Nonlinear Dyn (2021) 106:1525–1555
https://doi.org/10.1007/s11071-021-06836-y

ORIGINAL PAPER

Advances in the computational analysis of SARS-COV2 genome

J. A. Tenreiro Machado · J. M. Rocha-Neves · Filipe Azevedo · J. P. Andrade

Received: 30 December 2020 / Accepted: 15 August 2021 / Published online: 27 August 2021

Abstract Given a data-set of Ribonucleic acid (RNA) sequences we can infer the phylogenetics of the samples and tackle the information for scientific purposes. Based on current data and knowledge, the SARS-CoV-2 seemingly mutates much more slowly than the influenza virus that causes seasonal flu. However, very recent evolution poses some doubts about such conjecture and shadows the out-coming light of people vaccination. This paper adopts mathematical and computational tools for handling the challenge of analyzing the data-set of different clades of the severe acute respiratory syndrome virus-2 (SARS-CoV-2). On one hand, based on the mathematical paraphernalia of tools, the concept of distance associated with the Kolmogorov complexity and Shannon information theories, as well as with the Hamming scheme, are considered. On the other, advanced data processing computational techniques, such as, data compression, clustering and visualization, are borrowed for tackling the problem. The results of the synergistic approach reveal the complex time dynamics of the evolutionary process and may help to clarify future directions of the SARS-CoV-2 evolution.

Keywords COVID-19 · Kolmogorov complexity theory · Shannon information theory · Multidimensional scaling · Evolutionary dynamics

J. A. Tenreiro Machado (✉)· F. Azevedo
Department of Electrical Engineering, Institute of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249 – 015 Porto, Portugal
e-mail: jtm@isep.ipp.pt

F. Azevedo
e-mail: fta@isep.ipp.pt

J. M. Rocha-Neves
Department of Biomedicine – Unity of Anatomy, and Department of Physiology and Surgery, Faculty of Medicine of University of Porto, Porto, Portugal
e-mail: joaorochaneves@hotmail.com

J. P. Andrade
Department of Biomedicine – Unity of Anatomy, Faculty of Medicine of University of Porto and Center for Health Technology and Services Research (CINTESIS), Porto, Portugal
e-mail: jandrade@med.up.pt

1 Introduction

The severe acute respiratory syndrome virus (SARS-CoV-2) is a single-stranded Ribonucleic Acid (RNA) beta-coronavirus presenting a 29,903 nucleotides-long genome. It caused an outbreak in the Chinese city of Wuhan, in December 2019. Subsequently, the new coronavirus has spread worldwide in less than 4 months, and a pandemic situation was declared by the World Health Organization (WHO). The disease was named on February 2020, Coronavirus Disease 2019 (COVID-19), and the disease is now responsible for more than 145,000,000 confirmed cases, and 4,230,000 deaths reported worldwide as of August 3, 2021 (source: World Health Organization https://covid19.who.int). There are significant differences in

case fatality rates (proportion of deaths from a specific disease compared to the total number of individuals diagnosed with the disease for a particular period) between countries, possibly related to the efficacy of the measures adopted to limit viral spreading, the demographic pyramid, i.e., the distribution of various age groups in a particular country, and the screening, test and tracing strategy [1].

Using the data from the database of the Global Initiative on Sharing Avian Influenza Data (GISAID) Consortium, major clades of SARS-CoV-2 were identified [2]. The initial RNA genome (accession number: NC 045512.2), identified in Wuhan is named as clade 'L'. This root clade suffered mutations due to numerous factors, and some new clades emerged with stable mutations: clade 'G' (presenting an alteration of the spike protein S-D614G was the first dominant variant) and its two derivatives 'GH' (with ORF3a-Q57H mutation) and 'GR' (affected by RG203KR mutation); clade 'V' (variant of the ORF3a coding protein NS3-G251), and clade 'S' (variant ORF8-L84S). Other alleles or combinations different from the previously described clades are classified as clade 'O' [2]. These data confirmed, until now, the relatively low variability presented by SARS-CoV-2 compared to other respiratory viruses [3]. The different fatality rates, speed of transmission, and infectiousness profiles observed in different countries were probably not related to differences in virulence of the clades and their characteristic mutations. However, very recently, some new mutations were found with potential epidemiological consequences. For example, 12 human cases were identified in September 2020 in North Jutland with a unique variant called 'cluster 5', a combination of mutations that have been not previously described. All 12 cases were linked to the mink farming industry or the local community [4]. However, the clinical presentation, severity, and duration of COVID-19, and transmission among those infected were similar to that of other circulating SARS-CoV-2 viruses. The variant cluster 5 has not been detected since September despite extensive sequencing and data sharing and is thought to be a dead-end, owing to the very restricted spreading and infectiousness to humans [4].

More disturbing is the new variant strain of SARS-CoV-2 that contains 23 mutations, eight of which are in the spike protein the virus uses to bind to and enter human cells. The spike protein is also the focus of most COVID-19 vaccines that are now being administered in numerous countries. Moreover, the diagnostics tests of COVID-19 are also based on the protein sequence found on the Wuhan reference strain spike. Therefore, their efficacy can be changed by these genomic variations. The appearance of these mutations can also lead to immunological resistance and vaccine escape [5]. The new British variant was reported for the first time in December 2020 and has become highly prevalent globally and responsible for another COVID-19 wave in numerous countries [6]. Based on these mutations, this variant has been predicted to potentially be more quickly transmissible than other circulating strains of SARS-CoV-2. The variant is referred to as SARS-CoV-2 VUI 202012/01 (i.e., Variant Under Investigation, year 2020, month 12, variant 01), or B.1.1.7, as the lineage of the clade GR was classified on GISAID. Variant B.1.1.7 presents increased transmissibility and increased virus load. However, apparently, there is no association with more severe disease [7] although the increased infectiousness can lead to more deaths due to the strain of the health systems of the affected countries. In Mid-November in South Africa, a new lineage of the clade GH has emerged but shared one of the mutations described in the British variant. The South African virus variant (B.1.351), known as 'triple variant' is distinct from the UK variant, but both contain an unusually high number of mutations, with potential functional significance, compared to other SARS-CoV-2 lineages, and it can, apparently, partially escape to some available vaccines [8]. Several other variants were described more recently in several countries. Examples are the variants P1 and P2 in Brazil, variants B.1.429 and B.1.526 found in the USA, and yet more recently, a variant first sequenced in India (variant B.1.167) [9]. In April 2021, the South Africa, Brazil, and India variants caused or are causing large disease outbreaks in their respective countries with increased excess deaths due to rupture of the hospital capacities.

The SARS-CoV-2 B.1.1.7 variant was already detected at the end of December 2020 in France, Denmark, Holland, and Italy [10]. In 31 of May, the WHO has assigned simple, easy to say and remember labels for the most important variants of SARS-CoV-2, using letters of the Greek alphabet (see Tracking SARS-CoV-2 variants at https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/#:~:text=WHO%20and%20its%20international%20networks,variant%2c%20and%20prevent%20its%20spread).

The most important variants of concern according to this new classification are: Alpha, corresponding to lineage B.1.1.7, found in the United Kingdom in September 2020; Beta, corresponding to lineages B.1.351, B.1.351.2, B.1.351.3 present in South Africa in May 2020; Gamma, corresponding to lineages P.1, P.1.1, P.1.2, sequenced in Brazil, in November 2020; and, Delta, corresponding to lineages B.1.617.2, AY.1, AY.2 and AY.3 found in India in October 2020 (see Table 1).

In the Summer of 2021, the SARS-CoV-2 Delta variant is becoming dominant in Europe, North America and other parts of the world and is responsible for a new wave, mainly in the unvaccinated population [11]. It is highly transmissible and it also appears to affect vaccine effectiveness and breakthrough infections in vaccinated individuals appear to be more frequent with this variant [11]. It was reported that the viral load is 1000 times higher for Delta compared with previous variants of the initial wave of infections and may have a faster replication rate, a reduced incubation period, and greater viral shedding [12]. The Delta variant was found to be approximately 64% more transmissible than the Alpha variant that was dominant in the waves of the end of December 2020 and first months of 2021. On the other hand, Alpha was already estimated to be 50% more transmissible than the D614G strain, responsible for the first wave in the beginning of 2020 [13].

The so-called coronavirus 'waves' can be more contagious and spread faster than the initial ones due to the new variants that can also present other epidemiological problems. There is no definitive evidences that the various variants are associated with an higher disease severity. However, there is a clear risk that future epidemic 'waves' may be larger and, therefore, associated with greater burden for the health systems and society due to the lockdowns. Therefore, with this work we try to understand SARS-CoV-2 new variants and the relations with the already known virus clades using new strategies for finding the relationships among them [14–16]. It is important to understand the recent evolution of the virus partially responsible for the so-called 'waves' [17]. With this regard computational techniques associated with mathematical tools are a promising strategy to tackle genomic data-sets. This approach was tested in [18,19] using the Kolmogorov complexity and Shannon information theories, associated with clustering techniques [20,21] such as the Multidimensional Scaling (MDS). Given its successful application in a primary set of 133 items, encompassing a variety of virus, this paper extends the study to a more challenging problem, namely of analyzing and comparing the SARS-CoV-2 mutations. For that purpose a new data-set of 307 virus including RNA information from the beginning of the spread up to the day of writing this paper is selected. Furthermore, based on the aforementioned mathematical tools, we include a larger set of indices to allow a more complete comparison. In the case of the Kolmogorov complexity we consider four indices [22,23], normalized information distance, Compression-based Dissimilarity Measure, Chen-Li Metric and Compression-based Cosine (that are abbreviated by the acronyms NCD, CDM, CLM and $CosS$). In the scope of the Shannon information theory we consider also four metrics, namely the Jaccard, Jensen-Shannon, Jeffreys and Topsøe distances (denoted as d_{Ja}, d_{JS}, d_{Je} and d_{To}) [24,25]. A third type and distinct type of assessment is also included and consists of the Hamming distance (d_{Ha}), widely used in information theory [26].

Following these ideas the rest of the paper is organized as follows. Section 2 introduces the fundamental tools. The main mathematical concepts involved with distance, Kolmogorov complexity, Shannon information and Hamming metric are summarized. Additionally, the computational tools such as data compression and MDS are also included. Section 3 describes and analyses the data-set of very close, but distinct, information for a large number of RNA sequences. Section 4 develops a synergistic performance of a variety of measures associated with the MDS clustering and visualization. The results shed light on the dynamics of the evolutionary process of the SARS-CoV-2 lineages. Finally, Sect. 5 presents the conclusions.

2 Fundamental tools

2.1 Distance

A function $d(\cdot,\cdot)$ stands for the distance between two objects x and y if satisfies three axioms [27], namely *identity* $d(x,y) = 0$ if $x = y$, *symmetry* $d(x,y) = d(y,x)$ and *triangle inequality* $d(x,y) \leq d(x,z) + d(y,z)$. These axioms imply the non-negativity (or separation condition) $d(x,y) \geq 0$. On the other hand, they allow the definition of a plethora of different functions, with distinct pros and cons [24,25]. Based on these notions, several algorithms [28–30] were adopted for

Table 1 Currently designated SARS-CoV-2 variants of concern

WHO label	Lineages	Clade	Date of designation
Alpha	B.1.1.7	GRY	18-Dec-2020
Beta	B.1.351, B.1.351.2, B.1.351.3	GH/501Y.V2	18-Dec-2020
Gamma	P.1, P.1.1, P.1.2	GR/501Y.V3	11-Jan-2021
Delta	B.1.617.2, AY.1, AY.2, AY.3	G/478K.V1	11-May-2021

comparing data sequences [26,31–33]. However, users must have in mind that the selection of a set of distances for a given application requires some experience and that a number of numerical trials are usually necessary before finding the 'best' ones [34–37].

In the final part of the paper, the Appendix presents the mathematical and algorithmic fundamentals of the distances used in this paper for assessing the genetic information.

2.2 Data compression

Compression data algorithms can be classified as 'lossless' and 'lossy'. Lossless compression algorithms are typically used for archival or other high fidelity purposes and reduce the size of files without losing any information in the file, which means that we can reconstruct the original data from the compressed file. Lossy compression algorithms reduce the size of files by discarding the less important information in a file, which can significantly reduce file size but also affect file quality.

In this paper we used the `BZip2` compression algorithm which is based on Burrows-Wheeler transform [38]. This compressor has the extension BZ2 designating a pure data compression format not providing file archival feature. In this algorithm the speed is somewhat slower than for the compressor `LZW` (extension `.Z`) and `Deflate` (extension `.zip` and `.gz`) compression algorithms [39]. These employ the classic Deflate algorithm (even if correctly implemented `Bzip2` algorithm can be easily made parallel, and benefit of recent multi-core CPU), but faster than more powerful compression schemes as in `RAR` format, `7Z` format, and new `ZIPX` format. The compression ratio, also, is usually intermediate between the older Deflate-based `ZIP/GZ` files and modern `RAR`, `7Z` and `ZIPX` formats [40].

2.3 Multidimensional scaling

Let us consider a group of N objects $x_i, i = 1, \cdots, N$, in a q-dim space. The MDS is a computational method for [41] that re-organizes them in a structure where the objects are represented by points trying to highlight the similarities between them in the sense of a predefined distance [42]. The process starts by calculating a $N \times N$ dimensional matrix, $D = [d_{ij}]$, with $d_{ij} \in \mathbb{R}^+$ for $i \neq j$ and $d_{ii} = 0$, $(i, j) = 1, \cdots, N$, giving object to object distances [43]. In a second phase, the MDS calculates the point coordinates $\hat{x}_i$ in a $d < q$-dim space, trying to mimic the original distances. The MDS technique includes a numerical iterations for optimizing a cost function, often called $Stress$, that compares the distances $d_{ij} = |x_i - x_j|$ and $\hat{d}_{ij} = |\hat{x}_i - \hat{x}_j|$, so that the index $Stress = \sqrt{\sum_{i<j} \left(\hat{d}_{ij} - d_{ij}\right)^2}$ is minimized.

The MDS points $\hat{x}_i$ have coordinates that yield a symmetric matrix $\hat{D} = [\hat{d}_{ij}]$ of distances that approximate D. The MDS results are interpreted based on the clusters, and eventually of patterns, of points [18,44]. Therefore, similar objects are represented by nearby points, and the opposite for dissimilar objects. Different distances produce distinct MDS maps and it is up to the user to choose the metrics that reflect better the characteristics of the objects under analysis. By other words, the different distances are correct from the mathematical point of view, but the association of each metrics with the MDS algorithm may produce disparate patterns in the plots. In some cases the emerging patterns, although different, lead to similar conclusions. In other cases, some distances reflect better (or worst) the information embedded in the dataset and the selection of the 'best' metric depends on a trial and error set of tests based on the user experience.

