Phylogenetic interpretation during outbreaks requires caution

Ch. Julián Villabona-Arenas[1,2], William P. Hanage[3], Damien C. Tully[1,2]*

[1] Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2] Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK

[3] Harvard TH Chan School of Public Health, Boston, Massachusetts, US

*Corresponding author: damien.tully@lshtm.ac.uk

How viruses are related, and how they have evolved and spread over time, can be investigated using phylogenetics. Here, we set out how genomic analyses should be used during an epidemic and propose that phylogenetic insights from the early stages of an outbreak should heed all the available epidemiological information.

A goal of genomic epidemiology is to infer epidemiological and emergence dynamics from virus genome sequences obtained over short epidemic timescales [1]. Rapid in situ sequence generation and phylogenetic inference is based on detection of genetic changes in pathogen sequences. But during outbreaks there are many unknowns. The outbreak of coronavirus disease 2019 (COVID-19), which originated in Wuhan, China, was reported in December 2019 [2]. By January 2020, the genome of the causative novel coronavirus, named SARS-CoV-2, had been sequenced and made publicly available [2]. Virus sequences have underpinned development of diagnostics and vaccines and been used to assess patterns of transmission and spread. Although sequence data was used to answer crucial epidemiological questions during the Ebola and Zika outbreaks [3,4], the pace of generation of SARS CoV-2 genome data generation is unprecedented and is informing public health policy in real-time.

Importantly, it's not only sequences that inform phylogenies, and multiple factors contribute to the outputs including model assumptions, sampling density, timing of sample collection, portion of the viral genome sequenced, quality of sequencing data and the mutation rate of the virus itself. Although it is important to extract as much information as possible from sequence data as outbreaks unfold, it is imperative to bear in mind that the historical relationships of strains (phylogenies) are hypotheses that can be challenged as more data becomes available. Here, we highlight some of the challenges of genomic epidemiology during outbreaks such as SARS-CoV-2 and advise that interpretation of findings from phylogenies needs to assess all epidemiological and supporting information and consider sources of bias.

During outbreaks we want to know if cases are linked and if this implies transmission. Most viruses can be separated into strains and if two infections are caused by dissimilar strains one can rule out transmission. The oft-forgotten point is that phylogenies can rule out transmission, but if infections are caused by the same strains or identical viruses it does not decisively prove transmission. During an emerging outbreak, when pathogens have not yet diverged into different strains, phylogenetic information is too weak to hypothesize transmission linkage—which in turn can be used for geographic inference; even if the phylogenetic information is stronger, the same phylogeny is consistent with multiple transmission histories and there may missing links due to incomplete sampling [5]. Consequently, we need to combine phylogenetic findings with epidemiological and supporting information such as environmental factors and human air travel data before we draw any immediate conclusions regarding transmission. This was the case with Zika virus in Africa where epidemiological, human mobility and climatic data supported the phylogenetic hypothesis that the outbreak was likely imported from Brazil [6].

In the first stage of an outbreak, we can use phylogenetics to discern possible zoonotic sources, as in the case of the 2018 Lassa fever virus outbreak, where phylogenetic patterns indicated independent spillover events from rodent hosts [7]. The crucial observation was that the correct identification of the source of zoonotic transmission relies on the availability of viral genome sequences from potential animal reservoirs. If the source of any virus has not been sampled, it cannot be inferred, because phylogenetic linkage alone does not prove it. This is the reason for uncertainty surrounding the zoonotic source of SARS-CoV-2, because we have limited knowledge about the viral abundance from potential animal reservoirs [8]. The generation of additional viral genome sequences from an outbreak, coupled with virus-specific and epidemiological knowledge, provides insight into whether or not multiple 'jumps' occurred from a reservoir that might warrant appropriate control measures. Identical or nearly-identical virus genomes are expected from early transmission chains if a single spillover occurred recently, unless multiple zoonoses originated from the same low-genetic diversity virus pool. In contrast, higher diversity in the early-stage of human-to-human transmission is expected if multiple zoonoses have occurred or if there is significant within-host evolution [9].

Geographical inferences (where and when) are feasible as more representative viral genome data—in temporal and spatial scales—becomes available. We can hypothesize the location of common ancestors using ancestral reconstruction methods and infer phylogenies scaled to time, in order to date epidemiological events. Such analyses require a molecular clock, which models how the rate with which mutations accumulate with time, and how this varies across the branches of a phylogeny. However, early in an outbreak there may not be sufficient signal to accurately estimate clock rate. If this is the case, then it might be appropriate to apply an estimate from another closely related virus [10]. If temporal signal is present and a clock rate can be estimated, results need to be reported as credible intervals (instead of point estimates) to account for uncertainty in both the data (incomplete, biased, or improper sampling can lead to misleading phylogenies) and the many aspects of the methods.

When investigating the dissemination of an emerging virus the number of sequenced viral genomes may not be representative. Even as the outbreak unfolds, and more genomes are obtained, they only represent a snapshot of the underlying genetic diversity. If phylogenies are considered alone we cannot conclusively assert the geographical origins of the virus—or the extent of community transmission—as we cannot distinguish between local transmission events and multiple introductions of genetically similar viruses, from geographically distinct sources, if one of them has not been sampled. In this way uneven sampling can also lead to misleading conclusions on the geographical source, number of introductions and the size and duration of local transmission chains [11]. The significance of these associations is harder to ascertain when the phylogeny is reported without any assessment on the reliability of internal branches. Therefore, phylogenetic interpretation from ongoing outbreaks as is the case of SARS-CoV-2 needs to be done in the context of all available information such as temporal and spatial distribution of cases, travel patterns and any evidence of epidemiological linkage, sampling uncertainty and other sources of bias need to be carefully considered and reported.

