Characterizing geographical and temporal dynamics of novel coronavirus SARS-CoV-2 using informative subtype markers

Zhengqiao Zhao[1*], Bahrad A. Sokhansanj [2†], Gail L. Rosen[1+],

1 Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Department of Electrical and Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA, USA
2 Independent Researcher, Los Angeles, CA, USA

*** zz374@drexel.edu**
† bahrad@molhealtheng.com
+ glr26@drexel.edu

Abstract

Genetic subtyping of viruses and bacteria is a critical tool for visualizing and modeling their geographic distribution and temporal dynamics. Quantifying viral dynamics is of particular importance for the novel coronavirus responsible for COVID-19, SARS-CoV-2. Effective containment strategies and potential future therapeutic and vaccine strategies will likely require a precise and quantitative understanding of viral transmission and evolution. In this paper, we employ an entropy-based analysis to identify mutational signatures of SARS-CoV-2 strains in the GISAID database available as of April 5, 2020. Our analysis method identifies nucleotide sites within the viral genome which are highly informative of variation between the viral genomes sequenced in different individuals. These sites are used to characterize individual virus sequence with a characteristic Informative Subtype Marker (ISM). The ISMs provide signatures that can be efficiently and rapidly utilized to quantitatively trace viral dynamics through geography and time. We show that by analyzing the ISM of currently available SARS-CoV-2 sequences, we are able to profile international and interregional differences in viral subtype, and visualize the emergence of viral subtypes in different countries over time. To validate and demonstrate the utility of ISM-based subtyping: (1) We show the distinct genetic subtypes of European infections, in which early on infections are related to the viral subtypes that has become dominant in Italy followed by the development of local subtypes, (2) We distinguish subtypes associated with outbreaks in distinct parts of the United States, identify the development of a local subtype potentially due to community to transmission and distinguish it from the predominant subtype in New York, suggesting that the outbreak in New York is linked to imported cases from Europe. (3) We present results that quantitatively show the temporal behavior of the emergence of SARS-CoV-2 from localization in China to a pattern of distinct regional subtypes as the virus spreads throughout the world over time. Accordingly, we show that genetic subtyping using entropy-based ISMs can play an important complementary role to phylogenetic tree-based analysis, such as the Nextstrain [9] project, in efficiently quantifying SARS-CoV-2 dynamics to enable modeling, data-mining, and machine learning tools. Following from this initial study, we have developed a pipeline to dynamically generate ISMs for newly added SARS-CoV-2 sequences and generate updated visualization of geographical and temporal dynamics, and made it available on Github at https://github.com/EESI/ISM.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the novel coronavirus responsible for the Covid-19 pandemic, was first reported in Wuhuan, China in late December 2019. [13, 24]. In a matter of

weeks, SARS-CoV-2 infections have been detected in nearly every country. Powered by advances in rapid genetic sequencing, there is an expansive and growing body of data on SARS-CoV-2 sequences from individuals around the world. There are now central repositories accumulating international SARS-CoV-2 genome data, such as the Global Initiative on Sharing all Individual Data (GISAID) [22] (available at https://www.gisaid.org/). Because the viral genome mutates over time as the virus infects and then spreads through different populations, viral sequences have diverged as the virus infects more people in different locations around the world.

Researchers have sought to use traditional approaches, based on sequence alignment and phylogenetic tree construction, to study evolution of SARS-CoV-2 on a macro and micro scale. At a high level, for example, the Nextstrain group has created a massive phylogenetic tree incorporating sequence data, and applied a model of the time-based rate of mutation to create a hypothetical map of viral distribution [9] (available at https://nextstrain.org/ncov). Similarly, the China National Center for Bioinformation has established a "2019 Novel Coronavirus Resource", which includes a clickable world map that links to a listing of sequences along with similarity scores based on alignment (available at https://bigd.big.ac.cn/ncov?lang=en) [36].

In more granular studies, early work by researchers based in China analyzing 103 genome sequences, identified two highly linked single nucleotides, and suggested the development of two major strain sub-types, and "L" subtype that was predominantly found in the Wuhan area, and an "S" subtype that was derived from "S" and found elsewhere [25]. Subsequently, further diversity was recognized as the virus continued to spread, and researchers developed a consensus reference sequence for SARS-CoV-2, to which other sequences may be compared [30]. Some researchers are looking at international expansion, but the timeline and variant composition has been limited [31]. Studies have also been undertaken of sequences from passengers on the *Diamond Princess* cruise ship, including for US passengers by the CDC as well as by Japanese researchers [20].

Researchers are also seeking to analyze sequence variants to identify potential regions where selection pressure may result in phenotypic variation, such as in the ORF (open reading frame) coding for the spike (S) receptor-binding protein which may impact the development of vaccines and antivirals. Notably, a group studying sequence variants within patients reported limited evidence of intra-host variation, though they cautioned that the results were preliminary and could be the result of limited data [10, 21]. That study suggests an additional layer of complexity in evaluating viral variation that may have an influence on disease progression in an individual patient, or be associated with events that can generate sequence variation in other individuals that patient infects.

Given the broad importance in tracking and modeling genetic changes in the SARS-CoV-2 virus as the outbreak expands, however, there is a need for an efficient methodology to *quantitatively* characterize the virus genome. It has been proposed that phylogenetic trees obtained through sequence alignment may be utilized to map viral outbreaks geographically and trace transmission chains [8, 18]. These approaches are being demonstrated for SARS-CoV-2 by, e.g., the Nextstrain group as discussed above. However, phylogenetic trees are complex constructs that are not readily quantifiable. As exemplary implementations of using phylogenetic trees in epidemiology demonstrate, requiring additional complex processing such as through the use of clustering that are cumbersome and introduce potential error and bias [5, 33]. To generate highly informative signatures, we look to methods that have been successfully employed in the microbiome field for 16S ribosomal DNA (16S rDNA). 16S rDNA is a highly conserved sequence and therefore can be used for phylogenetic analysis in microbial communites [4, 7, 14, 15, 32]. To differentiate between closely related microbial taxa, Meren et al. introduced a novel method for identifying "oligotypes", which represent subgroups, using nucleotide positions that represent information-rich variation [6]. A more efficient framework, similar to oligotyping, to quantify viral subtypes can therefore help achieve important goals for understanding the progression of the COVID-19 pandemic, as well as ultimately contain and resolve the disease. Exemplary potential applications of quantitative subtyping include:

- Characterizing potentially emerging variants of the virus in different regions, which may ultimately express different phenotypes.
- Monitoring variation in the viral genome that may be important for vaccine, for example due to emerging structural differences in proteins encoded by different strains.

- Designing future testing methodology to contain disease transmission across countries and regions, for example developing specific tests that can characterize whether a COVID-19 patient developed symptoms due to importation or likely domestic community transmission.
- Identifying viral subtypes that may correlate with different clinical outcomes and treatment response in different regions (and potentially even patient subpopulations).

In this paper, we propose a method to define a signature for the viral genome that can be 1) utilized to define viral subtypes that can be quantified, and 2) efficiently implemented and visualized. In particular, to satisfy the latter need, we propose compressing the full viral genome to generate a small number of nucleotides that are highly informative of the way in which the viral genome dynamically changes. Based on such a signature, SARS-CoV-2 subtypes may thus be defined and then quantitatively characterized in terms of their geographic abundance, as well as their abundance in time — and, potentially also detect clinical variation in disease progression associated with viral subtypes. Taking inspiration from the aforementioned oligotyping approach [6], we propose and develop a pipeline to utilize entropy to identify highly informative nucleotide positions, and, in turn, identifying characteristic Informative Subtype Markers (ISM) that can be used to subtype individual SARS-CoV-2 virus genomes. We evaluate the pipeline by demonstrating the potential of ISMs to model and visualize the geographic and temporal patterns of the SARS-CoV-2 using sequences that are currently publicly available from the GISAID database. We have made the pipeline available on Github https://github.com/EESI/ISM, where it will be continuously updated as new sequences are uploaded to data repositories.

Methods

Data collection and preprocessing

SARS-CoV-2 (novel coronavirus) sequence data was downloaded from GISAID (http://www.gisaid.org) on April 5, 2020 which contains 4087 sequences. The preprocessing pipeline then begins by filtering out sequences that are less than 25000 base pairs (the same threshold used in Nextstrain project built for SARS-CoV-2[1]). We also included a reference sequence from National Center for Biotechnology Information[2] (NCBI Accession number: NC_045512.2). This resulted in an overall data set of 3981 sequences with sequence length ranging from 25342 nt to 30355 nt. We then align all remaining sequences after filtering together using MAFFT [11] using the "FFT-NS-2" method in XSEDE [28]. After alignment, the sequence length is extended (for the present data set, up to 35362 nt).

