Advances, Challenges and Opportunities of Phylogenetic and Social Network Analysis Using COVID-19 Data

Yue Wang[1,*], Yunpeng Zhao[1,*], and Qing Pan[2,*]

[1]School of Mathematical and Natural Sciences, Arizona State University, 4701 W Thunderbird Rd, Glendale, AZ 85306, USA
[2]Department of Statistics, George Washington University, 801 22nd St. NW, 20052, Washington DC, USA

Abstract

COVID-19 has attracted research interests from all fields. Phylogenetic and social network analyses based on connectivity between either COVID-19 patients or geographic regions and similarity between SARS-CoV-2 sequences provide unique angles to answer public health and pharmaco-biological questions such as relationships between various SARS-CoV-2 mutants, the transmission pathways in a community, and the effectiveness of prevention policies. This paper serves as a systematic review of current phylogenetic and social network analyses with applications in COVID-19 research. Challenges in current phylogenetic network analysis on SARS-CoV-2 such as unreliable inferences, sampling bias and batch effects are discussed as well as potential solutions. Social network analysis combined with epidemiology models helps to identify key transmission characteristics and measure the effectiveness of prevention and control strategies. Finally, future new directions of network analysis motivated by COVID-19 data are summarized.

Keywords: batch effects, control policy, epidemiology model with network topology, network characteristics, phylogenetic tree, sampling bias

*These authors contributed equally

1 Introduction

The global pandemic of coronavirus disease 2019 (COVID-19), caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has infected over 197 million people worldwide as of July 2021. Coronaviruses are single-stranded RNA viruses that cause respiratory, gastrointestinal and neurological diseases (Zhu et al., 2020). Coronaviruses have caused extremely infectious diseases with severe outcomes in the past 20 years including severe acute respiratory syndrome (SARS) in 2003 (Zhong et al., 2003), middle east respiratory syndrome (MERS) in 2012 (Zaki et al., 2012) and the current COVID-19 pandemic. The most common way of SARS-CoV-2 transmission is through respiratory droplets during face-to-face exposure or contaminated surfaces. Exposures to symptomatic patients are associated with higher risk for transmission but asymptomatic and presymptomatic carriers can also transmit SARS-CoV-2 (Ganyani et al., 2020). Among the infected people, approximately 17% requires intensive care units (ICU) services with impaired functions of the brain, heart, lung, liver, kidney or coagulation system (Docherty et al., 2020). The prevalence and prognosis of COVID-19 infections differ by race (Garg et al., 2020; Price-Haywood et al., 2020) and age (Richardson et al., 2020) with a 8700 times higher mortality rate in the $85+$ age group compared to that in the $5-17$ age group.

Network analysis has been a fast developing research field with diverse applications in public health (Harris and Clements, 2007), biology (Tringali et al., 2020), communication (Hagen et al., 2018), economics (Jackson, 2010), information theory (El Gamal and Kim, 2011), political science (Ward et al., 2011), and computer science (Getoor and Diehl, 2005) et al. Contrary to most statistical methods studying properties of individuals without considering impacts from others, network analysis study relationships (e.g. contact, interaction, transmission, similarity) between nodes (e.g. virus samples or patients or geographic regions). Researchers study underlying properties of networks such as network connectivity (McPherson et al., 2001), distributions of ties from a node (Opsahl et al., 2010), group structures among nodes (Holland et al., 1983).

Visualization tools (Freeman, 2000) are commonly used to illustrate the patterns and inferences from network analysis.

COVID-19 has attracted large amounts of attention in almost all research fields from all over the world in a short period of time. Among various research tools, network analysis provides a unique angle to study the relationships between regions, organizations, people, virus sequences, genes, proteins, molecules et al. It has developed into many research domains such as gene expression networks based on omics data (Horvath, 2011), network analysis using natural language processing tools to extract information from social media (Bail, 2016; Hung et al., 2020), ontology network analysis (Alani et al., 2003), protein-protein interaction studies with applications in drug design (Murakami et al., 2017), network based systems pharmacology (Zhao and Iyengar, 2012) and many more. This review focuses on network analysis using either virus genomics data or patient interaction data. Two main types are discussed - phylogenetic network analysis and social network analysis. Phylogenetic network analysis based on similarities of different SARS-CoV-2 genomes estimate the evolutionary relationship among various SARS-CoV-2 sequences. Social network analysis considers the interaction between individual patients or similarity between geographic regions, and discloses the underlying community structure or infectious pathway of COVID-19 transmissions in human society. Ultimately, network analysis based on COVID-19 data will provide evidence for policy makers in choosing effective prevention and control measures, help individuals avoid high risk events as well as shed light on proteins or RNA sequences that may serve as therapeutic targets in bio-pharmaceutical exploration of COVID-19 vaccines and treatments.

2 Challenges in Phylogenetic Network Analysis Using Virus Genomes from COVID-19 Patients

Recent advances on phylogenetic analysis of SARS-CoV-2 genome sequences gain insights into the evolutionary relationships of the SARS-CoV-2 strains identified worldwide (Wang et al., 2020; Bai et al., 2020; Worobey et al., 2020; Li et al., 2020; Mavian et al., 2020a; Forster et al., 2020; Kemenesi et al., 2020; Morel et al., 2021; Zehender et al., 2020). A selected review of the scientific findings of these studies is given in Table 1. These scientific findings are based on phylogenetic analyses, which construct phylogenetic trees or networks with nodes representing genome sequences and edges between nodes representing evolutionary relationships between sequences. Fig. 1 is a flowchart illustrating steps in a typical phylogenetic analysis of SARS-CoV-2 genome sequences.

The first step is to obtain a data set consisting of SARS-CoV-2 genome sequences of interest. This can be done by either wet-lab sequencing of virus samples from COVID-19 patients or retrieving existing COVID-19 genome sequences from public databases (e.g., the *gisaid* database). After a data set is assembled, the next step is to perform multiple sequence alignment (MSA) that arranges the sequences in a matrix to identify regions of homology. Existing tools for MSA are rich, including T-Coffee (Notredame et al., 2000), MUSCLE (Edgar, 2004), Cluster Omega (Sievers et al., 2011), MAFFT (Katoh and Standley, 2013), etc. However, different MSA strategies (e.g., whether or not to use outgroups) can impact downstream phylogenetic analyses differently; see the discussions in Morel et al. (2021) for more details. One can also refer to Kemena and Notredame (2009); Thompson et al. (2011); Chatzou et al. (2016) for more extensive reviews of MSA. Next, statistical methods are applied to determine the tree topology and calculate the branch lengths that best describe the phylogenetic relationships of the aligned sequences. Such statistical tools can be roughly divided into two categories: model-based methods (Bayesian or frequentist) and distance-based methods. Model-based methods use probabilistic models to as-

sign scores (likelihoods) to all possible trees. Then, the tree with the highest score or one among the top scored trees with biological significance is deemed the optimal choice. Distance-based methods measure pairwise genetic distances of the aligned sequences and generate a dendrogram from this distance matrix as an estimate of the phylogenetic tree. In cases where no dendrogram fits the distance perfectly, some optimality criterion, such as minimum evolution (ME) (Price et al., 2009), are employed to determine the optimal dendrogram. Model-based methods are generally more accurate but computationally intensive, whereas distance-based methods have opposite features. Potential complexities and issues exist in each of the steps, which may lead to spurious conclusions if not handled properly. In the following sections, we will first review popular statistical methods for phylogenetic inference and highlight challenges for each of them. Next, we will discuss potential data issues, including sampling bias, missing data, and batch effects. Finally, we discuss additional challenges in phylogenetic research on SARS-CoV-2 genomes, which arise from the molecular features of SARS-CoV-2 variants.

