VIEWPOINT: COVID-19

Insights from SARS-CoV-2 sequences

Analysis of viral sequences can tell us how SARS-CoV-2 spreads and adapts

By **Michael A. Martin[1,2], David VanInsberghe[1], Katia Koelle[1,3]**

As severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread across the globe, so have efforts to sequence its RNA genome. More than 260,000 sequences are now available in public databases, about a year after the viral genome was first sequenced (*1*). These sequences and their associated metadata have allowed researchers to estimate the timing of SARS-CoV-2 spillover into humans, characterize the spread of the virus, and gauge virus adaptation to its new host. Such analyses rely on interpreting patterns of nucleotide changes that have occurred in the virus population over time and are brought into focus through the reconstruction of genealogical relationships between sampled viruses that are depicted in phylogenetic trees.

Analysis of phylogenetic trees allows for conclusions to be drawn about epidemic and pandemic viruses that are important for public health (see the figure). After the emergence of a new virus, viral sequence data and sampling dates can be used to infer the rate at which the virus population evolves, along with the time of the most recent common ancestor (TMRCA) of all sampled viruses. For SARS-CoV-2, these analyses have revealed that the virus evolves at a rate of ~1.1×10^{-3} substitutions per site per year (*2*)—corresponding to one substitution every ~11 days—and a TMRCA of around late November 2019. However, owing to limited sampling early in the COVID-19 outbreak, this date likely lags behind the spillover into humans by several weeks.

Once virus circulation becomes widespread, phylodynamic analyses can give insight into how a virus spreads both spatially and temporally. Viruses from a given region can be placed in the context of those circulating globally, allowing for the number of independent virus introductions into a region to be estimated through phylogeography. These methods rely on assigning geographical states to unsampled ancestral viruses through a process called ancestral state reconstruction. Multiple applications of this approach have consistently shown that regional SARS-CoV-2 spread was ignited not by one, but by many independent introductions. This is likely due to slow or imperfect implementation of screening efforts at borders in the early stages of the pandemic. Similar types of analyses are conducted in local outbreak investigations, in what is often called genomic epidemiology. This has been used, for example, to link multiple transmission chains to a motorcycle rally in South Dakota (*3*).

Care must be taken, however, when interpreting phylogeographic analyses of this virus owing to the limited extent of its circulating genetic diversity. Given the evolutionary rate of SARS-CoV-2, viruses sampled weeks apart on different continents can have identical nucleotide sequences, making robust inferences more challenging. Additionally, sampling efforts are geographically heterogeneous, which can bias phylogeographic analyses. Methods have recently been developed to better accommodate the degree of undersampling across regions and the known travel history of sampled cases (*4*). Application of these methods has shown that later, rather than earlier, virus introductions established the first sustained transmission networks in the United States and Europe (*5*), highlighting the utility of genomic data in epidemiological investigations.

Once viral lineages have been introduced into a region, phylodynamic methods can also be used to infer the rate of viral spread through a host population and the basic reproduction number R_0, defined as the average number of infections generated by an infected host in a susceptible host population. This inference can be done using coalescent-based methods, which infer changes in the underlying population size of infected individuals using the time points at which viral lineages "coalesce" (merge) backward in time. Epidemiological models can also be fit directly to viral phylogenies based on the timing of coalescent events (*6*). Another approach to infer R_0 from sequence data relies on a forward-in-time birth-death process, with a "birth" corresponding to a transmission event (resulting in the birth of an infected individual) and a "death" corresponding to a removal of an infected individual (*7*). Both methods have been applied to SARS-CoV-2 sequence data, with relatively consistent results across methods and regions: At the beginning of the pandemic, R_0 fell between 2 and 3.5 (*8*, *9*) but decreased substantially after the implementation of nonpharmaceutical interventions. Although R_0 can be, and usually is, estimated from epidemiological case data, these estimates can be biased by changes in reporting rates. The relative contributions of new introductions versus local spread are often also indistinguishable in tabulated case data. Thus, analysis of sequence data provides an alternative approach to infer R_0 that may be particularly useful in the early stages of virus circulation when a large proportion of identified cases may be new introductions and when reporting rates are likely to be low. The development of methods that integrate both epidemiological time series and sequence data is an area of active research that will improve parameter estimation.

