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

We propose an efficient framework for genetic subtyping of a pandemic virus, with application to the novel coronavirus SARS-CoV-2. Efficient identification of subtypes is particularly important for tracking the geographic distribution and temporal dynamics of infectious spread in real-time. In this paper, we utilize an entropy analysis to identify nucleotide sites within SARS-CoV-2 genome sequences that are highly informative of genetic variation, and thereby define an Informative Subtype Marker (ISM) for each sequence. We further apply an error correction technique to the ISMs, for more robust subtype definition given ambiguity and noise in sequence data. We show that, by analyzing the ISMs of global SARS-CoV-2 sequence data, we can distinguish interregional differences in viral subtype distribution, and track the emergence of subtypes in different regions over time. Based on publicly available data up to April 5, 2020, we show, for example: (1) distinct genetic subtypes of infections in Europe, with earlier transmission linked to subtypes prevalent in Italy with later development of subtypes specific to other countries over time; (2) within the United States, the emergence of an endogenous U.S. subtype that is distinct from the outbreak in New York, which is linked instead to subtypes found in Europe; and (3) dynamic emergence of SARS-CoV-2 from

localization in China to a pattern of distinct regional subtypes in different countries around the world over time. Our results demonstrate that utilizing ISMs for genetic subtyping can be an important complement to conventional phylogenetic tree-based analyses of the COVID-19 pandemic. Particularly, because ISMs are efficient and compact subtype identifiers, they will be useful for modeling, data-mining, and machine learning tools to help enhance containment, therapeutic, and vaccine targeting strategies for fighting the COVID-19 pandemic. We have made the subtype identification pipeline described in this paper publicly available at https://github.com/EESI/ISM.

Author Summary

The novel coronavirus responsible for COVID-19, SARS-CoV-2, expanded to reportedly 1.3 million confirmed cases worldwide by April 7, 2020. The global SARS-CoV-2 pandemic highlights the importance of tracking dynamics of viral pandemics in real-time. Through the beginning of April 2020, researchers obtained genetic sequences of SARS-CoV-2 from nearly 4,000 infected individuals worldwide. Since the virus readily mutates, each sequence of an infected individual contains useful information linked to the individual's exposure location and sample date. But, there are over 30,000 bases in the full SARS-CoV-2 genome — so tracking genetic variants on a whole-sequence basis becomes unwieldy. We describe a method to instead efficiently identify and label genetic variants, or "subtypes" of SARS-CoV-2. Applying this method results in a compact, 17 base-long label, called an Informative Subtype Marker or "ISM." We define viral subtypes for each ISM, and show how regional distribution of subtypes track the progress of the pandemic. Major findings include (1) showing distinct viral subtypes of infections in Europe emanating from Italy to other countries over time, and (2) tracking emergence of a local subtype across the United States connected to Asia and distinct from the outbreak in New York, which is connected to Europe.

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. [17,29]. 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. 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. There are now central repositories accumulating international SARS-CoV-2 genome data, such as the Global Initiative on Sharing all Individual Data (GISAID) [27] (available at https://www.gisaid.org/).

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 [11] (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) [41].

In more granular studies, early work by researchers based in China analyzing 103 genome sequences, identified two highly linked single nucleotides, leading them to suggest that two major subtypes had emerged: one called "L," predominantly found in the Wuhan area, and "S," which derived from "L" and found elsewhere [30]. 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 [35]. Researchers have also begun to publish studies of the international expansion of specific variants, though as yet the timeline and variant composition of such systematic studies have been limited [36]. Studies have also been undertaken of specific localized infections out of context of the pandemic as a whole, such as analysis of sequences from passengers on the *Diamond Princess* cruise ship, including for U.S. passengers by the CDC as well as by Japanese researchers [25].

Researchers are also actively studying sequence variation in order 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 [14, 26]. 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 importance of tracking and modeling genetic changes in the SARS-CoV-2 virus as the outbreak expands, however, there is a need for an efficient and systematic methodology to *quantitatively* characterize the SARS-CoV-2 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 [10, 23].

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 [7, 38]. 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 [6, 9, 18, 19, 37]. To differentiate between closely related microbial taxa,nucleotide positions that represent information-rich variation may be identified [8]. This kind of approach has also been used in the reverse direction to find conserved sites as a way to assemble viral phylogenies [3].

We apply these principles to develop a more efficient framework to quantify viral subtypes, which can 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, our method compresses 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. We develop and implement a pipeline to utilize entropy

analysis 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.

The ISM pipeline further includes an error correction procedure. Losing sequence information would otherwise substantially hamper tracking of the full scope of the viral pandemic. In particular, even though the SARS-CoV-2 data set appears to be large, it represents only a small sample of the full scope of cases. Error correction of subtype labels, therefore, represents an essential component for an effective viral subtyping for real-time tracking of a pandemic, such as SARS-CoV-2. We describe and apply a methodology to account for the substantial amount of noise by in the data set by resolving base-call ambiguities in sequence data, which give rise to spurious ISMs that reflect those anomalies. We demonstrate that by including this error correction procedure in our pipeline, we can eliminate most such spurious ISMs and, thus, maximize utilization of sequence information.

We evaluate the pipeline as a whole 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.[1]

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[2]). We also included a reference sequence from National Center for Biotechnology Information[3] (NCBI Accession number: NC_045512.2). This resulted in an overall data set of 3982 sequences with sequence length ranging from 25342 nt to 30355 nt. We then align all remaining sequences after filtering together using MAFFT [15] using the "FFT-NS-2" method in XSEDE [33]. After alignment, the sequence length is extended (for the present data set, up to 35362 nt).

[1] The latest report at the time of paper submission, run on April 14, 2020 with data up to April 12, 2020, can be found in https://github.com/EESI/ISM/blob/master/ISM-report-20200412-with_error_correction.ipynb.

[2] https://github.com/nextstrain/ncov

[3] https://www.ncbi.nlm.nih.gov/

Entropy analysis and ISM extraction

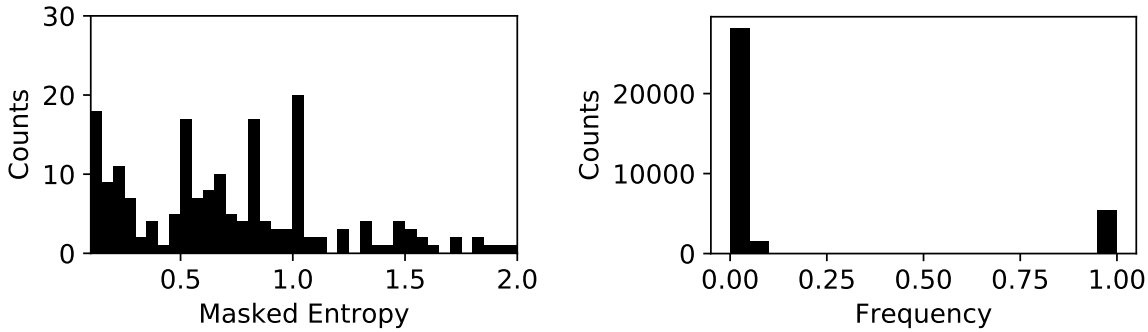

Figure 1. Left: histogram of masked entropy values; Right: histogram of percentages of `n` and `-`

For the aligned sequences, we merged the sequence with the metadata in Nextstrain project[4] as it is in April 6, 2020 based on identification number, **gisaid_epi_isl**, provided by GISAID [27]. 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 resulting final data set 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, which are listed in 1.[5]

Sites `N` and `-` (representing a fully ambiguous site and a gap respectively) are substantially 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

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

[5] 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.

Table 1. Sequence notation [1]

Symbol	Meaning
A	A
C	C
G	G
T	T
W	A or T
S	G or C
M	A or C
K	G or T
R	G or A
Y	T or C
B	G or T or C
D	G or A or T
H	A or C or T
V	G or C or A
N	G or A or T or C

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.

