requires the continual monitoring of clades being formed as well as the manual upkeep of lineage designations (O'Toole *et al.*, 2021). Another approach, as enacted by us, is to model such movement(s) between clades or clusters as a function of the known viral genomic landscape at that time. Therefore, analysis is focused on the changing viral landscape rather than the classification of each genome. In the directed network, such movement of samples between clusters occurs when a vertex at time t is the receiver of samples from both a split and a merge; that is, the vertex has an indegree greater than one and at least one of its connecting vertices from time t-1 has an outdegree greater than one. This can be seen between the B.1.1.7 cluster and the heterogenous cluster in IDS on

several occasions (January 31st, February 7th, March 14th, and 21st, April 4th, and July 25th, 2021, figure 4). In total, six samples were moved between the two clusters: EPI_ISL_944231 (B.1), EPI_ISL_1278513 (B.1.1.50), EPI_ISL_944228 (B.1), EPI_ISL_944230 (B.1.1.50), EPI_ISL_889113 (B.1.1.50), and EPI_ISL_1278514 (B.1.1). The ambiguity of their clustering is reflected in both the phylogenetic tree (figure 7A), where they share a most recent common ancestor with the B.1.1.7 clade, and the network (figure 7B), where they connect to several vertices in both clusters. In PDS, this also occurs once on January 17th, 2021. Sample EPI_ISL_1138825 (B.1) moves from the B.1.160 to the heterogeneous cluster. Similar to the example in IDS, the sample shares a most recent common ancestor with the B.1.160 clade (figure 7C) and has connections to both clusters in the network (figure 7D). Modeling this uncertainty not only gives an accurate description of the position of those genomes in the genomic space at a given time, but it also provides information on the degree of separation of the clusters that share samples. Although the relationship of clades or clusters can be inferred by the hierarchical structure of the phylogenetic tree, it is assumed that this relationship is not dynamic, *i.e.*, ancestor-descendant clades cannot move closer or further away from each other in time.