Entropy anlysis and ISM extraction

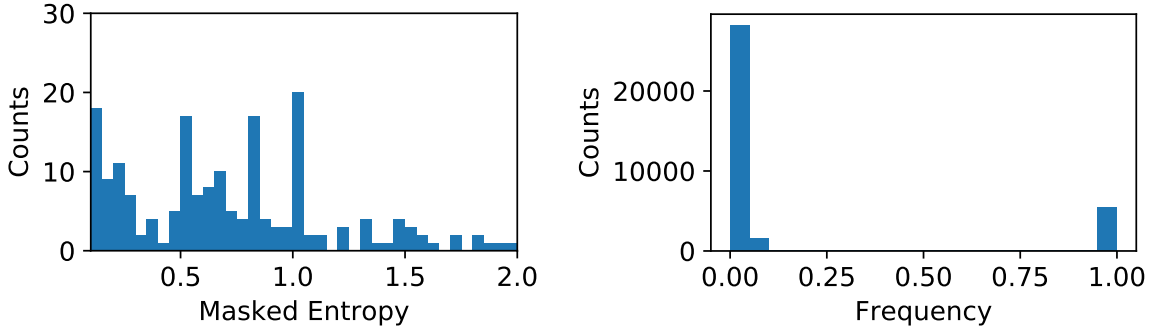

Figure 1. Left: histogram of masked entropy values; Right: histogram of percentages of `n` and `-`

[1] https://github.com/nextstrain/ncov
[2] https://www.ncbi.nlm.nih.gov/

For the aligned sequences, we merged the sequence with the metadata in Nextstrain project[3] as it is in April 6, 2020 based on identification number, **gisaid_epi_isl**, provided by GISAID [22]. We further filtered out sequences with incomplete date information in metadata (e.g, "2020-01"), so that our analysis can also incorporate temporal information with daily resolution, given the fast-moving nature of the pandemic. In addition, we filtered out sequences from unknown host or non-human hosts. The resultant final dataset contains 3832 sequences excluding the reference sequence. Then, we calculate the entropy by:

$$H = -\sum_{k \in L} p_k * log_2(p_k)$$

where L is a list of unique characters in all sequences and p_k is a probability of a character k. We estimated p_k from the frequency of characters. We refer to characters in the preceding because, in addition to the bases `a`, `c`, `g`, and `t`, the sequences include additional characters representing gaps and ambiguities: `-`, `b`, `d`, `h`, `k`, `m`, `n`, `r`, `s`, `v`, `w`, and `y`. (Note that the sequences are of cDNA derived from viral RNA, so there is a `t` substituting for the `u` that would appear in the viral RNA sequence.) However, sites `n` and `-` (representing an ambiguous site and a gap respectively) are less informative. Therefore, we further define a *masked entropy* as entropy calculated without considering sequences containing `n` and `-` in a given nucleotide position in the genome. Based on the entropy calculation, we developed a masked entropy calculation whereby we ignore the `n` and `-`. With the help of this masked entropy calculation, we can focus on truly informative positions, instead of positions at the start and end of the sequence in which there is substantial uncertainty due to artifacts in the sequencing process. Finally, high entropy positions are selected by two criteria: 1) entropy > 0.5, and 2) the percentage of `n` and `-` is less than 25%. This yielded 17 distinct positions along the viral genome sequence. We then extract Informative Subtype Markers (ISMs) at these 17 nucleotide positions from each sequence. Figure 1 shows how we identified these two criteria. The left hand side of the plot shows that there is a peak with entropy greater than 0.5, which we sought to retain. Looking to the right hand side of the plot, setting threshold to 0.25 will keep the peak on the left which represents the most informative group of sites in the genome.

Quantification and visualization of viral subtypes

At the highest level, we assess the geographic distribution of SARS-CoV-2 subtypes, and, in turn, we count the frequency of unique ISMs per location and build charts and tables to visualize the ISMs, including the pie charts, graphs, and tables shown in this paper. To improve visualization, ISMs with frequency less than 5% in a given location are collapsed into "OTHER" category per location. Our pipeline then creates pie charts for different locations to show the geographical distribution of subtypes. Each subtype is also labeled with the earliest date associated with sequences from a given location in the dataset. To study the progression of SARS-CoV-2 viral subtypes in the time domain, we group all sequences in a given location that are no later than a certain date together and compute the relative abundance of corresponding subtypes. Any subtypes with frequency less than 5% are collapsed into "OTHER" category per location. The following formula illustrates this calculation:

$$ISM_{(s,c)}(t) = \frac{N_{s,c}(t)}{N_c(t)}$$

where $ISM_{(s,c)}(t)$ is the relative abundance of a subtype, s, in location, c, at a date t, $N_{s,c}(t)$ is the total number of instances of such subtype, s, in location, c, that has been sequenced no later than date t and $N_c(t)$ is the total number of sequences in location, c, that has been sequenced no later than date t.

Results and Discussion

Identification and Mapping of subtype markers

Figure 2 shows the overall entropy at each nucleotide position, determined based on calculating the masked entropy for all sequences as described in the Methods section. Notably, at the beginning and the end of the

[3] https://github.com/nextstrain/ncov/blob/master/data/metadata.tsv

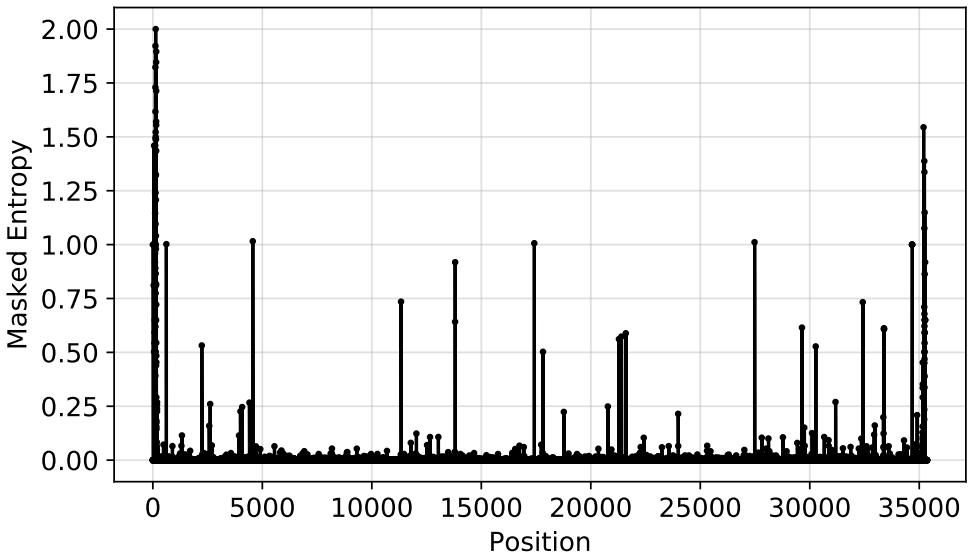

Figure 2. Overall entropy as a function of nucleotide position for all SARS-CoV-2 sequences in the data set

sequence, there is high level of uncertainty. This is because there are more `n` and `-` symbols, representing ambiguity and gaps, in these two regions (Gaps are likely a result of artifacts in MAFFT's alignment of the viruses or its genomic rearrangement [10], and both `N`'s and `-`'s may result due to the difficulty of accurately sequencing the genome at the ends). After applying filtering to remove low entropy positions and uncertain positions, we identified 17 informative nucleotide positions on the sequence to generate informative subtype markers (see filtering details in Methods section).

Importantly, even though the combinatorial space for potential ISMs is potentially very large due to the large number of characters that may present at any one nucleotide position, only certain ISMs occur in significantly large numbers in the overall sequence population. Figure 3 shows the rapid decay in the frequency of sequences with a given ISM, and shows that only the first nine ISMs represent subtypes that are significantly represented in the sequences available worldwide.

Some potential reasons for the rapid dropoff in the frequency relative to the diversity of ISMs may include the following: (1) Since the virus is transmitting and expanding so quickly, and the pandemic is still at a relatively early stage, there has not been enough time for mutations that would affect the ISM to occur and take root. In that case, we would expect the number of significant ISMs to rise over time. (2) The population of publicly available sequences is biased to projects in which multiple patients in a cluster are sequenced at once: For example, a group of travelers, a family group, or a group linked to a single spreading event. An example of this is the number of sequences from cruise vessels in the database. We expect that the impact of any such clustering will be diminished in time as more comprehensive sequencing efforts take place. (3) ISMs may be constrained by the fact that certain mutations may result in a phenotypic change that may be selected against. In this case, we may expect a steep change in a particular ISM or close relative in the event that there is selection pressure in favor of the corresponding variant phenotype. However, as described above, at the present time the high-entropy nucleotide sequences appear to be primarily in open reading frame regions that, at least in comparison to other SARS-related viruses, do not represent areas in which there would be high selection pressure (i.e., due to exposure to the human immune response or need to gain entry to host cells).