2.1 Inferences from phylogenetic analysis

Selecting an appropriate statistical method is fundamental to accurate phylogenetic inference. In any model-based phylogenetic analysis, the substitution model, a Markov model that describes evolutionary changes in genome sequences, plays a central role. Popular substitution models include the simple Jukes and Cantor's model (Jukes et al., 1969), the more complex General Time Reversible (GTR) model and its variants (Tavaré et al., 1986; Yang, 1994), the Hasegawa-Kishino-Yano model (HKY) (Hasegawa et al., 1985), and the unrestricted model (Zharkikh, 1994). In general, the complexity of the substitution model increases with the number of substitution parameters, which characterize heterogeneous substitution rates depending on the source and target nucleotide (Nascimento et al., 2017). However, fitting parameter-rich models is computationally intensive. Moreover, some substitution parameters may be unidentifiable, especially in the analysis of highly similar sequences (e.g., the COVID-19 genome sequences). The non-

identifiability may cause the iterative fitting process failing to converge . While a Bayesian procedure can alleviate this convergence issue by incorporating prior information, the resulting parameter estimation may be mainly driven by the prior but not the data, which will lead to misleading results if the prior does not match the data (Rannala, 2002). On the other hand, any simplistic model or under-parameterization can lead to incorrect inference of tree topology and biased estimates of branch lengths (Yang, 1996; Huelsenbeck and Rannala, 2004). Existing software for selecting a substitution model, such as jModelTest (Darriba et al., 2012) and Modelgenerator (Keane et al., 2006), examines standard goodness-of-fit statistics, e.g., the Akaike information criterion (AIC) (Akaike, 1998) and the Bayesian information criterion (BIC) (Schwarz et al., 1978). These statistics can, to varying degrees, measure how a model fits the data, but do not guarantee that the selected model is the optimal one (in terms of the trade-off between bias and computation expense). For example, when analyzing highly similar sequences, information in the sequences is too limited to fit any parameter-rich model. In this case, parameter-rich models may still yield slightly better goodness of fit compared over simpler models (e.g., the Jukes and Cantor's model). But given the computation expense and potential identifiability issues of parameter-rich models, simple models are often preferred in such cases. An additional challenge for model-based methods is the computational feasibility when the number of sequences and/or the number of genome sites queried per genome increase. This computational issue is in fact critical for COVID-19 phylogenetic research, because to date, more than 1.8 million COVID-19 genome sequences obtained by high-resolution sequencing technologies are available in the *gisaid* database, providing a unique opportunity for a comprehensive understanding of the evolution of COVID-19. However, since the number of possible trees grows super-exponentially with the number of sequences (Roch, 2006), an exhaustive search over all possible trees to find the optimal one is computationally infeasible even when analyzing hundreds of sequences. Previous efforts for efficient parallel computation and optimization (Aberer et al., 2014; Ogilvie et al., 2017) may help alleviate the computational burden. Moreover, since there are a large number

of invariant sites in the genome sequence, excluding less important ones (often called “tree thinning") can accelerate the computation, where the importance of genome sites may be inferred from molecular studies on SARS-CoV-2. Such tree thinning strategy has been adopted in many phylogenetic applications (Prosperi et al., 2011; Ragonnet-Cronin et al., 2013), but inappropriate implementation of thinning algorithms may compromise data quality, thus leading to incorrect phylogenetic inference (Morel et al., 2021).

Distance-based methods are fast alternatives to model-based methods, but they also have complexities in selecting appropriate pairwise genetic distance measures and efficient algorithms to infer the dendrogram. A popular genetic distance measure between two aligned sequences is the fraction of mismatches at aligned positions, also known as the Hamming distance (Mount and Mount, 2001; Norouzi et al., 2012). Other genetic distance measures, including Nei’s genetic distance (Nei, 1987), Cavalli-Sforza chord distance (Cavalli-Sforza and Edwards, 1967), and the classical Euclidean distance, also have varying degrees of success in phylogenetic applications. Nonetheless, any distance-based method can suffer from information loss because distance-based methods do not use data of individual genome sites directly. Moreover, since early changes in ancestral lineages may be erased by later changes (often referred to as back mutations, Ellis et al. (2001)), any pairwise genetic distance measure may underestimate the true phylogenetic distance. To alleviate this issue, one could correct such biased distances by either assigning more weights to distantly related sequences or using a substitution model (e.g., the aforementioned Jukes and Cantor’s model) to get corrected distances (Felsenstein and Felenstein, 2004). With a “good" distance correction, the next step is to use an efficient algorithm for phylogenetic inference. Popular algorithms include the unweighted or weighted pair group method with arithmetic mean (UPGMA or WPGMA) (Sokal, 1958), neighbor-joining (NJ) (Saitou and Nei, 1987), median-joining (MJ) (Bandelt et al., 1999), and the Fitch-Margoliash method (FM) (Fitch and Margoliash, 1967). All these methods can efficiently handle many sequences but still suffer from their own limitations. Specifically, the UPGMA and WPGMA assume an ultrametric

tree, i.e., a tree where all the path-lengths from the root to the tips are equal, which is seldom satisfied in real applications. The NJ lacks a tree search criterion, so its estimated tree is not guaranteed to best fit the distances. This issue was addressed by the Fitch-Margoliash method (FM) that uses the least-squares criterion to ensure the optimality of the estimated tree (Fitch and Margoliash, 1967). However, since finding the optimal least-squares tree is generally NP-complete (Day, 1987), the FM method can be less efficient than NJ. The MJ method has been one of the most popular methods for phylogenetic inference in recent decades, but it has been criticized as “neither phylogenetic nor evolutionary" because of its distance-based nature and the lack of rooting (Kong et al., 2016; Sánchez-Pacheco et al., 2020). However, as far as we understand, the primary difference between distance-based methods and model-based methods is whether data of individual genome sites are fit to the tree, which does not necessarily mean that distance-based methods are less phylogenetic. Also, even for phylogenetic trees inferred using model-based methods, we root them after the analysis by defining one leaf as an outgroup; such outgroup rooting can also be applied to MJ.

Many of the aforementioned model-based and distance-based methods have been successfully applied in existing phylogenetic research on SARS-COV-2 genomes; see Table 1 for a selected review. However, we notice that many of these studies were conducted using default software settings without carefully checking model assumptions, potentially leading to unreliable inference. For example, in maximum-likelihood-based inference, the likelihood function may exhibit a multitude of local optima. Thus, different initial values of the model parameters may yield different tree topology (Morel et al., 2021). In Bayesian phylogenetic inference, misspecified prior may lead to heavily biased estimates of branch lengths (Nascimento et al., 2017). Moreover, all the trees in these studies are provided without any associated uncertainty measures. Therefore, it is unclear to what confidence level, readers can trust the inferred trees.

Paper	Methods	Major Findings
Forster et al. (2020)	MJ: The Hamming distance was used.	Three SARS-CoV-2 types (A, B, and C) were identified: types A and C circulate in Europeans and Americans; type B circulate in East Asians; type A was identified as the ancestral type.
Zehender et al. (2020)	HKY: A proportion of invariant sites were included.	SARS-CoV-2 was present in Italy weeks before the first reported case of infection in China.
Bai et al. (2020)	GTR: Gamma distributed variation rate among sites was assumed.	A haplotype-based phylogenetic analysis suggested that the United States and Australia are the most likely places where SARS-CoV-2 originated.
Worobey et al. (2020)	GTR: Inverse Gaussian distributed variation rate among sites was assumed.	Introductions of the virus from China to both Italy and United States founded the earliest sustained European and North America transmission networks.
Li et al. (2020)	GTR and NJ: The two methods yielded consistent results.	The human SARS-CoV-2 virus, which is responsible for the recent outbreak of COVID-19, did not come directly from pangolins.

Table 1: A selected review of existing phylogenetic research on SARS-CoV-2 genomes: statistical methods and scientific findings.

2.2 Sampling Bias and Missing Data

Many existing phylogenetic studies were performed based on samples from the database (Bai et al., 2020; Mavian et al., 2020a; Kemenesi et al., 2020). Thus, sampling bias may arise, due to the lack of sampling from certain areas or during certain time period. Moreover, coronavirus strains from less developed areas with limited medical resources or access to sequencing equipments may have fewer number of records in the database. For example, according to the country submission data in the *gisaid* database (`https://www.gisaid.org/hcov19-variants/`), 75% of the genome sequences of the lineage B.1.617 (that is, the Delta variant), a variant of COVID-19 virus first detected in India, were submitted by European or North American countries, whereas only 0.15% were submitted by African countries. In fact, even for the lineage B.1.351, a variant first detected in South Africa, only 24.7% of the

genome sequences in the *gisaid* database were submitted by African countries, whereas European countries submitted more than 50% of the sequences. This indicates that there likely exist transmission lines which are never detected or recorded in the less represented areas with few sequence data, causing non-ignorable missingness in the samples. These data quality issues may strongly compromise the completeness and accuracy of phylogenetic inference (Vakulenko et al., 2019; Mavian et al., 2020b).