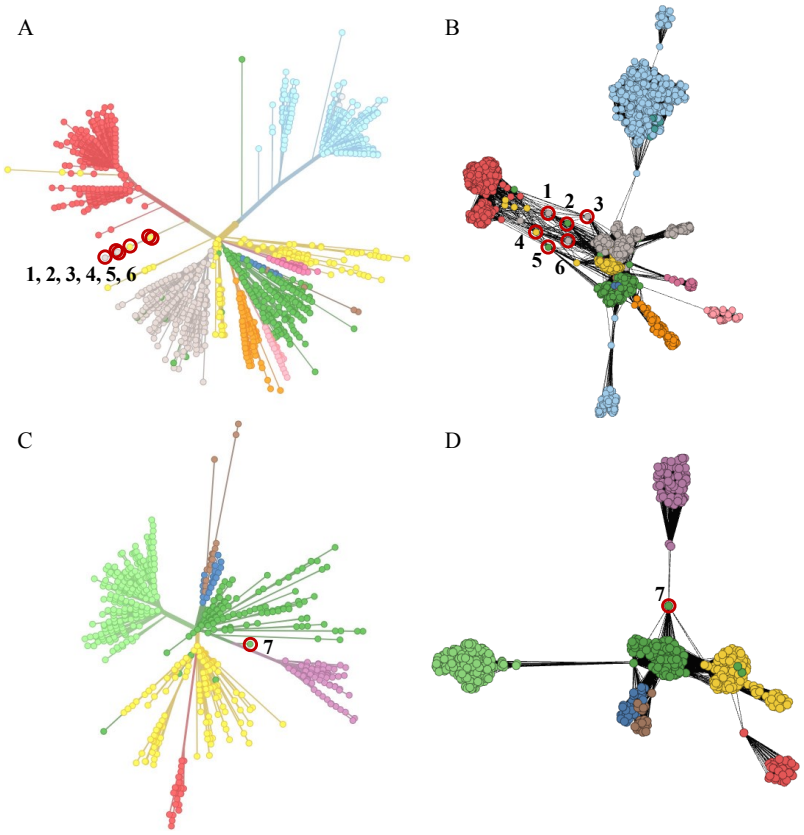

Figure 7. Phylogenetic trees of genomes sampled on July 25th, 2021 (A and B) in IDS, and January 24th, 2021 (C and D) in PDS. The circled and numbered vertices represent samples that were observed to move between clusters. Samples 1-6 (EPI_ISL_944231, EPI_ISL_1278513, EPI_ISL_944228, EPI_ISL_944230, EPI_ISL_889113, and EPI_ISL_1278514) moved between the B.1.1.7 cluster (cluster 7) and the heterogeneous cluster (cluster 4) from March 7th until July 25th, 2021. Sample 7 moved from the B.1.160 (cluster 5) to the heterogeneous cluster (cluster 10) on January 24th, 2021. In each dataset, phylogenetic trees show the labelled samples are situated between clades of the two corresponding clusters, respectively (B and D). Maximum Likelihood phylogenetic trees were built with the Nextstrain CLI (Hadfield et al., 2018).

4.4.2. Dynamic connections between clusters. In Phylogenetics, relationships between clades are established by estimating their most recent common ancestor, which is represented by an internal node connecting the clades in the tree. In our method, connecting vertices of different clusters establish the relationship between them. These vertices represent a "bridge" between two clusters. They are the vertex or vertices in a cluster, who have at least one *k*-nearest neighbor(s) in another cluster. Unlike phylogenetics, the connecting samples are actual genomes. This is beneficial because connections between clusters can be compared across time. The distinctions in groupings does lead to distinctions in interpreting the data. By focusing on the temporal dynamics of the connections between the heterogeneous cluster and the clusters that derive from it due to splitting events (beginning November 2020), different patterns emerge. In several instances, there is a temporal delineation between the connecting samples in the ancestral cluster and the split cluster. Across all time-points, the heterogeneous cluster samples connected to the B.1.617.2 cluster were collected when less than 1% (87/8,717) of B.1.617.2 samples were collected in IDS, and before any were collected in PDS. Lineages B.1.1.294 and B.1.351 became clusters on December 27th, 2020, and February 14th, 2021, respectively, and remained clusters until August 8th, 2021. All the while, they remained connected to the heterogeneous cluster by samples collected before May 2020. In other clusters, such as B.1.362, and B.1.160 and B.1.177, samples within each cluster and those they connected to in the heterogeneous cluster were collected contemporaneously.

Interestingly, connecting samples in the heterogeneous cluster can also be collected several months *after* the last sample in the split cluster was collected. All samples in the B.1.1.7 cluster were collected by July 11th, 2021, and August 8th, 2021, in IDS and PDS,

respectively. Initially, all connecting samples in the heterogeneous cluster were collected before May (IDS) and July (PDS) 2021. However, in January (PDS) and February (IDS) 2022, this changed. During this time, the first B.1.1.529 (Omicron) genome was sampled and was added to the heterogeneous cluster. This sample became the sole connecting sample to B.1.1.7 in both datasets. Up to three additional B.1.1.529 samples were added during the remaining of the tracked period and connected to the B.1.1.7 cluster. It was found that a combination of mutations was common to all connecting B.1.1.529 samples and a majority of samples in the B.1.1.7 clusters. These included C241T, C3037T, and C14408T in ORF1ab, A23063T, A23403G, and C23604A in the spike protein, and G28881A, G28882A, and G28883C in the nucleocapsid. The three nucleocapsid mutations, in particular, are characteristic of B.1.1.7 and are thought to be associated with increased transmissibility (Tao *et al.*, 2021). Individually, a portion of these mutations are found at a high frequency in both datasets (>99%). However, in combination, they are in 93.9% (124/132) and 100% (83/83) of B.1.1.529, 98.8% (5,151/5,213) and 98.8% (3,779/3,825) of B.1.1.7, and less than 0.1% (5/11,181 and 9/15,189) of the remaining samples of IDS and PDS, respectively. In the phylogenetic tree, B.1.1.529 forms a distinct clade that shares a most recent common ancestor with the B.1.1.7 clade approximately at the same time that other B lineages diverged (figure 8A and C). In actuality, the volume of mutations acquired by Omicron compared to other variants has made tracing its evolution difficult. Several theories exist, including recombination, evolution in an isolated population or a long-term infected immunocompromised individual before spreading through a larger population, and reintroduction into a human population after an epizootic (Khandia *et al*., 2022). A Jukes-Cantor neighbor joining phylogenetic model does show a common ancestor between B.1.17