Error correction to resolve ambiguities in sequence data and remove spurious ISMs

The focus of the error correction method is to correct an ISM that contains ambiguous symbols, i.e., a nucleotide identifier that represents an ambiguous base call (more details in 1 based on [1]), such as `N`, which represents a position that could either be `A`, `C`, `T`, or `G`. The basic idea for error correction of ISMs is to use ISMs with no or fewer ambiguous symbols to correct ISMs with such errors (ambiguities). Given an ISM with an error, we first find all ISMs that are identical to the subject ISM at all nucleotide positions in which there is no error in the subject ISM. We refer here to these nearly-identical ISMs as supporting ISMs. Then we iterate over all positions with an error that must be corrected in the subject ISM. For a given nucleotide position, if all other such supporting ISMs with respect to the said erroneous position contain the same *non-ambiguous* base (i.e., an `A`, `C`, `T`, or `G`), then we simply correct the ambiguous base by the *non-ambiguous* base found in the supporting ISMs. However, when the supporting ISMs disagree at a respective nucleotide position, the method generates an ambiguous symbol which represents all the bases that occurred in the supporting ISMs and compare this artificially generated nucleotide symbol with the original position in the subject ISM, if the generated nucleotide symbol identifies a smaller set of bases, e.g., `Y` representing `C` or `T`

rather than `N`, which may be any base, we use the generated symbol to correct the original one.

When we apply the foregoing error correction algorithm to all ISMs that have ambiguous nucleotide symbols resulting from the present data set, we find that 85.4% of erroneous ISMs are partially corrected (meaning at least one nucleotide position with ambiguity was corrected for that ISM but not all), and 41.6% of erroneous ISMs are fully corrected (meaning all positions with ambiguity are corrected to a *non-ambiguous* base (i.e., an `A`, `C`, `T`, or `G`)). Since one ISM my represent multiple sequences in the data set, overall the error correction algorithm is able to partially correct 90.4% of sequences identified by an erroneous ISM, and 47.0% of subjects with an erroneous ISM are fully corrected.

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. All visualizations in this paper and our pipeline are generated using Matplotlib [12]. To improve visualization, ISMs that occur with frequency of 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 were obtained no later than a certain date (as provided in the sequence metadata) together and compute the relative abundance (i.e., frequency) of corresponding subtypes. Any subtypes with a relative abundance that never goes above 2.5% for any date 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.

We also include a proof of concept for using hierarchical clustering to organize ISM subtypes in the pipeline. We picked the 50 most abundant ISMs and calculated the pairwise Hamming distance between them [13]. We then performed hierarchical clustering on these ISMs based on the Hamming distances using average linkage (specifically utilizing the UPGMA algorithm) [21]. The hierarchical clusters are then

visualized in a tree format, as shown below in Figure 16. The leaves in Figure 16 are labeled with the earliest date associated with sequences with the ISM subtypes in the dataset, and the corresponding countries or regions in which a sequence with that ISM was first observed.

Results and Discussion

Identification and Mapping of subtype markers

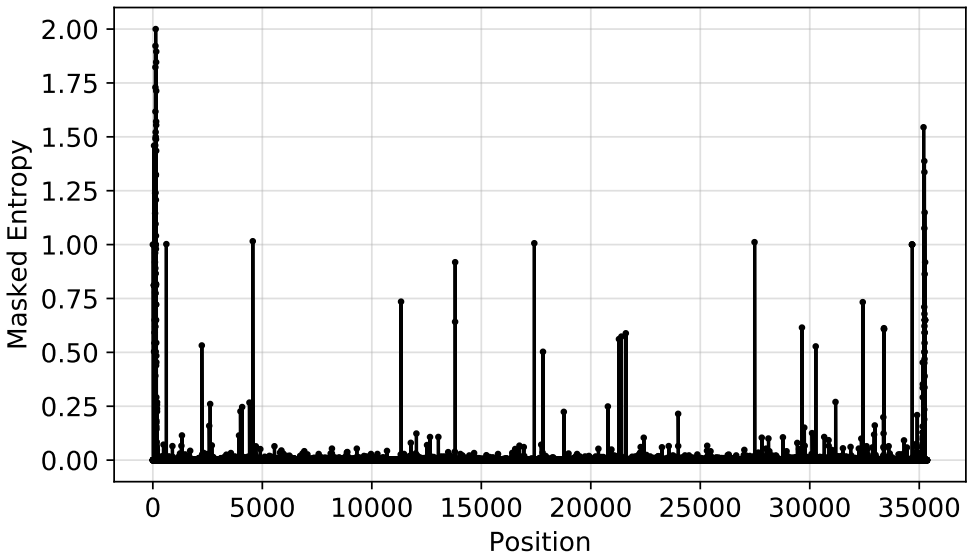

Figure 2. Overall entropy as a function of nucleotide position for all SARS-CoV-2 sequences in the data set

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 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 [14], 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

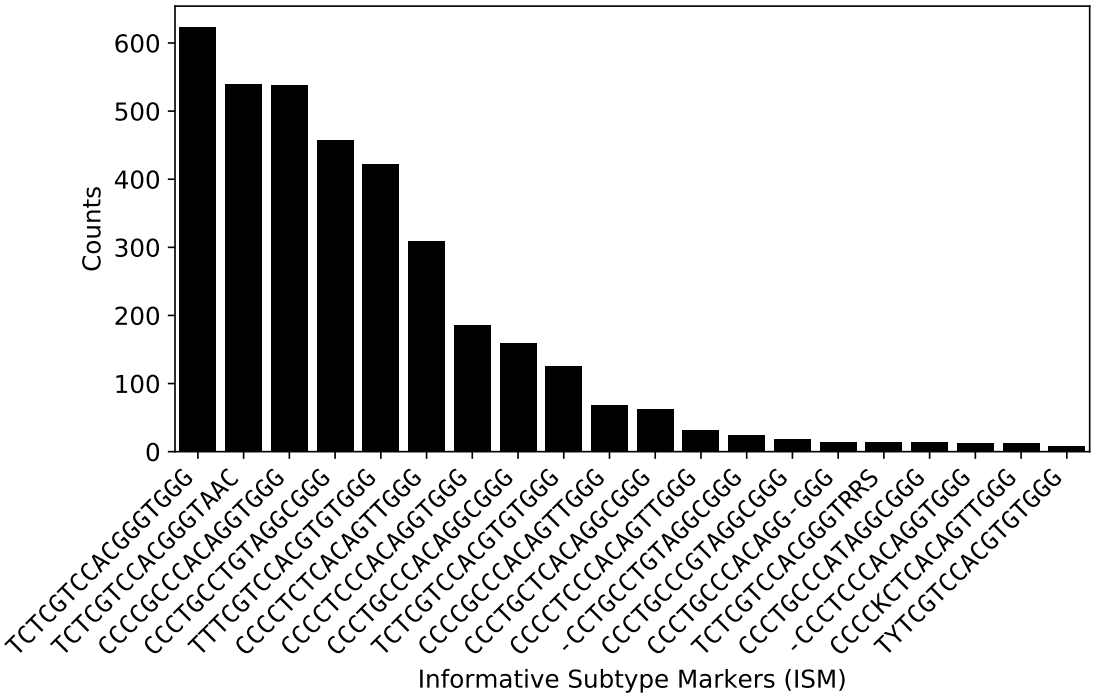

Figure 3. Number of sequences containing the 20 most abundant ISMs within the total data set (out of 3982 sequences).

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).

Table 2. 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

Figure 3 also shows that despite the application of the error correction method detailed in the the Methods section, some symbols representing ambiguously identified nucleotides, such as `S` and `K` still remain in the ISMs. These represent instances in which there was insufficient sequence information to fully resolve ambiguities. We expect that as the number of publicly available sequences increases, there will likely be additional samples that will allow resolution of base-call ambiguities. That said, it is possible that the ambiguity symbols in the ISMs reflect genomic regions or sites that are difficult to resolve using sequencing methods, in which case the ISMs will never fully resolve. Importantly, however, because of the application of the error correction algorithm, there are fewer spurious subtypes which are defined due to variants arising from sequencing errors, and all remaining ISMs are still usable as subtype identifiers.

After the informative nucleotide positions were identified, we then mapped those sites back to the annotated reference sequence for functional interpretation [35]. As shown in Table 2, 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 polymerase (i.e., synthesis) machinery [16]. 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 [22, 34]. High-entropy nucleotide positions were also found in the nucleocapsid formation protein, which is important for packaging the viral RNA. [39] A study has also shown that, like the spike protein, the internal

nucleoprotein of the virus is significant in modulating the antibody response. [32]

Additionally, Table 2 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 [5]. 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 [5, 40]. 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 [4, 28, 30]. 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 [5, 40].

