

# Haplotype Explorer: an infection cluster visualization tool for spatiotemporal dissection of the COVID-19 pandemic

Tetsuro Kawano-Sugaya\*, Koji Yatsu, Tsuyoshi Sekizuka, Kentaro Itokawa, Masanori Hashino, Rina Tanaka, and Makoto Kuroda

Pathogen Genomics Center, National Institute of Infectious Diseases, Toyama 1-23-1, Shinjuku, Tokyo, Japan

\*Corresponding author: Pathogen Genomics Center, National Institute of Infectious Diseases, Toyama 1-23-1, Shinjuku, Tokyo, Japan. Email: tks\_jp@seikai.org

## Abstract

The worldwide eruption of coronavirus disease 2019 (COVID-19) that began in Wuhan, China in late 2019 reached 10 million cases by late June 2020. In order to understand the epidemiological landscape of the COVID-19 pandemic, many studies have attempted to elucidate phylogenetic relationships between collected viral genome sequences using haplotype networks. However, currently available applications for network visualization are not suited to understand the COVID-19 epidemic spatiotemporally due to functional limitations that motivated us to develop Haplotype Explorer, an intuitive tool for visualizing and exploring haplotype networks. Haplotype Explorer enables to dissect epidemiological consequences via interactive node filters and provides the perspective on infectious disease dynamics depend on regions and time, such as introduction, outbreak, expansion, and containment. Here, we demonstrate the effectiveness of Haplotype Explorer by showing features and an example of visualization. The demo using severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes are available at <https://github.com/TKSjp/HaplotypeExplorer/blob/master/Example/>. There are several examples using SARS-CoV-2 genomes and Dengue virus serotype 1 E-genes sequence.

**Keywords:** SARS-CoV-2; COVID-19; haplotype network; infection clusters; infectious diseases; epidemiology

## Introduction

To control infectious diseases, it is important to quickly identify emerging infection clusters before they become critical issues. Many applications have been developed to assist researchers understanding the latest epidemiology. Indeed, the recent intensification of the coronavirus disease 2019 (COVID-19) pandemic, which began in late 2019 in Wuhan, China, has prompted development of new software to support investigations of this virus. For example, Nextstrain (Hadfield *et al.* 2018) is one of the most popular web services related to the COVID-19 pandemic which provides interactive molecular phylogenetic trees and geographic maps representing possible virus transmission routes. The COVID-19 Genome Tracker (Akther *et al.* 2020) is another unique application which shows the evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) using a haplotype network. This tool can dynamically display metadata, such as isolate conditions, locations, and mutations, compared to the reference genome. National Genomics Data Center in China also provides Viral Haplotype Network (Song *et al.* 2020). Although it is specialized for the COVID-19, the spatiotemporal eruption of that is visualized interactively.

So far, many phylogenetic trees and haplotype networks using the SARS-CoV-2 genome have been inferred because they are suited to interpret genetic and epidemiological relationships among sequences (Sekizuka *et al.* 2020a,b; Giovanetti *et al.* 2020).

In this time, haplotype networks are especially useful due to their potential for displaying short-term diversification of closely related genomes. Many available software programs for network inferring, such as TCS (Clement *et al.* 2000), PopART (Leigh and Bryant 2015), and Network (Bandelt *et al.* 1999), have supported these studies using haplotype networks of the SARS-CoV-2. Although these applications also work as network viewers, several alternatives are also available for additional annotation and exploration, including Cytoscape (Su *et al.* 2014), Gephi (Bastian *et al.* 2009), and tcsBU (Múrias dos Santos *et al.* 2016).