and B.1.1.529, however, other phylogenetic models do not (Kandeel *et al*., 2022). An inherent constraint in modeling a pathogenic viral population through common ancestry is that connections between samples must be made in a hierarchical manner backward in time. Thus, the presence of common mutations between Omicron and Alpha may not be obvious from a tree (figure 8 B and D). Our approach does not rely on estimating common ancestry. Instead, the viral population is tracked based on changes in connectivity between actual sampled genomes. In a case such as B.1.1.7, numerous studies have been published on the increased transmissibility of the lineage by the time the first B.1.1.529 sample was identified in November 2021 (Fisman and Tuite, 2021; Khandia *et al.*, 2022; Kumar *et al*., 2021; Salleh *et al*., 2021). Consequently, a change in connectivity between B.1.1.7 and the ancestral cluster by a recently collected sample in early 2022, months after the last connecting sample was collected, would alert to the new samples as variants to monitor.

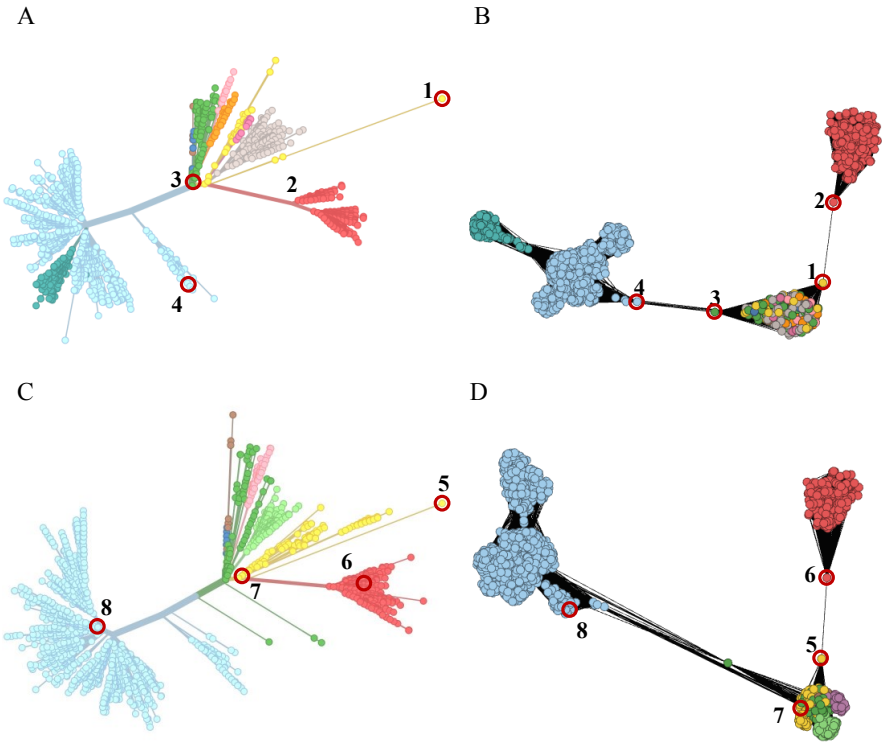

Figure 8. Phylogenetic trees and networks of genomes sampled on February 1st, 2022, in IDS (A and B) and on January 23rd, 2022, in PDS (C and D). The chosen dates represent the first week a B.1.1.529 genome was sampled. The most recently collected samples that connect clusters are numbered and highlighted with a red circle. In IDS, EPI_ISL_12093290 (1) was found to connect the ancestral cluster to the B.1.1.7 cluster by forming an edge with EPI_ISL_1762234 (2), while EPI_ISL_447277 (3) and EPI_ISL_7551993 (4) connected the heterogeneous and larger B.1.617.2 clusters, respectively. In PDS, EPI_ISL_9606418 (5) and EPI_ISL_1116719 (6) connected the heterogeneous cluster with the B.1.1.7 cluster, and EPI_ISL_453976 (7) and EPI_ISL_4107688 (8) connected the heterogeneous cluster with the B.1.617.2 cluster. Only the most recently collected connecting samples are shown for readability.

