A Call for Practical Phylogenetics for Public Health

Emma Hodcroft, Nextstrain, University of Bern (co-corresponding author)
Nicola De Maio, EMBL-European Bioinformatics Institute, Cambridge
Rob Lanfear, Australian National University and GISAID
Duncan R. MacCannell, CDC's Office of Advanced Molecular Detection, USA
Bui Quang Minh, Australian National University
Heiko Schmidt, Center for Integrative Bioinformatics Vienna
Alexandros Stamatakis, Heidelberg Institute for Theoretical Studies
Nick Goldman, EMBL-European Bioinformatics Institute, Cambridge (co-corresponding author)
Christophe Dessimoz, University of Lausanne, University College London & Swiss Institute of Bioinformatics (co-corresponding author)

What phylogenetics can and could do for public health

The sharing of a novel pathogen genome on 10 January 2020[1]—just seven days after the first report of COVID-19 to the World Health Organization—was a defining moment of the current pandemic: the value of knowing and analyzing the sequence of approximately 30,000 nucleotides comprising the RNA genome of SARS-CoV-2 was immediately recognised, and scientists globally began an unprecedented effort to contribute to this data collection, leading to the rapid creation of vast quantities of genome sequences from the population of viruses circulating and evolving in humans worldwide. By comparing these sequences, it is possible to infer their evolutionary history. The relationships between viral genomes are often represented as a *phylogenetic* tree, with the sampled genomes at the leaves and branches connecting them in a way which shows their shared ancestry.

Phylogenetic trees have long interested biologists as a way to investigate the evolution of species and other biological entities, and are becoming increasingly relevant to public health decision making (Fig. 1). Indeed, the SARS-CoV-2 pandemic has provided a unique window into the possibilities achievable through phylogenetics as a more widely-used public health tool, to help curb not only this pandemic, but future ones as well.

Phylogenetic analysis confirmed the virus's likely origins in China and was crucial to our understanding of how it radiated from there[2]. Phylogenies revealed that the virus had spread across Europe before the end of February 2020, and how it moved to, then within, the United States[3]. As the virus moved between countries, phylogenetics illuminated its routes—even when little information regarding prevalence and individual travel histories was available. Iran's rapidly worsening epidemic had no sequences shared from their domestic cases, but samples from travellers returning from Iran to countries all over the world illustrated the existence and global spread of the country's cases[4]. In both Iceland and the UK, high-coverage testing and sequencing has allowed estimates of the number and source of introductions[5]. A study in Europe highlighted the role of travel over the summer: a variant that originally emerged in Spain spread across Europe with holidaymakers[6], knowledge that informs public health recommendations and impacts travel restrictions. Phylogenetics has been critical to understanding within-country dynamics of SARS-CoV-2 as well. In New Zealand and Australia, for example, these techniques have allowed scientists to pinpoint outbreak sources in otherwise well-contained epidemics, enabling targeted tracing and interventions[7].

Phylogenetics also informs on past and potential zoonotic (between animals and humans) transmission of the virus. By looking for the virus's closest relatives and their hosts, we can better understand the process of zoonotic transmission and take measures to prevent future events. Inferring the position of SARS-CoV-2 within a phylogenetic tree of previously known coronaviruses permitted the identification of its closest relatives—coronaviruses RaTG13 and RmYN02 that both infect bats[8]. Furthermore, the relationship between phylogenies of viruses and those of their host species provide insights on where to search to find even closer relatives than RaTG13 and RmYN02. Recently, phylogenetic trees provided important evidence of transmission between minks and humans in several countries, which may have contributed to Denmark's decision to cull 17 million mink.

To date, however, these emerging applications of phylogenetics have largely come out of academic research projects, and are yet to be widely adopted by most public health organizations. Sequencing has become relatively easy and inexpensive, but analyzing the very large numbers of resulting genomes is posing acute methodological challenges to existing phylogenetic methods. We believe that solving these challenges will require new ways of working together among researchers and public health departments.

The methodological challenges of phylogenetics for public health

Practical phylogenetics faces three major methodological challenges: estimating very large trees, accounting for uncertainty in phylogenetic analyses, and making it easier to draw actionable public health insights from them.

How to infer very large pathogen phylogenies in real time?