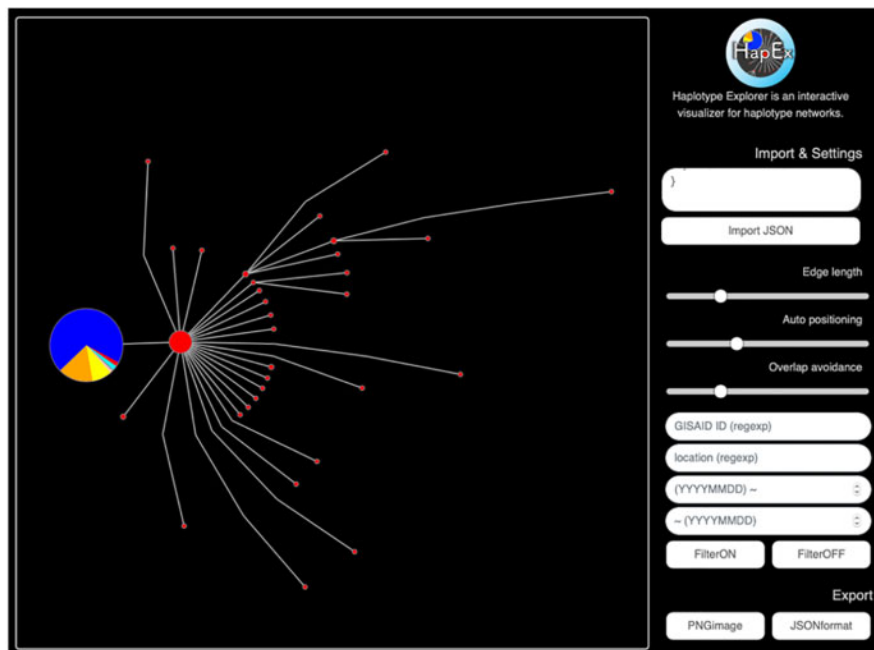
However, currently available tools are sometimes not the best to visualize infection clusters because they usually do not simultaneously fulfill the requirements essential to dissect epidemic situations: (1) nodes that can be dynamically filtered with metadata by complex search queries, (2) nodes can be indicated by real-time pie charts, which reflect sample size and content proportions at a given time span, and (3) creating interactive distribution files which require no external software installation. Hence, we endeavored to develop Haplotype Explorer, a specialized network viewer which assists onsite actions against emerging pathogens. Haplotype Explorer is a novel platform for network analysis displaying the network data from small scale to large scale and helping users to dissect the network by complex metadata filters. It also can export not only figure at a certain timepoint but continuous sections for constructing a movie to help people understanding the expansion of the pathogen at a glance.

Received: December 15, 2020. Accepted: March 25, 2021

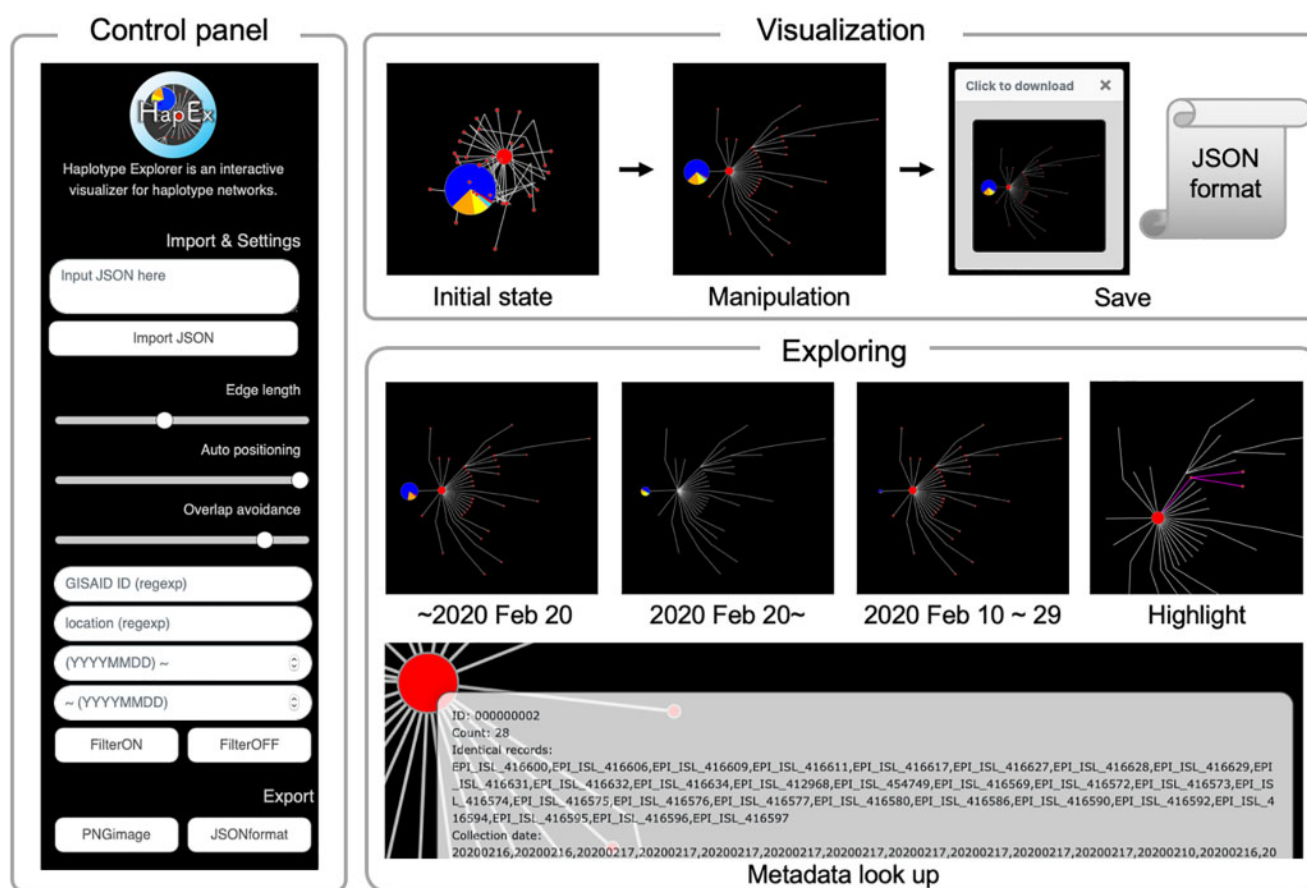
© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