Phylodynamic analyses can also be used to identify instances of viral adaptation. Adaptation is a particular concern because SARS-CoV-2 only recently spilled over into humans and thus may still adapt to its new host through substitutions that facilitate its spread. One emerging variant that has been implicated as being more transmissible is 614G, which replaces aspartic acid (D) with glycine (G) at amino acid site 614 in the cellular entry (spike) protein of SARS-CoV-2. This variant likely arose in China in January 2020 and has since become dominant worldwide. The D-to-G substitution results in more efficient infection and replication in vitro and enhances transmission in animal models (*10*). Coalescent-based phylodynamic analyses using densely sampled genomes from infections in London found trends toward higher transmissibility of 614G clusters relative to 614D clusters (*11*). More recently, a new SARS-CoV-2 lineage, B.1.1.7, has rapidly spread from southeast England around the globe, and early analyses indicate that it has a substantial fitness advantage over other currently circulating lineages. These recent evolutionary events indicate that SARS-CoV-2 still has the capacity to develop more efficient transmission between human hosts.

When considering viral adaptation, a commonly held fear is that a virus may evolve to become more virulent, that is, cause more severe disease and host mortality. However, natural selection acts on variation in viral transmission potential, not variation in virulence per se. More virulent strains could have

[1]Department of Biology, Emory University, Atlanta, GA, USA. [2]Population Biology, Ecology, and Evolution Graduate Program, Laney Graduate School, Emory University, Atlanta, GA, USA. [3]Emory–University of Georgia Center of Excellence for Influenza Research and Surveillance (CEIRS), Atlanta, GA, USA. Email: katia.koelle@emory.edu

higher transmission potential if infection with a more virulent strain resulted in infected hosts shedding more virus. However, a more virulent infection may reduce contact rates of infected individuals, limiting the opportunity for viral transmission. Therefore, it is not clear whether more virulent SARS-CoV-2 strains are likely to evolve.

As SARS-CoV-2 continues to spread, the virus will begin to face new evolutionary pressures. Other respiratory viruses can provide insight into how SARS-CoV-2 evolution may manifest. For example, the 2009 pandemic H1N1 influenza virus started to evolve antigenically a couple of years after its emergence (*12*) and recent work has identified signatures of adaptive evolution in the spike protein of seasonal coronaviruses, consistent with antigenic evolution (*13*). Emerging evidence suggests that some SARS-CoV-2 variants may already be exhibiting antigenic evolution, and this is likely to continue as population immunity (through natural infection or vaccination) builds. Efforts by public health agencies to monitor emergent SARS-CoV-2 lineages, particularly those that may escape vaccine or natural immunity, are under way. Additionally, widespread infection of farmed mink has recently been observed, and there is evidence of transmission of mink lineages back into humans (*14*). Animal reservoirs may therefore contribute to the dynamics of SARS-CoV-2 evolution and adaptation.

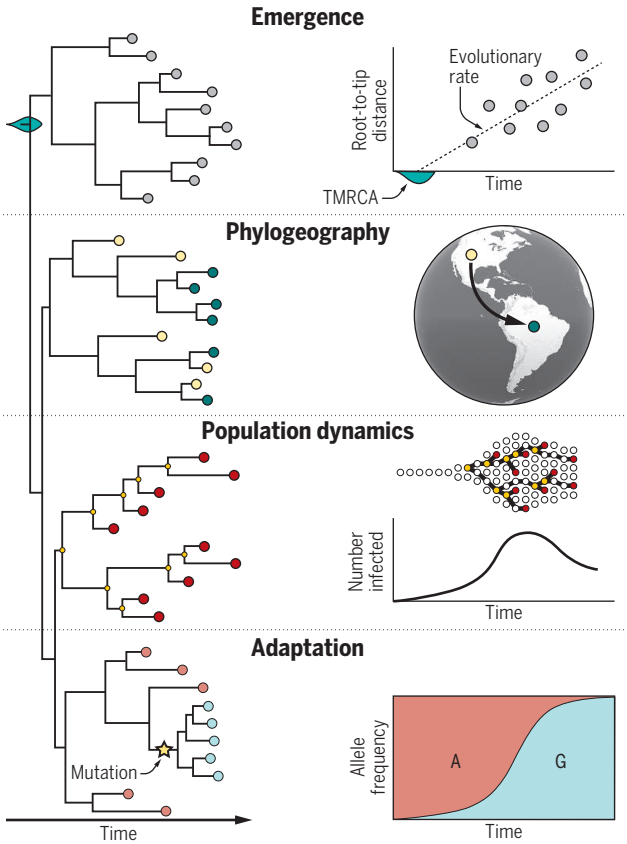
For viral adaptation to be possible, be it through the evolution of immune escape or other viral traits, genetic (and phenotypic) variation needs to be present. Although nucleotide substitutions are the primary source of genetic variation in SARS-CoV-2, insertions and deletions of nucleotides have also been observed. Furthermore, recombination is common in coronaviruses and may potentially give rise to new SARS-CoV-2 lineages. A small number of SARS-CoV-2 recombinant genomes have already been detected (*15*). Because of their potential phenotypic effects, circulation of genomic insertions, deletions, and recombinants should be monitored.