3 Dataset description

The RNA is commonly sequenced indirectly by copying it into the complementary DNA (cDNA). Then the cDNA is amplified and analyzed using a number of DNA sequencing methods. The sequences of the RNA are published in the databases presenting the bases, adenine (A), cytosine (C), guanine (G), and thymine (T). Some symbols, such as N (unspecified or unknown nucleoside), R, (unspecified purine nucleoside), Y (unspecified pyrimidine nucleoside) and others, permeate only a small percentage of the information and are not considered [45]. The information about the $N = 307$ GS was collected in the Global Initiative on Sharing Avian Influenza Data (GISAID) available at https://www.gisaid.org/. The information regarding the sequences, serial, clade/variant and country are listed in the Tables 5, 6, 7, 8 and 9. The genetic information is organized in 8 clades {GH, GR, O, GV, G, L, S, V} with 10 elements each, making a total of 80 cases. The recent advent of new variants correspond to the remaining 227 additional items as follows:

- South Africa, with 10 cases for the variant 'South Africa Triple Variant', denoted as TV-ZA
- Denmark, with 10 cases for the variant 'Mink Cluster V', denoted as CL5-DK
- England, with 10 cases for the variant VUI2020/01, denoted as VUI-GB
- Italy, with 10 cases for the variant VUI2020/01, denoted as VUI-IT
- Denmark, with 10 cases for the variant VUI2020/01, denoted as VUI-DK
- Portugal, with 10 cases for the variant VUI2020/01, denoted as VUI-PT
- USA, with 10 cases for the variant VUI2020/01, denoted as VUI-US
- a mixture of several cases scattered along the world to give an higher spatial diversity
  - Ireland, with 7 cases for the variant VUI2020/01, denoted as VUI-IE
  - Japan, with 6 cases for the variant VUI2020/01, denoted as VUI-JP
  - Australia, with 5 cases for the variant VUI2020/01, denoted as VUI-AU
  - Singapore, with 5 cases for the variant VUI 2020/01, denoted as VUI-SG
  - Israel, with 4 cases for the variant VUI2020/01, denoted as VUI-IL
  - South Korea, with 3 cases for the variant VUI2020/01, denoted as VUI-KR
  - Norway, with 3 cases for the variant VUI2020/01, denoted as VUI-NO
  - France, with 2 cases for the variant VUI2020/01, denoted as VUI-FR
  - Germany, with 2 cases for the variant VUI2020/01, denoted as VUI-DE
  - Spain, Gibraltar, Hong Kong, India, Luxembourg, Switzerland and Sweden, with 1 case each, for the variant VUI2020/01, denoted simply as VUI
- Japan with 10 cases for the variant B.1.1.28, denoted as B.1.1.28-JP
- Brazil with 10 cases for the variant B.1.1.28, denoted as B.1.1.28-BR
- Brazil with 10 cases for the variant P.1.1.28, denoted as P.1-BR
- Brazil with 10 cases for the variant P.2, denoted as P.2-BR
- South Africa with 10 cases for the variant B.1.351, denoted as B.1.351-ZA
- USA with 10 cases for the variant B.1.427, denoted as B.1.427-US
- California, USA, with 10 cases for the variant B.1.429, denoted as B.1.429 -US
- New York, USA, with 10 cases for the variant B.1.526, denoted as B.1.526-US
- India with 10 cases for the first variant VUI B.1.617, denoted as VUI B.1.617.1-IN
- India with 10 cases for the second variant VUI B.1.617, denoted as VUI B.1.617.2-IN
- India with 10 cases for the third variant VUI B.1.617, denoted as VUI B.1.617.3-IN.

In synthesis, we have collected a first set of 80 GS of the SARS-CoV-2 virus obtained in several countries during a first period of the outbreak. The second set includes 227 recent GS. The smaller number of cases the recent genomic data for some countries is limited to the data set available at the time of writing this paper. All ASCII files have approximately 30 kBytes.

The $N = 307$ GS exhibit very small differences and, therefore, are difficult to distinguish. We can first characterize them by their length L that varies between minimum and maximum values of $L_{min} = 28560$ and $L_{max} = 29900$ symbols. Moreover, we have an average and standard deviation of $L_{av} = 29773.5$ and $L_{sd} = 109.31$ symbols, respectively. This small

variability in the size is relevant for the reliability when comparing strings with the Kolmogorov- and Hamming-based metrics.

As mentioned before, the viral RNA information is represented by ASCII files with the four nitrogenous bases. Therefore, we can consider the grouping of $k_s = \{1, 2, 3\}$, consecutive symbols. For simplicity, we denote the corresponding sub-strings by $\left\{S^1, S^2, S^3\right\}$ and obtain the statistics listed in Tables 2, 3 and 4. The term 'others' stands for a small number of other symbols distinct of the four nucleotide bases. Therefore, occasionally we find the symbols N, M, S, R, Y, K, H, V, and W, that abbreviate aNy (A, C, G or T), aMino (A or C), Strong interaction (3 H-bonds, G or C), puRine (A or G), pYrimidine (C or T), Keto (G or T), not-G (A, C or T), not-T (A, C or G), and Weak interaction (2 H-bonds, A or T), respectively.

Figure 1 shows the cumulative numbers of cases for sub-strings including $k_s = \{1, 2, 3\}$ symbols in the $N = 307$ GS.

As a complementary analysis of the relationship between GS we use the VOSviewer software tool [46–50] for constructing and visualizing bibliometric networks (https://www.vosviewer.com/). The program was built having in mind scientometric applications, but can be used in the present case if we consider the associations of symbols as keywords in a standard technical text. For that purpose we construct k_s-tuples of consecutive symbols in the GS in order to have 'words' (i.e., sub-strings) with k_s consecutive symbols. We considered the VOSviewer options '`Full counting`', '`Minimum number of occurrences of a term = 5`' and '`Number of terms selected = 28`'. For example, Fig. 2 shows the network for the sub-strings with $k_s = 3$ consecutive symbols in the $i = 1$ genomic sequence. For the other virus the results are of the same type. We observe the very small relevance of 'phrases' with several triplets and, on the other hand, the complex network relationship between triplets.

From the Tables 2, 3 and 4 and Figs. 1 and 2 we verify that the case of $k_s = 3$ symbols represent a good compromise between complexity and accurate description of the information content. This conclusion follows previous observations with genetic data [19,51–53].

The existence of the symbols classified as 'other' and the variation of size in the GS can be considered as a kind of noise. However, as verified with the previous tests they reflect in very small numbers. Also, in most cases we considered about 10 GS for each type of virus. Therefore, we can proceed with the analysis of the genetic information knowing its robustness against possible volatility in the data.

The mathematical description of the viral information is based on the Kolmogorov, Shannon and Hamming perspectives implemented by the distances $\{NCD, CLM, CDM, CosS\}, \{d_{Ja}, d_{Js}, d_{Je}, d_{To}\}$ and d_{Ha} presented in the Appendix. The MDS clustering and visualization is used to unravel relationships between the data and to identify possible patterns. In the case of the Kolmogorov complexity we consider the compressor `BZip2` (https://www.zlib.net). In the case of the Shannon information, we start by calculating the 64-bin histograms (i.e., the triplets $\{AAA, AAC, \ldots, GGT, GGG\}$) for the triplets ($k_s = 3$) of the nitrogenous bases and then we calculate the distances between them. The resulting matrix D, 307×307 dimensional, that is processed by MDS using the Matlab command `cmdscale`.

4 Data-set analysis: clustering results

The distances following the Kolmogorov theory $\{NCD, CLM, CDM, CosS\}$ yield almost similar MDS plots. This behavior was also observed in previous studies with distinct data-sets [54]. In what concerns the distances based on the Shannon theory $\{d_{Ja}, d_{Js}, d_{Je}, d_{To}\}$ we note that d_{Ja} produces a slightly different plot from the group $\{d_{Js}, d_{Je}, d_{To}\}$, which return charts having just small differences. On the other hand, the distance d_{Ha} leads to a very different chart. Therefore, for parsimony, for these sets of distances we depict just the plots for the NCD, d_{Jacc}, d_{JS} and d_{Ha}.

The MDS charts produced by the four distances are represented in Fig. 3, 4, 5 and 6, respectively. In all cases the MDS plots require a careful rotation to get the correct 3-dimensional perspective and assessment, since the planar projections in the figures are not totally capable of depicting their structure.

Several MDS loci reveal different clusters, but, in general we do not have a clear group for each variant of the virus. The Kolmogorov- and the Shannon-based metrics show some clusters, but with a mixture of many variants, while the Hamming scheme gives the worst map in the perspective of clustering. Nonetheless, we can ask a different and more relevant question

Table 2 Statistics of 1-tuples of symbols in the $N = 307$ GS

S^1	S^1_{min}	S^1_{max}	S^1_{av}	S^1_{sd}	Symbol	S^1_{min}	S^1_{max}	S^1_{av}	S^1_{sd}
A	7352	8952	8805.5	204.5	T	7854	9609	9476.7	224.9
C	4497	5491	5409.8	126.9	G	4870	5863	5787.5	132.4
					Other	0	5037	293.0	666.6

Table 3 Statistics of 2-tuples of symbols in the $N = 307$ GS

S^2	S^2_{min}	S^2_{max}	S^2_{av}	S^2_{sd}	$S^2$1	S^2_{min}	S^2_{max}	S^2_{av}	S^2_{sd}	S^2	S^2_{min}	S^2_{max}	S^2_{av}	S^2_{sd}	$S^2$1	S^2_{min}	S^2_{max}	S^2_{av}	S^2_{sd}
AA	1207	1469	1405.3	43.8	CA	848	1053	1028.4	27	TA	945	1205	1174.3	32.4	GA	641	836	798.5	28.5
AC	834	1040	998.8	25	CC	353	448	434.1	12	TC	588	763	693.8	44.5	GC	489	606	579	17.9
AT	936	1200	1139.2	39.6	CT	829	1067	1020.3	29.6	TT	1312	1620	1589.3	40	GT	829	1030	989.4	30.3
AG	727	887	853.6	24.4	CG	180	235	219.2	11.2	TG	1073	1322	1278.6	32	GG	457	574	534	18.3
															other	0	2525	148.9	336.3

Table 4 Statistics of 3-tuples of symbols in the $N = 307$ GS

S^3	S^3_{min}	S^3_{max}	S^3_{av}	S^3_{sd}	S^3	S^3_{min}	S^3_{max}	S^3_{av}	S^3_{sd}	S^3	S^3_{min}	S^3_{max}	S^3_{av}	S^3_{sd}	S^3	S^3_{min}	S^3_{max}	S^3_{av}	S^3_{sd}
AAA	242	329	290.8	19.2	CAA	193	246	230.3	10.2	TAA	77	345	271.8	85.7	GAA	85	255	166.2	54.6
AAC	169	216	203.7	9.6	CAC	98	194	156.9	28.4	TAC	154	233	208.5	19.6	GAC	66	143	110.9	27.3
AAT	164	317	242.4	44.8	CAT	116	178	164.3	15.2	TAT	112	282	217	65.5	GAT	50	214	140.2	58.9
AAG	97	248	174.9	62.5	CAG	76	188	130.4	43.8	TAG	76	211	145	45.3	GAG	70	121	92.1	12.5
ACA	214	291	258.1	18.4	CCA	90	139	112.5	12.6	TCA	143	190	178.9	8.1	GCA	82	169	117.7	21.8
ACC	89	151	130.4	18.8	CCC	28	45	39.7	5	TCC	48	88	68.7	9.3	GCC	39	73	60.5	10.3
ACT	142	292	213.3	42	CCT	80	149	108.3	19.4	TCT	117	219	173.8	30.8	GCT	75	266	158.3	58.6
ACG	36	66	55.9	7.2	CCG	13	30	25.8	4.2	TCG	25	43	35	4.4	GCG	16	39	26.6	5.7
ATA	93	188	144.5	29.6	CTA	90	269	170.3	71	TTA	191	363	280.7	56.6	GTA	102	175	146.8	24.6
ATC	88	123	111.7	6.5	CTC	60	116	92.2	16.5	TTC	120	211	180.5	29.2	GTC	56	105	88.7	15.9
ATT	166	289	243.8	38.4	CTT	170	275	234.5	30	TTT	247	387	338.3	39.5	GTT	139	311	219.1	52.4
ATG	95	315	207.9	89.8	CTG	64	266	152.7	78.6	TTG	149	364	255.3	79.7	GTG	78	279	167.9	81.1
AGA	112	277	221	54.4	CGA	19	42	32.9	3.8	TGA	61	306	241.7	81.6	GGA	59	119	95.4	19.2
AGC	47	141	111.9	31	CGC	24	41	34	6.2	TGC	81	270	204.6	61.2	GGC	44	97	74.8	18.9
AGT	127	202	174.2	22.8	CGT	44	68	54.8	5.7	TGT	193	370	306.7	58.8	GGT	52	246	136.8	62.4
AGG	71	136	111.3	17.7	CGG	16	33	25.9	4.5	TGG	90	264	205.1	57.2	GGG	30	52	46.6	6.4
															other	0	1689	100.4	225.6

which is how to assess the 'dynamics' of the evolutionary process. We must note that the available information just reflects 'time samples' of the variants, that is, the date where the procedure for collecting, identifying and recording the GS took place. Consequently, we do not have a precise control of the time elapsed between the real mutation and the laboratory measurement. Nonetheless, we can have a good idea of the dynamical behavior if we include time information in the MDS plots, even if some 'noise' is present in the time information.

Figures 7, 8 , 9 and 10 depict the MDS plots with the time information represented by the colors of the marks. The colorbar with the interval $[0, 1]$ corresponds to the period between the dates 2020/Feb/24 and 2021/Apr/23. We observe some improvement over

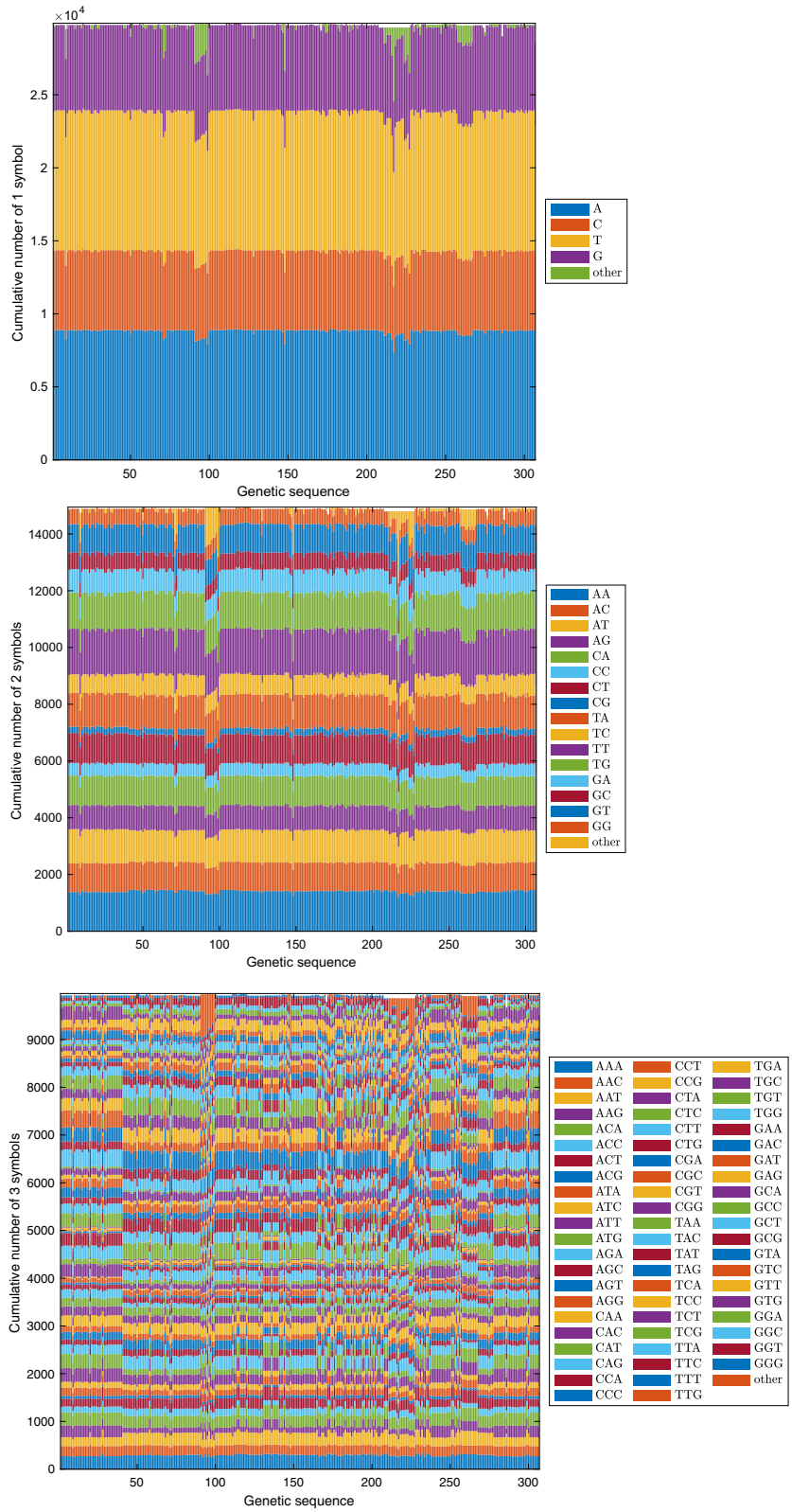

Fig. 1 Cumulative numbers of cases for sub-strings including $k_s = \{1, 2, 3\}$ symbols in the $N = 307$ GS