We have now passed 235,000 SARS-COV-2 sequences and may well exceed one million sequences by the end of the pandemic. Current state-of-the-art phylogenetic data analysis tools were largely conceived, optimised, and tested for static data sets of up to a few thousands of sequences. The orders-of-magnitude larger scale of pandemic genomic data, coupled with the need to update trees on a daily basis, requires substantially different approaches to tree inference. As the number of sequences increases, the number of potential trees relating them grows super-exponentially. Phylogenetic methods have long resorted to various shortcuts and heuristics for inference, but the scale of pandemic data precludes the use of anything but the fastest, coarsest methods. It has become impractical to re-run even these analyses every time new sequences are added, in particular because analyses of data sets containing hundreds of thousands of relatively short sequences are hard to execute in parallel on a large number of cores on high performance computing systems. In addition, executing well-established maximum likelihood methods such as RAxML or IQ-TREE on the current data requires dedicated numerical methods to cope with the limited phylogenetic signal contained in very many near-identical sequences[9], as well as very large amounts of memory—over 4 terabytes of RAM to process one million sequences.

The utility of any phylogenetic analysis is obviously impacted by the quality of input data. Since sample collection and surveillance coverage are often imperfect, sequence data may come with significant temporal, spatial and demographic biases. These biases have been notable in SARS-CoV-2 sequence data sharing, with the US and UK generating the most sequence data, and other parts of the world being substantially under-represented, including regions with high case numbers (e.g. Brazil, India, Iran). Even where there has been an

effort to sequence comprehensively, sequencing has not always kept up with all cases, with uneven coverage across regions (e.g. the well-resourced COVID-19 Genomics UK initiative had sequenced 9.4% of approximately 1,250,000 cases in the UK as of 2 November[10]). This has dramatic consequences for the quality of the inference of viral spread: for example, current phylogeographic models (i.e. phylogenetic models that are used to infer geographic spread) do not account for these sampling biases, and tend to infer the UK as a common source of spread because of its high rate of SARS-CoV-2 sequencing. Refinements to cope with biased sampling have been proposed, but the resulting methods tend to be computationally prohibitive on large data sets. Further, because they assume a stable population of viruses, they are not appropriate for recent and expanding outbreaks. To fully exploit the genomic resources available now and in the future, we need efficient models that are robust to sampling biases as well as their implementation in tools which allow users to gauge the impact of these biases on downstream analyses and inform on changes in sampling strategy to minimise these biases.

SARS-CoV-2 has also highlighted the limitations of relying solely on sequence data for inferring transmission. The high similarity of viral genomes collected from different patients makes it hard to reconstruct transmission histories from these sequences alone. Further, there can be inherent danger to interpreting phylogenetic trees alone for public health interventions: early and incomplete sequence inferences can be misleading without the support of other epidemiological data. SARS-CoV-2 evolves slowly, and has been sequenced more intensively than any other pathogen. Many of the genomes studied are identical, and no analysis based on genomes alone can infer transmission history in this situation. Even with a faster-evolving virus, genome sequencing will likely be even more prevalent in future; we can assume this problem will remain in any future pandemics. To increase our power for tracing transmission events, we need efficient phylogenetic methods that can accurately account for the difference between transmission history and phylogenetic tree, while at the same time incorporating information from time data (sampling time, location of sampling, onset of symptoms, etc), within-patient pathogen genetic diversity, patient contacts, and any other available epidemiological information. While these issues have been addressed piece-meal in the scientific literature, we still lack practical all-encompassing methods and infrastructure for the real-time analyses of large data sets.

How to deal with uncertainty in the trees?

Phylogenetic trees are statistical inferences and should always be interpreted in light of their uncertainty, and particularly so when the sequences being compared are very closely related. Even though there might be insufficient or incomplete information to infer the entire viral phylogenetic tree, it is typically still possible to infer parts of it with confidence. For instance, we might be able to assess that two clusters of cases are not related, despite not being able to reconstruct a detailed transmission history within each cluster. This is not conveyed by conventional measures of tree confidence, such as the phylogenetic bootstrap, that place an undue emphasis on individual branches. In addition, these measures of uncertainty do not convey the information needed in making public health decisions. For instance, how distant on a tree do two samples need to be to exclude that they resulted from the same superspreader event? Some methods address these issues within the framework of Bayesian phylogenetics—providing natural measures of phylogenetic uncertainty, and including explicit models of pathogen transmission[11]. However, these methods are typically

too computationally demanding to be practical on pandemic-scale data sets, and so the assessment of uncertainty thus remains a pressing open problem.

Another source of uncertainty is in the sequence data itself. The molecular phylogenetics community has usually been able to treat the data available to them as reliable in aggregate, as conventional analyses tend not to place too much reliance on individual data points. In contrast, individual data points are often the necessary focus of important public health decisions. In the SARS-CoV-2 pandemic, the urgency of sharing information meant that data became widely available before it was checked to the standard that the community (and the methods they use) expects, sometimes impacting the conclusions reached by papers written, circulated and translated to the media with consequences on public health policy and response. As we learn more about the kinds of errors that affect pandemic sequences, we can hope to devise methods that detect, tolerate or even correct for certain errors.