The number of available SARS-CoV-2 sequences is, like many things during this pandemic, unprecedented. Gaining understanding from these data does not come without challenges. Many current methods rely not just on the sequence itself but on associated metadata that provide additional information about viral samples. Research laboratories are often limited in the metadata that can be released owing to patient privacy regulations. For similar reasons, some government health agencies, such as the U.S. Centers for Disease Control and Prevention (CDC), receive only limited metadata from state health departments. Although privacy is an important concern, the routine release of more detailed metadata would improve the power of phylodynamic methods to describe viral dynamics and evolution. Notably, sequences are often released with only coarse sampling location data. However, viral dynamics may be heterogeneous even between locations that are geographically close. Furthermore, the incorporation of travel history can improve the accuracy of phylogenetic methods (*4*), but this information is not commonly reported. Because phylodynamic methods often require an assumption of random sampling, when samples from individuals belonging to the same transmission chain are sequenced, it is crucial that this information be labeled in the sequence metadata to avoid biasing analyses. Despite these challenges, a tremendous amount of SARS-CoV-2 sequence data is publicly available on GISAID's EpiCov database. Phylogenetic analyses of these sequences are conducted in near–real time and available on platforms such as Nextstrain and Microreact, allowing the ongoing evolution of SARS-CoV-2 genomes to be viewed in detail.

In many ways, the SARS-CoV-2 pandemic offers a distinct opportunity for the field of phylodynamics. Methods development over the past 10 to 15 years, the widespread availability of sequencing technologies, open data sharing, and the tireless efforts of clinicians and scientists who collect these data mean that more can be learned from viral genomes than ever before. As viral diversity continues to accumulate, SARS-CoV-2 sequence data and associated metadata can be used to answer questions focused on the longer-term evolution and adaptation of SARS-CoV-2. The volume of sequence data also presents an opportunity for methods development, because most current methods are computationally intractable when applied to hundreds of thousands of genomes. Continued efforts to collect viral sequence data and the development of efficient and scalable computational inference methods will help to further cement evolutionary analyses as a cornerstone of the public health response to viral spread and adaptation. ■

Uses for viral sequence data

Viral phylogenies, rooted at the most recent common ancestor (TMRCA), are inferred on the basis of genetic differences. These phylogenies can be used to estimate viral emergence, characterize the geographic spread of the virus, reconstruct epidemiological dynamics of viral spread within a region, and identify instances of adaptation.



GRAPHIC: A. KITTERMAN/*SCIENCE*

REFERENCES AND NOTES

1. F. Wu *et al.*, *Nature* **579**, 265 (2020).
2. S. Duchene *et al.*, *Virus Evol.* **6**, veaa061 (2020).
3. M. J. Firestone *et al.*, *MMWR* **69**, 1771 (2020).
4. P. Lemey *et al.*, *Nat. Commun.* **11**, 5110 (2020).
5. M. Worobey *et al.*, *bioRxiv* 10.1101/2020.05.21.109322 (2020).
6. E. M. Volz, I. Siveroni, *PLOS Comput. Biol.* **14**, e1006546 (2018).
7. T. Stadler *et al.*, *Mol. Biol. Evol.* **29**, 347 (2012).
8. D. Miller *et al.*, *Nat. Commun.* **11**, 5518 (2020).
9. S. A. Nadeau *et al.*, *medRxiv* 10.1101/2020.06.10.20127738 (2020).
10. Y. J. Hou *et al.*, *Science* **370**, 1464 (2020).
11. E. Volz *et al.*, *Cell* **184**, 64 (2021).
12. Y. C. F. Su *et al.*, *Nat. Commun.* **6**, 7952 (2015).
13. K. E. Kistler, T. Bedford, *bioRxiv* 10.1101/2020.10.30.352914 (2020).
14. B. B. Oude Munnink *et al.*, *Science* **371**, 172 (2021).
15. D. VanInsberghe *et al.*, *bioRxiv* 10.1101/2020.08.05.238386 (2020).

ACKNOWLEDGMENTS

This work was supported by National Institute of Allergy and Infectious Diseases (NIAID) Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant HHSN272201400004C and an Emory University MP3 seed grant. We thank G. Armstrong, T. Bedford, and two anonymous reviewers for feedback.

10.1126/science.abf3995