A



B



**Figure 1** Introduction of Haplotype Explorer. (A) The view of Haplotype Explorer. (B) Introduction of features of Haplotype Explorer. It can modify distances and attractive forces among nodes and avoid overlapping automatically using the slide bar. Nodes can be moved by dragging. After manipulation of the appearance of the network, the view can be exported into PNG and JSON formats. Nodes are easily hidden or visible depending on keyword filters; accession ID, location, collection date from YYYYMMDD, and until YYYYMMDD. In cases where users specify dates, the pie chart is redrawn according to metadata so as to match to the queues. The metadata is displayed by mouse-hover, making it easy to inspect the node of interest.

# Haplotype Explorer

## View example data

Open Result\_example.html

## View in-house data

### Install dependencies

1. All depending programs other than network analyzing software (TCS) can be installed via Anaconda. If you do not have conda, download and install Anaconda (<https://www.anaconda.com>).
2. Open terminal and execute following command:  
`conda install -c bioconda seqkit mafft cd-hit snp-sites`
3. Download TCS from Softpedia (<https://www.softpedia.com/get/Science-CAD/Posada-TCS.shtml>)

### Retrieve sequences

1. Create your account of GISAID (<https://www.gisaid.org>)
2. Log-in GISAID and open EpiCoV/Browse.
3. Specify "Host" as "Human" and check on for "complete", "high coverage", and "low coverage excl".
4. Check on sequences of interest and download them as "input0.fasta". (Note: currently GISAID restricts downloading over 10,000 records all at once)

### \*Step1.py

seqkit

MAFFT

snp-sites

cd-hit-est

\*parseCdhit.py

\*makePhy.py

1. Open terminal and execute following command:  
`python3 Step1.py`

### \*Step2.py

TCS

1. Open terminal and execute following command: `python3 Step2.py`
2. Run TCS and export the result as "TCS.gml" in GraphML

### \*Step3.py

1. Open terminal and execute following command:  
`python3 Step3.py`

Open Result\_in-house.html

**Figure 2** The workflow and dependencies of Haplotype Explorer. Users can visualize own data in Haplotype Explorer by running bundled python program (createHTML.py). After preparation of "input.fasta," the python script automatically processes sequences, runs analysis, and produces the html file (result.html).

