Informatics in Medicine Unlocked 21 (2020) 100442

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Virus database annotations assist in tracing information on patients infected with emerging pathogens

Akiko Nakashima [a,*], Mitsue Takeya [a], Keiji Kuba [b], Makoto Takano [a], Noriyuki Nakashima [a]

[a] Department of Physiology, Kurume University School of Medicine, Asahi-machi 67, Kurume, Fukuoka, 830-0011, Japan
[b] Department of Biochemistry and Metabolic Science, Akita University Graduate School of Medicine, 1-1-1 Hondo, Akita, 010-8543, Japan

ARTICLE INFO

Keywords:
Emerging viruses
SARS-CoV-2
Substrains
Clinical feature estimation
Transmission history
Medical annotations

ABSTRACT

The global pandemic of SARS-CoV-2 has disrupted human social activities. In restarting economic activities, successive outbreaks by new variants are concerning.

Here, we evaluated the applicability of public database annotations to estimate the virulence, transmission trends and origins of emerging SARS-CoV-2 variants. Among the detectable multiple mutations, we retraced the mutation in the spike protein. With the aid of the protein database, structural modelling yielded a testable scientific hypothesis on viral entry to host cells. Simultaneously, annotations for locations and collection dates suggested that the variant virus emerged somewhere in the world in approximately February 2020, entered the USA and propagated nationwide with periodic sampling fluctuation likely due to an approximately 5-day incubation delay. Thus, public database annotations are useful for automated elucidation of the early spreading patterns in relation to human behaviours, which should provide objective reference for local governments for social decision making to contain emerging substrains. We propose that additional annotations for past paths and symptoms of the patients should further assist in characterizing the exact virulence and origins of emerging pathogens.

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has disrupted social and economic activities worldwide since the first outbreak in China in 2019 [1]. COVID-19 presents varied symptomatic features [2–7] with a wide range of incubation periods and epidemic curves across the world [1,4,8], which is likely influenced by many factors, including age, region, preventive measures, health care infrastructure, and the fundamental nature of the virus [3,9,10]. Characterization of the pathogens and preparations for medication and preventive vaccines are being urgently and extensively explored worldwide [11–14].

On restarting economic activities, the second wave of SARS-CoV-2 cases is concerning [15,16]. In Japan, the initial wave of SARS-CoV-2 infection was suppressed by cluster tracking and quarantine [17]. Since the lifting of the emergency declaration, the number of cases has been rapidly increasing [16,18], which necessitates rapid decision making regarding policy updates [19] for deregulation balance in social activities including travelling and business. Currently, the suspected patients are checked for symptoms and tested using reverse-transcriptase polymerase chain reaction (RT-PCR) and antibody-associated immunological assays [20]. Diagnostic tests with high sensitivity and specificity are necessary [21–23]. To trace patient behaviours, local cluster tracking [17] is effective, but information accuracy depends on the reliabilities of directly interviewing the infected individuals. Mobile phone applications (apps) to trace past close contact with patients via Bluetooth technology are utilized in some areas, including Japan [24,25]. However, limitations of such mobile apps may exist for the global tracing outside of the service areas, the user number in all age groups [26], self-reporting reliability or the definition of a close-contact period and distance. Ethical problems may also exist in identifying the index case in small communities and exposing them to scrutiny and harsh judgment due to panic and anxiety [27,28].

Simultaneously, the emergence of mutated variants of SARS-CoV-2 has been confirmed [12,29,30]. RT-PCR, immune assays for diagnostics, medications for treatment or vaccines for prevention might be vulnerable to mutated substrains. Although technical advances are occurring [21,22], the pathogenicity and origins of the mutated substrains of SARS-CoV-2 should be available in real time to adopt early measures by authorities at the onset of emergence.

[*] Corresponding author.
E-mail address: nakashima_akiko@med.kurume-u.ac.jp (A. Nakashima).

https://doi.org/10.1016/j.imu.2020.100442
Received 31 July 2020; Received in revised form 20 September 2020; Accepted 3 October 2020
Available online 8 October 2020

In parallel with individual treatment at hospitals and clinics, specimens from infected patients are directly sequenced, and the genetic information of SARS-CoV-2 is being globally sampled and added to public databases [31–33]. The databases have been used to predict viral transmissibility, antibody affinities and drug efficacy [34]. The cross-disciplinary usability of databases should promote the feedback of accumulating raw data to predict the actual profiles of pathogenic diseases [35]. Simple real-time surveys with regional public assistance are fundamentally necessary in an internationally available format.

Here, we utilized these database annotations to detect virus variants and to estimate the virulence and transmission trajectories of the emerging substrains. We examined the nucleotide mutations and visualized the transmission trajectories of SARS-CoV-2 by consulting the world specimens registered in the virus data bank of the National Center for Biotechnology Information (NCBI) [32].

2. Methods

2.1. Data acquisition for SARS-CoV-2 and other specimens

Due to its accessibility to the raw data of nucleotides and proteins with multiple annotations in a simple FASTA format, we used the data deposited in the NCBI Virus-SARS-CoV-2 data hub [36]. In the “Refine results” window, we specified the data by release date 2019/1/11–2019/5/3 (From 11 Jan 2019 to 3 May 2020). The latest data at that point were deposited on 1 May 2020. In the “Results” window, we rearranged the “Length” in ascending order. Then, we obtained 23042-FASTA formatted data of Protein containing Accession, GenBank Title, Geo_Location, Host, Species and Nucleotide Completeness in order, and 2051-FASTA formatted data of Nucleotide. For another analysis, we also obtained 23042-table view result data of Protein and 2051-table view result data of Nucleotide in CSV formats containing Accession, Release_Date, Species, Genus, Family, Length, Sequence_Type, Nuc_Completeness, Genotype, Segment, Authors, Publications, Geo_Location, Host, Isolation_Source, Collection_Date, BioSample and GenBank_Title. The protein and nucleotide sequences in one letter codes were analysed by using the Excel filter function. Accession numbers for viral genomes are SARS-CoV-2, NC_045512.2; SARS-CoV, NC_004718.3; MERS, NC_019843.3; HCoV-NL63, KF530114.1; HCoV-229E, KF514433.1; HCoV-HKU1, NC_006577.2; and HCoV-OC43, KX344031.1. The homology alignment was performed using the online tool CLUSTALW [37]. See Supplementary notes 1 and 2 for the step-by-step procedures.

2.2. Entropy calculation

To evaluate the proportions (P) of mutations at a certain ith residue of the S protein, we used the information entropy [30] as a sum of $P_{AA}\ln P_{AA}$, where AA is 20 biological amino acids. $P_{AA}\ln P_{AA}$ was defined as zero when the mutated residue was not detected in the samples at the ith residue. The calculated entropy scores were plotted along the whole S protein structure to yield the spectrum view of the scores as an entropy spectrum. See Supplementary note 1 for how to use the program.

2.3. Protein conformational analysis

Model structures with D614G mutagenesis were constructed using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC., NY, USA). The structural models for coronavirus spike glycoprotein are 6vsb [38] (SARS-CoV-2), 5xlr [39] (SARS-CoV), 5x5c [40] (MERS), 6nzk [41] (HCoV-OC43) and 6u7h [42] (HCoV-229E) obtained from the Protein Data Bank [43] (www.rcsb.org). For the hydrophobicity search, we consulted the online resource portal of the Swiss Institute of Bioinformatics (https://web.expasy.org/protscale/) [44].

2.4. Program for sequential data analysis

We originally built a program to manipulate big data. Codes are provided as supplementary data by Excel Visual Basic (Office Professional 2016, Microsoft Corporation, WA, USA). All the source codes of the programs are provided with the annotation. Each program operates as follows: “prPCVcov2” aligns each single letter code of all the amino acid sequences separately in each cell for all the different protein datasets in a FASTA formatted data file; “priNuc” extracts a text string “GAT” or “GGT”, which comes after a text string “CCAGGTTGCTGTTCTTTATCAG”. See Supplementary notes 1 and 2 for how to use the program.