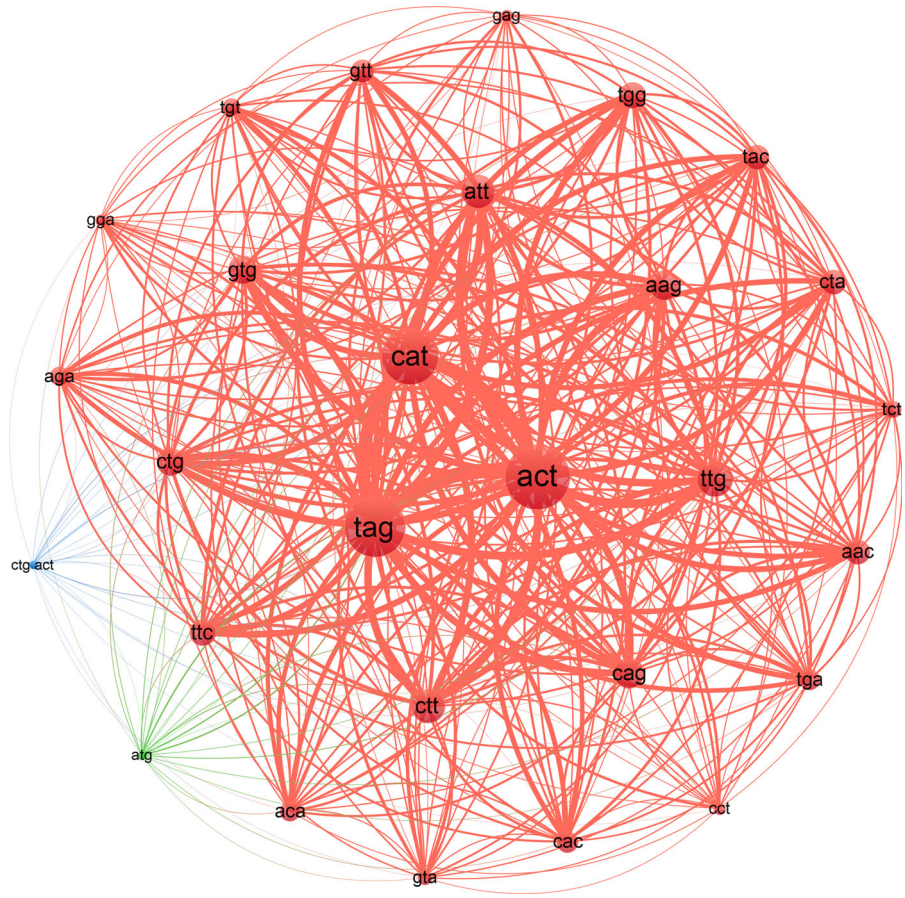

Fig. 2 Network of the sub-strings with $k_s = 3$ symbols extracted from the GS of case $i = 1$

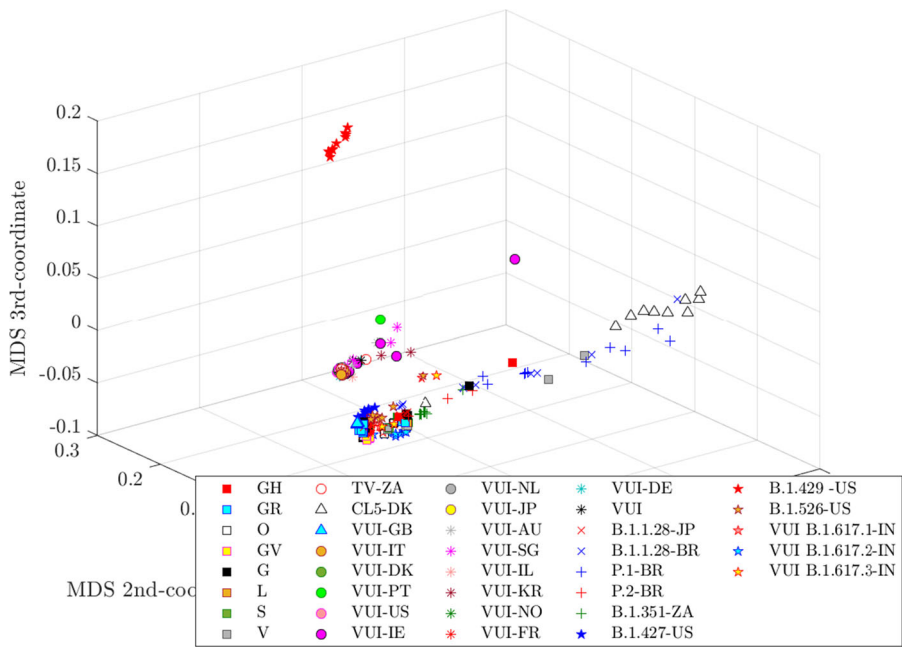

Fig. 3 One perspective of the MDS 3-dim locus for the set of $N = 307$ virus using the normalized information distance, NCD, and `BZip2`

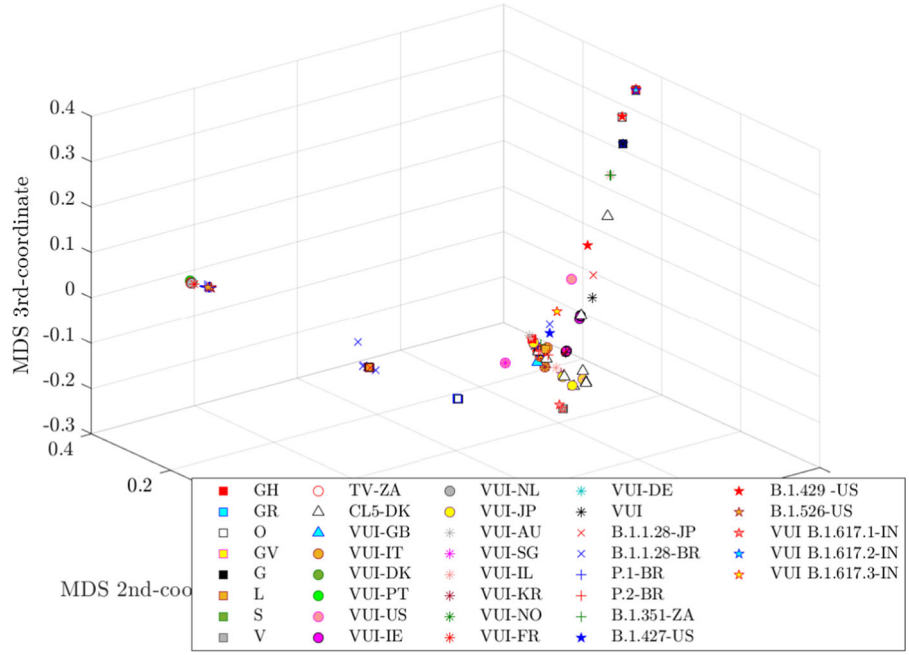

Fig. 4 One perspective of the MDS 3-dim locus for the set of $N = 307$ virus using the Jaccard distance, d_{Ja}, based on the sub-strings with $k_s = 3$ consecutive symbols

the initial set of experiments, but we have still some 'noisy' behavior of the time flow.

It is well known that each distance captures a given characteristic of the phenomenon under analysis. Therefore, we wander if some measure associating several distances could reveal better the patterns embedded in the datataset. In this line of thought we can design a 'generalized' distance by weighting several of the previous distances. If we consider the distances NCD, d_{Jacc}, d_{JS} and d_{Ha}, then we can define the new metric:

$$d_{Ge} = \mu_{NCD}\frac{NCD}{\max(NCD)} + \mu_{Ja}\frac{d_{Ja}}{\max(d_{Ja})} + \mu_{JS}\frac{d_{JS}}{\max(d_{JS})} + \mu_{Ha}\frac{d_{Ha}}{\max(d_{Ha})}, \quad (1)$$

where μ_{NCD}, μ_{Ja}, μ_{JS} and μ_{Ha} are weight factors for the distances NCD, d_{Jacc}, d_{JS} and d_{Ha}, respectively, so that $\mu_{NCD} + \mu_{Ja} + \mu_{JS} + \mu_{Ha} = 1$.

We can adjust the numerical values of the wight factor to reflect the importance of each distance for improving the MDS representation in the sense of providing a more clear visualization of the dynamical effect. In our case, after some experiments, and having in mind the dynamics in time, we consider the weights $\mu_{NCD} = 0.55$, $\mu_{Ja} = 0.2$, $\mu_{JS} = 0.2$ and $\mu_{Ha} = 0.05$. Figure 11 shows the MDS plot where we observe the emergence of five clusters denoted by the symbols $\mathcal{A}$ to $\mathcal{F}$. The time 'arrow' follows somehow the pattern $\mathcal{A} \to \mathcal{B} \to \mathcal{C} \to \{\mathcal{D}, \mathcal{E}\} \to \mathcal{F}$.

We verify (i) the first and second clusters, $\mathcal{A}$ and $\mathcal{B}$, exhibit a very low scattering in contrast with the others, (ii) the existence of some noise in the evolutionary process, (iii) that time flows in discrete steps in the MDS representation, that is, the time evolution is not continuous, (iv) that we can have some evolutionary bifurcations in time as it is visible for $\{\mathcal{D}, \mathcal{E}\}$, and (v) clusters for different time instants may be close in the MDS locus. A deep reflection upon these results seems consistent with our present knowledge of the evolutionary process. In fact, we can interpret and justify the previous assertions as follows:

- the initial strains of virus had very similar characteristics, contrary to the more recent variants that reveal an increasing variability, which agrees with the first observation
- evolution and, in particular, mutations, occur randomly which justifies the second conclusion
- mutations do not emerge continuously in time and, in fact, they are the result of a multitude of issues such as environmental, social, geographical and economical factors. Therefore, new variants that lead to a relevant number of infections are expected to emerge without a clear time pattern which is in accordance with the third consideration

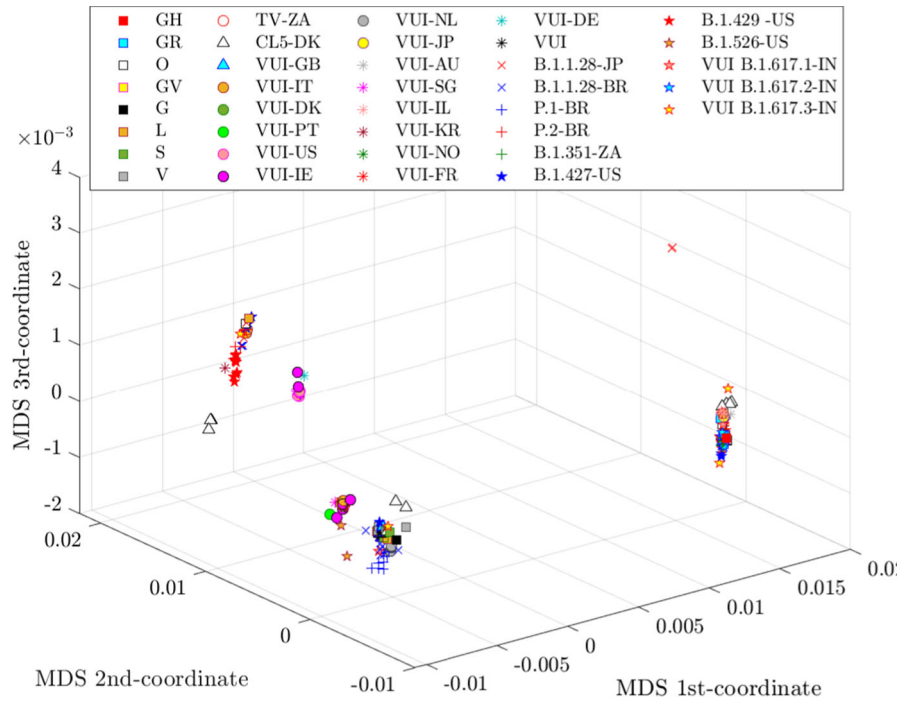

Fig. 5 One perspective of the MDS 3-dim locus for the set of $N = 307$ virus using the Jensen-Shannon divergence, d_{JS}, based on the sub-strings with $k_s = 3$ consecutive symbols

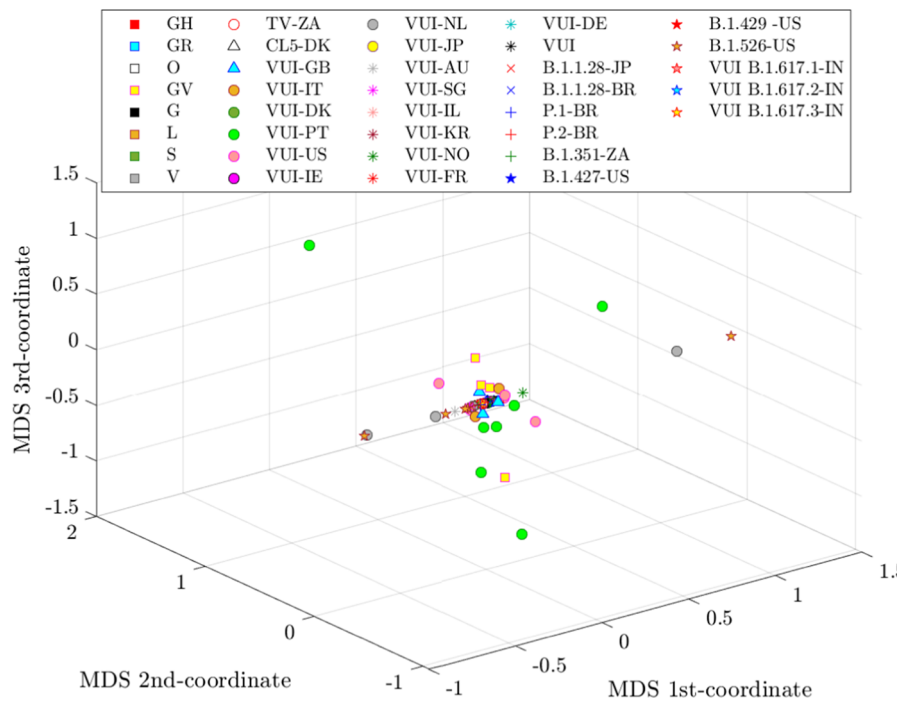

Fig. 6 One perspective of the MDS 3-dim locus for the set of $N = 307$ virus using the Hamming distance, d_{Ha}, based on the sub-strings with $k_s = 3$ consecutive symbols