While carefully balancing samples across different regions may alleviate these data quality issues, this may be unrealistic given the current situation of the pandemic. An alternative way is to increase the number of sequences in the analysis, which may be advantageous for phylogenetic inference (Pollock et al., 2002). However, this exacerbates the computational burden of phylogenetic inference because the number of possible tree topology grows super-exponentially with the number of sequence (Roch, 2006), as we discussed in the previous section. In addition, existing statistical methods may help reduce the sampling bias. For example, if some viral clades of the coronavirus are underrepresented and the degree of underrepresentation can be quantified via external data, then incorporating appropriate sample weights into phylogenetic inference may help reduce the bias (Huang et al., 2006). Popular weighting schemes include the inverse probability weighting (IPW) and its variants (Wooldridge, 2007; Seaman and White, 2013; Mansournia and Altman, 2016), which inflate the weight for underrepresented sequences. Theoretically speaking, IPW consists of two steps. In the first step, we estimate the propensity score, i.e., the probability of a unit being sampled, using statistical models or empirical estimates based on external data. For example, to quantify the sampling rate of the SARS-CoV-2 genome sequences in each country or region, one could first estimate the total number of COVID-19 cases by the ratio between the total number of reported COVID-19 cases and the estimated percentage of cases getting reported. Then, the sampling rate could be estimated by the ratio between the number of deposited SARS-CoV-2 genome sequences and the estimated total number of COVID-19 cases. In the second step, one could create a “representative" sample by assigning each sequence a weight equal to

the inverse of the sampling probability in the country or region where the sequence data was collected. Finally, one could construct a phylogenetic tree based on the weighted sample. However, IPW has limited applicability in the absence of external data quantifying levels of representation. In such cases, a broad class of distance-based weighting schemes that characterize distances among the sequences may be employed (e.g., Vingron and Argos (1989); Sibbald and Argos (1990), and Henikoff and Henikoff (1994)). Consider n sequences with $d(i, j)$ denoting some valid distance measure between sequence i and sequence j. A typical distance-based weighting scheme weights sequence i by $w_i(\lambda) = 1/\sum_{j=1}^{n} I\{d(i, j) \geq \lambda\}$ for some pre-specified threshold $\lambda > 0$, where $I\{A\} = 1$ if A is true and $I\{A\} = 0$ otherwise. Under this weighting scheme, highly unique sequences are given high weights, whereas sequences that are similar to others are assigned low weights (Hockenberry and Wilke, 2019). However, any distance-based weighting scheme should be used with caution because the distance may not be consistent with the intrinsic phylogenetic distance between sequences. Nonetheless, developing efficient methods for integrating weighting schemes into phylogenetic inference is a fruitful future research direction.

2.3 Batch effects

Non-negligible batch effects, i.e., measurements that behave differently under different conditions with potentials to confound the outcome of interest, reflect a common issue in high-throughput data analysis (Leek et al., 2010). Batch effects may be further aggravated when samples are obtained from multiple runs in different labs with different sequencing technologies and/or platforms. This is the case in many existing phylogenetic studies on COVID-19 (Bai et al., 2020; Kemenesi et al., 2020; Mavian et al., 2020a), in which samples were drawn directly from public databases where sequences were shared by various research institutes. Samples within a single lab may also suffer from batch effects due to changes in personnel, storage, or processing time (Leek et al., 2010). Published studies have demonstrated that batch effects can lead to increased variability, decreased power, or spurious biological conclusions in biomarker

detection (Petricoin III et al., 2002; Akey et al., 2007; Leek and Storey, 2007; Spielman et al., 2007). In particular, current research on SARS-CoV-2 genomes detected potential batch effects and highlighted the importance of addressing such batch effects to achieve scientifically meaningful outcomes (Wu et al., 2020; Song et al., 2020; Ravindra et al., 2020; Han et al., 2021). Though little research has examined to what extent batch effects may influence phylogenetic inference, intuitively, batch effects can mislead phylogenetic inference through inflated correlations within sequences from the same batch or attenuated correlations between sequences from different batches regardless of the phylogeny (Gu, 2016). Below we discuss several existing experimental and computational tools for the removal of batch effects.

While challenging to implement, standardizing experimental procedures across the whole COVID-19 research community can reduce batch effects. If changes in personnel, reagents, storage, or technology are inevitable, such information should also be recorded and shared to the public. However, even in a perfectly designed and documented study, it is impossible to record all potential sources of batch effects. Thus, statistical modeling solutions are needed to reduce the impact of both recorded and latent batch effects. The first step in a typical statistical analysis of batch effects is to identify batch effects using exploratory (unsupervised) tools, such as principal component analysis (Abdi and Williams, 2010), multi-dimensional scaling (Chen et al., 2007), and hierarchical clustering (Sneath et al., 1973). In particular, hierarchical clustering of sequences labelled with recorded sources of batch effects can reveal whether the major differences among sequences are due to biology or batch (Leek et al., 2010). One can further plot individual variants versus known batch variables to investigate which variant is correlated with certain batch. If strong batch effects exist, they should be accounted for in downstream phylogenetic analysis. As far as we know, no existing methods for removing batch effects are tailored to phylogenetic inference, but plenty of methods have been proposed for modeling batch effects in regression settings. The simplest approach to model known batch effects in regression models is to include them as covariates (Johnson et al., 2007; Scherer, 2009). When the true

sources of batch effects are largely unknown, one may instead use the surrogate variable analysis (SVA) (Leek and Storey, 2007; Leek et al., 2012) to estimate the sources of batch effects from the input data. These methods have been implemented in various sequencing studies (e.g., Sun et al. (2011); Jaffe et al. (2015); Gibbons et al. (2018)), but future work is needed to extend these methods to phylogenetic inference.

2.4 Additional challenges in phylogenetic analysis of SARS-CoV-2 genomes

In this section, we briefly discuss two additional challenges in COVID-19 phylogenetic research. First, the SARS-CoV-2 accumulates only two single-letter mutations per month in its genome, a rate of change about half the rate of influenza and one-quarter the rate of HIV (Callaway, 2020). Thus, genome sequences of SARS-CoV-2 variants are highly similar, introducing difficulties to the selection of substitution models (see Section “Inferences from phylogenetic analysis" for more details). Second, similar to influenza viruses, different SARS-CoV-2 genome segments can re-assort among related strains (Shafique et al., 2020). This indicates that different SARS-CoV-2 genome segments may have different phylogenetic tree topology. Therefore, it may be beneficial to perform phylogenetic analysis separately for each genome segment, which is often termed the partitioned analysis (Bull et al., 1993), accounting for the heterogeneity in the evolution of SARS-CoV-2.

3 Social Network Analysis of Covid Patients

3.1 Empirical study of COVID-19-related networks

We use the term *empirical study of networks* to refer to research that utilizes measures calculated from network topology, such as degrees and various centrality measures, to study transmissions of COVID-19. We list a few typical measures below and the readers are referred to part II and III

Category	Measure	Meaning in Infection Networks
Node characteristic	In-degree	The number of possible sources of infections a patient had contacted, which is one if the source was confirmed.
	Out-degree	The number of individuals infected by the patient, which measures the infectious power of the patient.
	Betweenness centrality	The number of chains of infection that pass through the patient.
	Closeness centrality	The average number of intermediate steps in infection chains from a patient to other patients in the network.
Network characteristic	Degree distribution	The fraction of patients in the network with a certain in/out-degree. The tail of the distribution of out-degrees measures the proportion of super-spreaders in the network.
	Average path length	The average number of intermediate steps in all infection chains.
	Diameter	The maximum number of intermediate steps in all infection chains.

