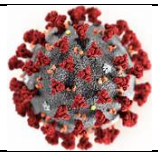
National COVID-19 Science Task Force (NCS-TF)



Type of document: update on previous policy brief	
In response to request from: Advisory Panel	**Date of request:** 7.6.2020
Expert groups involved: Data and modelling	**Date of response:** 15.6.2020
Contact person: Richard Neher richard.neher@unibas.ch, Tanja Stadler tanja.stadler@bsse.ethz.ch , Sebastian Bonhoeffer sebastian.bonhoeffer@env.ethz.ch and the Data and Modeling Group	
Comment on planned updates : None planned at the moment	

Title: Phylogenetic analysis in COVID-19 surveillance

Executive summary:

Small, often insignificant, changes in viral genomes allow reconstruction of the spread of SARS-CoV-2. Such analysis of viral genomes can complement contact tracing by establishing links between outbreak clusters or by linking new outbreak clusters to viruses circulating outside of Switzerland. While the temporal resolution of such genomic epidemiology is too low to establish direct transmission between two individuals with confidence, genomic information can often rule out that two cases are connected or suggest a linkage between clusters too far apart to be linked by contact tracing. In the current low-incidence situation, analysis of viral genomes can thus provide information to control COVID-19. To maximize impact, sequencing, analysis and sharing of viral genomes have to be fast. To reduce turn-around time, coordination between diagnostic labs, sequencing centers, and data deposition hubs should be improved. Real-time sequencing and phylogenetic analysis should be viewed as an integral part of the continued effort to contain COVID19.

Main text

As viruses replicate, mutations -- small changes -- accumulate in their genomes. These mutations usually do not change the virulence or transmissibility of the virus, but they can serve as markers that allow reconstructing the transmission history of an outbreak. This history can be reconstructed using phylogenetic methods. Such genomic epidemiology has proven to be highly successful in understanding the spread of emerging infectious diseases, such as Ebola or Zika [1,2], albeit often only retrospectively. In Switzerland, phylogenetic methods have been used to retrospectively track an outbreak of tuberculosis over many years [3]. In the current COVID-19 pandemic, however, genomic epidemiology is performed in near real-time due to unprecedented sequencing efforts and extremely timely data submission around the world. The SARS-CoV-2 genomes display global diversity with about a mutation occurring every 2 weeks on average.

Sequence data and phylogenetics provide insight into age, structure, and origin of an outbreak and can connect different outbreaks. The essence of how sequence data can be used to reconstruct the spread of a virus and connect different clusters is illustrated in Fig. 1: Similar genomes sharing many mutations likely belong to the same transmission cluster, while distinct genomes stem from different clusters. Phylogenetic methods also reveal hidden spread and allow the timing of introduction events [4]: the more diverse the sampled genomes are, the more time has elapsed since their common ancestor.

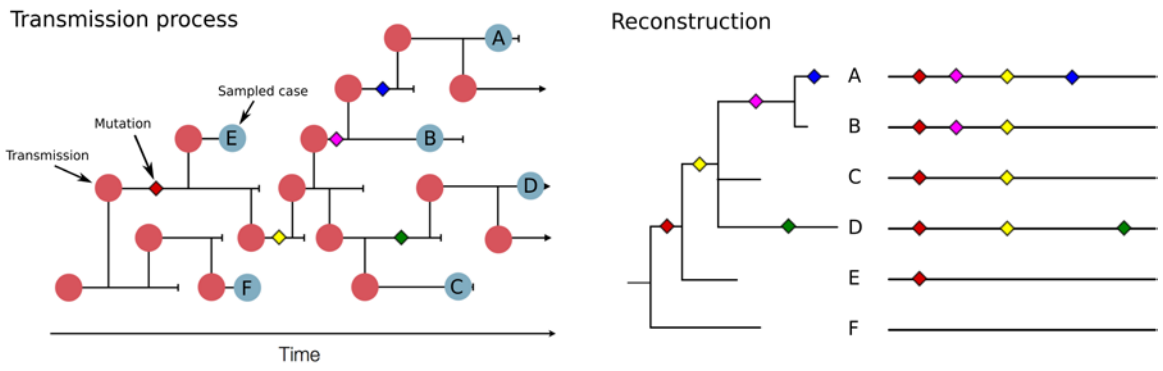


Fig 1: ***Mutations in viral genomes record a transmission tree.*** *As the virus spreads (left), random mutations (colored diamonds) in the genome are shared by all descendant viruses in a transmission chain. Consequently, we can use sequences (right, colored diamonds mark mutations) to reconstruct a transmission tree, albeit with limited resolution. In this hypothetical example, we could infer that samples A and B have a close recent common ancestor and might belong to the same cluster, while D is distinct. In most situations, only a fraction of cases are sampled (grey circles labelled with letters) and transmissions (red circles) connecting these cases are not observed.*

Earlier this year at the very beginning of the SARS-CoV-2 pandemic, all sampled genomes were very similar. This primarily showed that SARS-CoV-2 is the result of a single and very recent spill-over from animals to humans. By now, SARS-CoV-2 circulating around the globe has diversified considerably with many sampled genomes differing by about 10 mutations from each other. Distinct lineages are circulating around the globe, some localized, some cosmopolitan. As the viral population diversifies, our ability to detect novel introductions and identify distinct transmission chains and outbreaks increases. Genomic epidemiology could, therefore, be a valuable complement to classical contact tracing in that it allows connecting different transmission clusters that cannot be linked by contact tracing.