After the informative nucleotide positions were identified, we then mapped those sites back to the annotated reference sequence for functional interpretation [30]. As shown in Table 1, we found that all but one of the nucleotide positions that we identified were located in coding regions of the reference sequence. The majority of the remaining sites (9/16) were found in the *ORF1ab* polyprotein, which encodes a polyprotein replicase complex that is cleaved to form nonstructural proteins that are used as RNA

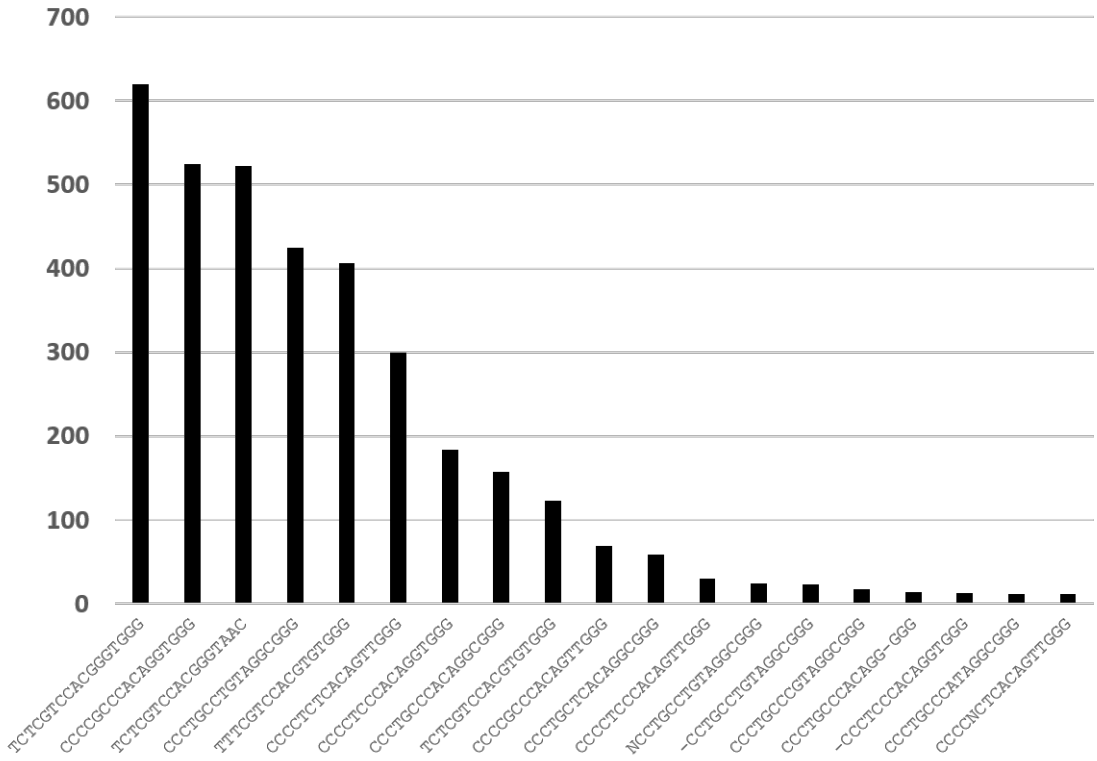

Figure 3. Number of sequences containing the 20 most abundant ISMs within the total data set.

Table 1. Mapping ISM sites to the reference viral genome

Site	Nucleotide Position	Entropy	Annotation
1	241	1.002215927	Non-coding Region
2	1059	0.531987853	ORF1ab
3	3037	1.015626172	ORF1ab
4	8782	0.735237992	ORF1ab
5	11083	0.641834959	ORF1a
6	14408	1.006329881	ORF1ab
7	14805	0.502722748	ORF1ab
8	17747	0.561489842	ORF1ab
9	17858	0.573208404	ORF1ab
10	18060	0.588462469	ORF1ab
11	23403	1.011257757	S surface glycoprotein
12	25563	0.614935552	ORF3a
13	26144	0.527628595	ORF3a
14	28144	0.732986643	ORF8
15	28881	0.612149979	nucleocapsid phosphoprotein
16	28882	0.608003271	nucleocapsid phosphoprotein
17	28883	0.608003271	nucleocapsid phosphoprotein

polymerase (i.e., synthesis) machinery [12]. One site is located in the reading frame encoding the S spike glycoprotein, which is responsible for viral entry and antigenicity, and thus represents an important target for understanding the immune response, identifying antiviral therapeutics, and vaccine design [17, 29]. High-entropy nucleotide positions were also found in the nucleocapsid formation protein, which is important for packaging the viral RNA. [34] A study has also shown that, like the spike protein, the internal nucleoprotein of the virus is significant in modulating the antibody response. [27]

Additionally, Table 1 shows a high-entropy, informative site in the predicted coding region *ORF8*. Based on structural homology analysis the *ORF8* region in SARS-CoV-2 does not have a known functional domain or motif [3]. In previously characterized human SARS coronavirus, *ORF8* has been associated with an enhanced inflammatory response, but that sequence feature does not appear to have been conserved in SARS-CoV-2, and, in general, SARS-CoV-2 *ORF8* appears divergent from other previously characterized SARS-related coronaviruses. [3, 35] Previous entropy-based analysis of earlier and smaller SARS-CoV-2 sequence data sets have suggested that there is a mutational hotspot in *ORF8*, including early divergence between sequences found in China, and large scale deletions found in patients in Singapore — which is consistent with the results we have found here on a much more comprehensive analysis of genomes [2, 23, 25]. Similarly, sites were identified in the *ORF3a* reading frame, which also appears to have diverged substantially from other SARS-related viruses. In particular, the SARS-CoV-2 variant in the predicted *ORF3* region appear also to not contain functional domains that were responsible for increased inflammatory response as they were in those viruses [3, 35] .

While the significance of *ORF8* to viral biology and clinical outcomes remains uncertain, the majority of high-entropy sites are in regions of the genome that may be significant for disease progression and the design of vaccines and therapeutics. Accordingly, ISMs derived from the corresponding nucleotide positions can be used for viral subtyping for clinical applications, such as identifying variants with different therapeutic responses or patient outcomes, or for tracking variation that may reduce the effectiveness of potential vaccine candidates.

Geographic distribution of SARS-CoV-2 subtypes

Figure 4 shows the distribution of ISMs, each indicating a different subtype, in the countries/regions with the relatively larger amount of available sequenced genomes. As shown therein, the ISMs are able to successfully identify and label viral subtypes that produce distinct patterns of distribution in different countries/regions. Beginning with Mainland China, the consensus source of SARS-CoV-2, we observe three dominant subtypes in Mainland China, as indicated by relative abundance of the ISM among available sequences: `CCCCGCCCACAGGTGGG` (as indicated on the plot, first seen in December 24, 2019 in sequences from Mainland China in the dataset), `CCCTGCCCACAGGCGGG` (first seen in January 5, 2020 in sequences from Mainland China in the dataset) and `CCCCTCCCACAGGTGGG` (first seen in January 18, 2020 in sequences from Mainland China in the dataset). These subtypes are found in other countries/regions, but in distinct patterns, which may likely correspond to different patterns of transmission of the virus. For example, sequences in Japan are dominated by `CCCCTCCCACAGGTGGG`, the third of the subtypes listed for Mainland China, and first observed in a sequence in Japan on February 10, 2020. However, this subtype is not prevalent in other countries/regions, with limited abundance in Australia, Canada, and Singapore — countries that are likely to have travel links to both Mainland China and Japan.. However, in Singapore a major subtype is `CCCTGCCCACAGG-GGG` , which is very close to subtype `CCCCTCCCACAGGTGGG` which was found in Mainland China. In sum, the subtype patterns in Asian countries/regions are related or shared, but as can be seen, Singapore and Japan appear to have drawn from distinct subtypes that were found in Mainland China, potentially leading to the hypothesis that there were multiple distinct travel-related transmissions from China to both countries by the time of this analysis.

The data further indicate that the United States has a distinct pattern of dominant subtypes. In particular the subtype with the highest relative abundance among US sequences is `CCCTGCCTGTAGGCGGG`, first seen in February 20, 2020. This subtype has also emerged as a major subtype in Canada, with the first sequence being found on March 5, 2020. A different pattern is found among sequences in Europe, in which dominant subtypes tend to be different from those found in most Asian countries/regions. Of particular note though, Japan includes a number of sequences, of a subtype that is found extensively in European countries,