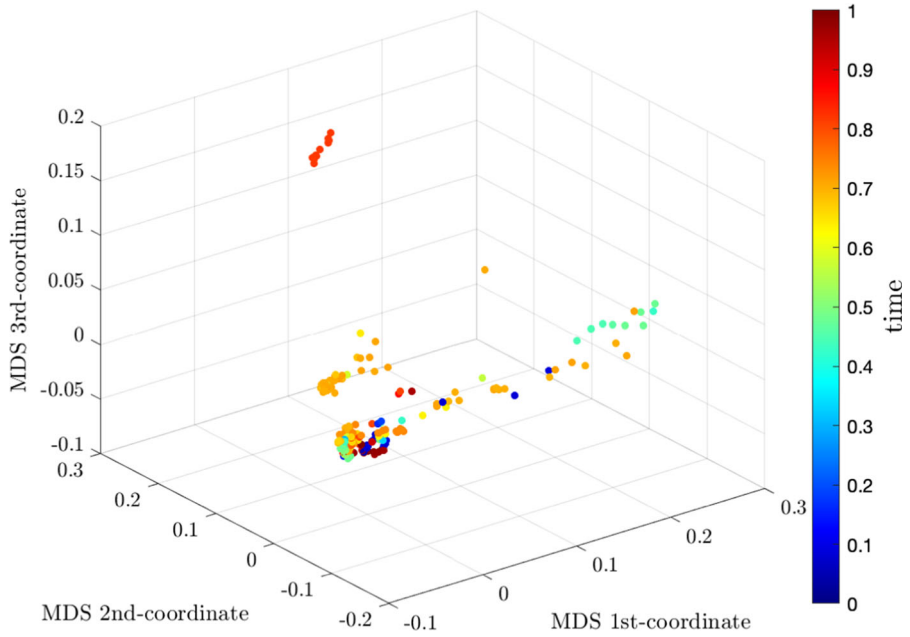

Fig. 7 MDS 3-dim locus exhibiting the dynamics in time of the virus evolution during the period 2020/Feb/24-2021/Apr/23 for the set of $N = 307$ virus using the normalized information distance, NCD, and `BZip2`

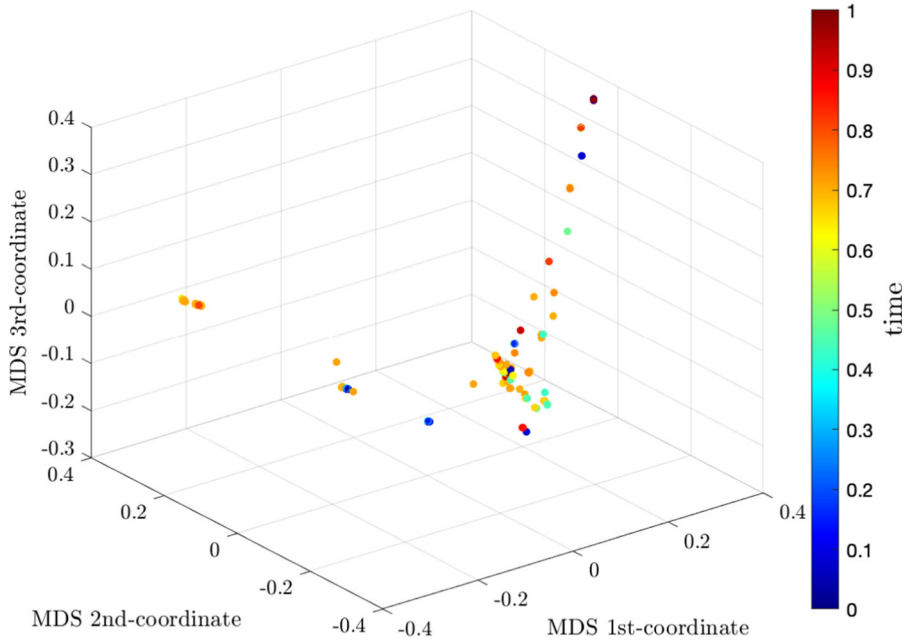

Fig. 8 MDS 3-dim locus exhibiting the dynamics in time of the virus evolution during the period 2020/Feb/24-2021/Apr/23 for the set of $N = 307$ virus using the Jaccard distance, d_{Ja}, based on the sub-strings with $k_s = 3$ consecutive symbols

- the infection spreads in very different space locations and it is reasonable to expect to have several important variants at the same time, particularly when the number of infected people grows considerably. These considerations support the forth remark
- the close location of some clusters for non-consecutive time samples (e.g., $\mathcal{C}$ and $\mathcal{E}$) can be interpreted as the MDS representation the infection 'waves', which explain the fifth observation.

It is important to analyze the effect of the clustering algorithm, the dimension of the visualization space and the type of representation. In this perspective, the Generalized distance (1) and, consequently, the same matrix D used for the MDS scheme, is now used with the set of programs Phylip [55,56]

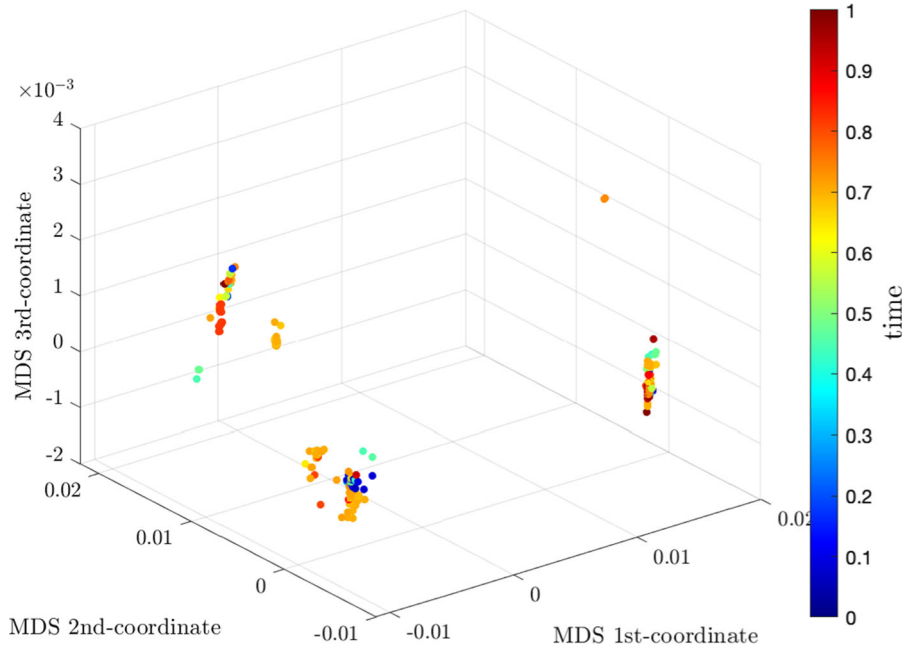

Fig. 9 MDS 3-dim locus exhibiting the dynamics in time of the virus evolution during the period 2020/Feb/24-2021/Apr/23 for the set of $N = 307$ virus using the Jensen-Shannon divergence, d_{JS}, based on the sub-strings with $k_s = 3$ consecutive symbols

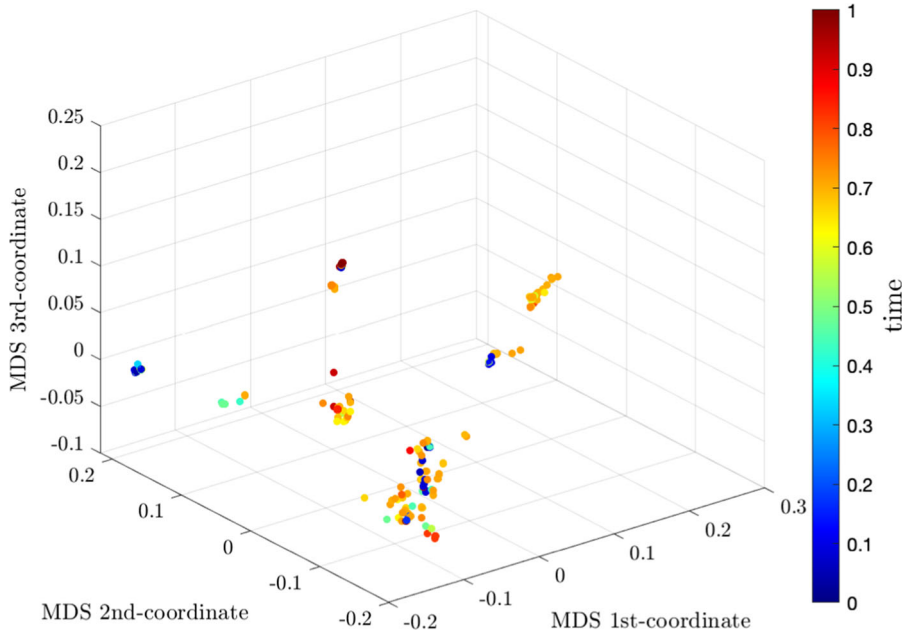

Fig. 10 MDS 3-dim locus exhibiting the dynamics in time of the virus evolution during the period 2020/Feb/24-2021/Apr/23 for the set of $N = 307$ virus using the Hamming distance, d_{Ha}, based on the sub-strings with $k_s = 3$ consecutive symbols

available at http://evolution.genetics.washington.edu/phylip.html. This program is used in in phylogenetics for displaying the evolutionary relationships among various biological objects. The program allows 2-dim representations where the final objects are the 'leafs' of some type of tree. The algorithm `neighbor` processes the matrix D and produces the data clustering. The programs `drawgram` and `drawtree` are used for obtaining the graphical representations in the form of dendrograms and trees, respectively.

Figures 12 and 13 shows the dendrogram and tree generated by Phylip for the generalized distance, d_{Ge}, respectively. The time evolution is represented by colors as before.

The 2-dim graphical representations are easier to visualize in the sense that the user does not needs to

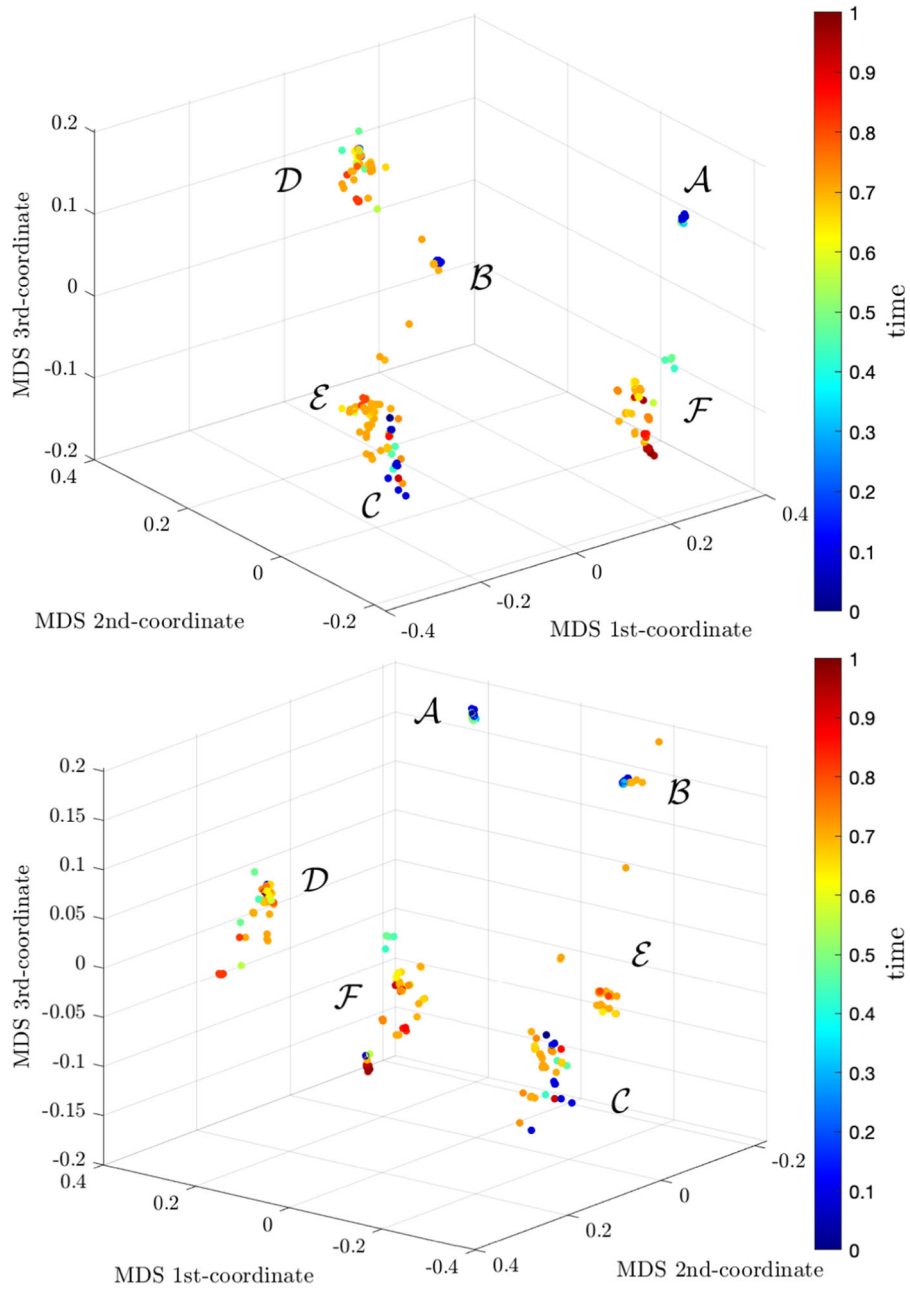

Fig. 11 Two perspectives of the MDS 3-dim locus exhibiting the dynamics in time of the virus evolution during the period 2020/Feb/24-2021/Apr/23 for the set of $N = 307$ virus using the Generalized distance, d_{Ge}. The time 'arrow' follows the pattern $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C} \rightarrow \{\mathcal{D}, \mathcal{E}\} \rightarrow \mathcal{F}$

rotate and shift the plot. However, the lack of the third dimension leads to a clustering somewhat inferior to the one performed by the 3-dim MDS. Nonetheless, we can still see some clusters. In the case of the dendrogram we observe a simple evolution from top (initial period) to bottom (final period) with 3 main clusters. The two clusters at the top do not show a precise separation of the periods of time since with have overlapping object for the initial 60% of the total period. In the case of a graphical output my means of a tree we have a more intricate pattern, with the objects placed in the form of two 'arcs'. The outer arc covers the period of time from beginning up to approximately 75%, while the inner arc corresponds to the rest, that is, to the more recent 25% GS. Moreover, the initial period of time of about 30% is place it the top of the outer arc.