2.5. Graph design

The processed metrics were visualized by Kaleida Graph 4 (HULINKS Inc., Tokyo, Japan), and artworks were originally created with Illustrator (Adobe Systems Incorporated, CA, USA). The SVG data were obtained from the public domain under the license of CC0 1.0 at https://en.wikipedia.org/wiki/File:BlankMap-World6-Equirectangular.svg. The world map originated from the United States Central Intelligence Agency’s World Fact Book.

2.6. Spectral analysis of sampling periodicity

The periodicity of sampling of the mutated specimens was analysed by power spectrum using Axograph X (Version 1.7.4).

3. Results

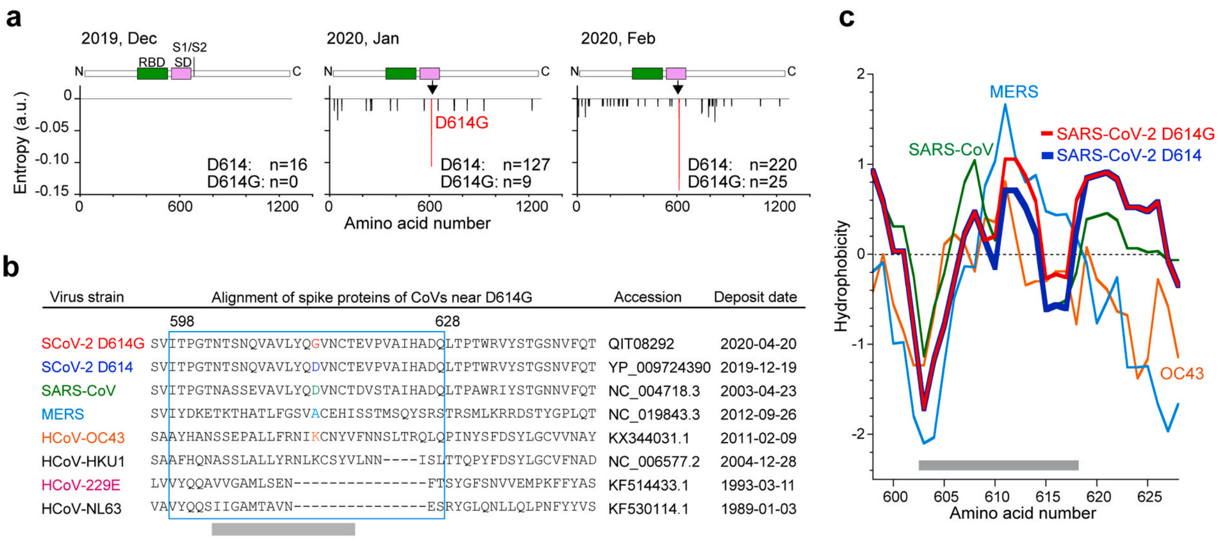
3.1. Annotation search detects multiple conversions across the proteins of SARS-CoV-2

Coronaviruses are unique RNA viruses equipped with proofreading machinery [45]. However, substantial mutations were expected, leading to the overestimation of substrains with unchanged genetic codons. On the other hand, amino acid mutations occur less frequently due to the wobble nature of codons [46].

Therefore, we consulted the NCBI database [36] and utilized a downloaded data table with all the applicable annotations (see also Methods). Among them, partial sequences or incomplete readouts were eliminated. We used 1500–2000 nucleotide and protein sequences with all applicable annotations, including sampling dates, locations, and genetic information of the virus.

We detected the accumulation of the same mutations or the branching to multiple amino acids at approximately 100 residues in several component proteins of SARS-CoV-2 (Supplementary Table 1). Despite the genome proofreading ability of coronaviruses [45], multiple random mutations in SARS-CoV-2 have been reported [47,48]. Any of these conversions might be attributable to increased or decreased virulence of viral particles [29]. In particular, the presence and increase of identical mutations at the same residues from different specimens could be due to the transmissible pathogenic substrains of SARS-CoV-2 [12]. Mutations in the amino acid sequences have indeed occurred in different phases of the COVID-19 pandemic and are probably fixed, inherited and dominantly spreading around the world.

However, the pathogenicity and exact origins of these variations are difficult to retrace only using this mutation profile. Among the proteins with frequent mutations, the surface glycoprotein, namely, the spike or S protein, contained a single eminent mutation from aspartate (D) to glycine (G) at 614 (D614G conversion; Supplementary Table 1). The relative mutation frequency at each residue was calculated as information entropy to digitize the variations across the S protein in the database [30] and visualized in a spectral view across the S protein: D614G appeared in the early stage of the COVID-19 pandemic and accumulated over time (Fig. 1a). Among the D614G substrain, additional major



Virus strain	Alignment of spike proteins of CoVs near D614G	Accession	Deposit date
SCoV-2 D614G	SVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQT	QIT08292	2020-04-20
SCoV-2 D614	SVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQT	YP_009724390	2019-12-19
SARS-CoV	SVITPGTNASSEVAVLYQDVNCTDVSTAIHADQLTPAWRIYSTGNNVFQT	NC_004718.3	2003-04-23
MERS	SVIYDKETKTHATLFGSVACEHISSTMSQYSRSTRSMLKRRDSTYGPLQT	NC_019843.3	2012-09-26
HCoV-OC43	SAAYHANSSEPALLFRNIKCNYVFNNSLTRQLQPINYSFDSYLGCVVNAY	KX344031.1	2011-02-09
HCoV-HKU1	SAAFHQNASSLALLYRNLKCSYVLNN----ISLTTQPYFDSYLGCVFNAD	NC_006577.2	2004-12-28
HCoV-229E	LVVYQQAVVGAMLSEN---------------FTSYGFSNVVEMPKFFYAS	KF514433.1	1993-03-11
HCoV-NL63	VAVYQQSIIGAMTAVN---------------ESRYGLQNLLQLPNFYYVS	KF530114.1	1989-01-03

Fig. 1. Comparison between highly pathogenic and common coronaviruses.
(**a**) Schematic of the SARS-CoV-2 spike (S) protein (upper) and the entropy spectra (lower) to visualize mutations in S protein at different dates. RBD = receptor-binding domain, SD = subdomains, S1/S2 = protease cleavage site between N-terminal S1/C-terminal S2 domains. N and C = amino and carboxyl termini. D614 is in SD. (**b**) Homology alignment of the residues around D614 between highly pathogenic coronaviruses and other human coronaviruses (HCoV-OC43, -HKU1, -229E and -NL63). HCoVs are less lethal and common cold viruses. Middle East respiratory syndrome coronavirus (MERS) contains A614, but the surrounding residues are also variable compared to SARS-CoV and SARS-CoV-2 (SCoV-2). Among HCoVs, only HCoV-OC43 possesses the equivalent residues with relative homology to SARS-CoV-2. Amino acid numbering is based on SARS-CoV-2. (**c**) Hydrophobicity between the 598–628 residues of different coronaviruses. The hydrophobicity presents a steep increase in SARS-CoV, SARS-CoV-2 with D614G conversion, and MERS compared to that in the initial SARS-CoV-2 with D614 between 603 and 618 residues indicated in the grey bars in (**b**) and (**c**).