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

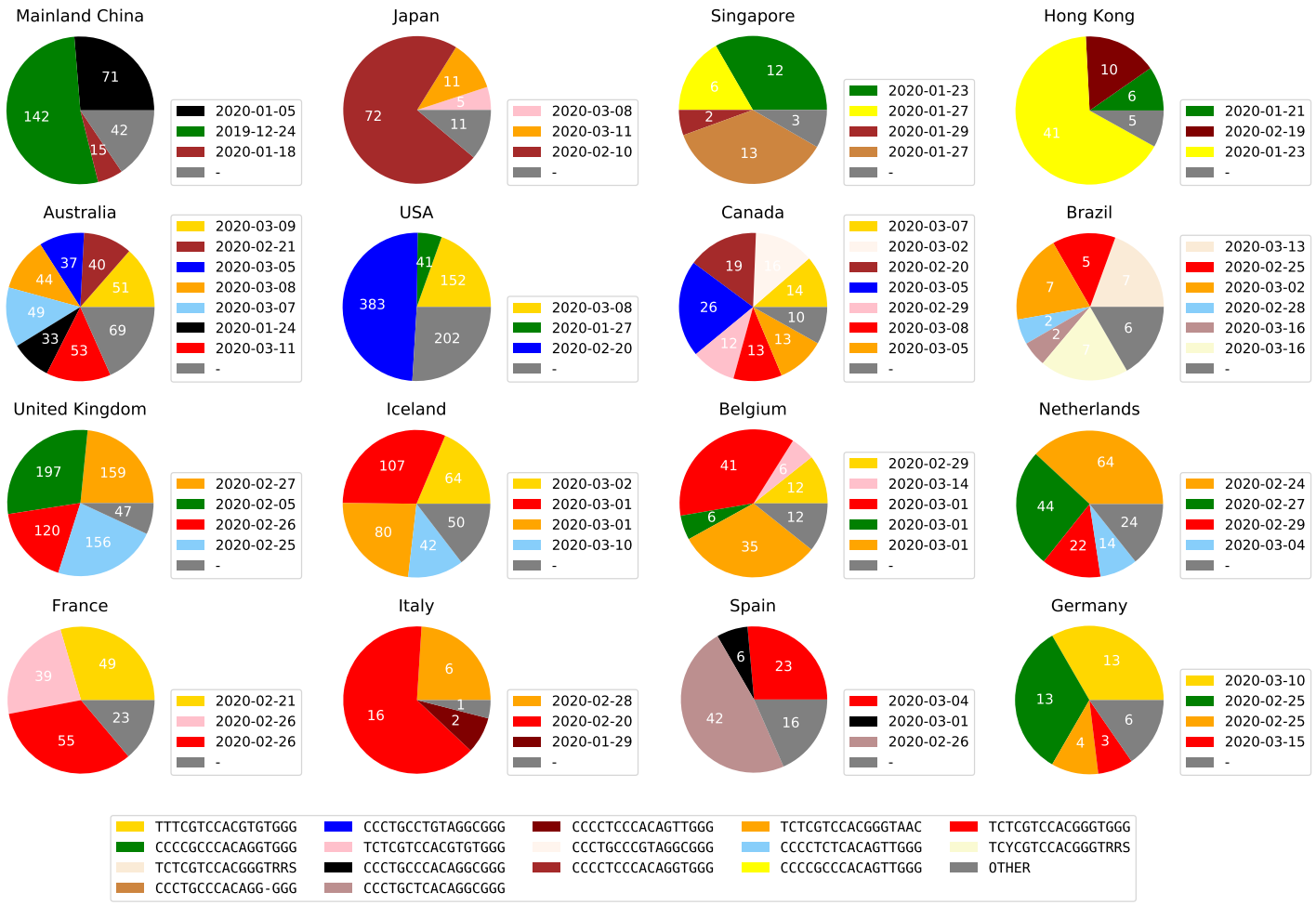

Figure 4. Major subtypes in countries/regions with the most sequences (indicating date subtype was first sequenced in that country/region). Subtypes with less than 5% abundance are plotted as "OTHER". The raw counts for all ISMs in each country/region, as well as the date each ISM was first found in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

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 which was found in Mainland China (CCCCTCCCACAGGTGGG).

Moreover, additional metadata provided for certain sequences for the country of exposure in cases involving travelers reveals that in other non-Asian countries, the CCCCTCCCACAGGTGGG subtype was frequently found in travelers from Iran, including multiple sequences from Canada and Australia, as well as sequences from the United States (specifically New York), New Zealand, and France. Given that the subtype was otherwise not abundant in these countries, it suggests that these traces are due to the subtype being prevalent in Iran at the time these travel cases were found. But in the absence of sequence data from Iran, it is not presently possible to confirm that hypothesis. In any event, the analysis shows that the ISM-based subtype patterns in Asian countries/regions are related or shared, though as can be seen, the larger portion of outbreaks Singapore and Japan appear to have drawn from distinct subtypes that were found earlier in Mainland China, potentially leading to the hypothesis that there were multiple distinct travel-related transmissions from China (or potentially from Iran) 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 U.S. 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, 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.

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 further 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 0.59% 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.

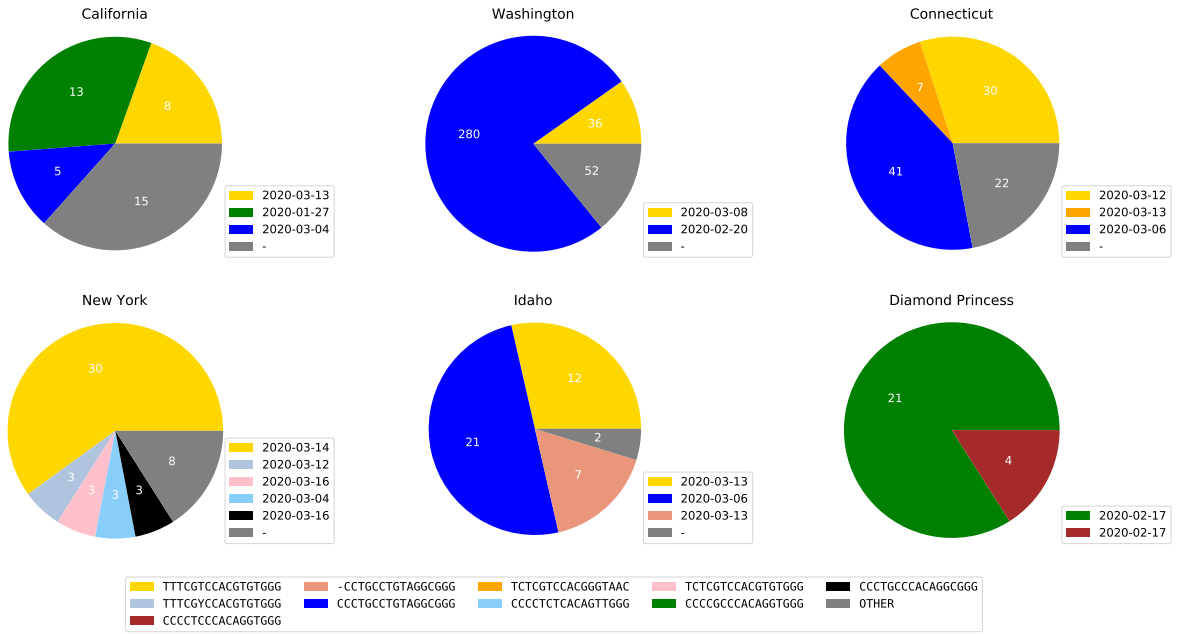

Figure 5. Viral subtype distribution in the United States, showing California (CA), New York (NY), Washington (WA), and U.S. passengers on the *Diamond Princess* cruise ship. Subtypes with less than 5% abundance are plotted as Other. The raw counts for all ISMs in each state, as well as the date each ISM was first found in a sequence in that state, are provided in Supplementary file 2 — ISM abundance table of 5 U.S. 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. [24]. 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 United States and the subtypes distributed in other regions. 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. In the United States, by contrast, the patterns of viral subtypes are distinctly different.