5. CONCLUSIONS

The application of phylogenetics analysis to the real-time sequencing of SARS-CoV-2 genomes has been crucial for the monitoring of potentially concerning variants during the ongoing pandemic. However, the static and rigid nature of the phylogenetic reconstruction imposes limitations that can have important consequences as discussed earlier in this paper. We have described a novel and general-purpose data representation, the dynamic epidemiological network along with algorithms for its construction, to address this problem. Compared with the batch-mode nature of most phylogenetic and phylodynamic methods, the proposed representation only requires incremental computation of the network for new data and its incorporation in the existing network based on cluster correspondences. As demonstrated by experimental results, the proposed framework is capable of automatically identifying and tracking variants of interest such as Alpha and Delta, along with other region-specific lineages. The experimental studies also

illustrated both the dynamic nature of variant assignment (for specific genomes), as well as the insights that can be gained from observing changes in the spectral partitioning patterns of temporally adjacent networks.

The proposed approach has certain limitations: large datasets were sub-sampled due to the computational requirements of spectral partitioning and subsequent network analysis. The effect of different sampling strategies remains an important question for further research. Currently in the method, identification of a variant group is based, in part, on the assumption that the variant grows substantially (in terms of collected samples) over time. This is a straightforward assumption. However, as the virus becomes endemic, new variant populations may be smaller than previous ones. One solution could involve preferential treatment of more recently collected samples as compared to legacy data. Another solution may consider factors such as infectiousness and virulence of the pathogen. Finally, like any representation method, the proposed framework critically depends on sample collection quality and coverage. For example, despite the rapid spread of the Omicron variant globally by early 2022 (Khandia *et al*., 2022), the variant is underrepresented in our networks due to the lack of Omicron samples from from Israel and Portugal at the time of our data collection (April 2022). This final point underscores the centrality of high-quality disease surveillance and data collection in molecular epidemiology.

6. AUTHORS' CONTRIBUTIONS

RS formulated the problem, designed the algorithmic framework, and provided technical guidance and mentoring. FS was responsible for data collection and coding. The experiments were designed by FS and RS and conducted by FS. The paper was written by RS and FS. Both authors read and approved the final manuscript

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the authors and their originating laboratories responsible for collecting and sharing the genetic sequences and associated metadata via the GISAID Initiative.

8. FUNDING

This research was funded by the National Science Foundation (NSF) grant IIS-1817239.

9. AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

10. REFERENCES

Alotaibi N and Rhouma D. A Review on Community Structures Detection in Time Evolving Social Networks. Journal of King Saud University - Computer and Information Sciences 2022;34(8):5646–5662; doi: 10.1016/j.jksuci.2021.08.016.

Attwood SW, Hill SC, Aanensen DM, et al. Phylogenetic and Phylodynamic Approaches to Understanding and Combating the Early SARS-CoV-2 Pandemic. Nat Rev Genet 2022;23(9):547–562; doi: 10.1038/s41576-022-00483-8.

Campo DS, Xia G-L, Dimitrova Z, et al. Accurate Genetic Detection of Hepatitis C Virus Transmissions in Outbreak Settings. J Infect Dis 2016;213(6):957–965; doi: 10.1093/infdis/jiv542.

Cazabet R and Rossetti G. Challenges in Community Discovery on Temporal Networks. In: Temporal Network Theory. (Holme P and Saramäki J. eds). Computational Social Sciences

Springer International Publishing: Cham; 2019; pp. 181–197; doi: 10.1007/978-3-030-23495-9_10.

Enright J and Kao RR. Epidemics on Dynamic Networks. Epidemics 2018;24:88–97; doi: 10.1016/j.epidem.2018.04.003.

Fiedler M. Algebraic Connectivity of Graphs. Czech Math J 1973;23(2):298–305; doi: 10.21136/CMJ.1973.101168.

Fisman DN and Tuite AR. Evaluation of the Relative Virulence of Novel SARS-CoV-2 Variants: A Retrospective Cohort Study in Ontario, Canada. CMAJ 2021;193(42):E1619–E1625; doi: 10.1503/cmaj.211248.

Fukunaga K and Hostetler L. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE Trans Inform Theory 1975;21(1):32–40; doi: 10.1109/TIT.1975.1055330.