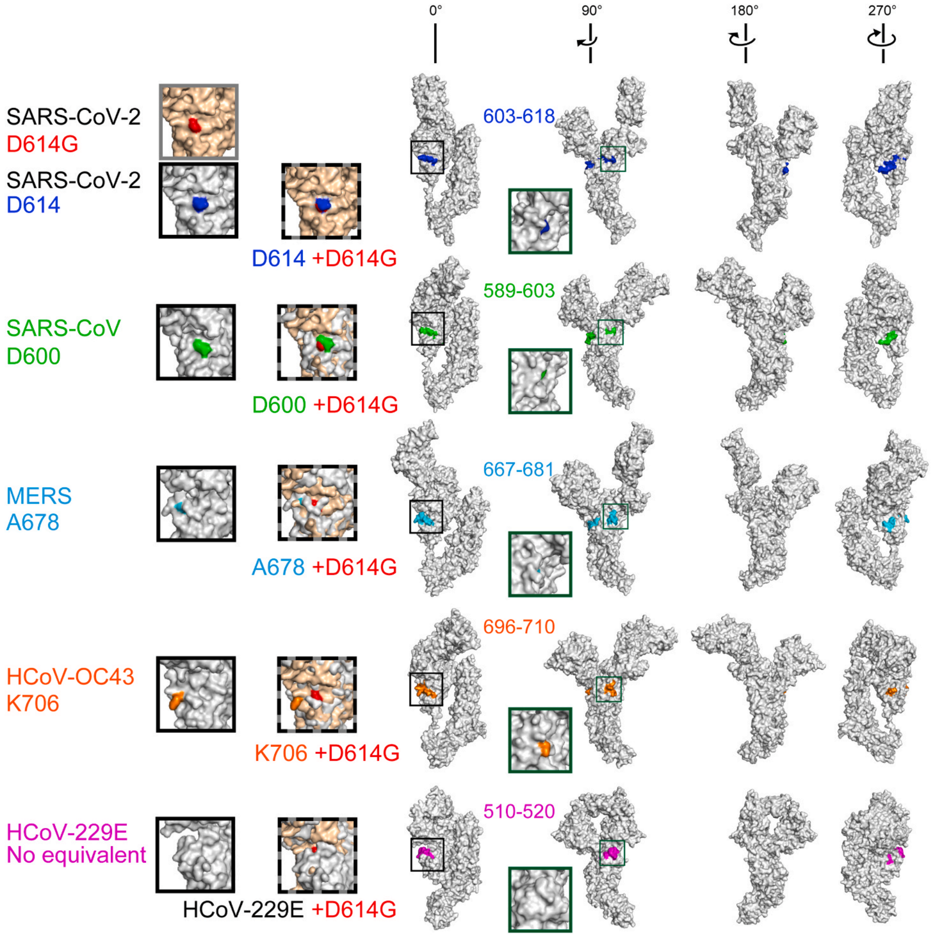


Fig. 2. Comparison of D614 and equivalent residues across coronaviruses.
The entire surface views of the available spike crystal structures of SARS-CoV-2, SARS-CoV, MERS, HCoV-OC43 and HCoV-229E. See also the grey bar in Fig. 1b and 1c. The SARS-CoV-2 S protein with D614G conversion is indicated in beige, and the other spikes are in light grey; the aligned structures are shown in mosaic colours. Homologous regions surrounding D614 of SARS-CoV-2 are highlighted with distinct colours. A box with grey solid lines shows a close-up view of D614G conversion in SARS-CoV-2. Boxes with black solid lines show the equivalent residues in other CoVs. Boxes with black-grey dotted lines show the structural alignment of equivalent regions between SARS-CoV-2:D614G and the other respective CoVs. HCoV-229E lacks the residue equivalent to D614 of SARS-CoV-2. Boxes in green indicate the close-up view of the equivalent residues rotated by 90°; not overlaid. The data for HCoV-HKU1 and HCoV–NL63 were not found in the database. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

mutations accumulated in other viral proteins in contrast to the D614 original strain (Supplementary Table 1, Supplementary Figure 1); the D614G could be an initial mutation for a more dominant substrain circulating in the world afterwards. Therefore, we next investigated the possible impact of the D614G mutation in the S protein of the converted substrain by the structural analysis and estimated the regional origins by the sampling periodicity analysis based on the obtained Excel data.

3.2. D614G conversion in the S protein may affect viral entry

We consulted the Protein Data Bank [43] and the Swiss Institute of Bioinformatics resource portal [44] for the subsequent structural analysis. The spike of SARS-CoV-2 forms a homotrimer. Each S protein, comprising approximately 1300 amino acids, is a large transmembrane protein containing two subdomains, S1 and S2, which are responsible for receptor binding and membrane fusion, respectively [49] (Fig. 1a). D614 in S protein is conserved in SARS-CoV in 2003 as well as in the initial isolates of SARS-CoV-2 in China in January 2020 [50].

Compared to less lethal human coronaviruses [51], highly pathogenic coronaviruses possess a large increase in hydrophobicity upstream of D614 in the subdomain (Fig. 1b and 1c). Moreover, D614 and the corresponding residues slightly deviated or did not exist in the equivalent positions among the other coronaviruses (Fig. 2).

Structurally, D614 is embedded in the S1 domain of the S protein, facing another protein unit within the trimer (Fig. 3a and 3b), but this aspartic acid residue is not accessible from the orifice for receptor binding. Thus, D614G conversion is expected to change the inter- and intramolecular properties of the spike trimers. Molecular simulation predicted that the single D614G replacement would increase the thermal fluctuation not only in the vicinity but also throughout S protein, especially in the S2 subunit near the viral membrane [52] (Fig. 3c–f). D614G conversion resulted in the deletion of the side chain of the aspartic acid residue, and the distance between D614 and T859 of another protein unit should expand from 4.4 to 6.4 angstroms (Fig. 3g and 3h). This estimation indicates that the D614G mutation should change the inter-subunit interaction in the subdomain and the conformational state of the receptor binding domain so that the mutated viral particles can effectively interact with its cognate receptor in host cells for viral entry [12,34,53].

3.3. GAU to GGU conversion can be traced back to February 2020

Next, we investigated the mutations at the genome level to retrace the transmission history of the D614G-converted virus. By analysing the nucleotide database, we detected the identical conversion from guanine-adenine-uracil to guanine-guanine-uracil (GAU to GGU) in all the D614G-converted cases despite 8 more convertible codons.

We then retraced the specimens with a GGU mutation. The GGU specimens exponentially increased in March 2020 worldwide (Fig. 4a). As of May 1, the conversion was found in more than half of the databank resource samples and similarly in almost all the regions (Fig. 4a). The SARS-CoV-2 virus originally isolated in Wuhan, China, was a non-converted GAU type [50,54] and so were all the specimens reported in China until the *de facto* termination of an emergency state in March. This GAU-to-GGU conversion was first detected in a specimen in Spain (MT292580) and next in the United States; there was also one from New England (MT276323) and two from Florida (MT276329, MT276330) collected on 28 February and one each from New Hampshire (MT304484) and Georgia (MT276327) on 29 February (Fig. 4b). Interestingly, the mutated specimens were sampled in distant cities in the USA on 1 March 2020, in Washington (MT415895), Connecticut (MT350239), and California (MT304491), implying that the D614G conversion might confer the virulence of SARS-CoV-2 in Europe and America [29,54]. This result suggests that the patient had close contact with another patient along their mobility history.