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 most abundant 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. The same `CCCTGCCTGTAGGCGGG` subtype is also found in substantial abundance in Canada as well. Consistent with the hypothesis that this subtype is endogenous to the US, the ISM was found in the sequence of one suspected case of exposure in the United States found in Canada and 11 sequences in Iceland associated with suspected exposure in the United States. (One travel case in Iceland was of the `TTTCGTCCACGTGTGGG` subtype dominant in New York.) But this Washington-dominant ISM has not been detected in sequences from New York, further suggesting that the outbreak of SARS-CoV-2 centered in New York may have distinct characteristics likely due to it arising from import in Europe rather than a connection to other U.S. cases on the West Coast. Interestingly, as shown in Figure 5, this Washington-dominant and suspected-endogenous U.S. subtype has been detected in Connecticut, which may suggest that there could be some unsequenced circulating cases of this viral subtype in New York as well.

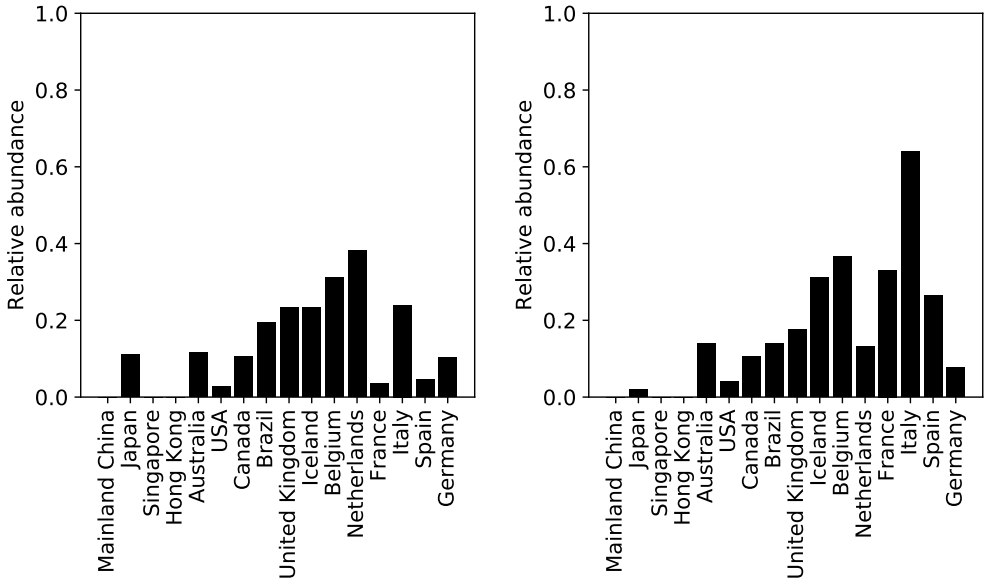

Figure 6. Relative abundance in other countries of the second-most abundant subtype in Italy, `TCGTCCACGGGTAAC` (left) and most abundant subtype in Italy `TCGTCCACGGGTGGG` (right).

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, `TCGTCCACGGGTAAC` and `TCGTCCACGGGTGGG`, as can be seen in the pie chart in Figure 4. (The third-most abundant subtype shown in the chart

(CCCCTCCCACAGTTGGG) corresponds to cases that were linked to original exposure from China, which is consistent with the ISM being in common with one found in Hong Kong.) Figure 6 shows the relative abundance (proportion of total sequences in that country/region) of each of these two dominant subtypes in Italy in other countries/regions. As the plot shows, the outbreak in other European countries have generally involved the same viral subtypes as those which are most abundant 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 [2]. 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 U.S. states and *Diamond Princess*). This 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, with the important caveat that the relatively low number of sequences available from Italy means that potential subtypes may not have been detected there. Indeed, the dominant subtype in New York (TTTCGTCCACGTGTGGG) was detected in one sequence in Iceland linked to exposure in Italy. However, that same subtype was found in Iceland in sequences from 27 cases linked to exposure in Austria, 2 from Denmark, and 1 from Germany. This further suggests that it was unlikely that the incidence of the TTTCGTCCACGTGTGGG subtype in New York is connected to Italy rather than elsewhere in Europe, but limited sequence coverage prevents more definitive inference.

The dominant subtypes in Italy are not found at all in locations in Asia, however, such as Mainland China and Singapore, as indicated in Figure 6. An additional observation to note from 6 is that the most abundant subtype in Italy (TCGTCCACGGGTGGG) is found proportionately much more in France and Spain than the second-most abundant subtype (TCGTCCACGGGTAAC), detected later. Conversely, the second-most abundant subtype is found more uniformly in other countries. These observations give rise to two potential hypotheses for further study: (1) An outbreak centered in Italy moved earlier to France and Spain, and cases elsewhere are connected to a combination of earlier and later exposures. Or, (2) the second-most abundant subtype in Italy is linked to potential exposure from another country in Europe. Overall, the foregoing 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 examination of the complex behavior of subtypes as infections expand in each country and the potential influence on regional outbreaks by subtypes imported from other regions.

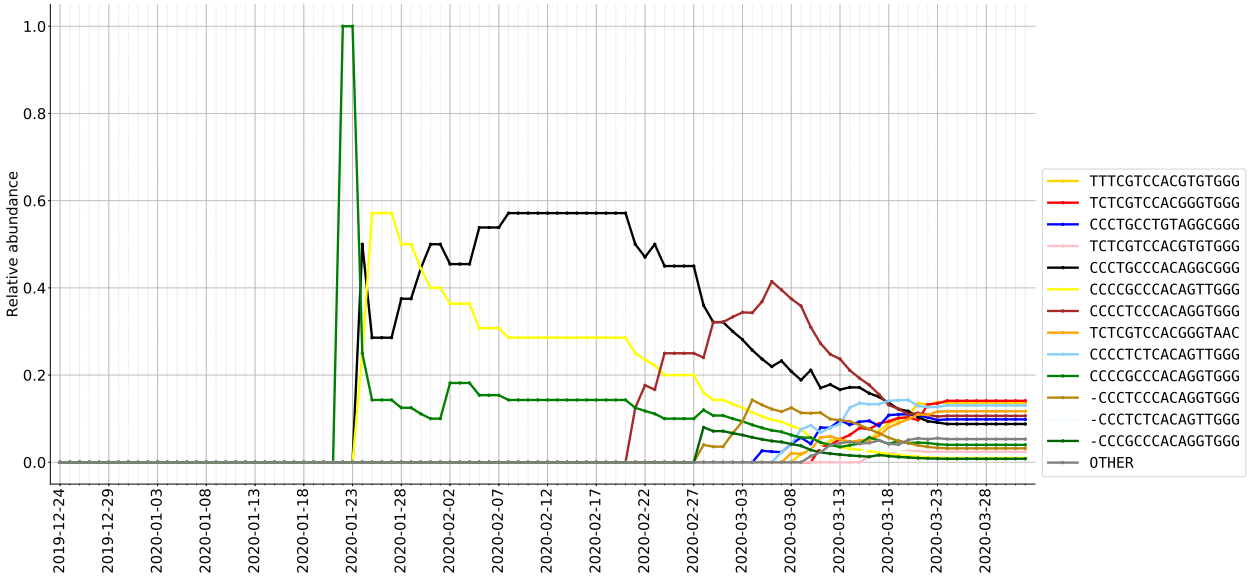

Figure 7. Relative abundance (%) of ISMs in DNA sequences from Australia as sampled over time.