Table 2: Commonly used measures in social network analysis and their meanings in infectious networks.

in Newman (2010) for a more comprehensive introduction. The networks considered in studies of infectious diseases are typically directed graphs, in which each edge is associated with a direction that indicates the order by which virus or infectious status was passed (Jo et al., 2021a). The measures listed below are defined for directed networks.

- The *in-degree* of a node is the number of arrows adjacent to the node, i.e., the number of incoming links to it. In an infection network, the in-degree of a patient is not necessarily equal to one if the patient had confirmed contact with more than one infectious patients and the source is uncertain (Saraswathi et al., 2020).

- The *out-degree* of a node is the number of outgoing links from the node, which can be used to measure the infectious power of a patient (Jo et al., 2021a; Saraswathi et al., 2020). Nodes with an out-degree above a certain threshold, for example five, are defined as a

super-spreader (Saraswathi et al., 2020).

- *Degree distribution* is the empirical probability distribution of node degrees over the entire network, which is one of the most fundamental network properties (Barabási and Albert, 1999; Newman, 2010). Studies on infection networks are particularly interested in the out-degree distribution as it impacts the infection status of a society (Meyers et al., 2005; Jo et al., 2021a).

- *Node centrality* measures the importance of each node in a network (Newman, 2010). There exist various versions of centrality, such as *degree centrality* (same as node degree), *betweenness centrality*, and *closeness centrality*, which measure different aspects of the word "importance". For example, the betweenness centrality of a node is the number of times that shortest paths pass through this node, which reflects its ability of forming bridges between other nodes. It is worth mentioning that in an infection network where all links are from confirmed infection routes (i.e., a tree network) (Jo et al., 2021a), betweenness centrality simply reflects the depth of a node. See Table 2 for more centrality measures and their meanings in the context of infection networks. It is worth mentioning that degree centrality as a centrality measure is a sub-category of node centrality, while node degree itself is a fundamental concept in graph theory.

- *Average path length* is the average of the shortest path lengths for all possible pairs of network nodes (Albert and Barabási, 2002). When a network is not fully-connected, which is the typical case if there exist multiple infection sources, the definition can be modified as the average of the shortest path lengths for all *connected pairs* (Jo et al., 2021a).

- *Network diameter* is the shortest path length between the two most distant nodes in a network, which can also be adjusted to only including pairs that are connected (Newman, 2010). The average path length and network diameter in an infection network can be used to measure the potential range of infection (Jo et al., 2021a).

Saraswathi et al. (2020) performed the network analysis of COVID-19 outbreak in Karnataka, India. The data were constructed using contact tracing details released online by the government of Karnataka, India. They analyzed various measures such as node degrees and betweeness centrality across different demographic groups (i.e., genders and ages) and concluded that geographic, demographic and community characteristics could influence the spread of COVID-19. For example, the paper reported that men had higher mean out-degree while women have higher mean betweenness centrality. Women therefore played a significant bridging role in connecting clusters.

Jo et al. (2021a) performed the analysis of an infection network in Seoul metropolitan areas, South Korea. The data were collected by the Seoul, Gyeonggi-do, and Incheon local governments in South Korea and publicly accessible. The analysis focused on the out-degree of each node and its distribution, the average path length, and the network diameter, and further studied the impact of removing the nodes with out-degrees above a certain threshold which varied from 51 to 1, and implementing different government policies. They concluded that out-degrees follow a power-law distribution, which is in line with the findings in other social network studies (Barabasi, 2005). Furthermore, removing nodes with high out-degrees can significantly decrease the size of the infection network and policies such as social distancing can reduce the infectious power.

Jo et al. (2021b) performed a regression analysis to study the spatial proliferation of COVID-19 at the county level in South Korea, using population density and four types of centrality measures including degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality as explanatory variables. The data are available in the Korean Public Data Portal, Korean Statistical Information Service, and Korea Transport Data Base.The study reported that degree centrality was more positively impacted by COVID-19 infection, measured by the number of cases or the number of cases per 10,000 residents, than population density, measured by the standardized coefficients of these two factors. They therefore suggested that mitigation strategies that take into account network structure might be helpful to control the outbreak of the disease.

Network visualization, which maps network topology onto a Euclidean space (usually 2D space), is another popular tool for exploratory analysis of networks (Komarek et al., 2015). A typical plot of a network consists of nodes connected by lines (with arrows if edges are directed). It is worth mentioning that the coordinates of the nodes are usually not a part of the raw data, but are determined by certain layouts. The most commonly-used layout algorithm is the Fruchterman-Reingold algorithm (Fruchterman and Reingold, 1991). Gephi (Bastian et al., 2009) in Java and igraph (Csardi and Nepusz, 2006) in R are open sources software packages for network analysis and visualization. A few research papers have used network visualization to understand networks related to COVID-19. Saraswathi et al. (2020) used various plots to show demographic information of nodes, sources of infection, centrality by different colors, shapes and node sizes, respectively. Furthermore, they visualized dynamic evolution of an infection network by series of plots, each for a different phase.

So et al. (2020) provided a visualization of the domestic and international spread of COVID-19, where nodes represent regions, such as countries at the international level and provinces in the national level, and the link between node i and j represents the correlation between the changes of case numbers in country/province i and country/province j.

3.2 Epidemic models on networks

In this section, we review model-based approaches to COVID-19-related dynamic processes on networks. The ultimate goal of studying networks is to better understand the behavior of the complex systems represented by networks (Newman, 2010). In the context of COVID-19 research, the focus is to understand disease transmission on networks with various topological structure and the impact of human behavior and policy implementation on the spread of SARS-CoV-2.

In traditional epidemiology theory, the majority of models for infectious diseases are population-based compartmental models (Anderson and May, 1992). For example, the famous susceptible-

infectious-recovered (SIR) model (Harko et al., 2014) partitions the population into three compartments: susceptible individuals (S), infectious individuals (I), and recovered or deceased individuals (R). The SIR model uses differential equations to characterize the changes of the number of individuals in these three compartments. Rigorously speaking, since the disease transmission is a random process, the numbers in the three compartments should be understood as the expected numbers. This idea is in line with the mean-field theory, originated from statistical physics (Kadanoff, 2009), which approximates the effect of many individuals by a single averaged effect to simplify the analysis.

The classical compartment models assume random mixing of the population; that is, each infectious individual has an equal chance of coming into contact with any other individual and transmitting the disease. In practice, it is more realistic to consider disease transmissions on social networks (Herrmann and Schwartz, 2020) with the observation that disease transmission between individuals being connected in the network is more likely than transmission between two random persons in the population. Researchers have pointed out that different network structures can result in very different transmission patterns even for diseases with the same R0 (basic reproduction number) (Meyers et al., 2005). The readers are referred to Keeling and Eames (2005); Wang et al. (2015); Britton (2020) for surveys on disease models on networks. Analytic solutions to late-time properties (i.e., as time goes to infinity), such as the fraction of people in the network being infected eventually, are available (Mollison, 1977; Grassberger, 1983; Newman, 2010) under simple model assumptions, such as the configuration model (Bender and Canfield, 1978) for network generation and a constant transmission rate for connected infectious and susceptible individuals. It is difficult or impossible, however, to solve more complicated models analytically, and computer simulation is usually the best feasible approach.