Current sequence data from Switzerland

We use Nextstrain [5], a phylogenetic analysis tool coupled to an interactive visualization, to provide an analysis of sequences from Switzerland in their international context. This analysis will be kept up-to-date and is available at the web address:

https://nextstrain.org/groups/swiss/ncov/switzerland

Currently (2020-06-15), this analysis includes 393 SARS-CoV-2 genomes from Switzerland covering most Cantons, see Fig. 2. These genomes were contributed by the Institute of Medical Virology at the University of Zürich, Laboratory of Virology at the University Hospitals of Geneva (HUG), the Clinical Virology and Clinical Microbiology units at the University Hospital Basel, and by the D-BSSE of the ETHZ in Basel (using samples provided by Viollier) and are shared via EpiCoV by the GISAID initiative.

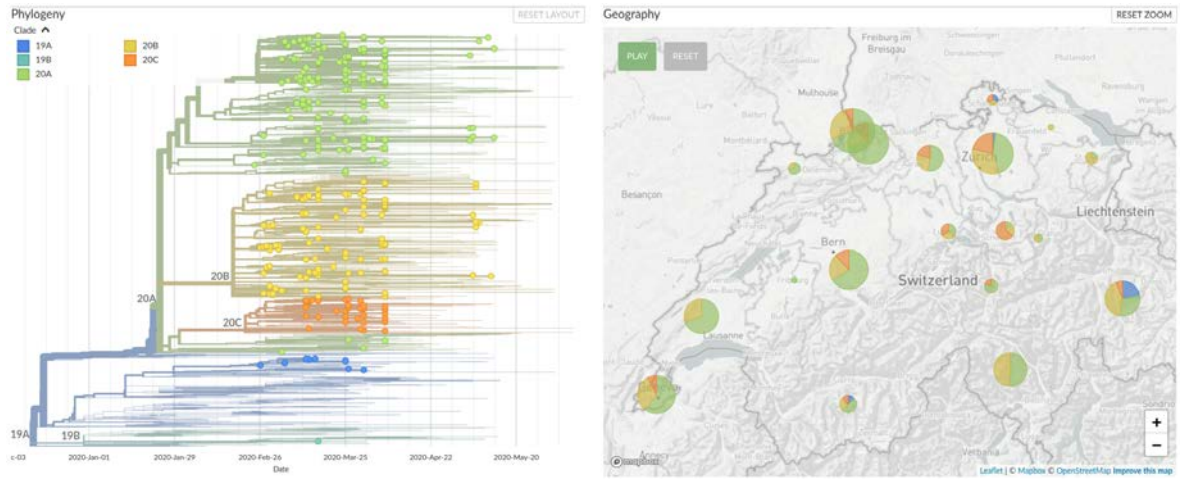


Fig 2: ***SARS-CoV-2 genomes in Switzerland.*** *The left panel shows a time scaled phylogenetic tree colored by the major genetic groups (clades) of the virus. Swiss samples are shown as colored circles, while international samples are hidden. The right panel shows the distribution of samples and their respective clades across Switzerland. The latest available samples from Switzerland are from early April 2020, while the international samples are available until mid May. This analysis will be updated continuously as more data become available.*

SARS-CoV-2 are currently categorized in five groups (clades) called 19A, 19B, 20A, 20B, and 20 C. Most of the Swiss sequences were sampled in late February or March and fall in clades 20A, 20B, or 20C. These genomes are closely related to other sequences from Europe, highlighting that SARS-CoV-2 across Europe is mostly linked to the large Italian outbreak that started late January or early February as clade 20A. Only a few samples fall into clade 19A (a subclade with mutation G26144T) that is less common than 20A but nevertheless observed in many European countries.

The 393 available genomes represent about 1.3% of known SARS-CoV-2 infections and an even smaller fraction of the total number of infections in Switzerland. Furthermore, these were mostly sampled before or shortly after border closures. Consistent with this, most Swiss sequences are nested among isolates from other European countries and do not show clear Switzerland-specific clustering. Nevertheless, we can identify a number of putative transmission clusters consisting of genomes that form tight clades in the tree without interspersed sequences from abroad. One example are sequences that share mutation A3116G, which were sampled in Zürich, Aargau and Schwyz between March 26 and March 31, see Fig. 3. This mutation has rarely been observed outside of Switzerland (there are samples from the UK), and the proximity of sampling locations supports the hypothesis that these sequences represent a local transmission cluster. Several other groups of identical genomes can be spotted in the tree, but they typically cannot be differentiated from international samples.