The methods for valid phylogenetic inference require multiple assumptions which are likely not met during emerging outbreaks. Examples (not exhaustive) include adequate phylogenetic signal, which is low when strains have not yet diverged; geographical representation and effective sampling time points with sufficient molecular clock signal, which only become feasible as the epidemic unfolds; and random mixing, which may be violated under certain circumstances, for instance when mitigation strategies are set in place. Estimates from phylogenies may be sensitive to one or more of these assumptions and conclusions need to be made and shared with caution. Another essential consideration during an epidemic is accurate rooting of the phylogeny as it determines the direction of transmission over time [12].

There are also genome features that are intrinsic to the biology of the virus that may impact the extent and applicability of phylogenetics during outbreaks. For instance, the presence of recombination/reassortment and low diversity (due to the rate of evolution, selective constraints and transmission bottlenecks) complicate the resolution of phylogenetic relationships, but the incorporation of within-host viral diversity may provide greater resolution in understanding transmission dynamics [13]. Moreover, some of mutations in the viral genome sequence can be due to the error rate of the sequencing technology, recurrent sequencing issues, hypermutability or contamination which warrant caution with interpretations and especially with those concerning selection and recombination.

Genomic epidemiology has supported public health outbreak responses. Indeed, the ability to exploit viral genome sequences has allowed us to characterise early patterns of SARS-CoV-2 transmission in China, New Zealand and Australia [14,15]. In the midst of an outbreak sharing data is both necessary and important for an effective response, but sharing the associated metadata is also necessary to aid interpretations (e.g. how representative is the data of the country-wide situation) and to avoid creating sampling bias by researchers that are not doing the sequencing themselves.

The emergence of SARS-CoV-2 has presented a series of challenges about how we reliably extract information from phylogenies to gain insights into virus transmission and spread, and how we responsibly present our findings. Owing to low genetic diversity and uneven sampling, several controversial hypotheses have already been put forward. One cautionary tale involves how an outbreak in Bavaria seeded the epidemic in northern Italy and the subsequent wider outbreak in Europe. This notion was based on a small sample of very similar sequences. However, it overlooked a more likely scenario in which this virus was already circulating in China and that European regions had multiple introductions from China. At this early stage conclusions about the impacts of mutations on transmission and disease (e.g. D614G mutation in the spike protein [16]) should not be made on the basis of phylogenies alone but with separate evidence supporting not only a phenotypic difference but the resulting consequences for epidemiology.

The SARS-CoV-2 pandemic has highlighted the importance of providing a comprehensive rationale for any conclusions about the spatio-temporal dispersal of the virus. Phylogenies represent hypotheses that encompass different sources of error and this uncertainty needs to be visualised and communicated far more transparently. Another challenge is how we facilitate the dissemination of metadata and integrate this with phylogenetic trees. Incorporating host characteristics (e.g. age, onset date, exposure history) to aid phylogenetic interpretation would undoubtedly results in more reliable inferences.

Now, more than ever, careful reporting of phylogenetic interpretations, while safeguarding the privacy of infected individuals, would ensure that both policymakers and the public have the best possible information during an outbreak. Failure to balance these issues could jeopardise both scientific integrity and public confidence in the field of genomic epidemiology.

REFERENCES

1. Grubaugh, N. D. et al. Nat Microbiol **4**, 10–19 (2019).

2. Wu, F. et al. Nature **579**, 265–269 (2020).

3. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. Nature **538**, 193–200

(2016).

4. Pollett, S. et al. J. Infect. Dis. (2019).

5. Hall, M. D. & Colijn, C. Mol. Biol. Evol. **36**, 1333–1343 (2019).

6. Hill, S. C. et al. Lancet Infect. Dis. **19**, 1138–1147 (2019).

7. Kafetzopoulou, L. E. et al. Science (80). **363**, 74–77 (2019).

8. Andersen, K. G., Rambaut, A., Ian Lipkin, W., Holmes, E. C. & Garry, R. F. Nat. Med. 1–3 (2020).

9. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Mol. Biol. Evol. **34**, 997–1007 (2017).

10. Fraser, C. et al. Science (80). **324**, 1557–1561 (2009).

11. Kraemer, M. U. G. et al. Epidemiol. Infect. **147**, (2019).

12. Dudas, G. & Rambaut, A. PLoS Curr. **6**, (2014).

13. Worby, C. J., Lipsitch, M. & Hanage, W. P. Am. J. Epidemiol. **186**, 1209–1216 (2017).

14. Lu, J. et al. Cell (2020). doi:10.1016/j.cell.2020.04.023

15. Eden, J.-S. et al. Virus Evol **6**, veaa027 (2020).

16. Korber, B. et al. bioRxiv 2020.04.29.069054 (2020). doi:10.1101/2020.04.29.069054

Contributions

D.C.T. conceived the commentary and wrote the first draft. C.J-V.A, D.C.T conceptualized the ideas with W.P.H. All authors edited the manuscript into its final form.

Competing Interests

The authors declare no competing interests.