Below we review research papers consisting of both a epidemic model component and a network component. For research primarily based on epidemic models, please refer to Gumel et al. (2021); Ren et al. (2020); Grimm et al. (2021); Bertozzi et al. (2020), etc. We focus on the

following aspects of each paper: (1) Which epidemic model is used? The classical SIR model serves as the backbone but researchers have added additional compartments to better characterize the disease, such as the "exposed" (E) status in the SEIR model (Hethcote, 2000), or the "asymptomatic" (A) status to characterize the significant proportion of asymptomatic COVID-19 patient. (2) Which network model is used? Different than transmission networks being discussed in the previous sub-section, where nodes represent patients and edges represent infections, the networks used in epidemic studies are ordinary social networks, which serves as the basis for dynamic process. Popular models for social networks include the small-world network (Watts–Strogatz model) (Watts and Strogatz, 1998), the configuration model (Bender and Canfield, 1978), the scale-free network (power-law degree distribution) (Barabási and Albert, 1999; Bollobás et al., 2011), etc. Variants of these models or more complicated setups have been used for studying disease transmissions. (3) Which human activities are modeled? The simplest epidemic model on networks assumes a constant transmission rate between two connected individuals. With the help of computer simulations, one can instead study more complicated and realistic human activities during the pandemic, such as non-uniform interaction within one's personal network and occasion long-distance interaction outside the personal networks (Block et al., 2020). (4) Are certain policies studied? In addition to studying transmission rates on networks with different topologies, researchers are also interested in the impact of imposing or lifting policies such as social distancing on disease transmission. (5) Whether or how real data have been used in the study? Due to the complexity level of computer-simulated models, it is difficult to conduct estimation or inference of the unknown parameters in a rigorous statistical sense even with real data. Therefore, how to gauge or calibrate a model using real data is an intriguing question. In addition, we summarize the major findings and policy recommendations in these papers in Table 3.

Karaivanov (2020) proposed a stochastic epidemic model consisting of five basic states: S for susceptible to the disease; E for exposed; I for infectious; R for recovered; F for dead; and

Paper	Major Findings and Policy Recommendations
Karaivanov (2020)	Disease transmissions over a network-connected population can be slower than transmissions modeled by SIR assuming random mixing; intermittent lockdown or distancing policies can effectively flatten the infection curve; lockdown or distancing policies, if lifted earlier, mostly shift the infection peak into the future.
Block et al. (2020)	Three social distancing strategies (limiting interaction to a few repeated contacts, seeking similarity across contacts, and strengthening communities via triadic strategies) can substantially slow the spread of the disease and the first strategy is particularly helpful.
Chang et al. (2021)	The magnitude of mobility reduction is at least as crucial as its timing; a minority of points of interest (POIs) are the cause of the majority of the infections; reopening with a reduced maximum occupancy that specifically targets high-risk POIs may be more effective than less targeted strategies.
Firth et al. (2020)	Contact tracing and quarantine might be most effective when contact rates are high; tracing contacts of contacts is a more effective strategy than tracing of only contacts, but can result in large numbers of individuals being quarantined at a single point in time; combining physical distancing with contact tracing can control the disease while reduce the number of quarantined individuals.
Della Rossa et al. (2020)	Understand of heterogeneity between regions is essential to study the spread of the disease and design effective policies; lockdown and interventions with feedback at the regional level are beneficial.

Table 3: Major findings and policy recommendations in papers on network-based epidemic models.

two additional states P for tested positive and L for lockdown. A key assumption of the model is that a person can get infected with a small probability from the general population and with a larger probability proportional to the fraction of infectious persons in his or her personal network. Gillespie's algorithm (Gillespie, 1977) was applied to simulate the continuous-time stochastic process. A modified Barabási-Albert model were used to simulate the social network. The paper further evaluated the impact of certain government responses and policies by simulations,

including testing, contact tracing, social distancing, quarantine, lockdown, etc. The paper only used simulated data.

Block et al. (2020) simulated a social network-based epidemic model to evaluate three different social distancing strategies: limiting interaction to a few repeated contacts, seeking similarity across contacts, and strengthening communities via triadic strategies. The epidemic model was a classical SEIR model and the network they considered consists of links between individuals who live close geographically, individuals who are similar on attributes, individuals who belong to common groups, and random connections in the population. They reported that all three distancing strategies can substantially slow the spread of the disease and the strategy of limiting interaction to a few repeated contacts is particularly helpful. Ohsawa and Tsubokura (2020) recommended a similar strategy that limits inter-community contacts. Block et al. (2020) did not use real data.

Chang et al. (2021) combined the SEIR model with a mobility network to simulate the spread of COVID-19. The mobility network defined in the paper is a bipartite graph containing two types of nodes – census block groups (CBGs) that are residential areas typically containing 600 – 3,000 people, and specific points of interest (POIs) that are non-residential locations such as restaurants and grocery stores. The time-varying weighted links represent the number of visitors from CBGs to POIs, estimated from data collected by SafeGraph – a company that aggregates location data from mobile applications. Each CBG has its own S, E, I and R states and the transition probabilities between states are governed by parameters such as transmission rates at CBGs or POIs as well as weights of links from CBGs to POIs. Most of the parameters were estimated from SafeGraph and US census data with a few being calibrated by minimizing the mean squared errors between daily numbers of confirmed cases reported by *The New York Times* and the corresponding predicted numbers by the model. The paper also studied demographic disparities in infections and evaluated various mobility reduction and reopening strategies, such as reopening with a reduced maximum occupancy, through simulated mobility networks.

Firth et al. (2020) simulated epidemic models on a real-world network to evaluate the effect of tracing the contacts of patients and secondary contacts. The dataset on human social interactions, which is publicly available (https://github.com/skissler/haslemere), was collected for modeling infectious disease but not specifically for COVID-19 (Kissler et al., 2020). The epidemic model, built on a previous branching-process model (Hellewell et al., 2020), included standard states such as susceptible, infectious and recovered, and also states *isolated* or *quarantined* to describe the tracing and quarantining strategies. The paper reported that tracing contacts of contacts was an effective strategy but can result in large numbers of individuals being quarantined at a single point in time.

Della Rossa et al. (2020) modeled Italy as a network of regions and proposed epidemic models at regional and national levels to evaluate the effectiveness of the regional lockdown and social distancing strategies. The nodes of the network represent twenty regions of Italy and the edges represent geographical adjacency between regions and long-distance transportation routes to capture fluxes of people traveling between regions. Each region was assigned an individual ordinary differential equation (ODE) model including six compartments: suspectible, infected, quarantined, hospitalized, recovered and deceased. The regional level models were then aggregated to a national level model by considering fluxes between regions. The parameters were estimated from official COVID-19 data collected by government (http://github.com/pcm-dpc/COVID-19/tree/master/dati-andamento-nazionale) and publicly available mobility data from Google (https://www.google.com/covid19/mobility/). Furthermore, various regional feedback intervention strategies were simulated and the main findings include that inter-regional fluxes have dramatic effects on recurrent epidemic waves and it is beneficial that each of the twenty regions individually strengthens or weakens local mitigating actions.

Besides the COVID-19 studies highlighted in this review, the readers are referred to Qian et al. (2021); Deng et al. (2021) and Azzimonti et al. (2020) for more research on network-based

epidemic models.

4 Opportunities in Phylogenetic and Social Network Analyses for COVID-19 Patients

The unprecedented crisis of COVID-19 may be the biggest disaster since World War II, which has caused huge economic loss and costed millions of lives. People need to learn from past experiences to prepare for the next crisis. From the research point of view, COVID-19 provides rich data resources different from previous data types (such as electronic health records or “omics" data collected in designed trials) in the sense that COVID-19 data are not restricted to one cohort or one region. Instead diverse types of data come in the form of COVID-19 data consortium from heterogeneous resources all over the world. Therefore, COVID-19 also presents unique opportunities for public health research as well as statistical methodology developments. New directions for future research are emerging and we summarize a few in the following.

1. Virus sequence data produced on different platforms, processed in different software, cleaned and normalized using different software are not directly comparable. Large amounts of sequencing data from different research labs in different countries are being produced and deposited into large consortia such as the Data and Computation Resources for COVID-19 at NIH (`https://datascience.nih.gov/COVID-19-open-access-resources`), COVID-19 data warehouse (`https://covidclinical.net/`) and the COVID-19 Host Genetics Initiative (`https://www.covid19hg.org/`), which formed a global network of researchers to generate, share, and analyze data to study the genetic determinants of COVID-19 susceptibility and severity. Besides, electronic health records of COVID-19 patients after de-identification can be compiled from heterogeneous resources such as insurance companies, hospitals, research institutes etc. The accumulation of data

on a specific virus has never been so rapid in such large amounts. However, heterogeneity in the data formats and processing methods make it difficult in comparing across or integrative analysis of these data. Methods to unify data in different formats will greatly expand the researchers' ability to pool heterogeneous information sources.