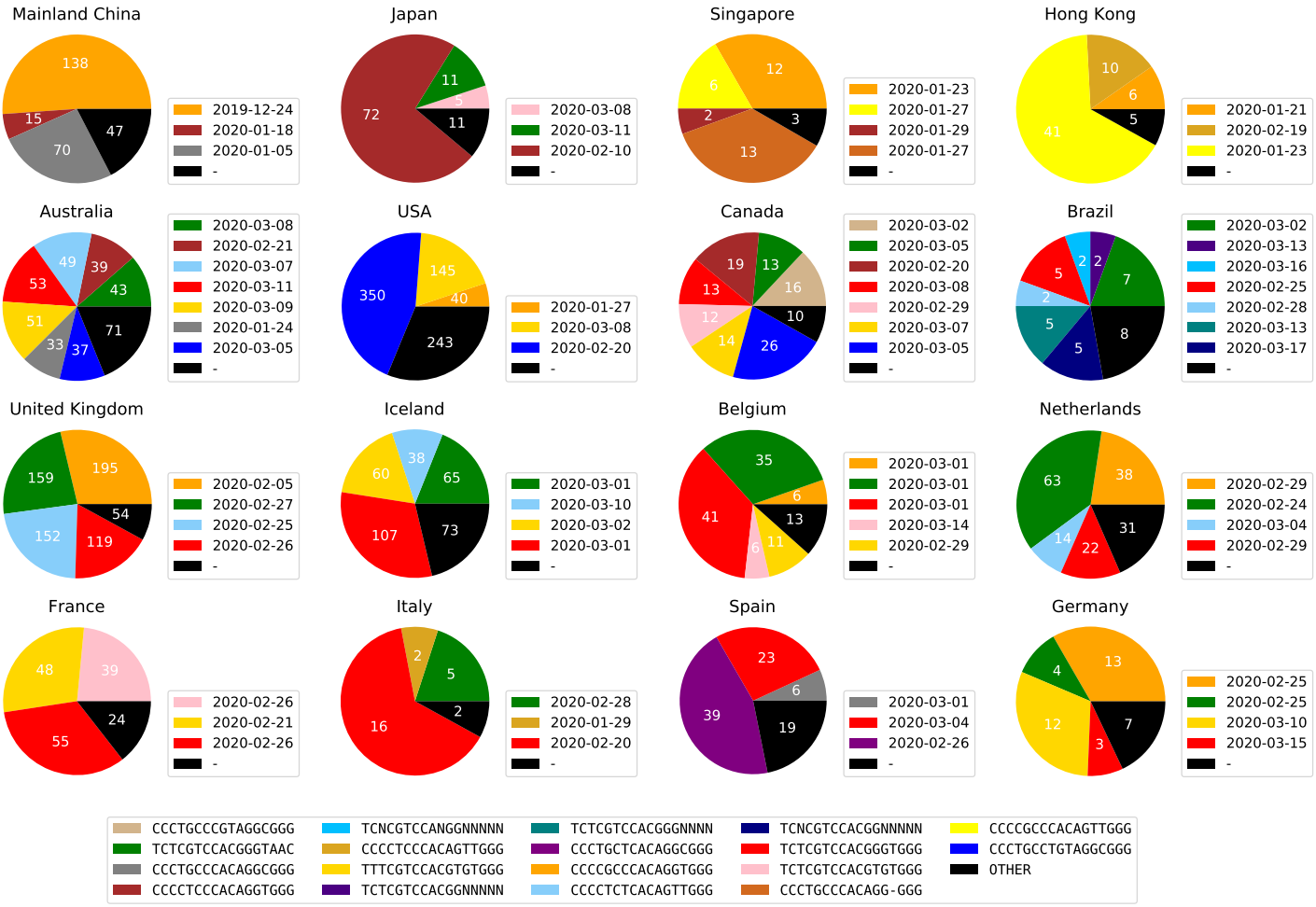

Figure 4. Major subtypes in countries/regions with the most sequences (indicating date subtype was first sequenced in that country/region). Subtypes with less than 5% abundance are plotted as "OTHER". The raw counts for all ISMs in each country, as well as the date each ISM was first found in a sequence in that Country, are provided in Supplementary file 1 — ISM abundance table of 16 countries/regions.

TCTCGTCCACGGGTAAC, first found in Japan on March 11, 2020. This subtype has also been found in Canada and Brazil, suggesting a geographical commonality between cases in these diverse countries with the progression of the virus in Europe.

Moreover, while many of the connections between shared subtypes in Figure 4 reflect the general understanding of how the virus has progressed between Asia, North America, and Europe, ISM-based subtyping suggests hypotheses of more granular linkages. For example, one of the most prevalent subtypes in sequences from France, TCTCGTCCACGTGTGGG, is also found in neighboring Belgium, but it is not a prevalent subtype in other European countries shown in Figure 4. Indeed, this subtype was found in 0.06% of sequences in the United Kingdom, and in only one sequence in Iceland, which has the largest per capita sample size in the data set (see Supplementary file 1 — ISM abundance table of 16 countries/regions). The subtype is found, however, in other countries like Canada, Australia, and Japan, suggesting a potential viral transmission due specifically to travel between France and those two countries.

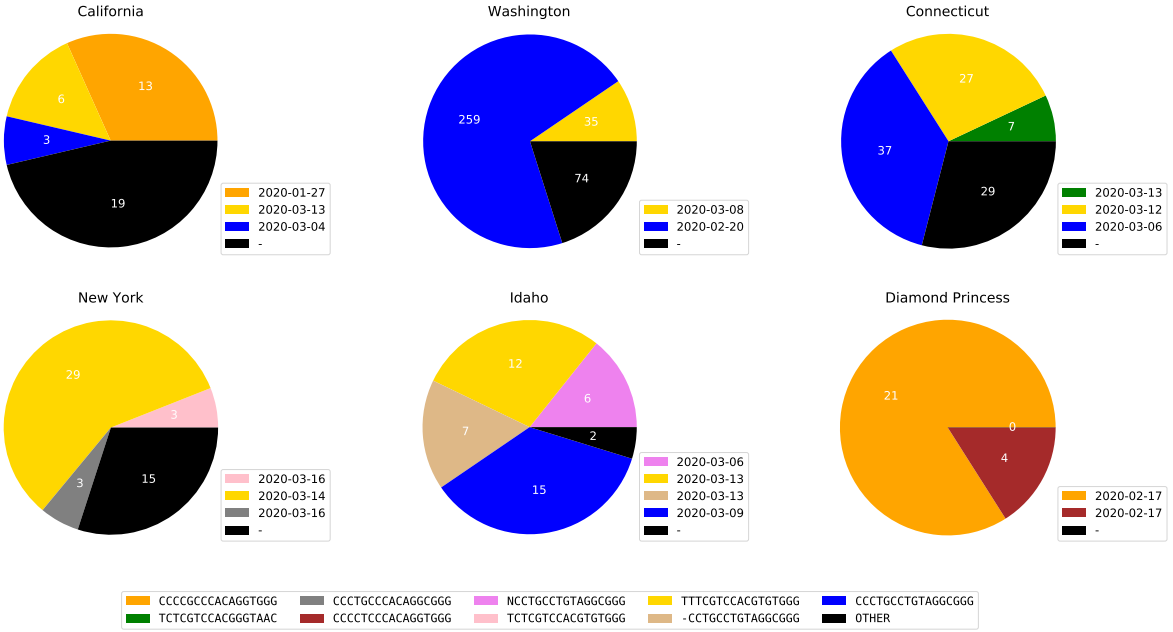

Figure 5. Viral subtype distribution in the United States, showing California (CA), New York (NY), Washington (WA), and U.S. passengers on the *Diamond Princess* cruise ship. Subtypes with less than 5% abundance are plotted as Other. The raw counts for all ISMs in each state, as well as the date each ISM was first found in a sequence in that state, are provided in Supplementary file 2 — ISM abundance table of 5 US states and *Diamond Princess*

We also found that different states within the United States have substantially different subtype distributions. Figure 5 shows the predominant subtype distributions in the states with the most available sequences. Figure 5 also shows the subtypes found among U.S. passengers on the *Diamond Princess* cruise, which experienced an outbreak that was traced back to a Hong Kong passenger who embarked on the vessel on January 21, 2020. [19]. The pie charts demonstrate subregional viral subtype diversity within the United States. The colors shown on the charts are also keyed to the colors used in Figure 4, which allows for the visualization of commonalities between the subregional subtypes in the US and the subtypes distributed in other regions. Most prominently, the sequences in New York are dominated a subtype, TTTCGTCCACGTGTGGG, which is also highly abundant among sequences from European countries, including France, Iceland, Germany, and Belgium. California, on the other hand, includes as a dominant subtype, CCCCGCCCACAGGTGGG, which is also a major subtype in Mainland China, as shown in Figure 4. The dominant subtype in Washington, CCCTGCCTGTAGGCGGG, is also the most abundant in the United States as a whole, likely as the result of Washington state having the most subtypes overall. This subtype is also found in substantial abundance in Canada as well. But, the ISM was not detected in sequences from New York, further

suggesting that the outbreak of SARS-CoV-2 centered in New York may have distinct characteristics. (That said, as shown in Figure 5, this subtype has been detected in nearby Connecticut.) Moreover, as expected, the viral subtypes on the *Diamond Princess* cruise ship are the same as those found in Mainland China early in the progress of the virus, which is consistent with the hypothesis of an outbreak resulting from an early exposure by a single source linked to China.