3.4. Transmission of D614G substrain within the USA after entry

Specimens of the original SARS-CoV-2 without mutations had already been reported in the United States in January 2020 (Fig. 5a). The sampling ratio of the GGU-mutated specimens with respect to the original GAU specimens suddenly increased at the end of February, followed by periodic fluctuations (Fig. 5b). The spectral analysis indicated that the predominant transmission interval ranged from 4 to 6 days. This period most likely corresponds to the approximate incubation delay at the early phase of transmission of the mutated substrain within the United States (Fig. 5c). When the database was consulted again on 22 June 2020, the deposited data increased not only in total number (from 1866 to 7596 specimens in the world) but also in the number of monthly specimens based on collection dates: January, from 84 to 129; February, from 78 to 189; March, from 1527 to 4155; April, from 161 to 2468 specimens. Therefore, many specimens were deposited after a

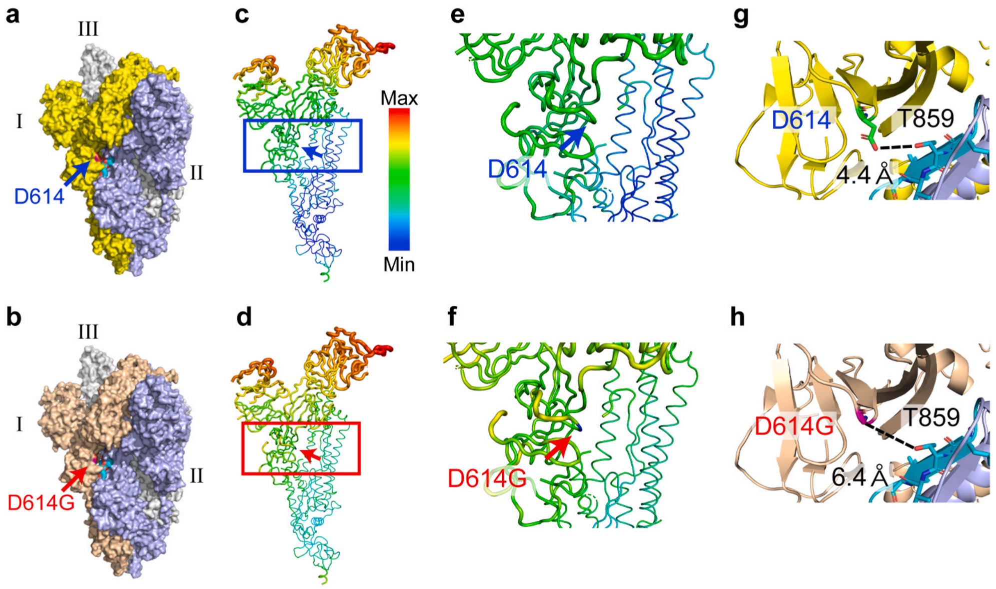


Fig. 3. D614G could affect the molecular stability of the S protein. (**a**, **b**) Spike structure of (**a**) D614 and (**b**) D614G. I-III indicate the trimeric S protein units shown in different colours. The D614 atoms of unit I are highlighted in blue and red. (**c**–**f**) Thermal fluctuation profiles by Debey-Waller factor (**c**, **d**) across the whole protein unit and (**e**, **f**) in the close-up views of the vicinities of D614 and D614G, respectively. The heat map represents the atomic displacement. (**g**, **h**) Estimated distance from (**g**) D614 or (**h**) D614G to T859 of the other protein unit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

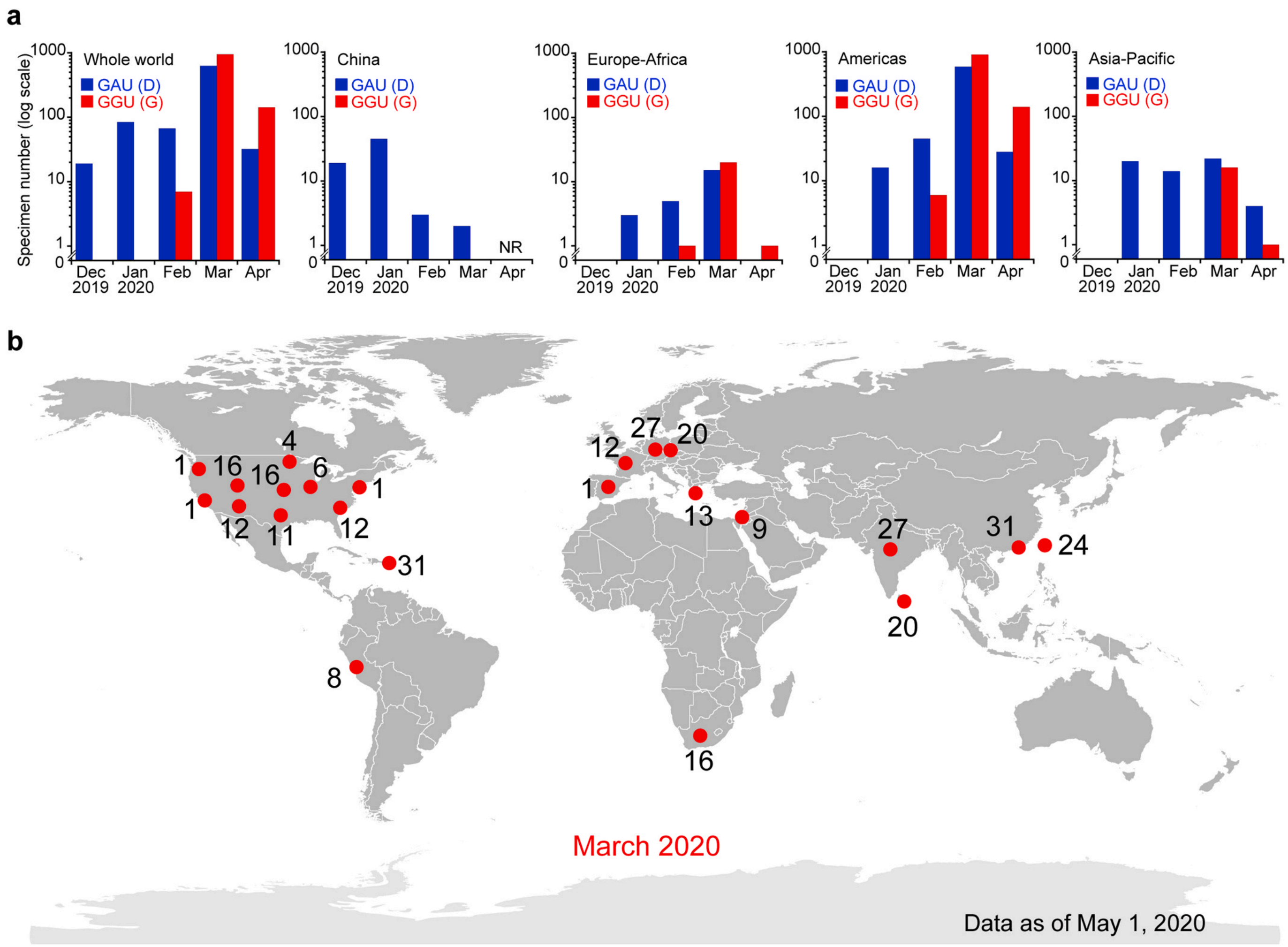


Fig. 4. GAU-to-GGU conversion was detected worldwide.
(**a**) Monthly timeline of the newly added specimens carrying GAU and GGU nucleotide mutations in the world and the continents. The data sampled in China are separately shown, where a state of emergency has been lifted. The data in April were not registered (NR) in China. GGU conversion was not registered in China. The specimen numbers are shown on a logarithmic scale. No other codons for D614G were found. (**b**) Geolocations and dates of the newly collected specimens in March 2020. Red circles indicate regions as follows: Spain, USA, Peru, Israel, France, Greece, South Africa, Sri Lanka, Czech Republic, Taiwan, India, Germany, Hong Kong and Puerto Rico. In the USA, only the representative states were listed: Washington, Connecticut, California, Minnesota, Illinois, Texas, Virginia, Arizona, Utah and Indiana. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

substantial delay because of the collection dates (for example, at around the end of February, Fig. 5a and 5b). Despite such a discrepancy, the trends in spectral analysis were mainly unchanged; the periodicity was slightly sharpened (Fig. 5c).