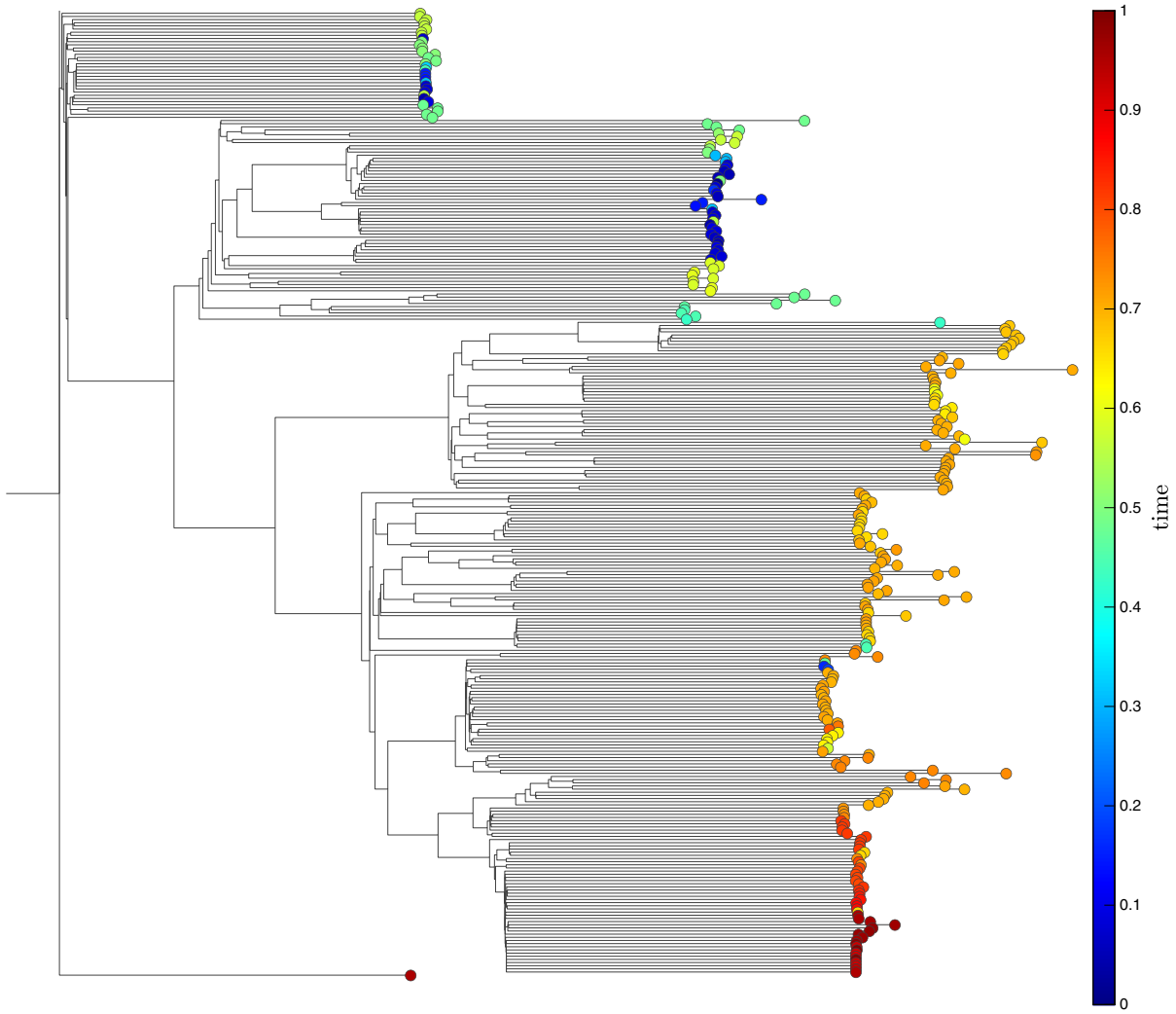

Fig. 12 Dendrogram representation of the SARS-COV2 time dynamics during the period 2020/Feb/24-2021/Apr/23 for $N = 307$ GS using the generalized distance, d_{Ge}

In summary, the strategy followed in this paper is consistent with present-day understanding about the SARS-CoV-2 genome and the adoption of several distinct distances allows users to have a complementary interpretation of the information embedded in the dataset.

5 Conclusions

The information of 307 RNA viruses available in a public database was explored by means of an association of mathematical and computational tools. The notions Kolmogorov complexity, Shannon information and Hamming distance, on the perspective of analytic tools, and the ideas of compression and MDS algorithms, on the point of view of computational tools, were considered. Three sets of indices of dissimilarity were adopted, allowing a broad comparison of the results, with four distances based on the `BZip2` compression, four distances using 2-dimensional histograms of consecutive triplets, and one metric using the Hamming distance between triplets of bases. From these, the MDS algorithm allowed an efficient clustering and 3-dim visualization. The MDS plots revealed pros and cons of the alternative distances adopted for

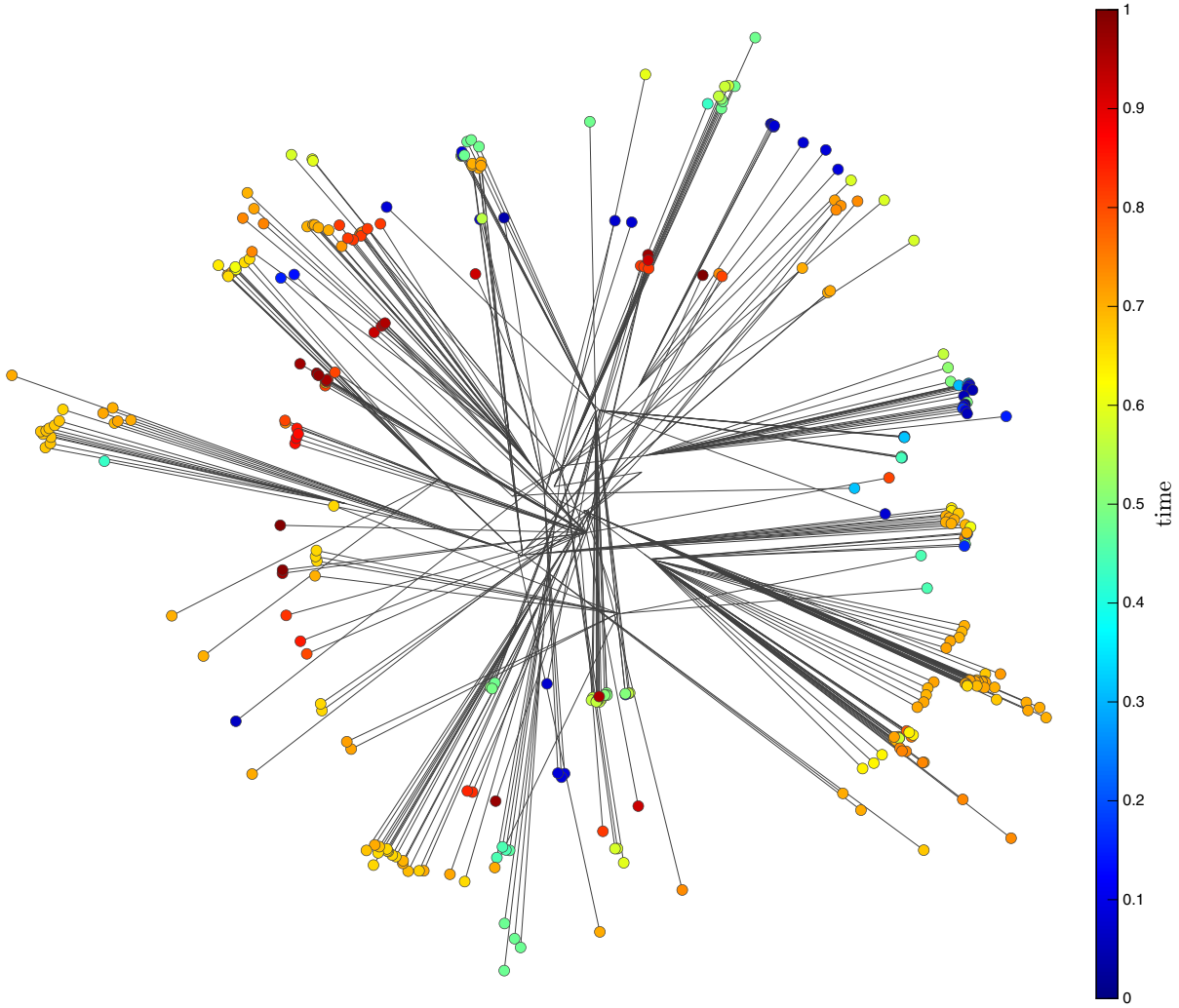

Fig. 13 Tree representation of the SARS-COV2 time dynamics during the period 2020/Feb/24-2021/Apr/23 for $N = 307$ GS using the generalized distance, d_{Ge}

assessing the set of viruses. The problem at stake proved to pose a considerable challenge and no clear clusters emerged for the virus variants included in the dataset. This motivated a new question, namely the relevance of clustering the variants when thinking about its evolutionary dynamics, since many variants have minor differences between themselves. The idea of assessing the dynamics in time lead to the design of a generalized distance taking advantage of the characteristics of distinct metrics and allowing a adjustment to the phenomenon by means of weight factors. The results showed interesting dynamical effects in terms of forming clear clusters in time. Besides the analytic tools, several algorithmic mechanisms were explored, such as the Matlab, VOSviwer and Phylip programs, which illustrate the diversity and richness of present day computational strategies. The synergistic perspectives provided by distinct processing tools and graphical representations allow comparing the genomic data and provide a computational strategy for exploring future viral outbreaks. The association of analytic and computational techniques may help interpreting the phylogeny of these new strain outbreaks, associate its dynamics and selective pressures, and give additional insight for the quick development and testing of tailored countermeasures. The computational analysis of the SARS-

Table 5 Information about the $N = 307$ GS

i	Serial	Clade	Variant	Country	Acronym
1	hCoV-19/England/MILK-ACC1C7/2020|EPI_ISL_629316|2020-10-21	GH		England	GH
2	hCoV-19/England/MILK-AC843B/2020|EPI_ISL_629334|2020-10-21	GH		England	GH
3	hCoV-19/England/MILK-AC7D31/2020|EPI_ISL_629358|2020-10-21	GH		England	GH
4	hCoV-19/England/MILK-ACE3A1/2020|EPI_ISL_629361|2020-10-21	GH		England	GH
5	hCoV-19/England/MILK-ACE4F9/2020|EPI_ISL_629366|2020-10-21	GH		England	GH
6	hCoV-19/England/MILK-ACC3D0/2020|EPI_ISL_629388|2020-10-21	GH		England	GH
7	hCoV-19/England/MILK-ACC2B5/2020|EPI_ISL_629399|2020-10-21	GH		England	GH
8	hCoV-19/England/MILK-ACCA9F/2020|EPI_ISL_629419|2020-10-21	GH		England	GH
9	hCoV-19/England/MILK-AC850B/2020|EPI_ISL_629450|2020-10-21	GH		England	GH
10	hCoV-19/USA/TX-DSHS-1223/2020|EPI_ISL_631683|2020-03-25	GH		USA	GH
11	hCoV-19/Germany/NW-HHU-269/2020|EPI_ISL_631357|2020-09-23	GR		Germany	GR
12	hCoV-19/Germany/NW-HHU-271/2020|EPI_ISL_631359|2020-09-23	GR		Germany	GR
13	hCoV-19/Germany/NW-HHU-274/2020|EPI_ISL_631362|2020-09-25	GR		Germany	GR
14	hCoV-19/Germany/NW-HHU-275/2020|EPI_ISL_631363|2020-09-25	GR		Germany	GR
15	hCoV-19/Germany/NW-HHU-276/2020|EPI_ISL_631364|2020-09-26	GR		Germany	GR
16	hCoV-19/Germany/NW-HHU-242/2020|EPI_ISL_631375|2020-09-15	GR		Germany	GR
17	hCoV-19/Germany/NW-HHU-248/2020|EPI_ISL_631381|2020-09-19	GR		Germany	GR
18	hCoV-19/Germany/NW-HHU-253/2020|EPI_ISL_631386|2020-09-27	GR		Germany	GR
19	hCoV-19/USA/WI-WSLH-200603/2020|EPI_ISL_631398|2020-07-13	GR		USA	GR
20	hCoV-19/Italy/APU-IZSPB-187PT/2020|EPI_ISL_722895|2020-08-28	GR		Italy	GR
21	hCoV-19/Denmark/DCGC-1430/2020|EPI_ISL_622474|2020-04-27	O		Denmark	O
22	hCoV-19/Denmark/DCGC-1934/2020|EPI_ISL_622528|2020-04-27	O		Denmark	O
23	hCoV-19/Denmark/DCGC-1948/2020|EPI_ISL_622531|2020-04-20	O		Denmark	O
24	hCoV-19/Denmark/DCGC-1896/2020|EPI_ISL_622563|2020-07-06	O		Denmark	O
25	hCoV-19/New Zealand/20CV0297/2020|EPI_ISL_622794|2020-03-27	O		New Zealand	O
26	hCoV-19/New Zealand/20CV0302/2020|EPI_ISL_622797|2020-03-29	O		New Zealand	O
27	hCoV-19/New Zealand/20CV0303/2020|EPI_ISL_622798|2020-03-30	O		New Zealand	O
28	hCoV-19/England/204290933/2020|EPI_ISL_622862|2020-10-17	O		USA	O
29	hCoV-19/USA/NY-NYCPHL-000063/2020|EPI_ISL_631557|2020-03-11	O		USA	O
30	hCoV-19/USA/NY-NYCPHL-000136/2020|EPI_ISL_631627|2020-04-02	O		USA	O
31	hCoV-19/Germany/NW-HHU-234/2020|EPI_ISL_631367|2020-09-14	GV		Germany	GV
32	hCoV-19/Germany/NW-HHU-235/2020|EPI_ISL_631368|2020-09-14	GV		Germany	GV
33	hCoV-19/Germany/NW-HHU-236/2020|EPI_ISL_631369|2020-09-15	GV		Germany	GV
34	hCoV-19/Germany/NW-HHU-237/2020|EPI_ISL_631370|2020-09-15	GV		Germany	GV
35	hCoV-19/Germany/NW-HHU-238/2020|EPI_ISL_631371|2020-09-15	GV		Germany	GV

CoV-2 genome may have a role in the early detection of potential variants of concern and help in the characterization of the risk posed to global public health. This is important as it may contribute to the global monitoring of SARS-CoV-2 variants and to improve the search for a more effective response to the COVID-19 pandemic.