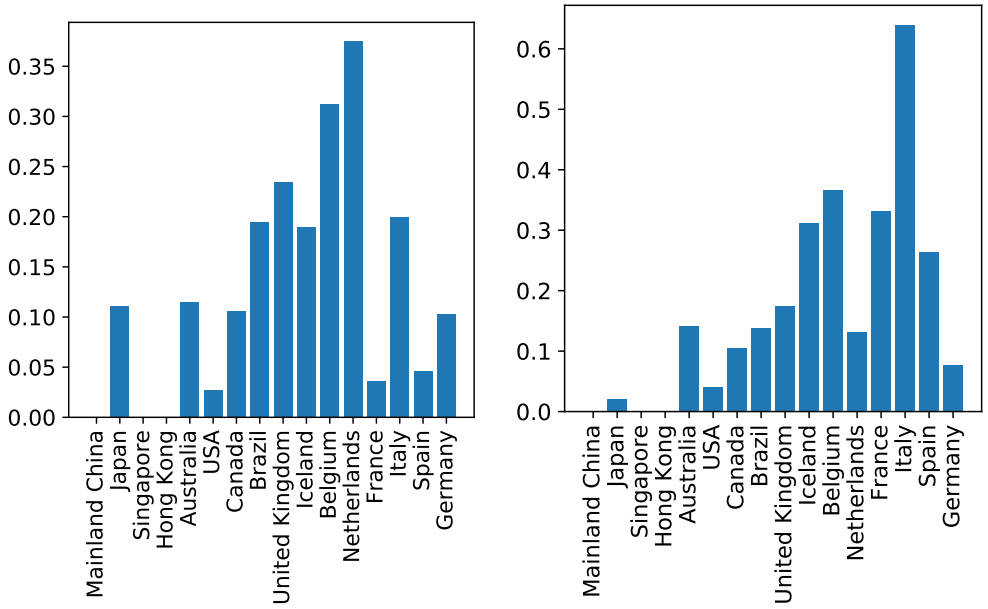

Figure 6. Worldwide distribution (%) of the most abundant subtypes in Italy, `TCGTCCACGGGTAAC` (left) and `TCGTCCACGGGTGGG` (right).

To further validate the utility of ISMs for subtyping, we focused on the analysis of the geographical distribution of the dominant subtypes in Italy. Based on publicly available sequence data from Italy, we found that Italy had two particularly abundant ISMs, as can be seen in the pie chart in Figure 4. Figure 6 shows the relative abundance (proportion of total sequences in that country/region) of each of these "Italy subtypes" in other countries/regions. As the plot shows, the outbreak in other European countries have generally involved the same viral subtypes as in Italy, as defined by ISM. Indeed, initial reports of cases in various other European countries in late February 2020 were linked to travellers from Italy [1]. The Italy subtypes are found, however, at lower yet still significant abundance in countries including Japan, Canada, the United States, and Australia. Somewhat surprisingly, though the Italy subtypes were found in other states, only 1 out of the 50 sequences from New York in the data set had the same ISM as a dominant subtype in Italy (see Supplementary file 2 — ISM abundance table of 5 US states and *Diamond Princess*). This further suggests that the outbreak in New York may not be linked directly to travel exposure directly from Italy, but rather from another location in Europe.

The Italy subtypes are not found at all in locations in Asia, however, such as Mainland China and Singapore, as indicated in Figure 6. Overall, the aforementioned results are consistent with phylogenetic tree-based analyses, such as that illustrated on NextStrain's website (`http://www.nextstrain.org/ncov`), which suggest a flow of the infection from Asia, to Italy, and then subsequently export from Italy to other countries in Europe. It is important to note, however, that the ISMs also resolve potentially significant differences in the subtype distributions in European countries outside of Italy as the virus continues to progress, indicated in Figure 4.

Temporal dynamics of SARS-CoV-2 subtypes

Temporal dynamics of viral subtypes within geographical regions

The present-time geographical distributions shown in Figures 4, 5, and 6 suggest that ISM subtyping may identify the temporal trends underlying the expansion of SARS-CoV-2 virus and the COVID-19 pandemic. To demonstrate the feasibility of modeling the temporal dynamics of the virus, we first analyzed the temporal progression of different ISMs on a country-by-country basis. This allows us to examine the complex behavior of subtypes as infections expand in each country and may be influenced by subtypes imported from other regions.

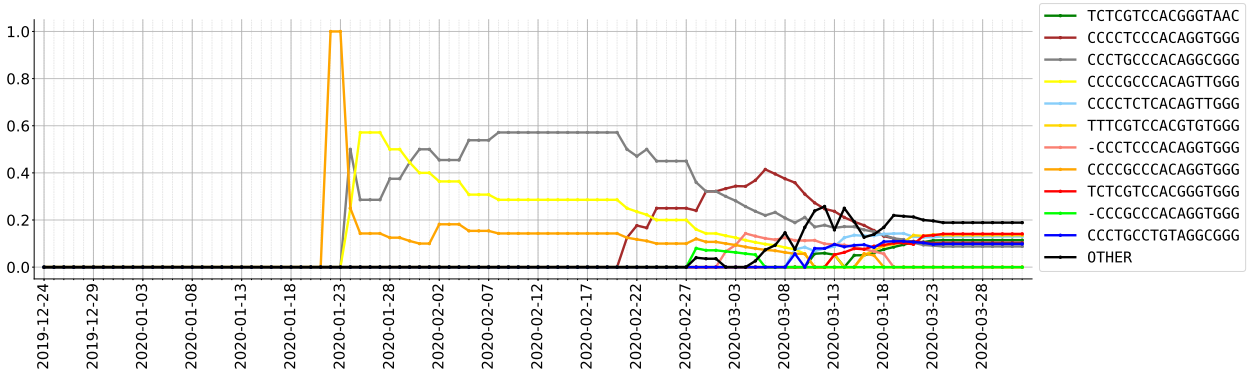

Figure 7. Relative abundance (%) of ISMs in DNA sequences from Australia as sampled over time.

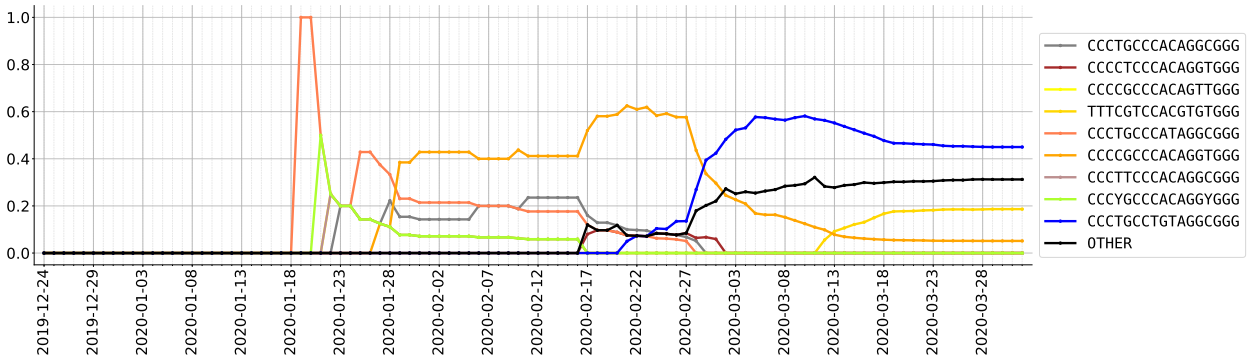

Figure 8. Relative abundance of ISMs in DNA sequences from USA as sampled over time.

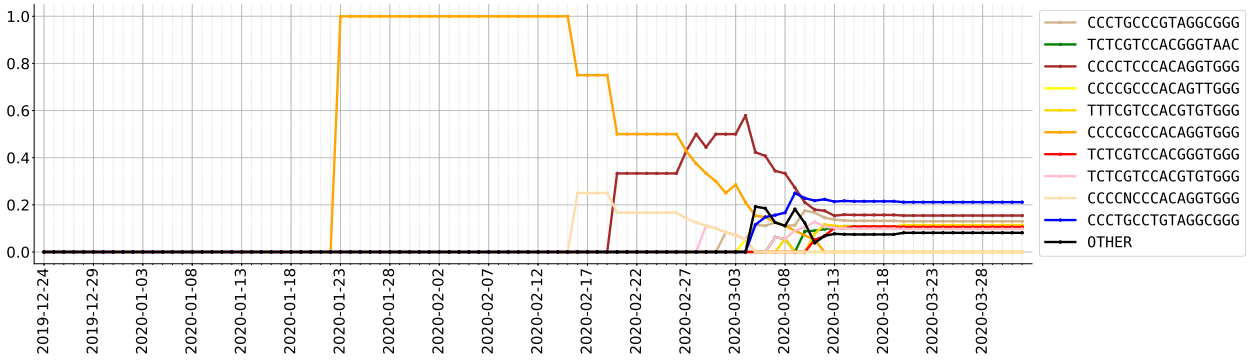

Figure 9. Relative abundance of ISMs in DNA sequences from Canada as sampled over time.