How to use the phylogenies to stay ahead of the virus?

Phylogenetic methods are routinely used to inform studies into the biology of living systems; for example, comparative genomics approaches, which often include a phylogenetic perspective, are powerful ways to identify genes and other functionally important sites in genomes, or detecting functionally constrained or non-constrained regions that might give clues to vaccine targets. Cancer is widely studied using evolutionary methods, for instance reconstructing ancestral sequences and to study the order of mutations and finding convergent change during tumour evolution[12]. Analysis of the variation exhibited by a population or between different species, including observing recurrent mutations, is a crucial part of this. For example, in the case of SARS-CoV-2, the D614G mutation in the spike protein has generated considerable interest because of its potential impact on virus infectivity and transmissibility[13]. In cases like this with a substantial signal, highly sophisticated analysis may not be necessary; however, few such examples have been found in SARS-CoV-2 and most instances of fitness differences are likely to be far more nuanced. How can we combine and strengthen existing models in order to detect reliable signals that give clues to SARS-CoV-2 biology against the sampled background of viral variation?

For example, observations of some mutations arising multiple times, others spreading rapidly within particular geographic regions, and still others associating with zoonotic transmission have sparked concerns about whether these variants could confer functional changes impacting transmission, treatment, or immunity (both infection- and vaccine-mediated). Considerable effort has been devoted to monitoring and quantifying the impact of both observed and hypothetical mutations: Datamonkey[14] scans SARS-CoV-2 phylogenies daily to identify sites with signatures of natural selection, and there have been experimental assessments of hundreds of receptor binding domain mutations for their impact on ACE2 binding and spike expression[15]. Similarly, multiple laboratories are using structural prediction and experimentation to characterise functional impacts of mutations. Following concerns about a new variant in Danish mink, attempts to track and monitor the spread of arising SARS-CoV-2 variants have increased. Finding a way to continuously and efficiently combine the information generated by each of these endeavors could help to more quickly direct focus onto mutations deserving deeper investigation and possible action. Understanding

how many times a mutation has arisen independently and how much it has spread is most easily accomplished by phylogenetic analyses.

Currently, direct actions that could be taken in response to such data, even when useful, may be limited. While it is perhaps unlikely that a worrying new mutation would be confidently identified quickly enough to curtail all spread, if serious enough, measures could be put in place to prevent the variant spreading nationwide or internationally. Even if, as we all hope, vaccination against COVID-19 starts to become widespread globally in the first half of 2021, knowledge about such mutations will be critical: as the number of immunologically naive individuals decreases through infection and vaccination, SARS-CoV-2 will be under increasing pressure to evade existing immunity. The global experience in mutation-monitoring systems for influenza provides an excellent starting point for similar structures working on SARS-CoV-2, but continuous monitoring networks and more information about mutational impacts is needed with this novel virus.

Another emerging use of phylogenetics is to improve the estimation of the effective reproduction number ("R_e")—the expected number of new infections caused by each infected individual. Traditional approaches estimate R_e from infection counts over time, whereas phylogenetic-based methods compare viral sequences and can thus potentially tease apart independent transmission clusters, making it possible to distinguish repeated imports of the virus from growing local cases and providing important information on how best to contain the spread of the virus. R_e values estimated from phylogenetic methods also have the potential to be more robust to sudden changes in testing and data reporting. Phylogenetics for R_e estimation has been used in regions with substantial sequencing (including Australia, New Zealand, and parts of the United States), but methods are still at the very forefront of development and will need to improve in speed and usability to become applicable widely.

Fostering genomic data sharing and collaboration

Besides the methodological challenges highlighted above, close collaboration both among scientists and between scientists and public health departments is of critical importance for phylogenetics to be routinely used to inform public health. This pandemic has already provided a tantalising glimpse into the benefits of putting aside differences and unifying against a common foe, but for this to become sustainable both in the current pandemic and in the future, new policies and incentives will also be needed.

How can pandemic genomic data be made available under open science principles while providing sufficient protection for data producers?

COVID-19 is the first epidemic where genome sequences of the infectious agent have become available to the research community in large numbers, in near real-time and on a near global scale—a tremendous boon for progress. This is a major advance over the shortcomings in data sharing observed during the West African Ebola outbreak in 2013-2016, where sequences were released only after publication, often months after being generated[16]. Yet data sharing and data reuse is still fraught. Because many sequences are currently generated by academic laboratories, the public health benefit of releasing sequences quickly and openly can be at odds with the need for sequence generators to be

the first to publish on the data they generate in order to receive full academic credit for their work.