Collectively, the annotations in the virus genome database are of fundamental use to hypothesize the pathogenicity and to trace the transmission route at the early phase of emergence of the new substrains. These results have elucidated the need for additional annotations on patients (Fig. 6a and 6b), which should reinforce the utility of virus genomic annotations by characterizing the symptomatic features (Fig. 6c).

4. Discussion

4.1. Annotations need further implementation

The current NCBI database can elucidate the weekly or monthly trends in the propagation of emerging virus variants. Estimations of the transmission trajectories and close contacts by multiple data comparisons can refine the current genomic and geological features of SARS-CoV-2 [25,30,54]. However, the exact origins and the pathogenicity of the virus variants need to be more refined [55,56]. The virus is closely linked to human behaviours and health conditions. If viral information is tagged with additional annotatable data on the patients, we can make the best of the limited number of specimens. In particular, travel history and medical records are critically useful. Such human-associated information should be tagged to the virus information.

4.2. Information on the origins of emerging viruses

To stop further economic loss on a global scale, the restrictions of international personal travelling will be mitigated in the future. If any outbreaks of other pathogenic substrains that require different medical treatments [57–59] occur during this ongoing pandemic, the restart of global traffic may result in the sequential attacks of variant viruses on human society. In regard to the urgent clinical necessity, it is also important to locate the origins of the emergence of new strains of fatal viruses to prepare medical facilities for the emergency [29]. The annotation tags for patients' mobility history linked to virus information should be useful to retrace the detailed transmission paths of virus variants using similar filtering functions on the Excel format. There should be a deposit delay after collection under conditions of social

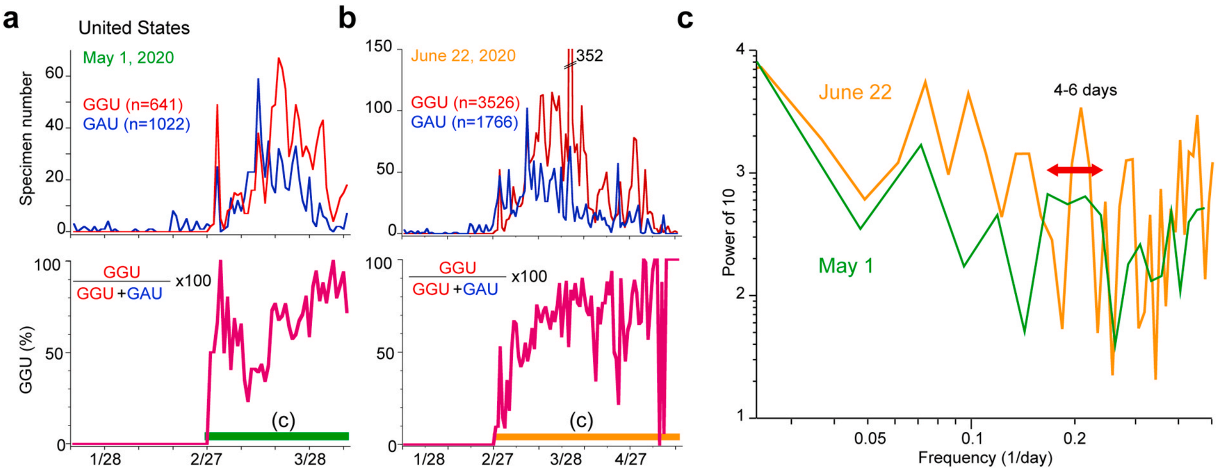


Fig. 5. The transmission in the United States was retraceable.
(**a**) The original SARS-CoV-2 (GAU) and the converted (GGU) specimens collected in the United States and deposited as of 1 May 2020 (upper), and the proportion of GGU in the daily deposits (GAU + GGU). (**b**) The GAU and GGU specimens collected as of 22 June 2020 (upper) and their proportions. (**c**) Spectral analysis of the ratio trend from 28 February through 8 April in (**a**) and from 28 February through 21 May in (**b**). Periodicities (red arrow) of infection expanding within the United States were elucidated. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

turmoil even with the aid of next-generation sequencing [15,60,61]. Thus, occasional updating on the same datasheet is important. Even though voluntary service is not always unlimited all over the world, international cooperation for fixed point surveys is necessary to reinforce the global monitoring and retracing of the transmission paths of emerging pathogens along with human mobility.

4.3. *Information on the pathogenicity of emerging viruses*

Structural modelling would be helpful for hypothesising pathogenicity. However, a portion of the conformational data downstream of D614 is also not in the PDB database. Such structural information which will add to further insights into the molecular mechanisms should be accompanied by symptomatic features [29,56] to ultimately understand pathogenicity in humans on the basis of experimental studies [62].

4.4. *Clinical review system at present*

The outbreak has been rather small in Japan [63], and the mechanisms remain unknown due to the lack of available information on the disease, namely, the patient symptoms.

Currently, individual case reports, case series, regional analyses, and meta-analyses are conducted under ethical regulations and structured protocols [17,63–65]. These close observations by trained clinical staff characterized the unique symptoms in COVID-19, including olfaction and gustatory impairments [66]. However, the meta-analysis on the symptoms will appear later [17,63]. Moreover, the majority of young patients with COVID-19 are suspected to be asymptomatic [67]. If the emergence of the new variant is traced and retraced in a real-time public platform with visualization [30,68] utilizing such medical data with human mobility history data [69], governments and other authorities can take swift and flexible actions to contain the virus.

4.5. *Medical record annotations are needed*

At present, little is known about the specific symptoms of COVID-19 [70]. An analysis of the trends of a pandemic is reinforced by open source, public databases with medical annotations. Medical records on past paths of human mobility should be used to refine the total profile of virus-human relationships with acceptable anonymity [70]. Since partial data are available at the initial time of deposit, the information may need continual updates.

SARS-CoV-2 and other emerging infectious diseases [71,72] are associated with human socioeconomic activities together with environmental and ecological factors. Compartmentalization of the world into monitorable regions based on human mobile trends [73] and sentinel surveillance including pathogen sampling with patient medical records is necessary. Local governments around the world should share real-time information on the changing nature of viruses and could conduct regional prevention measures, including caution procedures, travel restrictions and lockdowns [9].

Additionally, the susceptibility of animals to pathogens as the intermediate transmission source should be addressed [74]. Along with a risk assessment for animal-borne infections via domestic animals or animals in zoos, a simple tag for infected animals must be used to estimate the potential risks of zoonosis [75,76].

4.6. *Use of other medical bioinformatics*

Conventionally, the use of databases to predict other diseases has been developed as a disease mining method. The literature can be searched using MeSH terms [77], and the extraction of available data from the literature, including case reports using natural languages [78], can be conducted as a meta-analysis or evidence-based medicine (EBM). However, a lack of verbalized resources can be a barrier to EBM [79]. Diagnosis by digital data is especially powerful in evaluating electrocardiogram, gene-phenotype association, and pathological data [80,81] or radiographic images [78]. Composite phenotypes can also be assessed through multivariate correlations [82,83]. Automated clustering using digital annotations should decrease the substantial risk of overlooking a relevant prior study or finding. Artificial intelligence (AI) can further optimize the diagnostic accuracy [84]. However, AI may confront other risks in overlooking minor trends in rare cases by overfitting errors [84]. Virus and medical annotation tags in a simple and unified spreadsheet format are preferable for further analyses in the future. The empirical insights of medical staff are surely needed for detailed annotations, which is important for the emergence of unique pathogens.