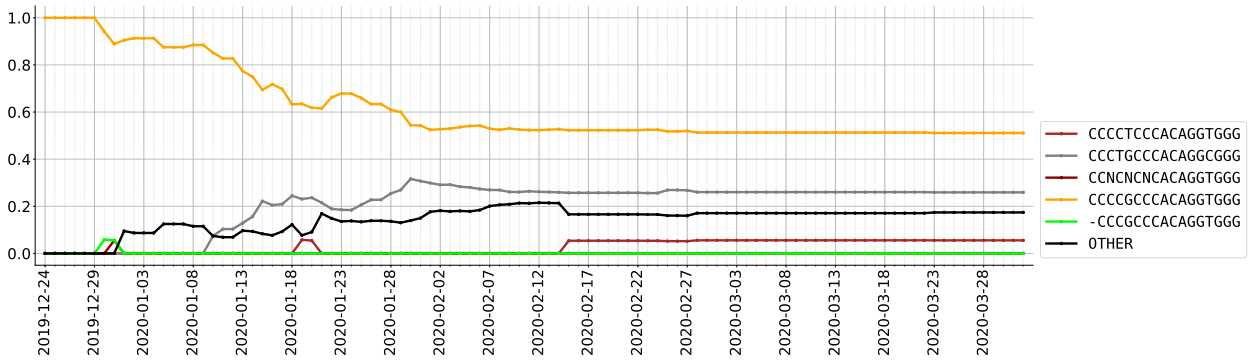

Figure 10. Relative abundance of ISMs in DNA sequences from Mainland China as sampled over time.

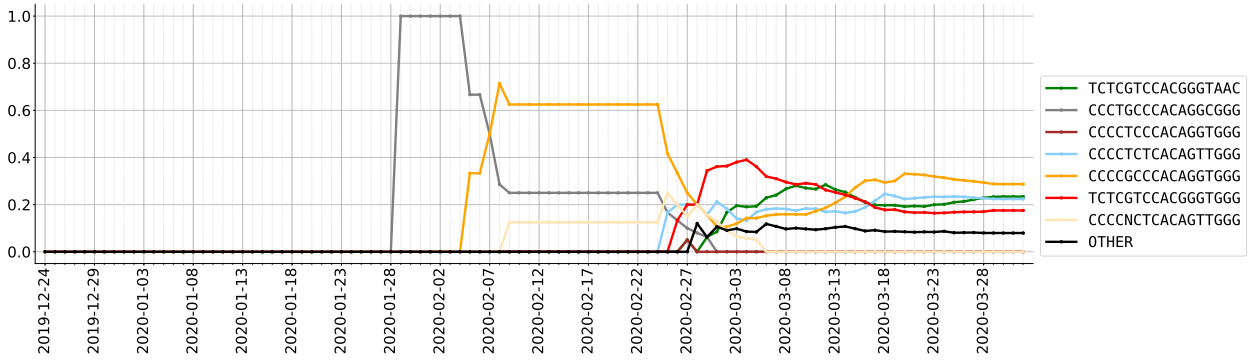

Figure 11. Relative abundance of ISMs in DNA sequences from the United Kingdom as sampled over time.

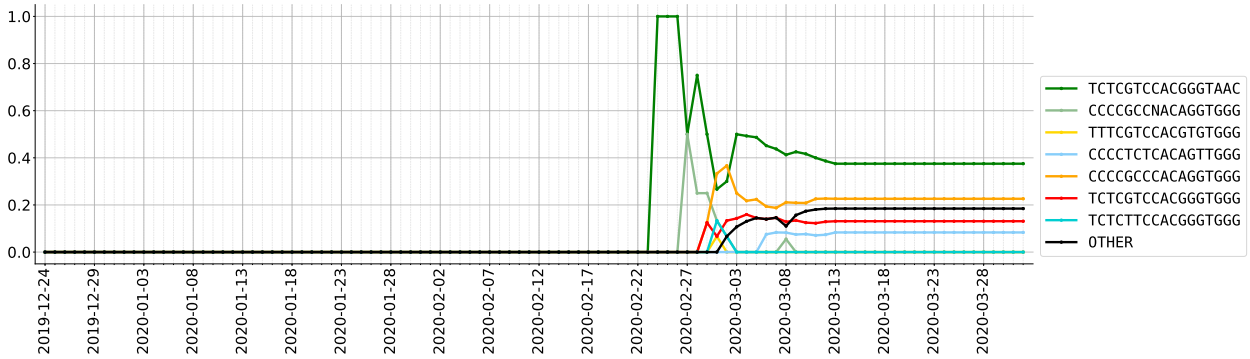

Figure 12. Relative abundance of ISMs in DNA sequences from the Netherlands as sampled over time.

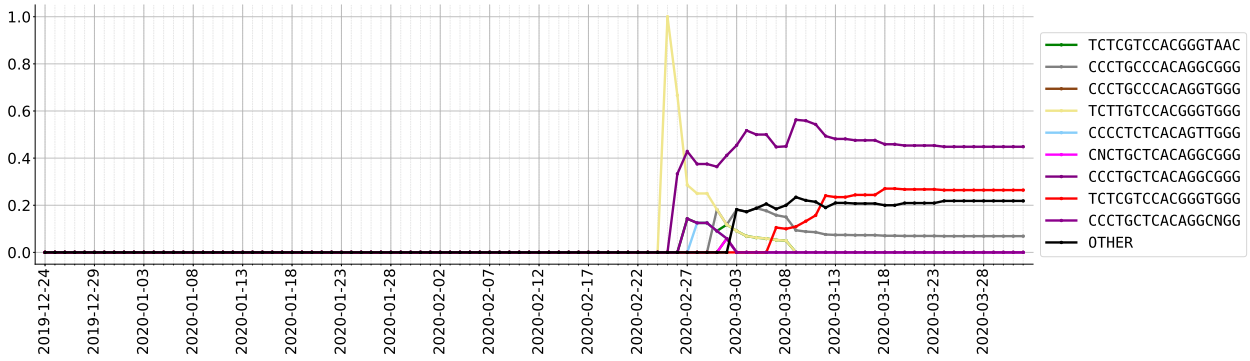

Figure 13. Relative abundance of ISMs in DNA sequences from Spain as sampled over time.

We focused our analysis on the temporal dynamics of the viral subtypes in Mainland China, Australia, Canada, the United States, the Netherlands, United Kingdom, and Spain. As Figure 4 shows, Australia, Canada, and the United States have a substantial level of geographical diversity in ISMs. By contrast, as discussed above and shown in Figure 6, at the present time one particular ISM is more uniformly dominant in European countries. Mainland China is shown for reference as the earliest site of viral detection. Following the methods described in the Methods section, we graph how viral subtypes are emerging and growing over time, by plotting the relative abundance of viral subtypes in a country (via the most frequently occurring ISMs over time), in Figs. 7–13. As discussed above, through the pipeline we have developed, these plots use a consistent set of colors to indicate different ISMs (and are also consistent with the coloring scheme in Figure 4.

As an initial matter, Figure 10 reflects Mainland China's containment of SARS-nCoV-2, as seen in the initial growth in viral genetic diversity, followed by a flattening as fewer new cases were found (and correspondingly fewer new viral samples were sequenced). Australia, on the other hand, shows growing subtype diversity as its cases increase over time. Initially, Australia's sequences were dominated by two subtypes that were also substantially abundant in Mainland China, and another subtype (`CCCCGCCCACAGTTGGG`) that was less relatively abundant in Mainland China but more highly abundant in sequences from Hong Kong and Singapore (see Figure 4). Later, another subtype that was found in Mainland China emerged in Australia, and then, starting with sequences obtained on February 27, 2020 and subsequently, more subtypes are seen to emerge in Australia that were not found in other Asian countries but were found in Europe. This pattern suggests a hypothesis that Australia may have had multiple independent viral transmissions from Mainland China, followed by potentially independent importation of the virus from Europe and North America. A similar pattern is seen in Canada. Figure 9 shows that the earliest viral sequences in Canada included mostly subtypes found in Mainland China, with the same pattern in which there was a second, later subtype in common with Mainland China, followed by a diversification of subtypes that including many in common in Europe and the United States. In sum, Australia and Canada show patterns that might be expected for smaller populations in countries with diverse and extensive travel connections.

In the United States, however, the most abundantly found subtype in the current subtype population, `CCCTGCCTGTAGGCGGG`, is not abundant in either Asia or Europe. However, the subtype has been found in substantial numbers of sequences in Canada and Australia. It is plausible, therefore, that this subtype has become abundant as the result of community transmission in the United States, and has been exported from the United States to these other countries. Interestingly, while this subtype has been found across the United States, as shown in Figure 5, it has not been found to be substantially abundant in New York. This is notable, as at the time of this study, within the United States initially been the state with the most significant outbreak of the virus as measured by positive tests, as well as COVID-19 hospitalizations and deaths [26]. The predominant subtype in New York is in fact the same as a major subtype in that found in European countries, such as France and Belgium. This suggests that the New York outbreak is the result of importation from Europe, as opposed to the subtype more characteristic of sequences elsewhere in the United States, particularly Washington state (see Figure 5).