## Materials and methods

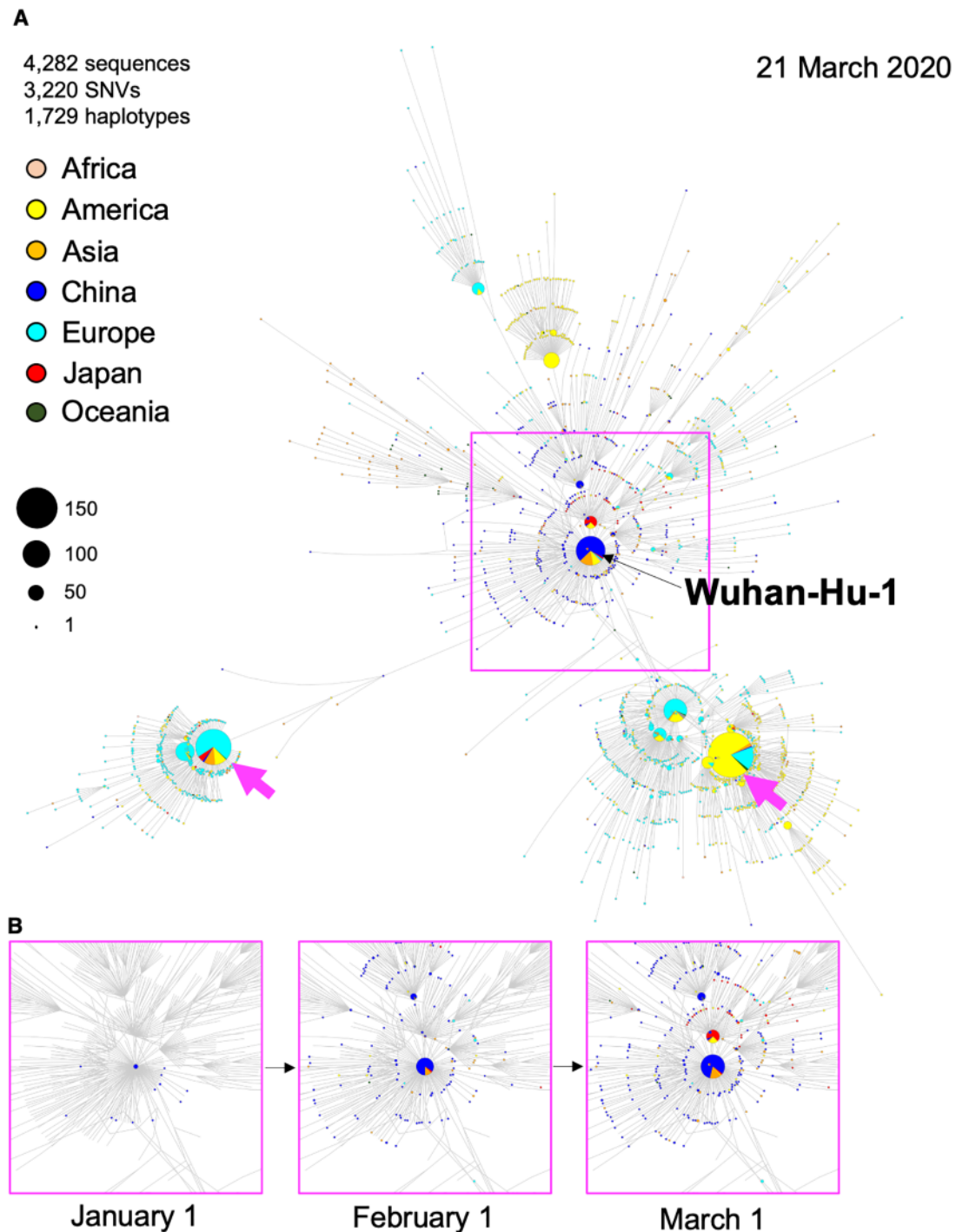
### Implementation and workflow

Haplotype Explorer is a JavaScript application executable in web browser, so it does not require uploading data to an external web server (Table 1). This allows users to analyze confidential data securely. It can produce distributions in HTML, which enables users to share originated networks with others easily. The network structure is written in JavaScript Object Notation (JSON) format which can be generated automatically from a multi-FASTA file with the provided python programs (createHTML.py). The production of network (result.html) from the raw multi-FASTA file

(input.fasta) is very straightforward like shown in Figure 2. In short, running "createHTML.py" will correct, curate, align sequences in input.fasta, execute TCS analysis, and convert data to an HTML file (result.html). We confirmed compatibilities of Haplotype Explorer and the bundled python scripts with the latest versions of Safari, Firefox, Edge, Chrome, and Python3 on macOS Catalina 10.15.3, respectively.