Goldberg Y, Mandel M, Bar-On YM, et al. Waning Immunity after the BNT162b2 Vaccine in Israel. N Engl J Med 2021;385(24):e85; doi: 10.1056/NEJMoa2114228.

Hadfield J, Megill C, Bell SM, et al. Nextstrain: Real-Time Tracking of Pathogen Evolution. Kelso J. ed. Bioinformatics 2018;34(23):4121–4123; doi: 10.1093/bioinformatics/bty407.

Kandeel M, Mohamed MEM, Abd El-Lateef HM, et al. Omicron Variant Genome Evolution and Phylogenetics. J Med Virol 2022;94(4):1627–1632; doi: 10.1002/jmv.27515.

Katoh K, Rozewicki J and Yamada KD. MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization. Briefings in Bioinformatics 2019;20(4):1160–1166; doi: 10.1093/bib/bbx108.

Khandia R, Singhal S, Alqahtani T, et al. Emergence of SARS-CoV-2 Omicron (B.1.1.529) Variant, Salient Features, High Global Health Concerns and Strategies to Counter It amid Ongoing COVID-19 Pandemic. Environ Res 2022;209:112816; doi: 10.1016/j.envres.2022.112816.

Khare S, Gurry C, Freitas L, et al. GISAID's Role in Pandemic Response. China CDC Weekly 2021;3(49):1049–1051; doi: 10.46234/ccdcw2021.255.

Konings F, Perkins MD, Kuhn JH, et al. SARS-CoV-2 Variants of Interest and Concern Naming Scheme Conducive for Global Discourse. Nat Microbiol 2021;6(7):821–823; doi: 10.1038/s41564-021-00932-w.

Kumar V, Singh J, Hasnain SE, et al. Possible Link between Higher Transmissibility of Alpha, Kappa and Delta Variants of SARS-CoV-2 and Increased Structural Stability of Its Spike Protein and HACE2 Affinity. Int J Mol Sci 2021;22(17):9131; doi: 10.3390/ijms22179131.

Leitch J, Alexander KA and Sengupta S. Toward Epidemic Thresholds on Temporal Networks: A Review and Open Questions. Appl Netw Sci 2019;4(1):105; doi: 10.1007/s41109-019-0230-4.

Lentz HHK, Koher A, Hövel P, et al. Disease Spread through Animal Movements: A Static and Temporal Network Analysis of Pig Trade in Germany. Boulinier T. ed. PLoS ONE 2016;11(5):e0155196; doi: 10.1371/journal.pone.0155196.

Little SJ, Kosakovsky Pond SL, Anderson CM, et al. Using HIV Networks to Inform Real Time Prevention Interventions. Harrigan PR. ed. PLoS ONE 2014;9(6):e98443; doi: 10.1371/journal.pone.0098443.

Longmire AG, Sims S, Rytsareva I, et al. GHOST: Global Hepatitis Outbreak and Surveillance Technology. BMC Genomics 2017;18(S10):916; doi: 10.1186/s12864-017-4268-3.

Melnyk A, Mohebbi F, Knyazev S, et al. From Alpha to Zeta: Identifying Variants and Subtypes of SARS-CoV-2 Via Clustering. Journal of Computational Biology 2021;28(11):1113–1129; doi: 10.1089/cmb.2021.0302.

Nunner H, Buskens V and Kretzschmar M. A Model for the Co-Evolution of Dynamic Social Networks and Infectious Disease Dynamics. Comput Soc Netw 2021;8(1):19; doi: 10.1186/s40649-021-00098-9.

O'Toole Á, Scher E, Underwood A, et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. Virus Evolution 2021;veab064; doi: 10.1093/ve/veab064.

Poon AFY, Gustafson R, Daly P, et al. Near Real-Time Monitoring of HIV Transmission Hotspots from Routine HIV Genotyping: An Implementation Case Study. The Lancet HIV 2016;3(5):e231–e238; doi: 10.1016/S2352-3018(16)00046-1.

Qin L, Ding X, Li Y, et al. Co-Mutation Modules Capture the Evolution and Transmission Patterns of SARS-CoV-2. Briefings in Bioinformatics 2021;22(6):bbab222; doi: 10.1093/bib/bbab222.