As shown in Figures Figs. 11–13, the subtype distribution in sequences in European countries differs significantly from that of North America and Australia. In particular, as detailed above, the European dynamics of SARS-CoV-2 appear to reflect the theory that in many European countries, the first cases were due to travel from Italy. In data from the United Kingdom, however, we observe the same initial subtypes shared with Mainland China that were also observed in Australia and Canada. It may be the case though that these subtypes would have been observed early on in the Netherlands and Spain as well, but were missed because sequencing only began with later cases. As expected, however, especially initially, but throughout, the predominant subtypes in Italy discussed above are represented among viral sequences in all three countries. But distinct subtypes are found in these countries as well. The `CCCCTCTCACAGTTGGG` subtype has emerged as a highly abundant subtype in United Kingdom data. This subtype has also been found in substantial numbers in the Netherlands, as well as in Australia, but not in Spain. As Figure 13 shows, in Spain, the latter subtype was found in an early sequence data but not thereafter. And, in Spain, a unique subtype has emerged that is not found in abundance in any other country.

Temporal dynamics of individual viral subtypes across different regions

We also include in the pipeline the generation of plots that show how the dynamics of a subset evolve over time in different geographical regions. We illustrate this analysis by tracing the progress of the subtype associated with the ISM obtained from the reference viral sequence [30]. Since this sequence appears to have arisen early in the international spread of the virus, it is a useful demonstration for this kind of comparative temporal analysis.This plot illustrates how the reference subtype, which was characterized in early sequences has progressed from being found entirely in Mainland China to being found in the United States, and then subsequently to a greater degree in Europe

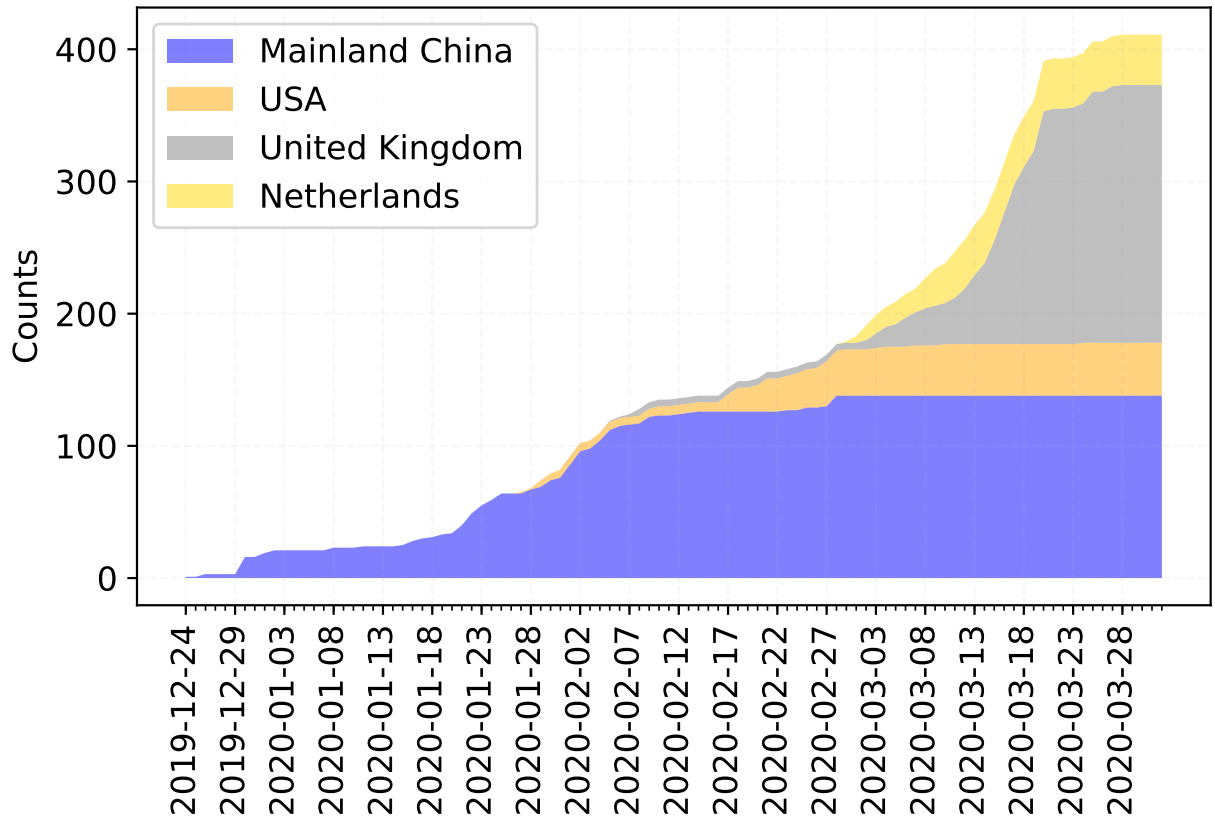

Figure 14. Stacked plot of the number of sequences of the reference sequence ISM subtype (CCCCGCCCACAGGTGGG).

Figure 14 shows, the reference genome subtype began to grow in abundance in Mainland China, before levelling off, and then being detected in the United States and Europe, and subsequently levelling off in those countries as well. In the case of Mainland China, that could be due to the substantial reduction in reported numbers of new infections and thus additional sequences being sampled. However, the other countries have continuing increases in reported infection as of the date of the data set, as well as substantially increasing numbers of sequences being sampled – making it less likely that the reference subtype is simply being missed. In those cases, it appears from Figures 8, 11, and 12 that in later times, other subtypes have emerged over time and are becoming increasingly abundant. One potential explanation is that because the SARS-CoV-2, is an RNA virus and thus highly susceptible to mutation as transmissions occur [16]. Therefore, as transmissions have continued, the ISM associated with the reference sequence has been replaced by different ISMs due to these mutations. Another plausible explanation for such levelling off in a region, which would be consistent with the pattern in data from the United States in particular, is that this leveling off represents containment of a transmission of the subtype with that ISM. And, meanwhile, other subtypes continue to

expand in that country or region. Further investigation and modeling of subtype distributions, as well as additional data, will be necessary to help resolve these questions — particularly in view of the caveats described below.

Important considerations for interpreting temporal trends based on viral subtyping

Inferences from the temporal trends in subtypes described in the foregoing must be limited by important caveats: Because the number of viral sequences is much smaller than the number of cases, there may be a lag before a sample is sequenced that includes a particular ISM. As a result, even though subtype `CCCTGCCTGTAGGCGGG` is first seen in the United States on February 20, 2020 and then a sequence with that ISM was obtained in Canada sequences about 14 days later, that does not necessarily mean that Canada acquired this subtype from US. However, given the amount of time that has lapsed, the general result that this subtype did not originate in Mainland China, for example, is more robust. We are also limited in that the depth of sequencing within different regions is highly variable. As an extreme case, Iceland, which has a small population, represents nearly 9% of all sequences in the complete data set. As a result, tracking the relative abundances of subtypes across different regions is complicated, because a region that does more sequencing may simply end up having a greater number of sequences of any given subtype. This problem is exacerbated in temporal analysis, because the extent of sequencing efforts in a region may also change over time.

Additionally, some ISMs include "`-`" and "`n`" symbols, which represent gaps and ambiguity in the sequence, which indicate the presence of noise is in the sequence data. For example, subtype `TCTCGTCCACGGGNNNN` could in fact be subtype `TCTCGTCCACGGGTAAC` and subtype `TCTCGTCCACGGGTGGG`. We are currently developing methodologies to improve the precision of ISM definition by accounting for technical sequencing errors and ambiguous base calls. We are also evaluating the potential to use epidemiological data for the growth in the number of cases, as a potential supplement for effectively calibrating temporal analysis of viral subtype dynamics.

Conclusions

In this paper, we propose to use short sets of nucleotides as markers to define subtypes of SARS-CoV-2 sequences (ISMs). We validate the utility of ISM distributions as a complement to phylogenetic tree-based approaches, e.g. as used in the Nextstrain project and by other investigators, by demonstrating that patterns of ISM-based subtypes similarly model the general understanding of how the outbreak has progressed through travel exposure and community transmission in different regions. Specifically, we show that the distribution of ISMs is an indicator of the geographical distribution of the virus as predicted by the flow of the virus from China, the initial European outbreak in Italy and subsequent development of local subtypes within individual European countries as well as interregional differences in viral outbreaks in the United States. In addition, we demonstrate that by using ISMs for subtyping, we can also readily visualization the geographic and temporal distribution of subtypes in an efficient and uniform manner. We have developed and are making available a pipeline to generate quantitative profiles of subtypes and the visualizations that are presented in this paper.

Overall, the entropy-based subtyping approach described in this paper represents a potentially efficient way for researchers to gain further insight on the evolution and diversity of SARS-CoV-2 sequences and their diversity over time. An important caveat of this approach, as with others based on analysis of viral genome sequence, is that it is limited by the sampling of viral sequences. Small and non-uniform samples of sequences may not accurately reflect the true diversity of viral subtypes within a given population. However, the ISM-based approach has the advantage of being scalable as sequence information grows, and as a result will be able to become both more accurate and precise as sequence information grows within different geographical and other subpopulations. Indeed, with the pipeline in place and access to continuously updating sequencing data, ISMs may be continuously identified as new sequences are sequenced and categorized to a subtype for further analysis. In the future, therefore, as data becomes available, ISM-based subtyping may be employed on subpopulations within regions, demographic groups, and groups of patients

with different clinical outcome. Efficient subtyping of the massive amount of SARS-CoV-2 sequence data will therefore enable quantitative modeling and machine learning methods to develop improved containment and potentially also therapeutic strategies against SARS-CoV-2.