Table 5 continued

36	hCoV-19/Germany/NW-HHU-239/2020|EPI_ISL_631372|2020-09-15	GV	Germany	GV
37	hCoV-19/Germany/NW-HHU-240/2020|EPI_ISL_631373|2020-09-15	GV	Germany	GV
38	hCoV-19/Germany/NW-HHU-241/2020|EPI_ISL_631374|2020-09-15	GV	Germany	GV
39	hCoV-19/Germany/NW-HHU-243/2020|EPI_ISL_631376|2020-09-15	GV	Germany	GV
40	hCoV-19/Germany/NW-HHU-255/2020|EPI_ISL_631388|2020-09-30	GV	Germany	GV
41	hCoV-19/England/QEUH-AD061C/2020|EPI_ISL_629979|2020-10-23	G	England	G
42	hCoV-19/England/MILK-ACD278/2020|EPI_ISL_630007|2020-10-21	G	England	G
43	hCoV-19/England/CAMC-AAFF66/2020|EPI_ISL_631149|2020-10-23	G	England	G
44	hCoV-19/England/MILK-ABBDC7/2020|EPI_ISL_631183|2020-10-21	G	England	G
45	hCoV-19/Germany/NW-HHU-244/2020|EPI_ISL_631377|2020-09-29	G	Germany	G
46	hCoV-19/USA/WI-WSLH-200700/2020|EPI_ISL_631457|2020-09-22	G	USA	G
47	hCoV-19/USA/WI-WSLH-200726/2020|EPI_ISL_631482|2020-07-02	G	USA	G
48	hCoV-19/USA/WI-WSLH-200727/2020|EPI_ISL_631483|2020-07-02	G	USA	G
49	hCoV-19/USA/WI-WSLH-200728/2020|EPI_ISL_631484|2020-07-02	G	USA	G
50	hCoV-19/USA/TX-DSHS-1219/2020|EPI_ISL_631679|2020-03-28	G	USA	G
51	hCoV-19/Lithuania/MR-LUHS-Eilnr30/2020|EPI_ISL_603090|2020-04-06	L	Lithuania	L
52	hCoV-19/USA/CA-QDX-2362/2020|EPI_ISL_604656|2020-03-16	L	USA	L
53	hCoV-19/USA/CA-QDX-2383/2020|EPI_ISL_604662|2020-03-15	L	USA	L
54	hCoV-19/USA/UT-QDX-2315/2020|EPI_ISL_604722|2020-03-16	L	USA	L
55	hCoV-19/Ireland/NVRL-20G31886/2020|EPI_ISL_605073|2020-09-22	L	Ireland	L
56	hCoV-19/Henan/HN04/2020|EPI_ISL_605930|2020-02-24	L	Henan	L
57	hCoV-19/Hong Kong/HKPU-0442/2020|EPI_ISL_610175|2020-03-26	L	Hong Kong	L
58	hCoV-19/USA/NM-UNM-00797/2020|EPI_ISL_610261|2020-05-06	L	USA	L
59	hCoV-19/USA/WI-WSLH-200589/2020|EPI_ISL_631391|2020-04-03	L	USA	L
60	hCoV-19/USA/TX-DSHS-1208/2020|EPI_ISL_631669|2020-03-18	L	USA	L
61	hCoV-19/Denmark/DCGC-1430/2020|EPI_ISL_622474|2020-04-27	S	Denmark	S
62	hCoV-19/Denmark/DCGC-1934/2020|EPI_ISL_622528|2020-04-27	S	Denmark	S
63	hCoV-19/Denmark/DCGC-1948/2020|EPI_ISL_622531|2020-04-20	S	Denmark	S
64	hCoV-19/Denmark/DCGC-1896/2020|EPI_ISL_622563|2020-07-06	S	Denmark	S
65	hCoV-19/New Zealand/20CV0297/2020|EPI_ISL_622794|2020-03-27	S	New Zealand	S
66	hCoV-19/New Zealand/20CV0302/2020|EPI_ISL_622797|2020-03-29	S	New Zealand	S
67	hCoV-19/New Zealand/20CV0303/2020|EPI_ISL_622798|2020-03-30	S	New Zealand	S
68	hCoV-19/England/204290933/2020|EPI_ISL_622862|2020-10-17	S	England	S
69	hCoV-19/USA/NY-NYCPHL-000063/2020|EPI_ISL_631557|2020-03-11	S	USA	S
70	hCoV-19/USA/NY-NYCPHL-000136/2020|EPI_ISL_631627|2020-04-02	S	USA	S

Table 6 Information about the $N = 307$ GS

i	Serial	Clade	Variant	Country	Acronym
71	hCoV-19/England/LIVE-1D2EC0/2020|EPI_ISL_612017|2020-03-31	V		England	V
72	hCoV-19/England/LIVE-1D2EDF/2020|EPI_ISL_612075|2020-03-30	V		England	V
73	hCoV-19/England/PORT-2D21A8/2020|EPI_ISL_613283|2020-03-22	V		England	V
74	hCoV-19/England/PORT-2D220F/2020|EPI_ISL_613285|2020-03-22	V		England	V
75	hCoV-19/USA/NY-NYCPHL-000032/2020|EPI_ISL_631528|2020-03-12	V		England	V
76	hCoV-19/USA/NY-NYCPHL-000133/2020|EPI_ISL_631624|2020-04-02	V		England	V
77	hCoV-19/England/LIVE-1D2BC9/2020|EPI_ISL_612414|2020-03-29	V		England	V
78	hCoV-19/England/LIVE-1D2C02/2020|EPI_ISL_612415|2020-03-29	V		England	V
79	hCoV-19/England/LIVE-1D2D2D/2020|EPI_ISL_612418|2020-03-30	V		England	V
80	hCoV-19/England/LIVE-1D2D4B/2020|EPI_ISL_612419|2020-03-31	V		England	V
81	hCoV-19/South Africa/N00352/2020|EPI_ISL_712084|2020-10-30	GH	South Africa Triple Variant	South Africa	TV-ZA
82	hCoV-19/South Africa/N00336/2020|EPI_ISL_712085|2020-10-30	GH	South Africa Triple Variant	South Africa	TV-ZA
83	hCoV-19/South Africa/N00334/2020|EPI_ISL_712086|2020-10-28	GH	South Africa Triple Variant	South Africa	TV-ZA
84	hCoV-19/South Africa/N00349/2020|EPI_ISL_712087|2020-11-03	GH	South Africa Triple Variant	South Africa	TV-ZA
85	hCoV-19/South Africa/N00337/2020|EPI_ISL_712088|2020-11-03	GH	South Africa Triple Variant	South Africa	TV-ZA
86	hCoV-19/South Africa/N00350/2020|EPI_ISL_712089|2020-10-28	GH	South Africa Triple Variant	South Africa	TV-ZA
87	hCoV-19/South Africa/N00348/2020|EPI_ISL_712090|2020-10-30	GH	South Africa Triple Variant	South Africa	TV-ZA
88	hCoV-19/South Africa/N00343/2020|EPI_ISL_712091|2020-10-28	GH	South Africa Triple Variant	South Africa	TV-ZA
89	hCoV-19/South Africa/N00338/2020|EPI_ISL_712095|2020-11-03	GH	South Africa Triple Variant	South Africa	TV-ZA
90	hCoV-19/South Africa/N00344/2020|EPI_ISL_712096|2020-11-04	GH	South Africa Triple Variant	South Africa	TV-ZA
91	hCoV-19/Denmark/DCGC-4467/2020|EPI_ISL_615652|2020-09-14	GR	Mink Cluster V	Denmark	CL5-DK
92	hCoV-19/Denmark/DCGC-4468/2020|EPI_ISL_615653|2020-09-14	GR	Mink Cluster V	Denmark	CL5-DK
93	hCoV-19/Denmark/DCGC-4472/2020|EPI_ISL_615657|2020-09-14	GR	Mink Cluster V	Denmark	CL5-DK
94	hCoV-19/Denmark/DCGC-4476/2020|EPI_ISL_615661|2020-09-14	GR	Mink Cluster V	Denmark	CL5-DK
95	hCoV-19/Denmark/DCGC-4306/2020|EPI_ISL_615743|2020-09-07	GR	Mink Cluster V	Denmark	CL5-DK
96	hCoV-19/Denmark/DCGC-3152/2020|EPI_ISL_615972|2020-08-31	GR	Mink Cluster V	Denmark	CL5-DK
97	hCoV-19/Denmark/DCGC-3542/2020|EPI_ISL_616335|2020-08-31	GR	Mink Cluster V	Denmark	CL5-DK
98	hCoV-19/Denmark/DCGC-3631/2020|EPI_ISL_616402|2020-08-31	GR	Mink Cluster V	Denmark	CL5-DK
99	hCoV-19/Denmark/DCGC-2934/2020|EPI_ISL_616695|2020-08-24	GR	Mink Cluster V	Denmark	CL5-DK
100	hCoV-19/Denmark/DCGC-2965/2020|EPI_ISL_616727|2020-08-24	GR	Mink Cluster V	Denmark	CL5-DK
101	hCoV-19/England/ALDP-C3B5E3/2020|EPI_ISL_731850|2020-12-08	GR	VUI2020/01	England	VUI-GB
102	hCoV-19/England/ALDP-C3BAF3/2020|EPI_ISL_731852|2020-12-10	GR	VUI2020/01	England	VUI-GB
103	hCoV-19/England/ALDP-C3AA4F/2020|EPI_ISL_731854|2020-12-09	GR	VUI2020/01	England	VUI-GB
104	hCoV-19/England/ALDP-C3C9F5/2020|EPI_ISL_731857|2020-12-10	GR	VUI2020/01	England	VUI-GB
105	hCoV-19/England/ALDP-C3CEAB/2020|EPI_ISL_731865|2020-12-10	GR	VUI2020/01	England	VUI-GB
106	hCoV-19/England/ALDP-C3D0D1/2020|EPI_ISL_731866|2020-12-10	GR	VUI2020/01	England	VUI-GB
107	hCoV-19/England/ALDP-C3C05A/2020|EPI_ISL_731872|2020-12-10	GR	VUI2020/01	England	VUI-GB
108	hCoV-19/England/ALDP-C3B10D/2020|EPI_ISL_731877|2020-12-08	GR	VUI2020/01	England	VUI-GB
109	hCoV-19/England/ALDP-C3BEAC/2020|EPI_ISL_731878|2020-12-10	GR	VUI2020/01	England	VUI-GB
110	hCoV-19/England/MILK-BFF537/2020|EPI_ISL_731893|2020-12-02	GR	VUI2020/01	England	VUI-GB

Table 6 continued

111	hCoV-19/Italy/LAZ-AMC3-5015/2020|EPI_ISL_717978|2020-12-14	GR	VUI2020/01	Italy	VUI-IT
112	hCoV-19/Italy/CAM-TIGEM-174/2020|EPI_ISL_736786|2020-12-21	GR	VUI2020/01	Italy	VUI-IT
113	hCoV-19/Italy/CAM-TIGEM-181/2020|EPI_ISL_736787|2020-12-21	GR	VUI2020/01	Italy	VUI-IT
114	hCoV-19/Italy/CAM-INMI-117/2020|EPI_ISL_736996|2020-12-20	GR	VUI2020/01	Italy	VUI-IT
115	hCoV-19/Italy/CAM-INMI-118/2020|EPI_ISL_736997|2020-12-20	GR	VUI2020/01	Italy	VUI-IT
116	hCoV-19/Italy/ABR-TE351971/2020|EPI_ISL_738045|2020-12-18	GR	VUI2020/01	Italy	VUI-IT
117	hCoV-19/Italy/ABR-TE353967/2020|EPI_ISL_738046|2020-12-18	GR	VUI2020/01	Italy	VUI-IT
118	hCoV-19/Italy/ABR-TE353968/2020|EPI_ISL_738047|2020-12-18	GR	VUI2020/01	Italy	VUI-IT
119	hCoV-19/Italy/ABR-TE353969/2020|EPI_ISL_738048|2020-12-18	GR	VUI2020/01	Italy	VUI-IT
120	hCoV-19/Italy/APU-IZSPB-400PT/2020|EPI_ISL_745192|2020-12-21	GR	VUI2020/01	Italy	VUI-IT
121	hCoV-19/Denmark/DCGC-9482/2020|EPI_ISL_668598|2020-11-09	GR	VUI2020/01	Denmark	VUI-DK
122	hCoV-19/Denmark/DCGC-9581/2020|EPI_ISL_668599|2020-11-09	GR	VUI2020/01	Denmark	VUI-DK
123	hCoV-19/Denmark/DCGC-9642/2020|EPI_ISL_668600|2020-11-09	GR	VUI2020/01	Denmark	VUI-DK
124	hCoV-19/Denmark/DCGC-13966/2020|EPI_ISL_711229|2020-11-30	GR	VUI2020/01	Denmark	VUI-DK
125	hCoV-19/Denmark/DCGC-14038/2020|EPI_ISL_711230|2020-11-30	GR	VUI2020/01	Denmark	VUI-DK
126	hCoV-19/Denmark/DCGC-14208/2020|EPI_ISL_711231|2020-11-30	GR	VUI2020/01	Denmark	VUI-DK
127	hCoV-19/Denmark/DCGC-14596/2020|EPI_ISL_711232|2020-11-23	GR	VUI2020/01	Denmark	VUI-DK
128	hCoV-19/Denmark/DCGC-15593/2020|EPI_ISL_713235|2020-11-23	GR	VUI2020/01	Denmark	VUI-DK
129	hCoV-19/Denmark/DCGC-15226/2020|EPI_ISL_713269|2020-11-23	GR	VUI2020/01	Denmark	VUI-DK
130	hCoV-19/Denmark/DCGC-18854/2020|EPI_ISL_748280|2020-12-07	GR	VUI2020/01	Denmark	VUI-DK
131	hCoV-19/Portugal/PT2101/2020|EPI_ISL_738101|2020-12-18	GR	VUI2020/01	Portugal	VUI-PT
132	hCoV-19/Portugal/PT2102/2020|EPI_ISL_738102|2020-12-18	GR	VUI2020/01	Portugal	VUI-PT
133	hCoV-19/Portugal/PT2104/2020|EPI_ISL_738103|2020-12-17	GR	VUI2020/01	Portugal	VUI-PT
134	hCoV-19/Portugal/PT2105/2020|EPI_ISL_738104|2020-12-17	GR	VUI2020/01	Portugal	VUI-PT
135	hCoV-19/Portugal/PT2106/2020|EPI_ISL_738105|2020-12-17	GR	VUI2020/01	Portugal	VUI-PT
136	hCoV-19/Portugal/PT2107/2020|EPI_ISL_738106|2020-12-15	GR	VUI2020/01	Portugal	VUI-PT
137	hCoV-19/Portugal/PT2108/2020|EPI_ISL_738107|2020-11-09	GR	VUI2020/01	Portugal	VUI-PT
138	hCoV-19/Portugal/PT2110/2020|EPI_ISL_738108|2020-12-07	GR	VUI2020/01	Portugal	VUI-PT
139	hCoV-19/Portugal/PT2112/2020|EPI_ISL_738109|2020-12-19	GR	VUI2020/01	Portugal	VUI-PT
140	hCoV-19/Portugal/PT2093/2020|EPI_ISL_738110|2020-12-20	GR	VUI2020/01	Portugal	VUI-PT