### Data analysis

Whole genome sequences were retrieved from global initiative on sharing all influenza data (GISAID) (Shu and McCauley 2017)



**Figure 3** Demonstration of spatiotemporal analysis of the SARS-CoV-2 genomic network using Haplotype Explorer. (A) An example of the exported network generated by Haplotype Explorer using 1729 of SNVs calculated from 4282 of world-wide SARS-CoV-2 genomes until March 21, 2020 obtained from the GISAID database. Each node size depends on sample size, and node colors differ by locations. The black arrow indicates the Wuhan-Hu-1 reference sequence. Magenta arrows indicate that distinct erosion of COVID-19 cases has occurred mainly in Europe or America. (B) Three snapshots of the SARS-CoV-2 genomic network around Wuhan, China from January 1 to March 1. Haplotype Explorer enabled us to dissect a haplotype network depending on metadata, giving significant insights into the epidemic.

on June 9, 2020 using the following options: (1) collection date was before March 21, 2020, (2) host was only human, (3) check was on for “complete,” “high coverage,” and “low coverage excl.” After retrieval of a total of 9583 sequences, they were curated using several external software [e.g., removing low-quality sequences, such as those containing spaces, gaps, degenerated bases, and ambiguous collection dates (i.e., month or date are absent in

collection date) using seqkit (Shen et al. 2016) and the Linux sed command]. Passing sequences were aligned by MAFFT (Katoh et al. 2002), clustered by CD-HIT (Fu et al. 2012; threshold: 100% identical, and SNVs were extracted by snp-sites; Page et al. 2016). The TCS analysis was run using extracted SNVs, and the resultant GraphML (.gml) file was converted into JSON format which is compatible to Haplotype Explorer. In following analyses, we

**Table 1** Comparison of features with other applications

	Haplotype Explorer	tcsBU (Múrias dos Santos et al. 2016)	COVID-19 Genome Tracker (Akther et al. 2020)	Cytoscape (Su et al. 2014)	Gephi (Bastian et al. 2009)	PopART (Leigh and Bryant 2015)	TCS (Clement et al. 2000)
Generate explorable HTML distribution	✓						
Time-dependent node pie charts	✓						
Filter by metadata	✓		✓	✓	✓	✓	
Resume operation	✓			✓	✓	✓	
Export image	✓	✓		✓	✓	✓	✓
Use in-house data	✓	✓		✓	✓	✓	✓
Real-time orchestration of node positions	✓	✓	✓		✓		
Run on web browser	✓	✓	✓				

The significance of Haplotype Explorer is that it enables users to generate an explorable HTML distribution, which includes several features all in one. It requires no external software other than a modern web browser to open the network, making it easy to share data. Furthermore, it can draw nodes as pie charts based on the specified span input into the search boxes, supporting spatiotemporal dissection of the network.

collected figures by applying the filters “~YYYYMMDD” from the initial day (December 31, 2019; Wuhan-Hu-1) to March 21, 2020.

### Data availability

All DNA sequences used in this study were downloaded from the GISAID. Source codes are available from the following URL: <https://github.com/TKSjp/HaplotypeExplorer/>. Supplemental Material available at figshare: <https://doi.org/10.25387/g3.14349563>.