From January 2020, the *de facto* standard data repository for SARS-COV-2 genomes rapidly became the GISAID database[17]. GISAID stands apart from the International Nucleotide Sequence Database Collaboration (INSDC—encompassing GenBank, the European Nucleotide Archive, and the DNA Database of Japan) by providing curation and offering sequences for use under conditions designed to provide greater protections to the data submitters, so that their sequences are not used in publications by others first. This protection has been critical in fostering sharing by data generators, but comes at some cost to reuse: unlike GenBank, publications involving sequences from GISAID require a good-faith effort to include all the data submitters, which can be impractical for analyses making use of tens of thousands of sequences. Further, the GISAID agreement against onward sharing of data, though intended to protect against third parties accessing the data without themselves signing up to the conditions of use, can limit even unpublished resources that contain enough information that sequence reconstruction might be possible, and hinders independent confirmation, peer review, and further development of some scientific work.

As long as sequencing is primarily achieved through researcher-lead initiatives, the sequence generators must feel that sharing sequences quickly and openly in pandemic scenarios will not endanger their own ability to publish the results of their work or gain other credit for their endeavours. However, it is not in the public health interest to incentivise a reliance on sequence generators to analyse data, or to suggest that scientists making use of others' data should do so only in unpublished analyses. How can we incentivise, help and protect those who generate genome sequence data while ensuring that their data is as widely available as possible? Many levels of nuance exist among data generators in what they feel is fair use. Expanding databases to allow authors to set varying levels of use—perhaps inspired by open-source licensing options—and 'embargos' for how long those restrictions are valid, could allow both data generators and re-users to feel more confident in data sharing.

As sequencing becomes ever cheaper and more accessible, routine and pandemic public-health sequencing could be moved out of the academic realm entirely. This would not only decouple sequencing production from grant cycles and reliance on interested academics, but should also result in fewer sharing restrictions.

How do we better align researchers and public health organisations going forward, to implement these things?

As the methodological barriers around pathogen phylogenetics are conquered and rapid open sequencing becomes the norm, integration of actionable phylogenetics within public health will rise—and with it, the need for phylogenetics expertise. The SARS-CoV-2 pandemic has seen inspiring instances of research and public health working closely together to wring maximum benefit from sequencing data, but many of these partnerships are informal, temporary, and not necessarily compensated either monetarily or in publishable output. Instead, long-term solutions to build both in-house phylogenetic expertise in public health and long-lasting collaborations between public health and research are needed.

During the current pandemic, many research-based scientists have been informally seconded to assist public health efforts and participate in government task forces. However, there are few ways to formally recognise such efforts by the metrics normally used to evaluate and assess academic output, which is largely based around publications and grant success. For early career researchers in particular, fear of such 'goodwill gaps' in their CV may discourage public health involvement, even when they likely lead to benefits in collaborations, implementation of new research methodologies, and better understanding of what is useful for public health.

The SARS-CoV-2 pandemic has also illustrated the importance of having routes to incorporate appropriate academic methods rapidly into public health frameworks. Rather than relying on sometimes *ad hoc* connections, building more formal bridges between public health agencies and academic research will allow faster implementation of useful analyses and techniques, hopefully co-developed to maximise public health benefit, to day-to-day public health use. For example, Nextstrain[18] has worked closely with SARS-CoV-2 consortia that include public health labs, such as SPHERES[19] in the US, to develop more accessible ways to run local phylogenetic analyses, and PHA4GE[20] has partnered with both sequence generators and public health agencies to provide guidelines on SARS-CoV-2 data standards. Further, supporting recommendations for interoperability, data management, and scalability, among others, will ensure that phylogenetic pipelines are accessible to use in public health settings[21].

One way to incentivise and foster connections between public health and research is to embrace work with groups such as the Declaration on Research Assessment (DORA), who are campaigning for better systems to evaluate and recognise the outputs of scholarly research[22]. For example, the successful implementation of a new phylogenetic outbreak investigation system in a local public health system might carry the same weight as a reputable publication. Academic institutions and funding agencies would also likely join in aiding to build such bridges if the resulting outputs were incorporated into existing metrics.

Calling phylogeneticists and public health practitioners to action

The challenges and losses brought about by the SARS-CoV-2 pandemic have been enormous, and the road still remains long. One bright spot has been the unparalleled production, sharing, and analysis of viral sequences for public health—with phylogenetics playing a central role driving this process. Still, for phylogenetics to mature from an intermittent, researcher-led effort to an integral part of the health system, we must refine data production, sharing, and analysis in ways that maximise public health benefit, and solidify connections between the researchers who develop the tools and the public health experts who will use them to actualise interventions and treatments. Progress toward such practical phylogenetics for public health will not only help end the COVID-19 pandemic sooner, but also help detect and counter future ones as effectively as possible.