Genomic epidemiology can complement contact tracing

With time, the SARS-CoV-2 population accumulates diversity. This diversity increases our ability to assign sequences to transmission clusters. The temporal resolution of phylogenetic analysis of SARS-CoV-2 genomes is, however, fundamentally limited to about

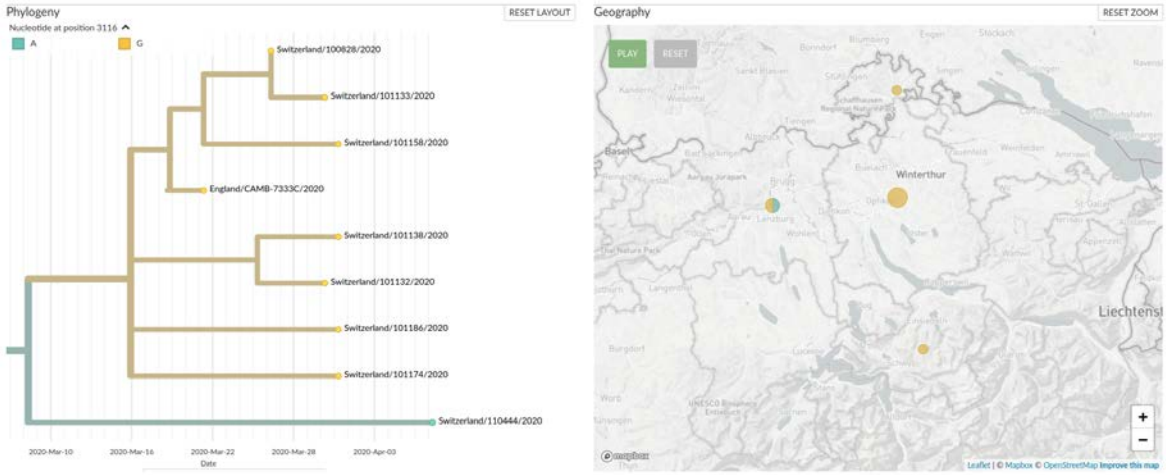


Fig 3: ***A putative transmission cluster 19B/A3116G.*** *The left panel shows zoom into clade 20C/A3116G that is dominated by sequences from the North-Eastern Switzerland (right panel).*

4 weeks and cannot establish direct epidemiological links. This limitation is due to the fact that mutations arise stochastically with an average rate of two mutations per month. Most transmission events will therefore involve identical genomes. Only after several steps in a transmission chain do we expect to find differences in the genomes. Despite this limitation, phylogenetic analysis can often rule out epidemiological links with confidence (if sequences are sufficiently different) and it can group cases in transmission clusters that cannot be linked by directly via contact tracing. Phylogenetic analysis is therefore complementary to classical contact tracing, in particular when a large fraction of all known cases are sequenced as is possible in a low incidence situation.

In addition to insights into the local circulation and transmission patterns, genomic epidemiology can help elucidate the sources of the reintroduction of the virus from abroad, which will become increasingly relevant as cross-border traffic resumes. Other countries in Europe are also sequencing SARS-CoV-2 genomes and share those data publicly. At present, the database includes 390 genomes from France, 215 from Germany, 132 from Italy, 250 from Austria, 18400 from the United Kingdom, 840 from the Netherlands, 617 from Belgium, 985 from Spain, etc. To enable SARS-CoV-2 tracking across borders, such data need to be produced rapidly and shared openly.

Conclusions and recommendations

Currently, the most recent publicly available sequence data from Switzerland is one month old. Reducing this lag is critical to maximizing the utility of sequence data. Once sequence data are available within 2 weeks of sampling (a turn-around often achieved in the UK or the Netherlands), phylogenetic analysis can assist investigation of ongoing outbreaks. Furthermore, with similar efforts underway in our European neighbors, we can contribute to pan-European genomic surveillance of SARS-CoV-2.

The added value of sequence information is particularly high in the current low incidence environment since moderate sequencing effort will cover a large fraction of cases and

labor-intensive case-based interventions (TTIQ - test, trace, isolate, quarantine) are feasible. Sequence information can complement these efforts, suggest transmission links not identified via contact tracing, improve the general understanding of remaining transmission chains, and link new transmissions chains to circulation in Europe and elsewhere.

Pipelines for sequencing and bioinformatic analyses are in place in different laboratories across Switzerland. We recommend that the different diagnostic laboratories coordinate sample collection and sequencing efforts to improve turn-around and ensure adequate coverage across Switzerland. In particular, with the current low daily case numbers, it is feasible to sequence and characterize all new cases across Switzerland.

References

[1] G. Dudas et al.Nature544:309–315 (2017).
[2] N. D. Grubaugh et al.Cell178:1057–1071.e11 (2019).
[3] Stucki D, et al. The Journal of infectious diseases. 2015 Apr;211(8):1306–16.
[4] E. M. Volz et al.PLOS Computational Biology9:e1002947 (2013).
[5] J. Hadfield et al.Bioinformatics(2018)