Table 7 Information about the $N = 307$ GS

i	Serial	Clade	Variant	Country	Acronym
141	hCoV-19/USA/CO-CDPHE-2100156850/2020|EPI_ISL_751800|2020-12-24	GR	VUI2020/01	USA	VUI-US
142	hCoV-19/USA/CA-SEARCH-5574/2020|EPI_ISL_751801|2020-12-29	GR	VUI2020/01	USA	VUI-US
143	hCoV-19/USA/FL-CDC-STM-P012/2020|EPI_ISL_755593|2020-12-19	GR	VUI2020/01	USA	VUI-US
144	hCoV-19/USA/CA-CDC-STM-P017/2020|EPI_ISL_755594|2020-12-20	GR	VUI2020/01	USA	VUI-US
145	hCoV-19/USA/CA-CDC-STM-P019/2020|EPI_ISL_755595|2020-12-21	GR	VUI2020/01	USA	VUI-US
146	hCoV-19/USA/CA-CDC-STM-P025/2020|EPI_ISL_755596|2020-12-20	GR	VUI2020/01	USA	VUI-US
147	hCoV-19/USA/CA-CDPH-UC301/2020|EPI_ISL_755940|2020-12-20	GR	VUI2020/01	USA	VUI-US
148	hCoV-19/USA/CA-CDPH-UC302/2020|EPI_ISL_755941|2020-12-20	GR	VUI2020/01	USA	VUI-US
149	hCoV-19/Ireland/D-NVRL-20IRL10513/2020|EPI_ISL_735384|2020-12-20	GR	VUI2020/01	Ireland	VUI-IE
150	hCoV-19/Ireland/KK-NVRL-20IRL10079/2020|EPI_ISL_735385|2020-12-19	GR	VUI2020/01	Ireland	VIU-IE
151	hCoV-19/Ireland/D-NVRL-20IRL09344/2020|EPI_ISL_735386|2020-12-18	GR	VUI2020/01	Ireland	VIU-IE
152	hCoV-19/Ireland/D-NVRL-20IRL10553/2020|EPI_ISL_735387|2020-12-20	GR	VUI2020/01	Ireland	VIU-IE
153	hCoV-19/Ireland/D-NVRL-20IRL10860/2020|EPI_ISL_735388|2020-12-21	GR	VUI2020/01	Ireland	VIU-IE
154	hCoV-19/Ireland/D-NVRL-20IRL10527/2020|EPI_ISL_735389|2020-12-20	GR	VUI2020/01	Ireland	VIU-IE
155	hCoV-19/Ireland/D-NVRL-20IRL10603/2020|EPI_ISL_735390|2020-12-20	GR	VUI2020/01	Ireland	VIU-IE
156	hCoV-19/Netherlands/NH-RIVM-20227/2020|EPI_ISL_728566|2020-12-05	GR	VUI2020/01	Netherlands	VUI-NL
157	hCoV-19/Netherlands/NH-RIVM-20432/2020|EPI_ISL_728568|2020-12-13	GR	VUI2020/01	Netherlands	VUI-NL
158	hCoV-19/Netherlands/ZH-EMC-1148/2020|EPI_ISL_734166|2020-12-19	GR	VUI2020/01	Netherlands	VUI-NL
159	hCoV-19/Netherlands/NH-RIVM-20631/2020|EPI_ISL_747521|2020-12-16	GR	VUI2020/01	Netherlands	VUI-NL
160	hCoV-19/Netherlands/NH-RIVM-20635/2020|EPI_ISL_747522|2020-11-29	GR	VUI2020/01	Netherlands	VUI-NL
161	hCoV-19/Netherlands/NH-RIVM-20652/2020|EPI_ISL_747524|2020-12-21	GR	VUI2020/01	Netherlands	VUI-NL
162	hCoV-19/Japan/IC-0413/2020|EPI_ISL_735439|2020-12	GR	VUI2020/01	Japan	VUI-JP
163	hCoV-19/Japan/IC-0419/2020|EPI_ISL_735440|2020-12	GR	VUI2020/01	Japan	VUI-JP
164	hCoV-19/Japan/IC-0421/2020|EPI_ISL_735441|2020-12	GR	VUI2020/01	Japan	VUI-JP
165	hCoV-19/Japan/IC-0422/2020|EPI_ISL_735442|2020-12	GR	VUI2020/01	Japan	VUI-JP
166	hCoV-19/Japan/IC-0423/2020|EPI_ISL_735443|2020-12	GR	VUI2020/01	Japan	VUI-JP
167	hCoV-19/Japan/IC-0424/2020|EPI_ISL_736891|2020-12	GR	VUI2020/01	Japan	VUI-JP
168	hCoV-19/Australia/NSW3456/2020|EPI_ISL_678386|2020-11-30	GR	VUI2020/01	Australia	VUI-AU
169	hCoV-19/Australia/NSW3647/2020|EPI_ISL_717711|2020-12-09	GR	VUI2020/01	Australia	VUI-AU
170	hCoV-19/Australia/SAP593/2020|EPI_ISL_732961|2020-12-19	GR	VUI2020/01	Australia	VUI-AU
171	hCoV-19/Australia/NSW1242/2020|EPI_ISL_740873|2020-12-09	GR	VUI2020/01	Australia	VUI-AU
172	hCoV-19/Australia/SAP595/2020|EPI_ISL_752598|2020-12-26	GR	VUI2020/01	Australia	VUI-AU
173	hCoV-19/Singapore/1453/2020|EPI_ISL_728189|2020-12-08	GR	VUI2020/01	Singapore	VUI-SG
174	hCoV-19/Singapore/1468/2020|EPI_ISL_754075|2020-12-21	GR	VUI2020/02	Singapore	VUI-SG
175	hCoV-19/Singapore/1489/2020|EPI_ISL_754076|2020-12-22	GR	VUI2020/03	Singapore	VUI-SG
176	hCoV-19/Singapore/1467/2020|EPI_ISL_754082|2020-12-19	GR	VUI2020/04	Singapore	VUI-SG
177	hCoV-19/Singapore/1466/2020|EPI_ISL_754083|2020-12-17	GR	VUI2020/05	Singapore	VUI-SG
178	hCoV-19/Israel/CVL-7075/2020|EPI_ISL_733498|2020-12-16	GR	VUI2020/01	Israel	VUI-IL
179	hCoV-19/Israel/CVL-46880/2020|EPI_ISL_737202|2020-12-20	GR	VUI2020/01	Israel	VUI-IL

Table 7 continued

180	hCoV-19/Israel/CVL-46754/2020|EPI_ISL_737203|2020-12-20	GR	VUI2020/01	Israel	VUI-IL
181	hCoV-19/Israel/CVL-46879/2020|EPI_ISL_737204|2020-12-20	GR	VUI2020/01	Israel	VUI-IL
182	hCoV-19/South Korea/KDCA0001/2020|EPI_ISL_738139|2020-12-22	GR	VUI2020/01	South Korea	VUI-KR
183	hCoV-19/South Korea/KDCA0002/2020|EPI_ISL_738141|2020-12-22	GR	VUI2020/01	South Korea	VUI-KR
184	hCoV-19/South Korea/KDCA0003/2020|EPI_ISL_738142|2020-12-22	GR	VUI2020/01	South Korea	VUI-KR
185	hCoV-19/Norway/7115/2020|EPI_ISL_738313|2020-12-21	GR	VUI2020/01	Norway	VUI-NO
186	hCoV-19/Norway/7129/2020|EPI_ISL_738314|2020-12-13	GR	VUI2020/01	Norway	VUI-NO
187	hCoV-19/France/CVL-SC719/2020|EPI_ISL_735391|2020-12-18	GR	VUI2020/01	France	GR-FR
188	hCoV-19/France/HDF-IPP11311/2020|EPI_ISL_754842|2020-12-21	GR	VUI2020/01	France	GR-FR
189	hCoV-19/Germany/NW-RKI-I-0026/2020|EPI_ISL_751799|2020-12-07	GR	VUI2020/01	Germany	GR-DE
190	hCoV-19/Germany/DE-BW-ChVir21528/2020|EPI_ISL_754174|2020-12-21	GR	VUI2020/01	Germany	GR-DE
191	hCoV-19/Spain/MD-H12-02-5372/2020|EPI_ISL_737930|2020-12-23	GR	VUI2020/01	Spain	VUI
192	hCoV-19/Gibraltar/205000662/2020|EPI_ISL_709957|2020-12	GR	VUI2020/01	Gibraltar	VUI
193	hCoV-19/Hong Kong/CM20000424/2020|EPI_ISL_733573|2020-12-07	GR	VUI2020/01	Hong Kong	VUI
194	hCoV-19/India/KA-NIMH-SEQ-8/2020|EPI_ISL_747244|2020-12-22	GR	VUI2020/01	India	VUI
195	hCoV-19/Luxembourg/LNS2043044/2020|EPI_ISL_755589|2020-12-24	GR	VUI2020/01	Luxembourg	VUI
196	hCoV-19/Switzerland/ZH-UZH-IMV130/2020|EPI_ISL_751193|2020-12-18	GR	VUI2020/01	Switzerland	VUI
197	hCoV-19/Sweden/20-53787/2020|EPI_ISL_738133|2020-12-20	GR	VUI2020/01	Sweden	VUI
198	hCoV-19/Japan/IC-0500/2020|EPI_ISL_768709|2020-12	GR	B.1.1.28	Japan	B.1.1.28-JP
199	hCoV-19/Japan/IC-0524/2020|EPI_ISL_779207|2020-12	GR	B.1.1.28	Japan	B.1.1.28-JP
200	hCoV-19/Japan/IC-0526/2020|EPI_ISL_779209|2020-12	GR	B.1.1.28	Japan	B.1.1.28-JP
201	hCoV-19/Japan/IC-0533/2020|EPI_ISL_779216|2020-12	GR	B.1.1.28	Japan	B.1.1.28-JP
202	hCoV-19/Japan/IC-0566/2020|EPI_ISL_779245|2020-09	GR	B.1.1.28	Japan	B.1.1.28-JP
203	hCoV-19/Japan/IC-0567/2020|EPI_ISL_779246|2020-09	GR	B.1.1.28	Japan	B.1.1.28-JP
204	hCoV-19/Japan/IC-0561/2021|EPI_ISL_792680|2021-01-02	GR	B.1.1.28	Japan	B.1.1.28-JP
205	hCoV-19/Japan/IC-0562/2021|EPI_ISL_792681|2021-01-02	GR	B.1.1.28	Japan	B.1.1.28-JP
206	hCoV-19/Japan/IC-0563/2021|EPI_ISL_792682|2021-01-02	GR	B.1.1.28	Japan	B.1.1.28-JP
207	hCoV-19/Japan/IC-0564/2021|EPI_ISL_792683|2021-01-02	GR	B.1.1.28	Japan	B.1.1.28-JP

Table 8 Information about the $N = 307$ GS

i	Serial	Clade	Variant	Country	Acronym
208	hCoV-19/Brazil/AM-20890117JP/2020|EPI_ISL_801401|2020-09-15	GR	B.1.1.28	Brazil	B.1.1.28-BR
209	hCoV-19/Brazil/AM-20890261MV/2020|EPI_ISL_801402|2020-05-03	GR	B.1.1.28	Brazil	B.1.1.28-BR
210	hCoV-19/Brazil/AM-20892948LS/2020|EPI_ISL_801403|2020-05-12	GR	B.1.1.28	Brazil	B.1.1.28-BR
211	hCoV-19/Brazil/AM-L70-CD1731/2020|EPI_ISL_804817|2020-12-17	GR	B.1.1.28	Brazil	B.1.1.28-BR
212	hCoV-19/Brazil/AM-L70-CD1729/2020|EPI_ISL_804818|2020-12-17	GR	B.1.1.28	Brazil	B.1.1.28-BR
213	hCoV-19/Brazil/AM-L70-CD1715/2020|EPI_ISL_804825|2020-12-15	GR	B.1.1.28	Brazil	B.1.1.28-BR
214	hCoV-19/Brazil/AM-L70-CD1717/2020|EPI_ISL_804826|2020-12-16	GR	B.1.1.28	Brazil	B.1.1.28-BR
215	hCoV-19/Brazil/AM-L70-CD1735/2020|EPI_ISL_804830|2020-12-21	GR	B.1.1.28	Brazil	B.1.1.28-BR
216	hCoV-19/Brazil/AM-L71-CD1748/2020|EPI_ISL_804834|2020-12-23	GR	B.1.1.28	Brazil	B.1.1.28-BR
217	hCoV-19/Brazil/AM-L71-CD1741/2020|EPI_ISL_804841|2020-12-22	GR	B.1.1.28	Brazil	B.1.1.28-BR
218	hCoV-19/Brazil/AM-L70-CD1716/2020|EPI_ISL_804815|2020-12-16	GR	P.1	Brazil	P.1-BR
219	hCoV-19/Brazil/AM-L70-CD1718/2020|EPI_ISL_804819|2020-12-16	GR	P.1	Brazil	P.1-BR
220	hCoV-19/Brazil/AM-L70-CD1739/2020|EPI_ISL_804820|2020-12-21	GR	P.1	Brazil	P.1-BR
221	hCoV-19/Brazil/AM-L70-CD1733/2020|EPI_ISL_804821|2020-12-21	GR	P.1	Brazil	P.1-BR
222	hCoV-19/Brazil/AM-L70-CD1722/2020|EPI_ISL_804823|2020-12-17	GR	P.1	Brazil	P.1-BR
223	hCoV-19/Brazil/AM-L70-CD1721/2020|EPI_ISL_804824|2020-12-17	GR	P.1	Brazil	P.1-BR
224	hCoV-19/Brazil/AM-L70-CD1723/2020|EPI_ISL_804832|2020-12-17	GR	P.1	Brazil	P.1-BR
225	hCoV-19/Brazil/AM-L71-CD1743/2020|EPI_ISL_804833|2020-12-22	GR	P.1	Brazil	P.1-BR
226	hCoV-19/Brazil/AM-L70-CD1727/2020|EPI_ISL_804835|2020-12-18	GR	P.1	Brazil	P.1-BR
227	hCoV-19/Brazil/AM-L71-CD1740/2020|EPI_ISL_804843|2020-12-21	GR	P.1	Brazil	P.1-BR
228	hCoV-19/Brazil/AP-IEC-177409/2021|EPI_ISL_918561|2021-01-10	GR	P.2	Brazil	P.2-BR
229	hCoV-19/Brazil/SP-1076/2021|EPI_ISL_940628|2021-01-22	GR	P.2	Brazil	P.2-BR
230	hCoV-19/Brazil/RS-CEVS-83356/2020|EPI_ISL_943604|2020-11-20	GR	P.2	Brazil	P.2-BR
231	hCoV-19/Brazil/RS-CEVS-82920/2020|EPI_ISL_943607|2020-11-20	GR	P.2	Brazil	P.2-BR
232	hCoV-19/Brazil/RS-CEVS-82973/2020|EPI_ISL_943608|2020-11-18	GR	P.2	Brazil	P.2-BR
233	hCoV-19/Brazil/RS-CEVS-83139/2020|EPI_ISL_943610|2020-11-18	GR	P.2	Brazil	P.2-BR
234	hCoV-19/Brazil/RS-CEVS-83150/2020|EPI_ISL_943612|2020-11-19	GR	P.2	Brazil	P.2-BR
235	hCoV-19/Brazil/TO-1185/2020|EPI_ISL_943984|2020-10-26	GR	P.2	Brazil	P.2-BR
236	hCoV-19/Brazil/TO-1187R2/2020|EPI_ISL_943986|2020-12-22	GR	P.2	Brazil	P.2-BR
237	hCoV-19/Brazil/GO-1190R2/2020|EPI_ISL_943989|2020-12-18	GR	P.2	Brazil	P.2-BR
238	hCoV-19/South Africa/KRISP-K008261/2021|EPI_ISL_944169|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
239	hCoV-19/South Africa/KRISP-K008262/2021|EPI_ISL_944170|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
240	hCoV-19/South Africa/KRISP-K008263/2021|EPI_ISL_944171|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
241	hCoV-19/South Africa/KRISP-K008265/2021|EPI_ISL_944172|2021-01-05	GH	B.1.351	South Africa	B.1.351-ZA
242	hCoV-19/South Africa/KRISP-K008267/2021|EPI_ISL_944173|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
243	hCoV-19/South Africa/KRISP-K008269/2021|EPI_ISL_944174|2021-01-01	GH	B.1.351	South Africa	B.1.351-ZA
244	hCoV-19/South Africa/KRISP-K008271/2021|EPI_ISL_944175|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
245	hCoV-19/South Africa/KRISP-K008274/2021|EPI_ISL_944176|2021-01-04	GH	B.1.351	South Africa	B.1.351-ZA
246	hCoV-19/South Africa/KRISP-K008276/2021|EPI_ISL_944177|2021-01-05	GH	B.1.351	South Africa	B.1.351-ZA
247	hCoV-19/South Africa/UFS-VIRO-NGS-163/2020|EPI_ISL_949632|2020-12-23	GH	B.1.351	South Africa	B.1.351-ZA
248	hCoV-19/USA/CA-CZB-17547/2020|EPI_ISL_955600|2020-12-16	GH	B.1.427	USA	B.1.427-US