## Results and discussion

### Principal features of Haplotype Explorer toward epidemic dissection

The primary feature of Haplotype Explorer is a vibrant and interactive visualization function utilizing D3.js (Bostock et al. 2011) and metadata, including sample size, accession number, collected location, and collection date, which are important clues for understanding the epidemic. Each node is represented by differently sized pie charts calculated from sample number and location proportion described in the metadata. Nodes and related edges can be interactively highlighted when a specific node is left-clicked, making it easy to dissect a crowded network with large samples (Figure 1). Users can quickly look into the node of interest by zooming with the scroll-wheel, and show metadata by mousing-over the tool-tip window. The application has four text boxes for filtering nodes: Sequence ID, location, YYYYMMDD~, and ~YYYYMMDD. Filters can be combined, and the Sequence ID and location can be specified by regular expressions. The current view of the network can be exported in a JSON format file, and users can resume it by importing the JSON. Finally, the current SVG view can be converted into a high-resolution PNG image using the export button. We also provide python scripts for assisting haplotype network construction with in-house data. Details are shown in the flow-diagram (Figure 2).

### Demonstration of Haplotype Explorer: spatiotemporal dissection of multimodal epidemics

The epidemic context of the SARS-CoV-2 from January 1 to March 21 was visualized by Haplotype Explorer (Figure 3). We began by capturing a snapshot for March 21, 2020 as an overall view. Haplotype Explorer effectively discerned significantly large, but distinctly invaded clusters consisting of a dozen to over one

hundred genome collections formed by late March (Figure 3A; magenta arrowhead). In order to understand epidemics in a time-dependent manner, Haplotype Explorer can also generate snapshots for specified dates (Figure 3B).

### Author contributions

Manuscript preparation: T.K.-S. M.K. Data analysis: T.K.-S., M.K., K.Y., and T.S. Data collection: K.I., M.H., and R.T.

### Acknowledgements

We would like to thank Editage ([www.editage.com](http://www.editage.com)) for English language editing.

### Funding

This study was supported by a Grant-in Aid from the Japan Agency for Medical Research and Development (AMED) under Grant number JP19fk0108103, JP19fk0108104, and JP20fk0108063. The funding agencies had no role in the study design, data collection or analysis, decision to publish, or manuscript preparation. We would like to thank all authors who have deposited genome sequences in the GISAID EpiCoV Database.

### Conflicts of interest

The authors declare no conflicts of interest associated with this manuscript.

### Literature cited

- Akther S, Bezrucenkovas E, Sulkow B, Panlasigui C, Li L, et al. 2020. CoV Genome Tracker: tracing genomic footprints of Covid-19 pandemic. *bioRxiv*. doi: 10.1101/2020.04.10.036343 (Preprint posted April 14, 2020).
- Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16:37–48.
- Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. In: *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17–20, 2009*.



- Bostock M, Ogievetsky V, Heer J. 2011. D<sup>3</sup> data-driven documents. *IEEE Trans Vis Comput Graph.* 17:2301–2309.
- Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol.* 9:1657–1659.
- Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics.* 28:3150–3152.
- Giovanetti M, Angeletti S, Benvenuto D, Ciccozzi M. 2020. A doubt of multiple introduction of SARS-CoV-2 in Italy: a preliminary overview. *J Med Virol.* 92:1634–1636.
- Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, et al. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* 34:4121–4123.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acid Res.* 30:3059–3066.
- Leigh JW, Bryant D. 2015. PopART: Full-feature software for haplotype network construction. *Methods Ecol Evol.* 6:1110–1116.
- Múrias dos Santos A, Cabezas MP, Tavares AI, Xavier R, Branco M. 2016. tcsBU: a tool to extend TCS network layout and visualization. *Bioinformatics.* 32:627–628.
- Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2:e000056.
- Sekizuka T, Itokawa K, Hashino M, Kawano-Sugaya T, Tanaka R, et al. 2020a. A genome epidemiological study of SARS-CoV-2 introduction into Japan. *mSphere.* 5(6):e00786-20.
- Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, et al. 2020b. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. *Proc Natl Acad Sci USA.* 117:20198–20201.
- Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One.* 11:e0163962.
- Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data—from vision to reality. *Euro Surveill.* 22:30494.
- Song S, Ma L, Zou D, Tian D, Li C, et al. 2020. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoV. *bioRxiv.* doi:10.1101/2020.08.30.273235 (Preprint posted September 3, 2020).
- Su G, Morris JH, Demchak B, Bader GD. 2014. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics.* 47:8.13.1–8.13.24.

Communicating editor: B. J. Andrews