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 and Canada have many different subtypes with similar relative abundance showing a substantial level of geographical diversity in ISMs, and by contrast, the United States and larger European countries have a smaller number of predominant subtypes, which may be due to the relative importance of cases linked to travel exposure and endogenous transmission. We include Mainland China in the analysis, as it was 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/region (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

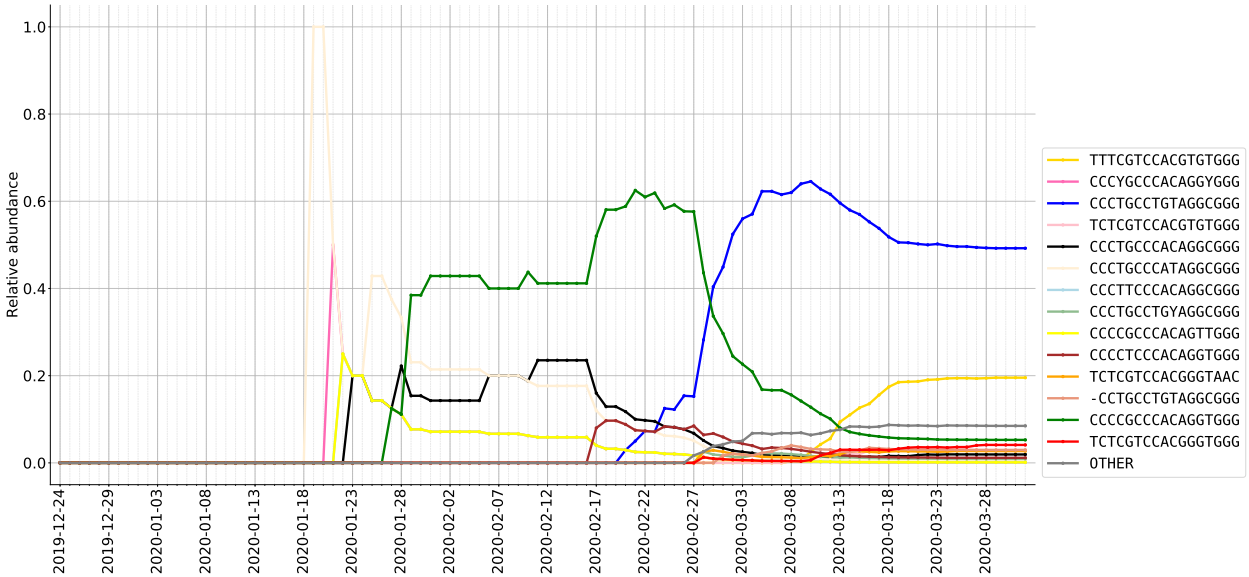

Figure 8. Relative abundance of ISMs in DNA sequences from USA as sampled over time.

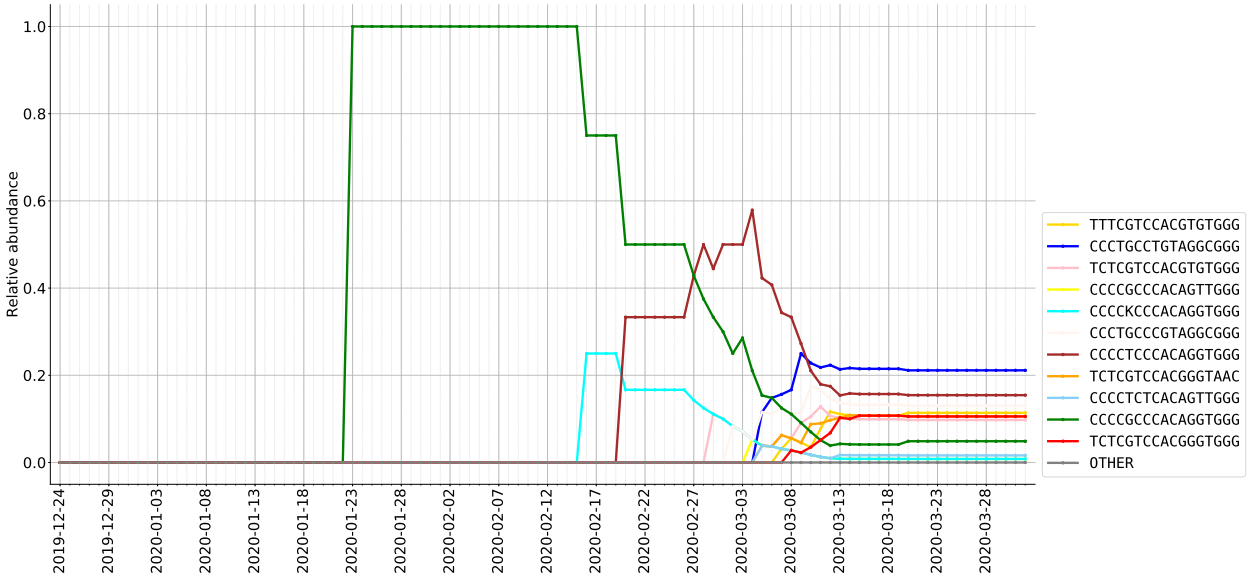

Figure 9. Relative abundance of ISMs in DNA sequences from Canada as sampled over time.

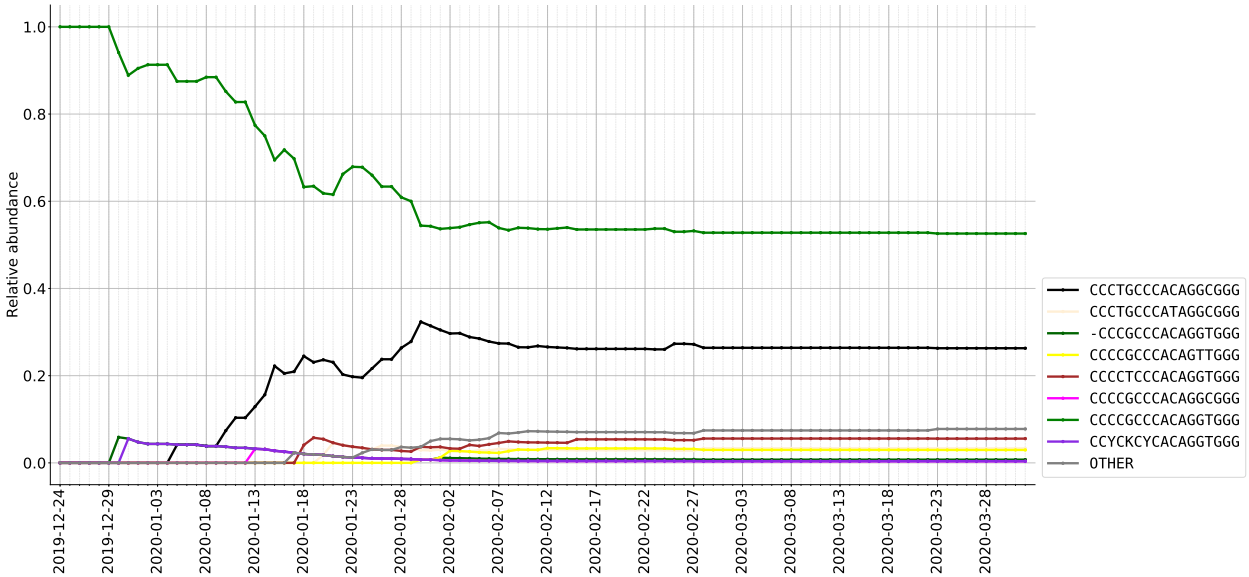

Figure 10. Relative abundance of ISMs in DNA sequences from Mainland China as sampled over time.