References

1. Zhang, Y. Z. Novel 2019 coronavirus genome. *Virological. Org* (2020).
2. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
3. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
4. Eden, J.-S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol* **6**, veaa027 (2020).
5. du Plessis, L. *et al.* Establishment & lineage dynamics of the SARS-CoV-2 epidemic in the UK. *medRxiv* 2020.10.23.20218446 (2020) doi:10.1101/2020.10.23.20218446.
6. Hodcroft, E. B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv* (2020) doi:10.1101/2020.10.25.20219063.
7. Geoghegan, J. L. *et al.* Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat. Commun.* **11**, 6351 (2020).
8. Zhou, H. *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Current Biology* vol. 30 2196–2203.e3 (2020).
9. Morel, B. *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *Cold Spring Harbor Laboratory* 2020.08.05.239046 (2020) doi:10.1101/2020.08.05.239046.
10. COG--UK geographic coverage of Sars--Cov--2 sample sequencing. https://www.cogconsortium.uk/wp-content/uploads/2020/12/COG-UK-geo-coverage_2020-11-23_summary.pdf.
11. Wang, L. *et al.* Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nat. Commun.* **11**, 5006 (2020).
12. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
13. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).
14. Datamonkey. http://covid19.datamonkey.org/.
15. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.e20 (2020).
16. Yozwiak, N. L., Schaffner, S. F. & Sabeti, P. C. Data sharing: Make outbreak research open access. *Nature* **518**, 477–479 (2015).
17. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data--from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
18. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
19. CDC. SPHERES. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html (2020).
20. Griffiths, E. J. *et al.* The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. *other* (2020) doi:10.20944/preprints202008.0220.v1.
21. Black, A., MacCannell, D. R., Sibley, T. R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.* **26**, 832–841 (2020).
22. DORA. https://sfdora.org/ (2020).

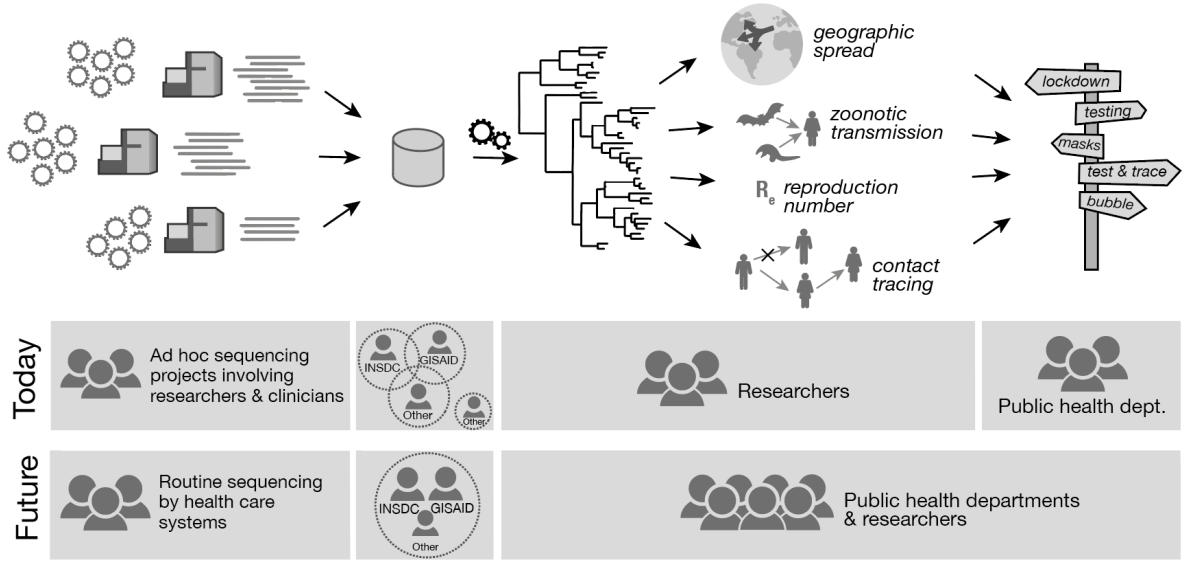

Figure 1. Phylogenetics has the potential to become an integral part of public health decision making, but this will require methodological improvements and tighter collaboration among sequence producers, phylogenetic method developers, and public health practitioners.