2. Besides the heterogeneity in data production, there exist diverse choices for the analysis methods to construct patients or geographic clusters and phylogenetic trees. Different similarity or connectivity measures and ad-hoc choices of thresholds for clustering may lead to quite different results and inferences. As the academic world is raising more and more emphasis on the "reproducibility" of scientific studies, standard benchmark datasets or simulation studies to compare different methods and validate inferences in different genetics or epidemiology papers would help people evaluate the reliability of their conclusions.

3. Meta-analysis has been extremely useful in clinical studies to pool studies carried by independent researchers and get comprehensive conclusions with higher accuracy. (`https://www.cdc.gov/coronavirus/2019-nCoV/index.html`). For example, the flagship paper Initiative (2021) of the COVID-19 Host Genetics Initiative, published in *Nature* recently, described the results of three genome-wide association meta-analyses comprised of around 50,000 patients from 46 studies across 19 countries. The paper reported 13 genome-wide significant loci associated with COVID-19. In addition, the paper reported four of these loci have a stronger link to susceptibility to SARS-CoV-2 than to severity and 9 are associated with increased risk of severe symptoms. Several of these loci reportedly correspond to lung or autoimmune and inflammatory diseases. Furthermore, the analysis in the paper suggested a causal role for smoking and body mass index for severe COVID-19. However, it is unclear how to carry out meta-analysis to estimate network characteristics (such as degree, centrality, distribution and length) or phylogenetic trees when so many papers using network analysis on COVID-19 data are being published at

the same time. Either individual level or summary level meta-analysis to pool similar network analyses on different COVID-19 data would be a research topic with great potential in real applications.

4. The sequence by which SARS-CoV-2 mutations occurred is a key question in the construction of phylogenetic trees and infection pathways. Most of the SARS-CoV-2 genomic sequences are accompanied with the collection dates and locations. Besides similarity or distances between SARS-CoV-2 genomes and closeness between locations, the collection dates may provide the timeline different mutations showed up and facilitate the construction of phylogenetic trees.

5. Currently, phylogenetic network analysis and social network analysis are carried out separately, which are seemly unrelated at all. However, the transmission of SARS-CoV-2 within social groups leads to similar patterns in the similarity between virus genome sequences. Virus from socially close individuals with direct contact tend to have similar sequences. Social network and transmission pathways would provide additional evidence or validation for the clustering of individual COVID-19 genomes. Joint clustering of SARS-CoV-2 sequence data and COVID-19 patients' connections may provide cluster estimates with higher accuracy.

5 Acknowledgments

This work is supported in part by funds from the National Science Foundation (NSF: # 1636933 and # 1920920).

References

Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 2020.

NS Zhong, BJ Zheng, YM Li, LLM Poon, ZH Xie, KH Chan, PH Li, SY Tan, Q Chang, JP Xie, et al. Epidemiology and cause of severe acute respiratory syndrome (sars) in guangdong, people's republic of china, in february, 2003. *The Lancet*, 362(9393):1353–1358, 2003.

Ali M Zaki, Sander Van Boheemen, Theo M Bestebroer, Albert DME Osterhaus, and Ron AM Fouchier. Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. *New England Journal of Medicine*, 367(19):1814–1820, 2012.

Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25(17):2000257, 2020.

Annemarie B Docherty, Ewen M Harrison, Christopher A Green, Hayley E Hardwick, Riinu Pius, Lisa Norman, Karl A Holden, Jonathan M Read, Frank Dondelinger, Gail Carson, et al. Features of 20 133 uk patients in hospital with covid-19 using the isaric who clinical characterisation protocol: prospective observational cohort study. *bmj*, 369, 2020.

Shikha Garg, Lindsay Kim, Michael Whitaker, Alissa O'Halloran, Charisse Cummings, Rachel Holstein, Mila Prill, Shua J Chai, Pam D Kirley, Nisha B Alden, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—covid-net, 14 states, march 1–30, 2020. *Morbidity and mortality weekly report*, 69(15):458, 2020.

Eboni G Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization

and mortality among black patients and white patients with covid-19. *New England Journal of Medicine*, 382(26):2534–2543, 2020.

Safiya Richardson, Jamie S Hirsch, Mangala Narasimhan, James M Crawford, Thomas McGinn, Karina W Davidson, Douglas P Barnaby, Lance B Becker, John D Chelico, Stuart L Cohen, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area. *Jama*, 323(20):2052–2059, 2020.

Jenine K Harris and Bruce Clements. Using social network analysis to understand missouri's system of public health emergency planners. *Public health reports*, 122(4):488–498, 2007.

Angela Tringali, David L Sherer, Jillian Cosgrove, and Reed Bowman. Life history stage explains behavior in a social network before and during the early breeding season in a cooperatively breeding bird. *PeerJ*, 8:e8302, 2020.

Loni Hagen, Thomas Keller, Stephen Neely, Nic DePaula, and Claudia Robert-Cooperman. Crisis communications in the age of social media: A network analysis of zika-related tweets. *Social Science Computer Review*, 36(5):523–541, 2018.

Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

Abbas El Gamal and Young-Han Kim. *Network information theory*. Cambridge university press, 2011.

Michael D Ward, Katherine Stovel, and Audrey Sacks. Network analysis and political science. *Annual Review of Political Science*, 14:245–264, 2011.

Lise Getoor and Christopher P Diehl. Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12, 2005.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251, 2010.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Linton C Freeman. Visualizing social networks. *Journal of social structure*, 1(1):4, 2000.

Steve Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.

Christopher Andrew Bail. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, 113(42):11823–11828, 2016.

Man Hung, Evelyn Lauren, Eric S Hon, Wendy C Birmingham, Julie Xu, Sharon Su, Shirley D Hon, Jungweon Park, Peter Dang, and Martin S Lipsky. Social network analysis of covid-19 sentiments: Application of artificial intelligence. *Journal of medical Internet research*, 22(8): e22590, 2020.

Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying communities of practice through ontology network analysis. *IEEE Intelligent Systems*, 18(2):18–25, 2003.

Yoichi Murakami, Lokesh P Tripathi, Philip Prathipati, and Kenji Mizuguchi. Network analysis and in silico prediction of protein–protein interactions with applications in drug discovery. *Current opinion in structural biology*, 44:134–142, 2017.

Shan Zhao and Ravi Iyengar. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. *Annual review of pharmacology and toxicology*, 52:505–521, 2012.

Pei Wang, Jun-an Lu, Yanyu Jin, Mengfan Zhu, Lingling Wang, and Shunjie Chen. Statistical and network analysis of 1212 covid-19 patients in henan, china. *International Journal of Infectious Diseases*, 95:391–398, 2020.

Yunmeng Bai, Dawei Jiang, Jerome R Lon, Xiaoshi Chen, Meiling Hu, Shudai Lin, Zixi Chen, Xiaoning Wang, Yuhuan Meng, and Hongli Du. Comprehensive evolution and molecular characteristics of a large number of sars-cov-2 genomes reveal its epidemic trends. *International Journal of Infectious Diseases*, 100:164–173, 2020.

Michael Worobey, Jonathan Pekar, Brendan B Larsen, Martha I Nelson, Verity Hill, Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and Philippe Lemey. The emergence of sars-cov-2 in europe and north america. *Science*, 370(6516):564–570, 2020.

Xingguang Li, Junjie Zai, Qiang Zhao, Qing Nie, Yi Li, Brian T Foley, and Antoine Chaillon. Evolutionary history, potential intermediate animal host, and cross-species analyses of sars-cov-2. *Journal of medical virology*, 92(6):602–611, 2020.

Carla Mavian, Simone Marini, Mattia Prosperi, and Marco Salemi. A snapshot of sars-cov-2 genome availability up to april 2020 and its implications: data analysis. *JMIR public health and surveillance*, 6(2):e19170, 2020a.

Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243, 2020.

Gábor Kemenesi, Safia Zeghbib, Balázs A Somogyi, Gábor Endre Tóth, Krisztián Bányai, Norbert Solymosi, Peter M Szabo, István Szabó, Ádám Bálint, Péter Urbán, et al. Multiple sars-cov-2 introductions shaped the early outbreak in central eastern europe: comparing hungarian data to a worldwide sequence data-matrix. *Viruses*, 12(12):1401, 2020.

Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey M Kozlov, et al. Phylogenetic analysis of sars-cov-2 data is difficult. *Molecular biology and evolution*, 38(5): 1777–1791, 2021.

Gianguglielmo Zehender, Alessia Lai, Annalisa Bergna, Luca Meroni, Agostino Riva, Claudia Balotta, Maciej Tarkowski, Arianna Gabrieli, Dario Bernacchia, Stefano Rusconi, et al. Genomic characterization and phylogenetic analysis of sars-cov-2 in italy. *Journal of medical virology*, 92(9):1637–1640, 2020.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

Carsten Kemena and Cedric Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, 2009.

Julie D Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. A comprehensive

benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3):e18093, 2011.

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26 (7):1641–1650, 2009.

Thomas Hughes Jukes, Charles R Cantor, HN Munro, et al. Mammalian protein metabolism. 1969.

Simon Tavaré et al. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

Ziheng Yang. Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111, 1994.

Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174, 1985.

Andrey Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–329, 1994.

Fabrícia F Nascimento, Mario Dos Reis, and Ziheng Yang. A biologist's guide to bayesian phylogenetic analysis. *Nature ecology & evolution*, 1(10):1446–1454, 2017.

Bruce Rannala. Identifiability of parameters in mcmc bayesian inference of phylogeny. *Systematic biology*, 51(5):754–760, 2002.

Ziheng Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

John P Huelsenbeck and Bruce Rannala. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic biology*, 53(6):904–913, 2004.

Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. jmodeltest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8):772–772, 2012.

Thomas M Keane, Christopher J Creevey, Melissa M Pentony, Thomas J Naughton, and James O Mclnerney. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary biology*, 6(1):1–17, 2006.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Gideon Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.

Sebastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1):92–94, 2006.

Andre J Aberer, Kassian Kobert, and Alexandros Stamatakis. Exabayes: massively parallel bayesian tree inference for the whole-genome era. *Molecular biology and evolution*, 31(10): 2553–2556, 2014.

Huw A Ogilvie, Remco R Bouckaert, and Alexei J Drummond. Starbeast2 brings faster species

tree inference and accurate estimates of substitution rates. *Molecular biology and evolution*, 34(8):2101–2114, 2017.

Mattia CF Prosperi, Massimo Ciccozzi, Iuri Fanti, Francesco Saladini, Monica Pecorari, Vanni Borghi, Simona Di Giambenedetto, Bianca Bruzzone, Amedeo Capetti, Angela Vivarelli, et al. A novel methodology for large-scale phylogeny partition. *Nature communications*, 2(1):1–10, 2011.

Manon Ragonnet-Cronin, Emma Hodcroft, Stéphane Hué, Esther Fearnhill, Valerie Delpech, Andrew J Leigh Brown, and Samantha Lycett. Automated analysis of phylogenetic clusters. *BMC bioinformatics*, 14(1):1–10, 2013.

David W Mount and David W Mount. *Bioinformatics: sequence and genome analysis*, volume 1. Cold spring harbor laboratory press Cold Spring Harbor, NY, 2001.

Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. In *Advances in neural information processing systems*, pages 1061–1069, 2012.

Masatoshi Nei. Genetic distance between populations. In *Molecular Evolutionary Genetics*, pages 208–253. Columbia University Press, 1987.

Luigi L Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233, 1967.

Nathan Ellis, Susan Ciocci, and James German. Back mutation can produce phenotype reversion in bloom syndrome somatic cells. *Human genetics*, 108(2):167–173, 2001.

Joseph Felsenstein and Joseph Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

Robert R Sokal. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438, 1958.

Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

Hans-Jurgen Bandelt, Peter Forster, and Arne Röhl. Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution*, 16(1):37–48, 1999.

Walter M Fitch and Emanuel Margoliash. Construction of phylogenetic trees. *Science*, 155 (3760):279–284, 1967.

William HE Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of mathematical biology*, 49(4):461–467, 1987.

Sungsik Kong, Santiago J Sánchez-Pacheco, and Robert W Murphy. On the use of median-joining networks in evolutionary biology. *Cladistics*, 32(6):691–699, 2016.

Santiago J Sánchez-Pacheco, Sungsik Kong, Paola Pulido-Santacruz, Robert W Murphy, and Laura Kubatko. Median-joining network analysis of sars-cov-2 genomes is neither phylogenetic nor evolutionary. *Proceedings of the National Academy of Sciences*, 117(23):12518–12519, 2020.

Yulia Vakulenko, Andrei Deviatkin, and Alexander Lukashev. The effect of sample bias and experimental artefacts on the statistical phylogenetic analysis of picornaviruses. *Viruses*, 11 (11):1032, 2019.

Carla Mavian, Sergei Kosakovsky Pond, Simone Marini, Brittany Rife Magalis, Anne-Mieke Vandamme, Simon Dellicour, Samuel V Scarpino, Charlotte Houldcroft, Julian Villabona-Arenas, Taylor K Paisie, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of sars-cov-2 infections unreliable. *Proceedings of the National Academy of Sciences*, 117(23):12522–12523, 2020b.

David D Pollock, Derrick J Zwickl, Jimmy A McGuire, and David M Hillis. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic biology*, 51(4):664, 2002.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

Jeffrey M Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301, 2007.

Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

Mohammad Ali Mansournia and Douglas G Altman. Inverse probability weighting. *Bmj*, 352, 2016.

Martin Vingron and Patrick Argos. A fast and sensitive multiple sequence alignment algorithm. *Bioinformatics*, 5(2):115–121, 1989.

Peter R Sibbald and Patrick Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of molecular biology*, 216(4):813–818, 1990.

Steven Henikoff and Jorja G Henikoff. Position-based sequence weights. *Journal of molecular biology*, 243(4):574–578, 1994.

Adam J Hockenberry and Claus O Wilke. Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. *Entropy*, 21(10):1000, 2019.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, et al. Use of proteomic patterns in serum to identify ovarian cancer. *The lancet*, 359(9306):572–577, 2002.

Joshua M Akey, Shameek Biswas, Jeffrey T Leek, and John D Storey. On the design and analysis of gene expression studies in human populations. *Nature genetics*, 39(7):807–808, 2007.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

Richard S Spielman, Laurel A Bastone, Joshua T Burdick, Michael Morley, Warren J Ewens, and Vivian G Cheung. Common genetic variants account for differences in gene expression among ethnic groups. *Nature genetics*, 39(2):226–231, 2007.

Fuqing Wu, Amy Xiao, Jianbo Zhang, Katya Moniz, Noriko Endo, Federica Armas, Richard Bonneau, Megan A Brown, Mary Bushman, Peter R Chai, et al. Sars-cov-2 titers in wastewater foreshadow dynamics and clinical presentation of new covid-19 cases. *Medrxiv*, 2020.

Hanbing Song, Bobak Seddighzadeh, Matthew R Cooperberg, and Franklin W Huang. Expression of ace2, the sars-cov-2 receptor, and tmprss2 in prostate epithelial cells. *BioRxiv*, 2020.

Neal G Ravindra, Mia Madel Alfajaro, Victor Gasque, Jin Wei, Renata B Filler, Nicholas C Huston, Han Wan, Klara Szigeti-Buck, Bao Wang, Ruth R Montgomery, et al. Single-cell longitudinal analysis of sars-cov-2 infection in human bronchial epithelial cells. *BioRxiv*, 2020.

Mi Seon Han, Jung-Hyun Byun, Yonggeun Cho, and John Hoon Rim. Rt-pcr for sars-cov-2: quantitative versus qualitative. *The Lancet Infectious Diseases*, 21(2):165, 2021.

Xun Gu. Understanding tissue expression evolution: from expression phylogeny to phylogenetic network. *Briefings in bioinformatics*, 17(2):249–254, 2016.

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. *Handbook of data visualization*. Springer Science & Business Media, 2007.

Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Andreas Scherer. *Batch effects and noise in microarray experiments: sources and solutions*, volume 868. John Wiley & Sons, 2009.

Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

Zhifu Sun, High Seng Chai, Yanhong Wu, Wendy M White, Krishna V Donkena, Christopher J Klein, Vesna D Garovic, Terry M Therneau, and Jean-Pierre A Kocher. Batch effect correction for genome-wide methylation data with illumina infinium platform. *BMC medical genomics*, 4(1):1–12, 2011.

Andrew E Jaffe, Thomas Hyde, Joel Kleinman, Daniel R Weinbergern, Joshua G Chenoweth, Ronald D McKay, Jeffrey T Leek, and Carlo Colantuoni. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. *BMC bioinformatics*, 16 (1):1–10, 2015.

Sean M Gibbons, Claire Duvallet, and Eric J Alm. Correcting for batch effects in case-control microbiome studies. *PLoS computational biology*, 14(4):e1006102, 2018.

Ewen Callaway. The coronavirus is mutating-does it matter? *Nature*, 585(7824):174–177, 2020.

Laiba Shafique, Awais Ihsan, Qingyou Liu, et al. Evolutionary trajectory for the emergence of novel coronavirus sars-cov-2. *Pathogens*, 9(3):240, 2020.

JJ Bull, John P Huelsenbeck, Clifford W Cunningham, David L Swofford, and Peter J Waddell. Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397, 1993.

M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2010.

Wonkwang Jo, Dukjin Chang, Myoungsoon You, and Ghi-Hoon Ghim. A social network analysis of the spread of covid-19 in south korea and policy implications. *Scientific Reports*, 11(1):1–10, 2021a.

Sakranaik Saraswathi, Amita Mukhopadhyay, Hemant Shah, and TS Ranganath. Social network analysis of COVID-19 transmission in Karnataka, India. *Epidemiology & Infection*, 148, 2020.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Lauren Ancel Meyers, Babak Pourbohloul, Mark EJ Newman, Danuta M Skowronski, and Robert C Brunham. Network theory and sars: predicting outbreak diversity. *Journal of theoretical biology*, 232(1):71–81, 2005.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435 (7039):207–211, 2005.

Yun Jo, Andy Hong, and Hyungun Sung. Density or connectivity: What are the main causes of the spatial proliferation of COVID-19 in Korea? *International Journal of Environmental Research and Public Health*, 18(10):5084, 2021b.

Ales Komarek, Jakub Pavlik, and Vladimir Sobeslav. Network visualization survey. In *Computational Collective Intelligence*, pages 275–284. Springer, 2015.

Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. URL `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`.

Mike KP So, Agnes Tiwari, Amanda MY Chu, Jenny TY Tsang, and Jacky NL Chan. Visualizing covid-19 pandemic risk through network connectedness. *International Journal of Infectious Diseases*, 96:558–561, 2020.

Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Tiberiu Harko, Francisco SN Lobo, and MK Mak. Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184–194, 2014.

Leo P Kadanoff. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137(5):777–797, 2009.

Helena A Herrmann and Jean-Marc Schwartz. Why covid-19 models should incorporate the network of social interactions. *Physical Biology*, 17(6):065008, 2020.

Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

Zhen Wang, Michael A Andrews, Zhi-Xi Wu, Lin Wang, and Chris T Bauch. Coupled disease–behavior dynamics on complex networks: A review. *Physics of life reviews*, 15:1–29, 2015.

Tom Britton. Epidemic models on social networks—with inference. *Statistica Neerlandica*, 74 (3):222–241, 2020.

Denis Mollison. Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(3):283–313, 1977.

Peter Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983.

Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

Abba B Gumel, Enahoro A Iboi, Calistus N Ngonghala, and Elamin H Elbasha. A primer on using mathematics to understand covid-19 dynamics: Modeling, analysis and simulations. *Infectious Disease Modelling*, 6:148–168, 2021.

Jinchang Ren, Yijun Yan, Huimin Zhao, Ping Ma, Jaime Zabalza, Zain Hussain, Shaoming Luo, Qingyun Dai, Sophia Zhao, Aziz Sheikh, et al. A novel intelligent computational approach to model epidemiological trends and assess the impact of non-pharmacological interventions for covid-19. *IEEE journal of biomedical and health informatics*, 24(12):3551–3563, 2020.

Veronika Grimm, Friederike Mengel, and Martin Schmidt. Extensions of the seir model for the analysis of tailored social distancing and tracing approaches to cope with covid-19. *Scientific Reports*, 11(1):1–16, 2021.

Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(29):16732–16738, 2020.

Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. In *The Structure and Dynamics of Networks*, pages 384–395. Princeton University Press, 2011.

Per Block, Marion Hoffman, Isabel J Raabe, Jennifer Beam Dowd, Charles Rahal, Ridhi Kashyap, and Melinda C Mills. Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, 4(6):588–596, 2020.

Alexander Karaivanov. A social network model of covid-19. *Plos one*, 15(10):e0240878, 2020.

Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

Josh A Firth, Joel Hellewell, Petra Klepac, Stephen Kissler, Adam J Kucharski, and Lewis G Spurgin. Using a real-world network to model localized covid-19 control strategies. *Nature medicine*, 26(10):1616–1622, 2020.

Fabio Della Rossa, Davide Salzano, Anna Di Meglio, Francesco De Lellis, Marco Coraggio, Carmela Calabrese, Agostino Guarino, Ricardo Cardona-Rivera, Pietro De Lellis, Davide Liuzza, et al. A network model of italy shows that intermittent regional strategies can alleviate the covid-19 epidemic. *Nature communications*, 11(1):1–9, 2020.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

Yukio Ohsawa and Masaharu Tsubokura. Stay with your community: Bridges between clusters trigger expansion of covid-19. *Plos one*, 15(12):e0242766, 2020.

Stephen M Kissler, Petra Klepac, Maria Tang, Andrew JK Conlan, and Julia R Gog. Sparking" the bbc four pandemic": Leveraging citizen science and mobile phones to model the spread of disease. *bioRxiv*, page 479154, 2020.

Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496, 2020.

Xinwu Qian, Lijun Sun, and Satish V Ukkusuri. Scaling of contact networks for epidemic spreading in urban transit systems. *Scientific reports*, 11(1):1–12, 2021.

Ou Deng, Kiichi Tago, and Qun Jin. An extended epidemic model on interconnected networks for covid-19 to explore the epidemic dynamics. *arXiv preprint arXiv:2104.04695*, 2021.

Marina Azzimonti, Alessandra Fogli, Fabrizio Perri, and Mark Ponder. Pandemic control in econ-epi networks. Technical report, National Bureau of Economic Research, 2020.

COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of covid-19 by worldwide meta-analysis. *Nature*, 2021.

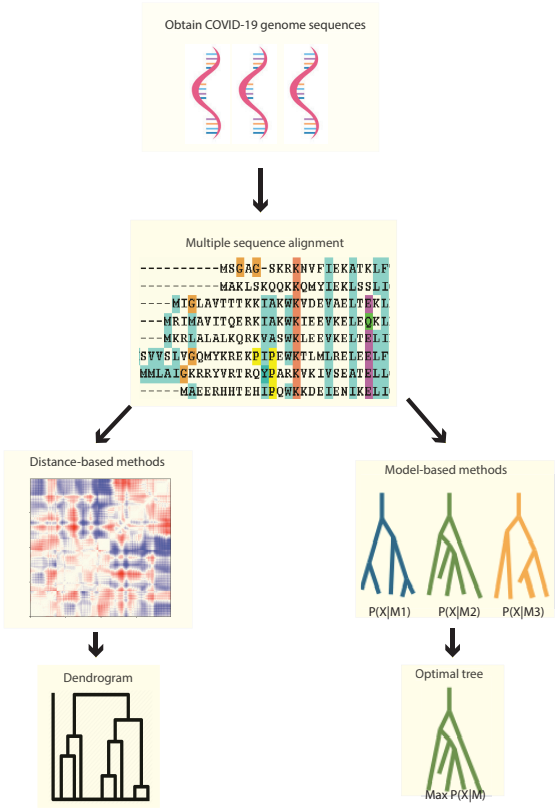

Figure 1: Flow chart of basic steps in COVID-19 phylogenetic analysis