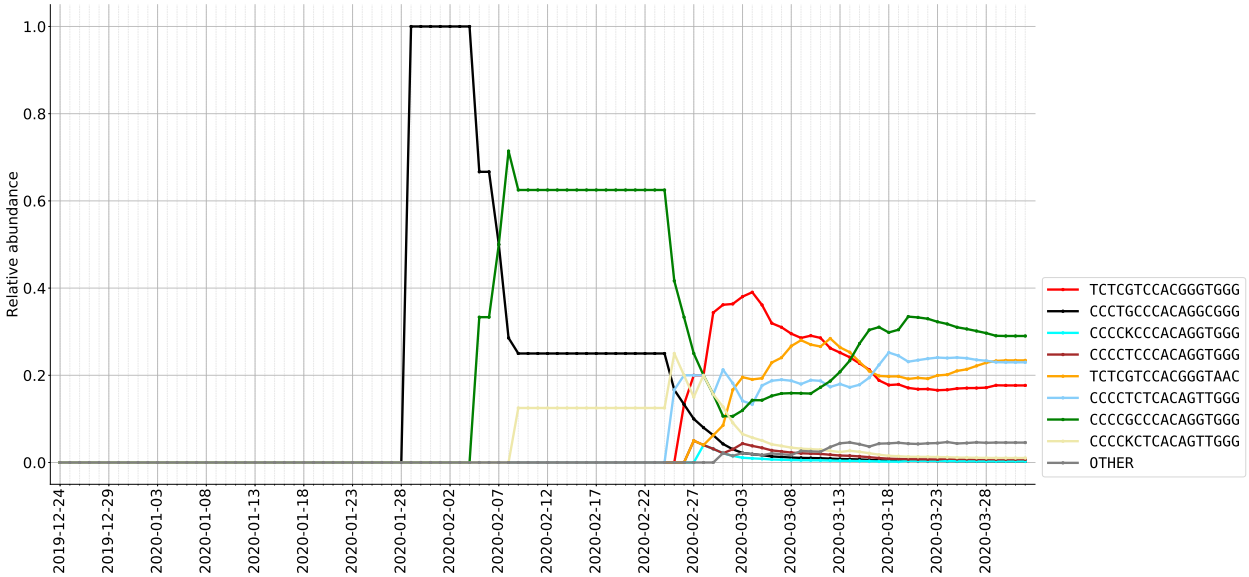

Figure 11. Relative abundance of ISMs in DNA sequences from the United Kingdom as sampled over time.

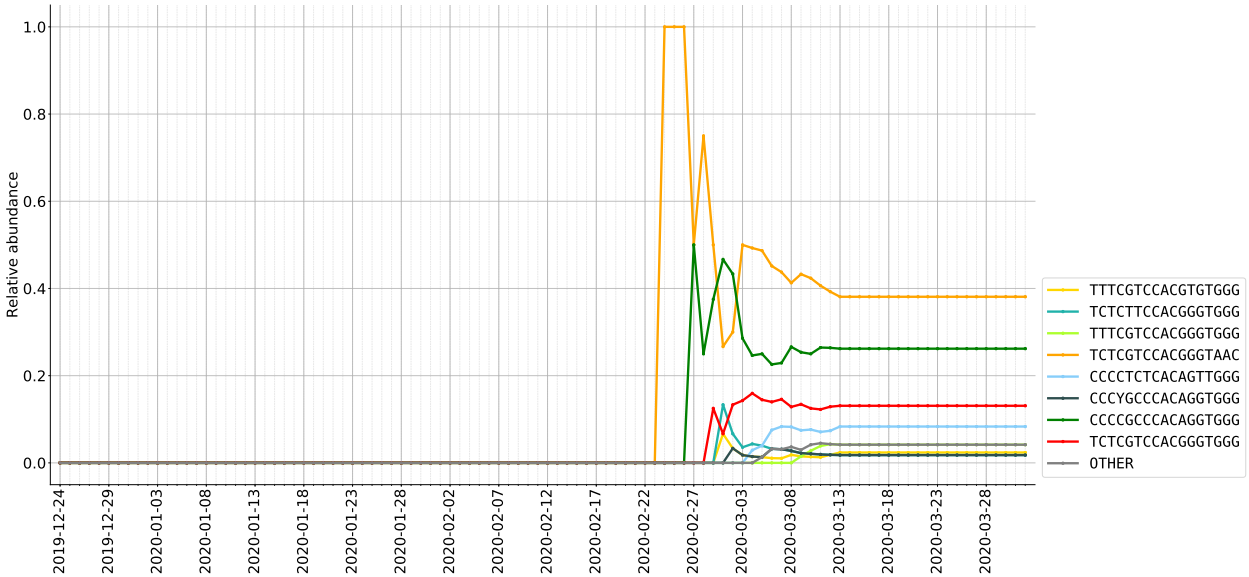

Figure 12. Relative abundance of ISMs in DNA sequences from the Netherlands as sampled over time.

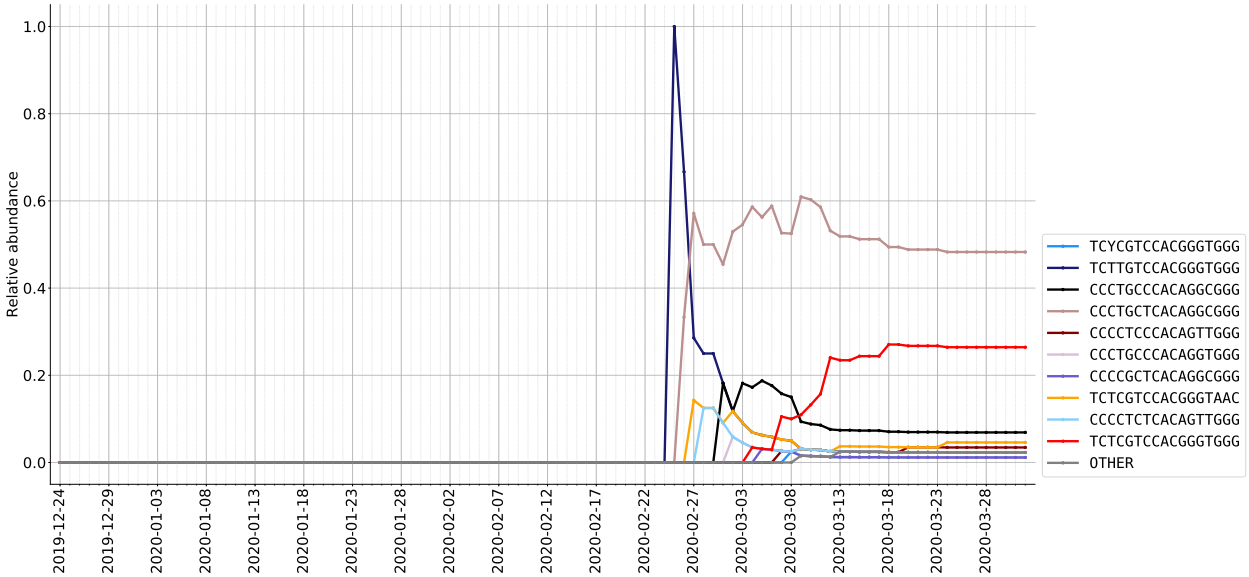

Figure 13. Relative abundance of ISMs in DNA sequences from Spain as sampled over time.

(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 (`CCCCGCCCACAGGTGGG` and `CCCTGCCCACAGGCGGG`), 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 (and linked to Iran as well) was found in Australia (`CCCCTCCCACAGGTGGG`). 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 — or, as noted in the previous discussion, potentially through transmissions from Iran — 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, which was also found in travel exposure from Iran (`CCCCGCCCACAGTTGGG`). And, like in Australia, in Canada these few initial viral sequences were 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 the `CCCTGCCTGTAGGCGGG` 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 `CCCTGCCTGTAGGCGGG` 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, New York is the state with the most significant outbreak of the virus as measured by positive tests, as well as COVID-19 hospitalizations and deaths [31]. As discussed above, the predominant subtype in New York (`TTTCGTCCACGTGTGGG`) is in fact the same as a major subtype found in European countries. such as France, suggesting a link between the New York outbreak and Europe, to a link to the putative endogenous

subtype in the United States (CCCCGCCCACAGTTGGG). The plot in Figure 8 further shows that the putative endogenous U.S. subtype was expanding prior to the first detection of the New York major subtype, further supporting the theory that New York's outbreak is not linked to the dominant subtype 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, i.e., CCCTGCCCACAGGCGGG amd CCCCGCCCACAGGTGGG. 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, TCGTCCACGGGTAAC and TCGTCCACGGGTGGG, 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 CCCCTCTCACAGTTGGG subtype was also 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

Our pipeline also includes 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 [35]. 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. A critical point to keep in mind while interpreting these time series data is that they are based on the reported date of sequences. The sequences of individuals who test SARS-CoV-2 positive will likely lag the actual infection date, since many of those individuals will be tested because they have symptoms – although some were almost certainly tested while presymptomatic due to contacts, such as passengers on the *Diamond Princess* cruise ship.