Ramachandran S, Thai H, Forbi JC, et al. A Large HCV Transmission Network Enabled a Fast-Growing HIV Outbreak in Rural Indiana, 2015. EBioMedicine 2018;37:374–381; doi: 10.1016/j.ebiom.2018.10.007.

Rambaut A, Holmes EC, O'Toole Á, et al. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. Nat Microbiol 2020;5(11):1403–1407; doi: 10.1038/s41564-020-0770-5.

Roemer C, Hodcroft, EB, Neher RA, et al. Nextstrain. 2022. Available from: https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022.

Ruget A-S, Rossi G, Pepler PT, et al. Multi-Species Temporal Network of Livestock Movements for Disease Spread. Appl Netw Sci 2021;6(1):15; doi: 10.1007/s41109-021-00354-x.

Saban M, Myers V and Wilf-Miron R. Changes in Infectivity, Severity and Vaccine Effectiveness against Delta COVID-19 Variant Ten Months into the Vaccination Program: The Israeli Case. Preventive Medicine 2022;154:106890; doi: 10.1016/j.ypmed.2021.106890.

Salleh MZ, Derrick JP and Deris ZZ. Structural Evaluation of the Spike Glycoprotein Variants on SARS-CoV-2 Transmission and Immune Evasion. Int J Mol Sci 2021;22(14):7425; doi: 10.3390/ijms22147425.

Senchyna F and Singh R. Analysis of SARS-CoV-2 Temporal Molecular Networks Using Global and Local Topological Characteristics. In: Computational Advances in Bio and Medical Sciences. (Bansal MS, Măndoiu I, Moussa M, et al. eds). Lecture Notes in Computer Science Springer International Publishing: Cham; 2022; pp. 149–162; doi: 10.1007/978-3-031-17531-2_12.

Singh R, Beasley R, Long T, et al. Algorithmic Mapping and Characterization of the Drug-Induced Phenotypic-Response Space of Parasites Causing Schistosomiasis. IEEE/ACM Trans Comput Biol Bioinform 2018;15(2):469–481; doi: 10.1109/TCBB.2016.2550444.

Tao K, Tzou PL, Nouhin J, et al. The Biological and Clinical Significance of Emerging SARS-CoV-2 Variants. Nat Rev Genet 2021;22(12):757–773; doi: 10.1038/s41576-021-00408-x.

Torgerson WS. Multidimensional Scaling: I. Theory and Method. Psychometrika 1952;17(4):401–419; doi: 10.1007/BF02288916.

Valdano E, Ferreri L, Poletto C, et al. Analytical Computation of the Epidemic Threshold on Temporal Networks. Phys Rev X 2015;5(2):021005; doi: 10.1103/PhysRevX.5.021005.

Villabona-Arenas ChJ, Hanage WP and Tully DC. Phylogenetic Interpretation during Outbreaks Requires Caution. Nat Microbiol 2020;5(7):876–877; doi: 10.1038/s41564-020-0738-5.

Volz EM, Koelle K and Bedford T. Viral Phylodynamics. PLoS Comput Biol 2013;9(3):e1002947; doi: 10.1371/journal.pcbi.1002947.

Weng S, Shang J, Cheng Y, et al. Genetic Differentiation and Diversity of SARS-CoV-2 Omicron Variant in Its Early Outbreak. Biosaf Health 2022;4(3):171–178; doi: 10.1016/j.bsheal.2022.04.004.

Wertheim JO, Kosakovsky Pond SL, Forgione LA, et al. Social and Genetic Networks of HIV-1 Transmission in New York City. Bonhoeffer S. ed. PLoS Pathog 2017;13(1):e1006000; doi: 10.1371/journal.ppat.1006000.

World Health Organization (WHO). Tracking SARS-CoV-2 Variants. 2022. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants. [Last accessed: 11/27/2022]

Wu F, Zhao S, Yu B, et al. A New Coronavirus Associated with Human Respiratory Disease in China. Nature 2020;579(7798):265–269; doi: 10.1038/s41586-020-2008-3.

Zarrabi N, Prosperi M, Belleman RG, et al. Combining Epidemiological and Genetic Networks Signifies the Importance of Early Treatment in HIV-1 Transmission. Khudyakov YE. ed. PLoS ONE 2012;7(9):e46156; doi: 10.1371/journal.pone.0046156.