5. Conclusions

The current databases are already powerful and useful and can evolve based on the needs of the implementation of sociomedical science. We propose the use of additional annotation tags for patients that are anonymized with maximum privacy protection and informed

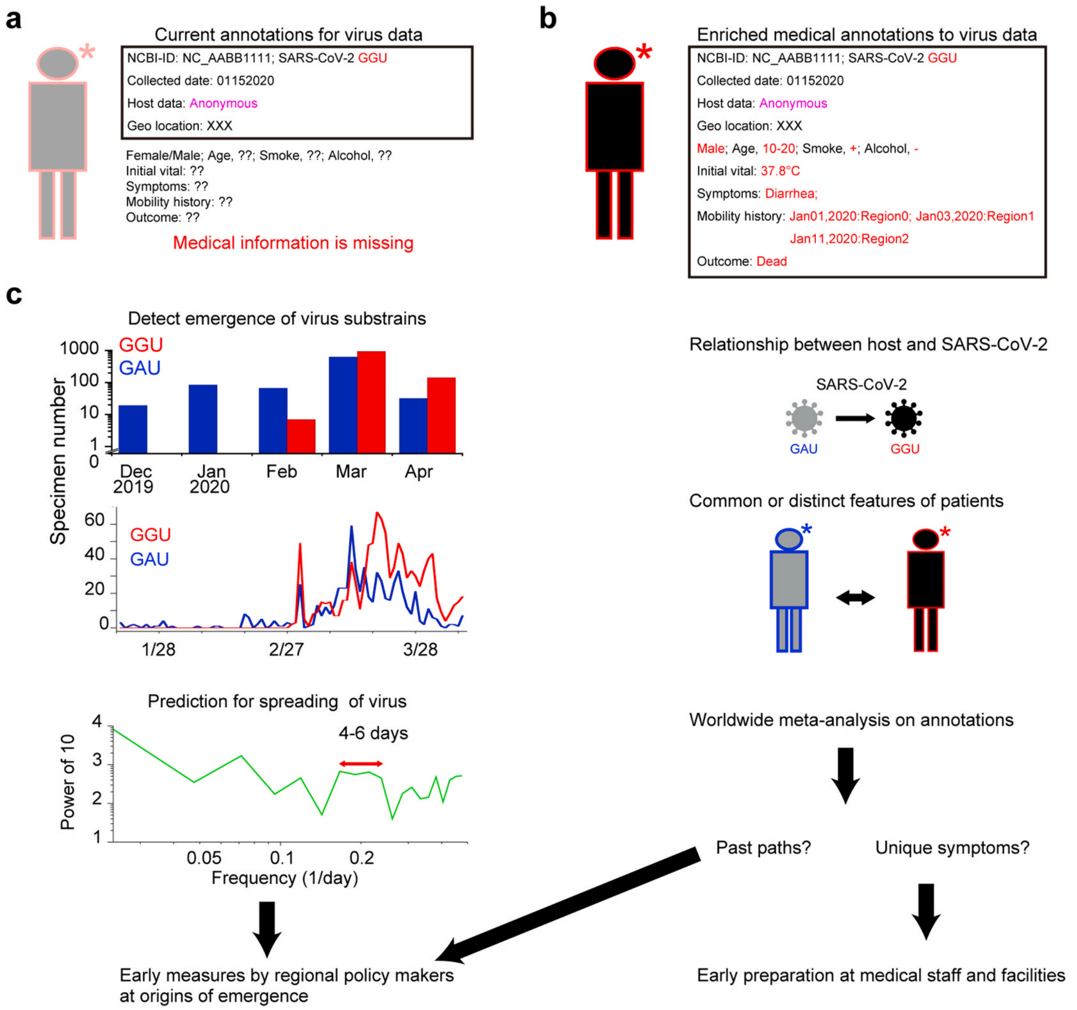

a
Current annotations for virus data
NCBI-ID: NC_AABB1111; SARS-CoV-2 GGU
Collected date: 01152020
Host data: Anonymous
Geo location: XXX
Female/Male; Age, ??; Smoke, ??; Alcohol, ??
Initial vital: ??
Symptoms: ??
Mobility history: ??
Outcome: ??
Medical information is missing
b
Enriched medical annotations to virus data
NCBI-ID: NC_AABB1111; SARS-CoV-2 GGU
Collected date: 01152020
Host data: Anonymous
Geo location: XXX
Male; Age, 10-20; Smoke, +; Alcohol, -
Initial vital: 37.8°C
Symptoms: Diarrhea;
Mobility history: Jan01,2020:Region0; Jan03,2020:Region1
Jan11,2020:Region2
Outcome: Dead
c
Detect emergence of virus substrains
Specimen number
1000
100
10
1
0
GGU
GAU
Dec 2019
Jan 2020
Feb
Mar
Apr
60
40
20
0
GGU
GAU
1/28
2/27
3/28
Prediction for spreading of virus
Power of 10
4
3
2
1
4-6 days
0.05
0.1
0.2
Frequency (1/day)
Early measures by regional policy makers at origins of emergence
Relationship between host and SARS-CoV-2
SARS-CoV-2
GAU
GGU
Common or distinct features of patients
Worldwide meta-analysis on annotations
Past paths?
Unique symptoms?
Early preparation at medical staff and facilities

Fig. 6. Annotation tags for medical and mobility history enable worldwide real-time meta-analysis.
(**a**) Current annotation list associated with virus data lacks medical information on the host patients. (**b**) Example format of digitized annotations based on medical interviews, laboratory data and medical treatment with privacy protection. (**c**)Analyses from the virus annotation data (left) reveal the emergence of the new substrains, predicted spreading span and cycles of the viruses. Additional information on medical records enables refinement of the relationship between the host and SARS-Cov-2 in general (right). The origins for the cases with similar clinical features and the same viral information can be retraced back with past-path annotation tags. Furthermore, the annotations with follow-ups and outcomes will update the profiles of COVID-19 with substrains including severity, morbidity or unique symptomatic trends.

consent on sampling virus genetic data around the world without borders. Additionally, a cooperative system of international databases [32, 33] in a single platform might also be helpful during this global emergency. Urgent international discussion is needed.

Author contributions

AN and NN analysed the downloaded data. AN and NN discussed the results and wrote the manuscript with the other authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financially supported by the Ishibashi Foundation for the Promotion of Science and the Kaibara Morikazu Medical Science Promotion Foundation to NN, and by KAKENHI JP20K16125 to AN. AN thanks Dr. Satoshi Matsuoka for his advice in programming (Department of Physiology, University of Fukui School of Medical Sciences), Dr. Kenji Shigemi (Department of Anaesthesiology, University of Fukui School of Medical Sciences), Dr Kiyonao Sada (Department of Genome Science and Microbiology, University of Fukui School of Medical Sciences), Dr Yoshifumi Yokota (Department of, Biochemistry, University of Fukui School of Medical Sciences), Dr. Harunori Ohmori (Department of Physiology, Kyoto University), Dr. Hisato Jingami (The Graduate Courses for Integrated Research Training, Kyoto University) and Dr. Takayuki Tokimasa (Department of Physiology, Kurume University School of Medicine) for their advice in conducting the research and discussion. For their assistance with documentary filing and manuscript proofreading, we thank Hideko Yoshitake and Akemi Sakamoto at Kurume University School of Medicine.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2020.100442.

References

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4. https://doi.org/10.1016/S1473-3099(20)30120-1.

[2] Tsang TK, Wu P, Lin Y, Lau EHY, Leung GM, Cowling BJ. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. Lancet Public Heal 2020;5:e289–96. https://doi.org/10.1016/S2468-2667(20)30089-X.

[3] Eurosurveillance Editorial Team. Updated rapid risk assessment from ECDC on coronavirus disease 2019 (COVID-19) pandemic: increased transmission in the EU/EEA and the UK. Euro Surveill 2020;25:2003121. https://doi.org/10.2807/1560-7917.ES.2020.25.12.2003261.