Figure 14 shows, the reference genome subtype began to grow in abundance in Mainland China, before

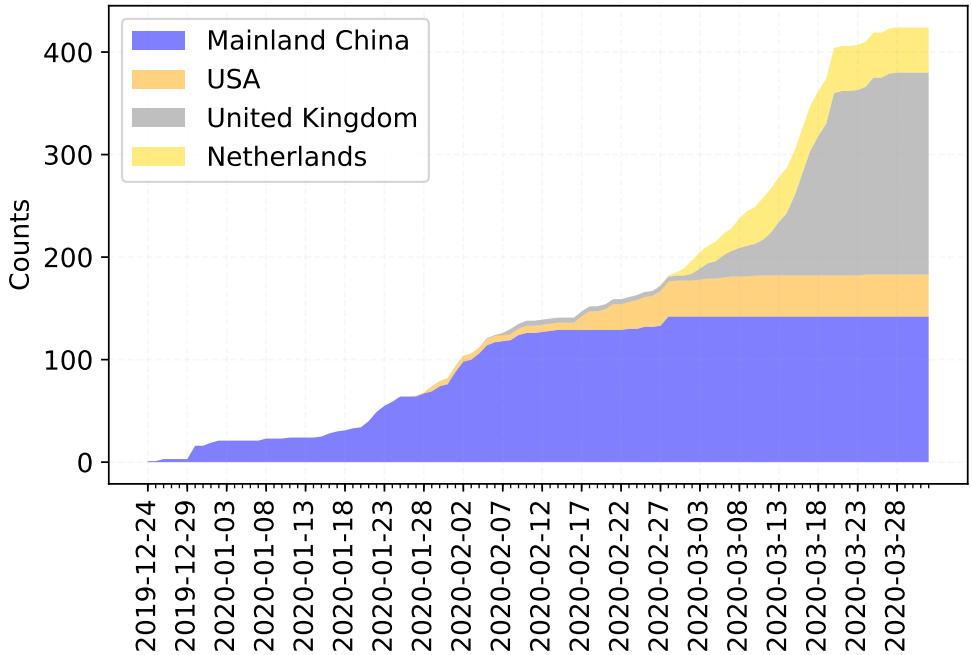

Figure 14. Stacked plot of the number of sequences of the reference sequence ISM subtype (`CCCCGCCCACAGGTGGG`).

leveling 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 (`CCCCGCCCACAGGTGGG`) 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 [20]. 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 leveling off in a region is that the leveling off in relative abundance of the subtype represents containment of that subtype's transmission while other subtypes continue to expand in that country or region. The latter could plausibly explain the pattern observed in the United States, where earlier subtypes connected to Asia did not increase in abundance while a putative endogenous subtype, as well as the dominant New York subtype, have significantly increased in abundance (see Figure 8 and accompanying discussion above). 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.

All the inferences from the temporal trends in subtypes described in the foregoing, however, 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 the United States. While the presence or absence of this subtype in sequences obtained from travelers may shed light on this question, it cannot be definitive because so few cases due to travel have been sequenced. 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.

Our approach, like all sequence-based interpretation of the COVID-19 pandemic, is further 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. Italy, on the other hand, had a large and early outbreak but has disproportionately less sequencing coverage. 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. Some ISMs also include gaps and ambiguity, which indicate the presence of noise in some sequence data. For example, subtype `TCTCGTCCACGGGNNNN` could in fact be subtype `TCTCGTCCACGGGTAAC` and subtype `TCTCGTCCACGGGTGGG`. Although we have implemented our "error" correction algorithm to improve the precision of ISM definition by accounting for technical sequencing errors and ambiguous base calls, there are still bases that can not be fully corrected. 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.

Increasing ISM subtype diversity over time and hierarchical clustering of related ISM subtypes

While keeping in mind the foregoing limitations on the data and our analysis, we do observe that tracking ISM based on the reported date of sequences does reflect the dynamics of the progress of SARS-CoV-2. This is evinced by the results discussed above, which show that the spatiotemporal trends of ISMs are consistent with our general understanding of how the virus has spread, as well as sequence data from travel cases.

Moreover, we observe that the diversity of ISMs has increased over time in a manner. We expect this to occur, because as the number of transmissions of the virus between people increases over time, the probability of changes to the sequence that will lead to a different ISM will increase.

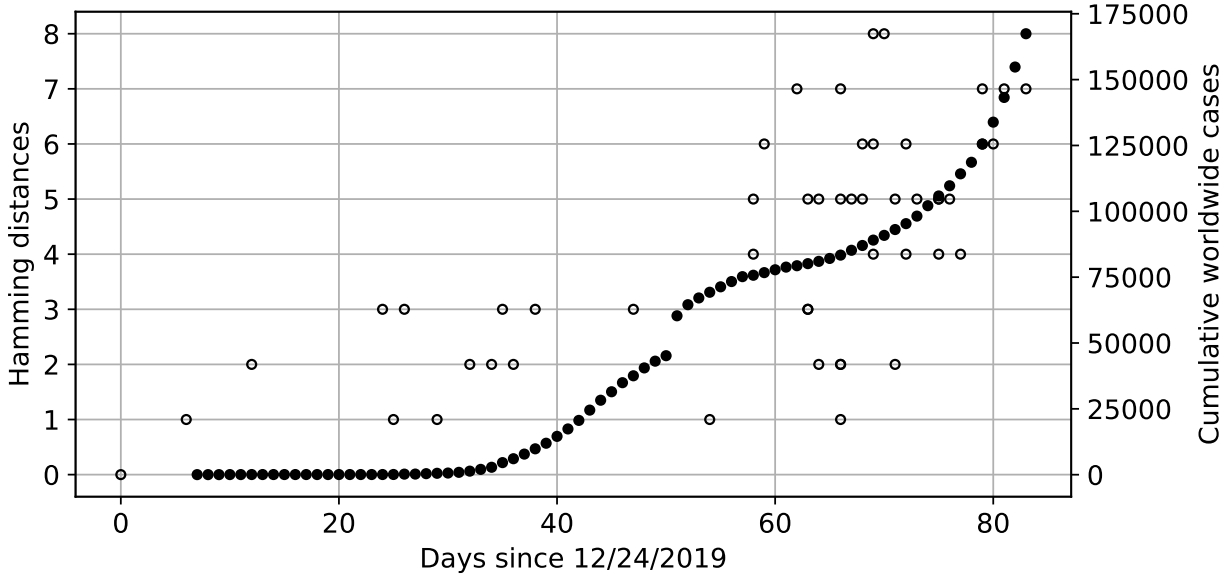

Figure 15. Plots of the Hamming Distance of sequences obtained after the date of the reference sequence (dated December 24, 2019) (open circles), and the cumulative reported number of worldwide COVID-19 cases over days after December 24, 2019 (filled circles).

To evaluate the increase in ISM diversity, we computed the Hamming distance between each of the 50 most prevalent ISMs and the ISM of the reference sequence [13]. These distances, although naïve as they are based on an equally probably nucleotide changes, provide a rough measure of how much each ISM has changed relative to the reference. We then plotted the distance of each ISM along the dates of the first sequence in which the ISM was first detected. As Figure 15 illustrates, even though there is significant variability in this rough measure of the difference between ISMs, there is a clearly accelerating trend towards greater differences over time based on the sequence dates. This provides additional confidence in the robustness of subtyping by ISM according to sequence date. By way of further illustrating this point, Figure 15 superimposes the cumulative reported number of confirmed COVID-19 cases over time.[6] While these data provide only a limted picture of the actual number of SARS-CoV-2 infections (including because of limited testing), it further demonstrates that ISM subtypes following sequence dates are capable of tracking the actual temporal progress of the virus.

As the growing diversity shown in Figure 15 illustrates, the number of ISMs is increasing, thus motivating

[6]The cumulative reported Worldwide cases were obtained from the Our World In Data Project, which reported them from the European Center for Disease Prevention and Control, made available at https://ourworldindata.org/coronavirus-source-data.