Acknowledgments

We downloaded all SARS-Cov-2 sequences available from and acknowledge the contributions of the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu database, which has made accessible novel coronavirus sequencing data, including from the NIH Genbank resource [22]. This work was partially supported by NSF grant #1919691. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number #ACI-1548562.

References

1. Coronavirus: Outbreak spreads in europe from italy. *BBC News*.
2. C. Ceraolo and F. M. Giorgi. Genomic variance of the 2019-ncov coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.
3. J. F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K. K.-W. To, S. Yuan, and K.-Y. Yuen. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting wuhan. *Emerging Microbes & Infections*, 9(1):221–236, 2020. PMID: 31987001.
4. J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1):D633–D642, 11 2013.
5. C. Colijn and J. Gardy. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*, 2014(1):96–108, 06 2014.
6. A. M. Eren, L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin. Oligotyping: differentiating between closely related microbial taxa using 16s rrna gene data. *Methods in Ecology and Evolution*, 4(12):1111–1119, 2013.
7. J. Gregory Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D Bushman, E. K Costello, N. Fierer, A. Gonzalez Peña, J. Goodrich, J. I Gordon, G. Huttley, S. T Kelley, D. Knights, J. E Koenig, R. Ley, C. Lozupone, D. Mcdonald, B. D Muegge, M. Pirrung, and R. Knight. Qiime allows analysis of high-throughput community sequencing data. nat met 7: 335-336. *Nature methods*, 7:335–6, 04 2010.
8. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology*, 4(1):10–19, 2019.
9. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 05 2018.
10. T. Karamitros, G. Papadopoulou, M. Bousali, A. Mexias, S. Tsiodras, and A. Mentis. Sars-cov-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *bioRxiv*, 2020.
11. K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, 01 2013.
12. R. N. Kirchdoerfer and A. B. Ward. Structure of the sars-cov nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications*, 10(1):2342, 2019.

13. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, and Z. Feng. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, 2020.

14. D. McDonald, A. Birmingham, and R. Knight. Context and the human microbiome. *Microbiome*, 3(1):52, Nov 2015.

15. D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, L. DeRight Goldasich, P. C. Dorrestein, R. R. Dunn, A. K. Fahimipour, J. Gaffney, J. A. Gilbert, G. Gogul, J. L. Green, P. Hugenholtz, G. Humphrey, C. Huttenhower, M. A. Jackson, S. Janssen, D. V. Jeste, L. Jiang, S. T. Kelley, D. Knights, T. Kosciolek, J. Ladau, J. Leach, C. Marotz, D. Meleshko, A. V. Melnik, J. L. Metcalf, H. Mohimani, E. Montassier, J. Navas-Molina, T. T. Nguyen, S. Peddada, P. Pevzner, K. S. Pollard, G. Rahnavard, A. Robbins-Pianka, N. Sangwan, J. Shorenstein, L. Smarr, S. J. Song, T. Spector, A. D. Swafford, V. G. Thackray, L. R. Thompson, A. Tripathi, Y. Vázquez-Baeza, A. Vrbanac, P. Wischmeyer, E. Wolfe, Q. Zhu, , and R. Knight. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3), 2018.

16. A. Moya, E. C. Holmes, and F. González-Candelas. The population genetics and evolutionary epidemiology of rna viruses. *Nature Reviews Microbiology*, 2(4):279–288, 2004.

17. X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, X. Chen, J. Chen, K. Hu, Q. Jin, J. Wang, and Z. Qian. Characterization of spike glycoprotein of sars-cov-2 on virus entry and its immune cross-reactivity with sars-cov. *Nature Communications*, 11(1):1620, 2020.

18. E. R. Robinson, T. M. Walker, and M. J. Pallen. Genomics and outbreak investigation: from sequence to consequence. *Genome Medicine*, 5(4):36, 2013.

19. J. Rocklöv, H. Sjödin, and A. Wilder-Smith. COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *Journal of Travel Medicine*, 02 2020. taaa030.

20. T. Sekizuka, K. Itokawa, T. Kageyama, S. Saito, I. Takayama, H. Asanuma, N. Naganori, R. Tanaka, M. Hashino, T. Takahashi, H. Kamiya, T. Yamagishi, K. Kakimoto, M. Suzuki, H. Hasegawa, T. Wakita, and M. Kuroda. Haplotype networks of sars-cov-2 infections in the diamond princess cruise ship outbreak. *medRxiv*, 2020.

21. Z. Shen, Y. Xiao, L. Kang, W. Ma, L. Shi, L. Zhang, Z. Zhou, J. Yang, J. Zhong, D. Yang, L. Guo, G. Zhang, H. Li, Y. Xu, M. Chen, Z. Gao, J. Wang, L. Ren, and M. Li. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *Clinical Infectious Diseases*, 03 2020. ciaa203.

22. Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), 2017.

23. Y. C. Su, D. E. Anderson, B. E. Young, F. Zhu, M. Linster, S. Kalimuddin, J. G. Low, Z. Yan, J. Jayakumar, L. Sun, G. Z. Yan, I. H. Mendenhall, Y.-S. Leo, D. C. Lye, L.-F. Wang, and G. J. Smith. Discovery of a 382-nt deletion during the early evolution of sars-cov-2. *bioRxiv*, 2020.

24. W. Tan, X. Zhao, X. Ma, W. Wang, P. Niu, W. Xu, G. Gao, and G. Wu. A novel coronavirus genome identified in a cluster of pneumonia cases—wuhan, china 2019- 2020. *China CDC Weekly*, 2(4):61–2, 2020.

25. X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, J. Cui, and J. Lu. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, 03 2020. nwaa036.

26. J. H. Tanne. Covid-19: New york city deaths pass 1000 as trump tells americans to distance for 30 days. *BMJ*, 369, 2020.

27. K. K.-W. To, O. T. yin Tsang, W. shing Leung, A. R. Tam, T. chiu Wu, D. C. Lung, C. C.-Y. Yip, J. piao Cai, J. M.-C. Chan, T. S.-H. Chik, D. P.-L. Lau, C. Y.-C. Choi, L.-L. Chen, W.-M. Chan, K. hung Chan, J. D. Ip, A. C.-K. Ng, R. W.-S. Poon, C. Luo, V. W.-S. Cheng, J. F.-W. Chan, I. F. N. Hung, Z. Chen, H. Chen, and K.-Y. Yuen. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by sars-cov-2: an observational cohort study. *The Lancet. Infectious diseases*, 2020.

28. J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74, Sep. 2014.

29. A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 2020/04/06 XXXX.

30. C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, T. He, and Z. Zhang. The establishment of reference sequence for sars-cov-2 and variation analysis. *Journal of Medical Virology*, n/a(n/a).

31. M. Wang, M. Li, R. Ren, A. Brave, S. v. d. Werf, E.-Q. Chen, Z. Zong, W. Li, and B. Ying. International expansion of a novel sars-cov-2 mutant. *medRxiv*, 2020.

32. W. G. Weisburg, S. M. Barns, D. A. Pelletier, and D. J. W. Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of bacteriology*, 173 2:697–703, 1991.

33. R. J. F. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, 2013.

34. I.-M. Yu, C. L. T. Gustafson, J. Diao, J. W. I. Burgner, Z. Li, J. qiang Zhang, and J. Chen. Recombinant severe acute respiratory syndrome (sars) coronavirus nucleocapsid protein forms a dimer through its c-terminal domain. *The Journal of biological chemistry*, 280 24:23280–6, 2005.

35. K.-S. Yuen, Z. W. Ye, S.-Y. Fung, C.-P. Chan, and D.-Y. Jin. Sars-cov-2 and covid-19: The most important research questions. *Cell & Bioscience*, 10(1):40, 2020.

36. W. Zhao, S. Song, M. Chen, D. Zou, L. Ma, Y.-K. Ma, R. Li, L. Hao, C. Li, D. Tian, B. Tang, Y.-Q. Wang, J. Zhu, H. Chen, Z. Zhang, Y. Xue, and Y. Bào. The 2019 novel coronavirus resource. *Yi chuan = Hereditas*, 42 2:212–221, 2020.

Supplementary Files

Supplementary file 1 — ISM abundance table of 16 countries/regions

The raw counts for all ISMs in each of 16 countries/regions, as well as the date each ISM was first found in a sequence in that country/region.

Supplementary Files

Supplementary file 2 — ISM abundance table of 5 US states and *Diamond Princess*

The raw counts for all ISMs in each of 5 US states and *Diamond Princess*, as well as the date each ISM was first found in a sequence in that location.