[4] Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med 2020;26:506–10. https://doi.org/10.1038/s41591-020-0822-7.

[5] Oxley TJ, Mocco J, Majidi S, Kellner CP, Shoirah H, Singh IP, et al. Large-vessel stroke as a presenting feature of covid-19 in the young. N Engl J Med 2020;382:e60. https://doi.org/10.1056/NEJMc2009787.

[6] Esper F, Shapiro ED, Weibel C, Ferguson D, Landry ML, Kahn JS. Association between a novel human coronavirus and kawasaki disease. J Infect Dis 2005;191:499–502. https://doi.org/10.1086/428291.

[7] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci Unit States Am 2020;117:9241–3. https://doi.org/10.1073/pnas.2004999117.

[8] He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med 2020;26:672–5. https://doi.org/10.1038/s41591-020-0869-5.

[9] Yen M-Y, Schwartz J, Chen S-Y, King C-C, Yang G-Y, Hsueh P-R. Interrupting COVID-19 transmission by implementing enhanced traffic control bundling: implications for global prevention and control efforts. J Microbiol Immunol Infect 2020;53:377–80. https://doi.org/10.1016/j.jmii.2020.03.011.

[10] Hirano T, Murakami M. COVID-19: a new virus, but a familiar receptor and cytokine release syndrome. Immunity 2020;52:731–3. https://doi.org/10.1016/j.immuni.2020.04.003.

[11] Bar-Zeev N, Moss WJ. Encouraging results from phase 1/2 COVID-19 vaccine trials. Lancet 2020;396:448–9. https://doi.org/10.1016/S0140-6736(20)31611-1.

[12] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182:812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

[13] Ishige T, Murata S, Taniguchi T, Miyabe A, Kitamura K, Kawasaki K, et al. Highly sensitive detection of SARS-CoV-2 RNA by multiplex rRT-PCR for molecular diagnosis of COVID-19 by clinical laboratories. Clin Chim Acta 2020;507:139–42. https://doi.org/10.1016/j.cca.2020.04.023.

[14] Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses 2020;12:254. https://doi.org/10.3390/v12030254.

[15] López L, Rodó X. The end of social confinement and COVID-19 re-emergence risk. Nat. Hum. Behav. 2020;4:746–55. https://doi.org/10.1038/s41562-020-0908-8.

[16] Nishiura H, Oshitani H, Kobayashi T, Saito T, Sunagawa T, Matsui T, et al. Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). MedRxiv 2020:2020. https://doi.org/10.1101/2020.02.28.20029272. 02.28.20029272.

[17] Tabata S, Imai K, Kawano S, Ikeda M, Kodama T, Miyoshi K, et al. Clinical characteristics of COVID-19 in 104 people with SARS-CoV-2 infection on the *Diamond Princess* cruise ship: a retrospective analysis. Lancet Infect Dis 2020;20:1043–50. https://doi.org/10.1016/S1473-3099(20)30482-5.

[18] COVID-19 dashboard by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. accessed 24 July 2020.

[19] Fischhoff B. Making decisions in a COVID-19 world. JAMA 2020;324:139–40. https://doi.org/10.1001/jama.2020.10178.

[20] Sethuraman N, Jeremiah SS, Ryo A. Interpreting diagnostic tests for SARS-CoV-2. JAMA 2020;323:2249–51. https://doi.org/10.1001/jama.2020.8259.

[21] Sidhu G, Schuster L, Liu L, Tamashiro R, Li E, Langaee T, et al. A single variant sequencing method for sensitive and quantitative detection of HIV-1 minority variants. Sci Rep 2020;10:8185. https://doi.org/10.1038/s41598-020-65085-y.

[22] Li Z, Yi Y, Luo X, Xiong N, Liu Y, Li S, et al. Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. J Med Virol 2020;92:1518–24. https://doi.org/10.1002/jmv.25727.

[23] Premkumar L, Segovia-Chumbez B, Jadi R, Martinez DR, Raut R, Markmann A, et al. The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. Sci Immunol 2020;5:eabc8413. https://doi.org/10.1126/sciimmunol.abc8413.

[24] Bay J, Kek J, Tan A, Hau CS, Yongquan L, Tan J, et al. BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders. Singapore: Government Technology Agency; 2020. https://bluetrace.io/.

[25] Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020;368:eabb6936. https://doi.org/10.1126/science.abb6936.

[26] Mobile Fact Sheet, Pew Research Center. n.d. https://www.pewresearch.org/internet/fact-sheet/mobile/ (accessed 24 July 2020)

[27] Islam MS, Ferdous MZ, Potenza MN. Panic and generalized anxiety during the COVID-19 pandemic among Bangladeshi people: an online pilot survey early in the outbreak. J Affect Disord 2020;276:30–7. https://doi.org/10.1016/j.jad.2020.06.049.

[28] Shi L, Lu Z-A, Que J-Y, Huang X-L, Liu L, Ran M-S, et al. Prevalence of and risk factors associated with mental health symptoms among the general population in China during the coronavirus disease 2019 pandemic. JAMA Netw Open 2020;3:e2014053. https://doi.org/10.1001/jamanetworkopen.2020.14053.

[29] Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 2020;74:e13525. https://doi.org/10.1111/ijcp.13525.

[30] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407.

[31] Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, et al. Global spread of SARS-CoV-2 subtype with spike protein mutation D614G is shaped by human genomic variations that regulate expression of *TMPRSS2* and *MX1* genes. BioRxiv 2020;2020.05.04.075911, https://doi.org/10.1101/2020.05.04.075911; 2020. 2020.

[32] Severe acute respiratory syndrome coronavirus 2 data hub. n.d. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Protein&VirusLineage_ss=Severe acute respiratory syndrome coronavirus 2. taxid:2697049.

[33] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Challenges. 2017;1:33–46. https://doi.org/10.1002/gch2.1018.

[34] Robson B. COVID-19 Coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance. Comput Biol Med 2020;121:103749. https://doi.org/10.1016/j.compbiomed.2020.103749.

[35] Wensing M, Sales A, Armstrong R, Wilson P. Implementation science in times of Covid-19. Implement Sci 2020;15:42. https://doi.org/10.1186/s13012-020-01006-x.

[36] Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource-improved response to emergent viral outbreaks. Nucleic Acids Res 2017;45:D482–90. https://doi.org/10.1093/nar/gkw1065.

[37] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999;27:29–34. https://doi.org/10.1093/nar/27.1.29.

[38] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–3. https://doi.org/10.1126/science.abb2507.

[39] Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res 2017;27:119–29. https://doi.org/10.1038/cr.2016.152.

[40] Yuan Y, Cao D, Zhang Y, Ma J, Qi J, Wang Q, et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nat Commun 2017;8:15092. https://doi.org/10.1038/ncomms15092.

[41] Tortorici MA, Walls AC, Lang Y, Wang C, Li Z, Koerhuis D, et al. Structural basis for human coronavirus attachment to sialic acid receptors. Nat Struct Mol Biol 2019;26:481–9. https://doi.org/10.1038/s41594-019-0233-y.

[42] Li Z, Tomlinson AC, Wong AH, Zhou D, Desforges M, Talbot PJ, et al. The human coronavirus HCoV-229E S-protein structure and receptor binding. eLife 2019;8:e51230. https://doi.org/10.7554/eLife.51230.

[43] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.

[44] Wilkins MR, Gasteiger E, Bairoch A, Sanchez J, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. In: Link AJ, editor. 2-D proteome analysis Protocols. New Jersey: Humana Press; 1999. p. 531–52. https://doi.org/10.1385/1-59259-584-7:531.

[45] Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol 2011;8:270–9. https://doi.org/10.4161/rna.8.2.15013.