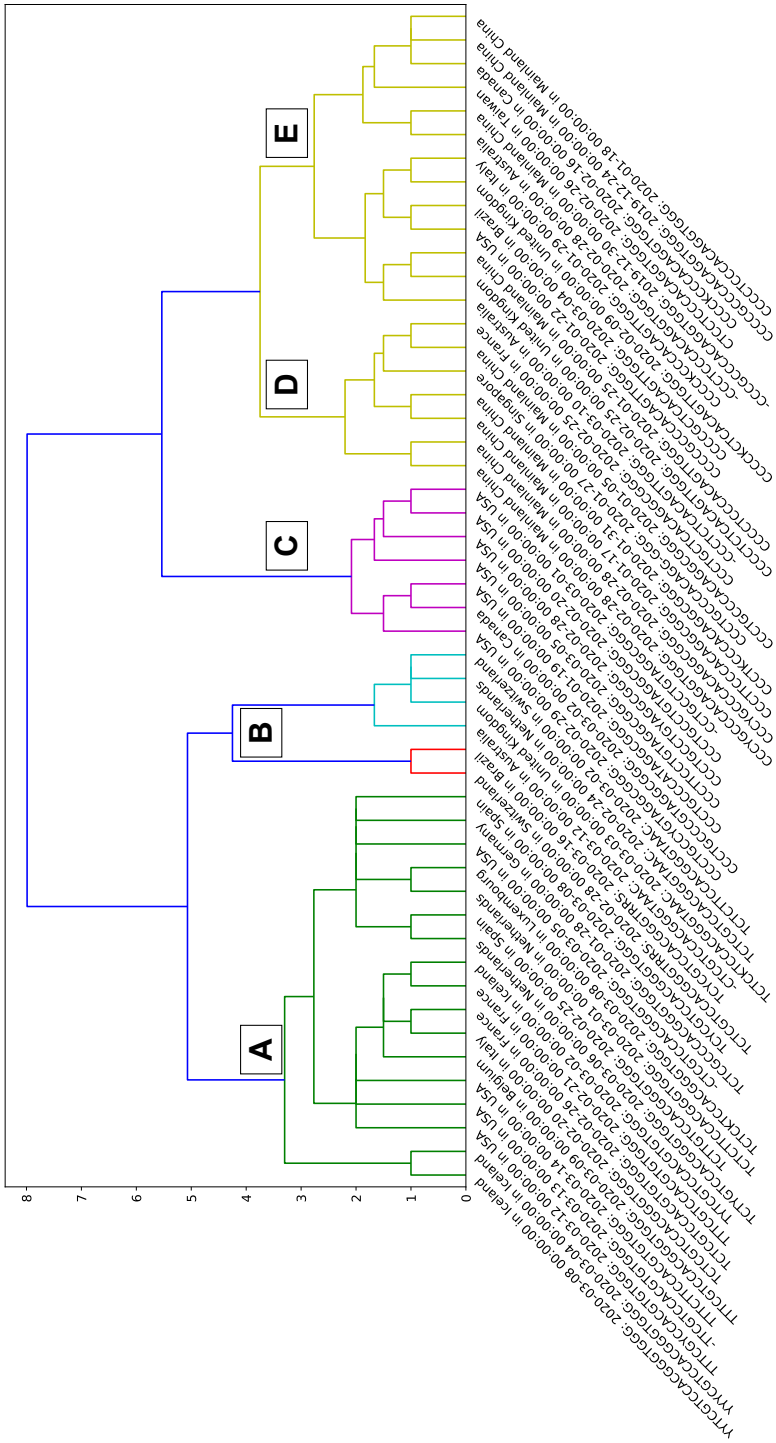

Figure 16. Hierarchical clustering of 50 most frequently-observed ISM subtypes

the development of methods to group and organize ISMs into different levels of subtyping. As an initial illustration of this approach, the tree shown in Figure 16 illustrates relationships between geographical and temporal subtypes, which suggests that subtyping can be done using "consensus ISMs" that include common subunits of ISMs beneath them in the hierarchy.

For example, the subtree colored dark-yellow (labeled D and E in Figure 16 contains the subtypes that are close to the reference sequence ISM subtype (`CCCCGCCCACAGGTGGG`). Other subtypes in this subtree are also found in Asia, as well as in other countries with links to travel exposure to Asia and, as discussed above, Iran (i.e. with the subtype `CCCCTCCCACAGGTGGG`). The closest subtree to the dark-yellow (labeled D and E) is the purple-colored subtree (labeled C). This cluster "C" includes many subtypes that first were observed in the United States, including the endogenous U.S. subtype described above, `CCCTGCCTGTAGGCGGG`. (This subtype is labeled as being first observed in Canada, though that first sequence was found from a case linked to U.S. exposure.) Clusters labeled A (green) and B (cyan and red), on the other hand, primarily include subtypes first found in Europe or (as indicated for Brazil) due to exposure from travelers to Europe.

Interestingly, cluster C (purple) with putative endogenous U.S. subtypes is closer to the cluster with predominantly cases from Asia than cluster A (colored green), which ontains subtypes first seen in Europe, including the subtype that is dominant in New York state (`TTTCGTCCACGTGTGGG`). Overall, ISM clustering further supports the hypothesis that the emergent putative endogenous U.S. subtypes is linked to the original U.S. outbreaks linked to Asia, and that subtypes in the New York outbreak are conversely linked more directly to Europe. We are currently following up on the proof-of-concept clustering results shown here to develop more sophisticated machine learning approaches to identify patterns in viral subtype distribution.

Conclusions

In this paper, we propose to use short sets of nucleotides, based on error corrected entropy-based identification of highly informative nucleotide sites in the viral genome, 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 visualize 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 on Github at http://github.com/EESI/ISM.

Overall, the entropy-based, and error corrected, subtyping approach described in this paper represents a potentially efficient way for researchers to gain further insight on the diversity of SARS-CoV-2 sequences and their evolution 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 ISM subtyping pipeline in place and access to continuously updating sequencing data, we are capable of (and are presently) updating subtype identification 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. Moreover, the ISM-based subtyping scheme and the computational pipeline described here for SARS-CoV-2 are directly applicable to other viruses and, therefore, can be utilized for efficient subtyping and real-time tracking of potential viral pandemics that may emerge in the future as well.

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 [27]. We would also like to acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database on which this research is based (a list is detailed in Supplementary file 3 — Acknowledgements of sequences this research is based on) and all future SARS-CoV-2 sequence contributors in GISAID's EpiFlu Database. 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.

Author Contributions

ZZ contributed to the conceptualization of the problem and solution, data curation, methodology, software, validation, visualization, and original draft preparation. GLR contributed to the project administration, conceptualization, methodology development, acquiring resources, validation, visualization, and original draft preparation. BAS contributed to the conceptualization, data curation, methodology, software, validation, and visualization.

References

1. Nomenclature for incompletely specified bases in nucleic acid sequences. recommendations 1984. nomenclature committee of the international union of biochemistry (nc-iub). *Proceedings of the National Academy of Sciences*, 83(1):4–8, 1986.

2. Coronavirus: Outbreak spreads in europe from italy, available at https://www.bbc.com/news/world-europe-51638095, last accessed 2020-04-05. *BBC News*, February 26, 2020.

3. M. V. Batista, T. A. Ferreira, A. C. Freitas, and V. Q. Balbino. An entropy-based approach for the identification of phylogenetically informative genomic regions of papillomavirus. *Infection, Genetics and Evolution*, 11(8):2026 – 2033, 2011.

4. C. Ceraolo and F. M. Giorgi. Genomic variance of the 2019-ncov coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.

5. 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.

6. 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.

7. C. Colijn and J. Gardy. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health*, 2014(1):96–108, 06 2014.

8. 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.

9. 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.

10. 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.

11. 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.

12. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

13. N. C. Jones, P. A. Pevzner, and P. Pevzner. *An introduction to bioinformatics algorithms*. MIT press, 2004.

14. 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.

15. 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.

16. 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.

17. 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.

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

19. 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.

20. 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.

21. D. Müllner. Modern hierarchical, agglomerative clustering algorithms, 2011.

22. 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.

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

24. 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.

25. 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.

26. 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.

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

28. 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.

29. 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.

30. 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.

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

32. 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.

33. 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.

34. 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.

35. 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), 2020.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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 file 2 — ISM abundance table of 5 U.S. states and *Diamond Princess*

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

Supplementary file 3 — Acknowledgements of sequences this research is based on

A list of sequences from GISAID's EpiFlu Database on which this research is based and corresponding authors and laboratories.