[46] Agris PF, Eruysal ER, Narendran A, Väre VYP, Vangaveti S, Ranganathan SV. Celebrating wobble decoding: half a century and still much is new. RNA Biol 2018;15:537–53. https://doi.org/10.1080/15476286.2017.1356562.

[47] Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 2020;11:1800. https://doi.org/10.3389/fmicb.2020.01800.

[48] Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ 2020;98:495–504. https://doi.org/10.2471/BLT.20.253591.

[49] Kuba K, Imai Y, Rao S, Gao H, Guo F, Guan B, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. Nat Med 2005;11:875–9. https://doi.org/10.1038/nm1267.

[50] Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3. https://doi.org/10.1038/s41586-020-2012-7.

[51] Abdul-Rasool S, Fielding BC. Understanding human coronavirus HCoV-NL63. Open Virol J 2010;4:76–84. https://doi.org/10.2174/1874357901004010076.

[52] Yang T, Chang Y, Ko T, Draczkowski P, Chien Y, Chang Y-C, et al. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans. Proc Natl Acad Sci Unit States Am 2020;117:1438–46. https://doi.org/10.1073/pnas.1908898117.

[53] Jemimah S, Gromiha MM. Insights into changes in binding affinity caused by disease mutations in protein-protein complexes. Comput Biol Med 2020;123: 103829. https://doi.org/10.1016/j.compbiomed.2020.103829.

[54] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. BioRxiv 2020. https://doi.org/10.1101/2020.04.29.069054. 2020.04.29.069054.

[55] Caroline B. Improving epidemic surveillance and response : big data is dead , long live big data. Lancet Digit Heal 2020;2:e218–20. https://doi.org/10.1016/S2589-7500(20)30059-5.

[56] Thorlund K, Dron L, Park J, Hsu G, Forrest JI, Mills EJ. A real-time dashboard of clinical trials for COVID-19. Lancet Digit Heal 2020;2:e286–7. https://doi.org/10.1016/S2589-7500(20)30086-8.

[57] Xu J, Jia W, Wang P, Zhang S, Shi X, Wang X, et al. Antibodies and vaccines against Middle East respiratory syndrome coronavirus. Microb Infect 2019;8:841–56. https://doi.org/10.1080/22221751.2019.1624482.

[58] Wang C, Li W, Drabek D, Okba NMA, van Haperen R, Osterhaus ADME, et al. A human monoclonal antibody blocking SARS-CoV-2 infection. Nat Commun 2020; 11:2251. https://doi.org/10.1038/s41467-020-16256-y.

[59] Dong L, Hu S, Gao J. Discovering drugs to treat coronavirus disease 2019 (COVID-19). Drug Discov Ther 2020;14:58–60. https://doi.org/10.5582/ddt.2020.01012.

[60] Drew DA, Nguyen LH, Steves CJ, Menni C, Freydin M, Varsavsky T, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science 2020;368:1362–7. https://doi.org/10.1126/science.abc0473.

[61] Kwok KO, Tang A, Wei VWI, Park WH, Yeoh EK, Riley S. Epidemic models of contact tracing: systematic review of transmission studies of severe acute respiratory syndrome and Middle East respiratory syndrome. Comput Struct Biotechnol J 2019;17:186–94. https://doi.org/10.1016/j.csbj.2019.01.003.

[62] Chan JF, Zhang AJ, Yuan S, Poon VK, Chan CC, Lee AC. Simulation of the clinical and pathological manifestations of Coronavirus Disease 2019 (COVID-19) in golden Syrian hamster model: implications for disease pathogenesis and transmissibility. Clin Infect Dis 2020. https://doi.org/10.1093/cid/ciaa325. In press.

[63] Sekizuka T, Kuramoto S, Nariai E, Taira M, Hachisu Y, Tokaji A, et al. SARS-CoV-2 genome analysis of Japanese travelers in nile river cruise. Front Microbiol 2020;11: 1316. https://doi.org/10.3389/fmicb.2020.01316.

[64] Pranata R, Huang I, Lim MA, Wahjoepramono EJ, July J. Impact of cerebrovascular and cardiovascular diseases on mortality and severity of COVID-19–systematic review, meta-analysis, and meta-regression. J Stroke Cerebrovasc Dis 2020;29: 104949. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.104949.

[65] Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis 2020;94:91–5. https://doi.org/10.1016/j.ijid.2020.03.017.

[66] Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodriguez A, et al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. Eur Arch Oto-Rhino-Laryngol 2020;277:2251–61. https://doi.org/10.1007/s00405-020-05965-1.

[67] Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection. Ann Intern Med 2020;173:362–7. https://doi.org/10.7326/m20-3012.

[68] Kolajo T, Daramola O, Adebiyi A. Big data stream analysis: a systematic literature review. J Big Data 2019;6:47. https://doi.org/10.1186/s40537-019-0210-7.

[69] Ruktanonchai NW, Ruktanonchai CW, Floyd JR, Tatem AJ. Using Google Location History data to quantify fine-scale human mobility. Int J Health Geogr 2018;17:28. https://doi.org/10.1186/s12942-018-0150-z.

[70] Callahan A, Steinberg E, Fries JA, Gombar S, Patel B, Corbin CK, et al. Estimating the efficacy of symptom-based screening for COVID-19. NPJ Digit Med 2020;3:95. https://doi.org/10.1038/s41746-020-0300-0.

[71] Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature 2008;451:990–3. https://doi.org/10.1038/nature06536.

[72] NIH. NIAID Emerging Infectious Diseases/Pathogens. n.d. https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens. accessed 24 July 2020

[73] Huang Z, Ling X, Wang P, Zhang F, Mao Y, Lin T, et al. Modeling real-time human mobility based on mobile phone and transportation data fusion. Transport Res C Emerg Technol 2018;96:251–69. https://doi.org/10.1016/j.trc.2018.09.016.

[74] Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. Science 2020; 368:1016–20. https://doi.org/10.1126/science.abb7015.

[75] Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. Trends Mol Med 2020;26:483–95. https://doi.org/10.1016/j.molmed.2020.02.008.

[76] Confirmed cases of SARS-CoV-2 in Animals in the United States, United States Dep. Agric. Anim. Plant Heal. Insp. Serv. (n.d.). https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/sa_one_health/sars-cov-2-animals-us (accessed 24 July 2020).

[77] NIH, Medical Subject Headings, (n.d.). https://www.nlm.nih.gov/mesh/meshhome.html (accessed 24 July 2020).

[78] Goff DJ, Loehfelm TW. Automated radiology report summarization using an open-source natural language processing pipeline. J Digit Imag 2018;31:185–92. https://doi.org/10.1007/s10278-017-0030-2.

[79] Sadeghi-Bazargani H, Tabrizi JS, Azami-Aghdash S. Barriers to evidence-based medicine: a systematic review. J Eval Clin Pract 2014;20:793–802. https://doi.org/10.1111/jep.12222.

[80] Jia J, An Z, Ming Y, Guo Y, Li W, Li X, et al. PedAM: a database for pediatric disease annotation and medicine. Nucleic Acids Res 2018;46:D977–83. https://doi.org/10.1093/nar/gkx1049.

[81] Il Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA 2007;104:8685–90. https://doi.org/10.1073/pnas.0701361104.

[82] Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms-disease network. Nat Commun 2014;5:4212. https://doi.org/10.1038/ncomms5212.

[83] Conesa A, Bro R, García-García F, Prats JM, Götz S, Kjeldahl K, et al. Direct functional assessment of the composite phenotype through multivariate projection strategies. Genomics 2008;92:373–83. https://doi.org/10.1016/j.ygeno.2008.05.015.

[84] Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit Med 2020;3:30. https://doi.org/10.1038/s41746-020-0229-3.