Table 8 continued

249	hCoV-19/USA/CA-CZB-17562/2020|EPI_ISL_955609|2020-12-16	GH	B.1.427	USA	B.1.427-US
250	hCoV-19/USA/CA-CZB-17583/2020|EPI_ISL_955624|2020-12-18	GH	B.1.427	USA	B.1.427-US
251	hCoV-19/USA/CA-CZB-17584/2020|EPI_ISL_955625|2020-12-18	GH	B.1.427	USA	B.1.427-US
252	hCoV-19/USA/CA-CZB-17586/2020|EPI_ISL_955626|2020-12-18	GH	B.1.427	USA	B.1.427-US
253	hCoV-19/USA/CA-CZB-17587/2020|EPI_ISL_955627|2020-12-18	GH	B.1.427	USA	B.1.427-US
254	hCoV-19/USA/CA-CZB-19323/2020|EPI_ISL_955713|2020-12-28	GH	B.1.427	USA	B.1.427-US
255	hCoV-19/USA/CA-CZB-19357/2020|EPI_ISL_955742|2020-12-30	GH	B.1.427	USA	B.1.427-US
256	hCoV-19/USA/CA-CZB-19363/2020|EPI_ISL_955748|2020-12-29	GH	B.1.427	USA	B.1.427-US
257	hCoV-19/USA/CA-CZB-19394/2021|EPI_ISL_955775|2021-01-04	GH	B.1.427	USA	B.1.427-US
258	hCoV-19/USA/CA-RD581-CV0013/2021|EPI_ISL_1068862|2021-02-05	GH	B.1.429	California	B.1.429 -US
259	hCoV-19/USA/CA-RD581-CV0016/2021|EPI_ISL_1068865|2021-02-05	GH	B.1.429	California	B.1.429 -US
260	hCoV-19/USA/CA-RD581-CV0017/2021|EPI_ISL_1068866|2021-02-05	GH	B.1.429	California	B.1.429 -US
261	hCoV-19/USA/CA-RD581-CV0018/2021|EPI_ISL_1068867|2021-02-05	GH	B.1.429	California	B.1.429 -US
262	hCoV-19/USA/CA-RD581-CV0019/2021|EPI_ISL_1068868|2021-02-05	GH	B.1.429	California	B.1.429 -US
263	hCoV-19/USA/CA-RD581-CV0020/2021|EPI_ISL_1068869|2021-02-05	GH	B.1.429	California	B.1.429 -US
264	hCoV-19/USA/CA-RD581-CV0021/2021|EPI_ISL_1068870|2021-02-05	GH	B.1.429	California	B.1.429 -US
265	hCoV-19/USA/CA-RD581-CV0024/2021|EPI_ISL_1068871|2021-02-05	GH	B.1.429	California	B.1.429 -US
266	hCoV-19/USA/CA-RD581-CV0026/2021|EPI_ISL_1068873|2021-02-05	GH	B.1.429	California	B.1.429 -US
267	hCoV-19/USA/CA-RD581-CV0027/2021|EPI_ISL_1068874|2021-02-05	GH	B.1.429	California	B.1.429 -US
268	hCoV-19/USA/NY-NP-3987/2020|EPI_ISL_1080799|2020-12	GH	B.1.526	New York	B.1.526-US
269	hCoV-19/USA/NY-NP-4053/2020|EPI_ISL_1080800|2020-12	GH	B.1.526	New York	B.1.526-US
270	hCoV-19/USA/NY-NP-4698/2021|EPI_ISL_1080801|2021-01	GH	B.1.526	New York	B.1.526-US
271	hCoV-19/USA/NY-NP-4916/2021|EPI_ISL_1080802|2021-02	GH	B.1.526	New York	B.1.526-US
272	hCoV-19/USA/NY-NP-4974/2021|EPI_ISL_1080803|2021-01	GH	B.1.526	New York	B.1.526-US
273	hCoV-19/USA/NY-NP-5102/2021|EPI_ISL_1080804|2021-02	GH	B.1.526	New York	B.1.526-US
274	hCoV-19/USA/NY-NP-5107/2021|EPI_ISL_1080805|2021-02	GH	B.1.526	New York	B.1.526-US
275	hCoV-19/USA/NY-NP-5113/2021|EPI_ISL_1080806|2021-02	GH	B.1.526	New York	B.1.526-US
276	hCoV-19/USA/NY-NP-5140/2021|EPI_ISL_1080807|2021-02	GH	B.1.526	New York	B.1.526-US
277	hCoV-19/USA/NY-NP-5197/2021|EPI_ISL_1080808|2021-02	GH	B.1.526	New York	B.1.526-US

Table 9 Information about the $N = 307$ GS

278	hCoV-19/India/MH-NEERI-NGP-32956/2021|EPI_ISL_1360352|2021-01-30	G	VUI B.1.617	India	VUI B.1.617.1-IN
279	hCoV-19/India/MH-NEERI-NGP-34282/2021|EPI_ISL_1360359|2021-02-08	G	VUI B.1.617	India	VUI B.1.617.1-IN
280	hCoV-19/India/MH-NEERI-NGP-34605/2021|EPI_ISL_1360361|2021-02-10	G	VUI B.1.617	India	VUI B.1.617.1-IN
281	hCoV-19/India/MH-NEERI-NGP-34972/2021|EPI_ISL_1360364|2021-02-13	G	VUI B.1.617	India	VUI B.1.617.1-IN
282	hCoV-19/India/MH-NEERI-NGP-36241/2021|EPI_ISL_1360370|2021-02-19	G	VUI B.1.617	India	VUI B.1.617.1-IN
283	hCoV-19/India/MH-NEERI-NGP-36444/2021|EPI_ISL_1360371|2021-02-20	G	VUI B.1.617	India	VUI B.1.617.1-IN
284	hCoV-19/India/MH-NEERI-NGP-37270/2021|EPI_ISL_1360375|2021-02-24	G	VUI B.1.617	India	VUI B.1.617.1-IN
285	hCoV-19/India/MH-NEERI-NGP-39017/2021|EPI_ISL_1360382|2021-03-03	G	VUI B.1.617	India	VUI B.1.617.1-IN
286	hCoV-19/India/ILSGS00182/2021|EPI_ISL_1372011|2021-02-26	G	VUI B.1.617	India	VUI B.1.617.1-IN
287	hCoV-19/India/ILSGS00308/2020|EPI_ISL_1372093|2020-12-01	G	VUI B.1.617	India	VUI B.1.617.1-IN
i	Serial	Clade	Variant	Country	Acronym
288	hCoV-19/India/MH-ILSGS01660/2021|EPI_ISL_2342964|2021-04-04	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
289	hCoV-19/India/JH-ILSGS01687/2021|EPI_ISL_2342972|2021-04-07	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
290	hCoV-19/India/JH-ILSGS01694/2021|EPI_ISL_2342973|2021-04-08	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
291	hCoV-19/India/JH-ILSGS01695/2021|EPI_ISL_2342974|2021-04-10	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
292	hCoV-19/India/JH-ILSGS01701/2021|EPI_ISL_2342975|2021-04-14	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
293	hCoV-19/India/JH-ILSGS01706/2021|EPI_ISL_2342976|2021-04-14	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
294	hCoV-19/India/JH-ILSGS01715/2021|EPI_ISL_2342977|2021-04-13	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
295	hCoV-19/India/JH-ILSGS01718/2021|EPI_ISL_2342978|2021-04-12	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
296	hCoV-19/India/JH-ILSGS01736/2021|EPI_ISL_2342979|2021-04-08	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
297	hCoV-19/India/OR-ILSGS01781/2021|EPI_ISL_2342980|2021-04-20	G	VUI B.1.617.2	India	VUI B.1.617.2-IN
298	hCoV-19/India/MH-SEQ-221_S66_R1_001/2021|EPI_ISL_1939891|2021-03-22 _S66_R1_001/2021|EPI_ISL_1939891|2021-03-22	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
299	hCoV-19/India/GJ-ICMR-NIV-INSAGOG-GSEQ-3117/2021|EPI _ISL_1970498|2021-04-23	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
300	hCoV-19/India/MH-NCCS-119281/2021|EPI_ISL_2107058|2021-03-22	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
301	hCoV-19/India/WB-1931300249962/2021|EPI_ISL_2132715|2021-03-23	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
302	hCoV-19/India/AP-CCMB-CIA1346/2021|EPI_ISL_2231661|2021-04-21	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
303	hCoV-19/India/MH-ILSGS01543/2021|EPI_ISL_2342916|2021-03-24	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
304	hCoV-19/India/MH-ILSGS01566/2021|EPI_ISL_2342925|2021-03-24	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
305	hCoV-19/India/MH-ILSGS01585/2021|EPI_ISL_2342933|2021-03-29	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
306	hCoV-19/India/MH-ILSGS01643/2021|EPI_ISL_2342949|2021-04-03	G	VUI B.1.617.3	India	VUI B.1.617.3-IN
307	hCoV-19/India/MH-ILSGS01658/2021|EPI_ISL_2342962|2021-04-04	G	VUI B.1.617.3	India	VUI B.1.617.3-IN

Acknowledgements The authors thank all those who have contributed and shared sequences to the GISAID database available at https://www.gisaid.org/.

Data availability The reformatted data that support the findings of this study are available in http://ave.dee.isep.ipp.pt/jtm/FILES/datasets/Archive.zip.

Appendix

This appendix describes the metrics adopted in the paper. The Kolmogorov complexity and Shannon information theories are revisited and several distances are recalled. Additionally, the Hamming distance is also included.

Kolmogorov complexity

The Kolmogorov complexity tackles the assessment of information without considering probabilistic notions [57]. Given a string x, the Kolmogorov complexity can be defined as 'the length of the shortest program that, given an empty string ψ, can compute x and then halts'. Consequently, the Kolmogorov complexity can be somehow viewed as the length of the compressed form of the string x.

The information distance between two strings [22, 23] leads to concepts such as, normalized information distance (NCD) [58,59], Chen-Li Metric (CLM) [60–62], Compression-based Dissimilarity Measure (CDM) [63] and Compression-based Cosine ($CosS$) [54], given by:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{2a}$$

$$CLM(x, y) = 1 - \frac{C(x) - C(x|y)}{C(xy)}, \tag{2b}$$

$$CDM(x, y) = \frac{C(xy)}{C(x) + C(y)}, \tag{2c}$$

$$CosS(x, y) = 1 - \frac{C(x) + C(y) - C(xy)}{\sqrt{C(x)C(y)}}. \tag{2d}$$

The operator $C(x)$ represents the length (measured as a number of bits) of string x after being compressed by a given algorithm $C(\cdot)$. The concatenation of two strings, x and y, is denoted as xy and, therefore, $C(xy)$ gives the number of bits when compressing together x and y. Moreover, the notation $C(x|y)$ stands for the length of x when conditionally compressed with a model built on string y.

Sculley and Brodley [54] noted that the compression has an associated feature space, $\mathbb{N}$. Therefore, they map a string x into a vector $\bar{x} \in \mathbb{N}$ and $C(x) = ||\bar{x}||_1$. After some simplifications, we verify that the four compression distances (2) use the expression $||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1$ since we can write

$$NCD(x, y) = 1 - \frac{||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1}{\max\left(||\bar{x}||_1, ||\bar{y}||_1\right)}, \tag{3a}$$

$$CLM(x, y) = 1 - \frac{||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1}{||\bar{x} + \bar{y}||_1}, \tag{3b}$$

$$CDM(x, y) = 1 - \frac{||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1}{||\bar{x}||_1 + ||\bar{y}||_1}, \tag{3c}$$

$$CosS(x, y) = 1 - \frac{||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1}{\sqrt{||\bar{x}||_1 \, ||\bar{y}||_1}}. \tag{3d}$$

The four indices reduce to the form $1 - \frac{||\bar{x}||_1 + ||\bar{y}||_1 - ||\bar{x} + \bar{y}||_1}{f(\bar{x}, \bar{y})}$, where the normalizing term $f(\bar{x}, \bar{y})$ varies for each case.

Shannon information

In the scope of the Information theory [64], the so-called information content I of an event x_i with probability $P(X = x_i)$ is defined as:

$$I(X = x_i) = -\ln(P(X = x_i)), \tag{4}$$

where X denotes a discrete random variable.

The expected value of the information, $E(I)$, named Shannon entropy [65,66], becomes:

$$H(X) = E(-\ln(X)) = -\sum_i P(X = x_i) \ln(P(X = x_i)), \tag{5}$$

that follows the four Khinchin axioms [67,68].

In the case of a two-dimensional distribution (X_1, X_2) we have the joint entropy $H(X_1, X_2)$ and mutual information $MI(X_1, X_2)$ given by:

$$H(X_1, X_2) = \sum_i \sum_j P\left(X_1 = x_i, X_2 = x_j\right) \ln\left(P\left(X_1 = x_i, X_2 = x_j\right)\right), \tag{6}$$

$$MI(X_1, X_2) = \sum_i \sum_j P\left(X_1 = x_i, X_2 = x_j\right) \ln\left(\frac{P\left(X_1 = x_i, X_2 = x_j\right)}{P(X_1 = x_i) \, P\left(X = x_j\right)}\right). \tag{7}$$

Several distances can be defined from these mathematical tools. In the follow-up we adopt the Jaccard and Jensen-Shannon distances, $d_{Ja}(X_1, X_2)$ and $d_{JS}(X_1 \parallel X_2)$ respectively.

The Jaccard distance represents a set-theoretic interpretation of information based on the normalized distance $\frac{MI(X_1, X_2)}{H(X_1, X_2)}$ and is formulated as:

$$d_{Ja}(X_1, X_2) = 1 - \frac{MI(X_1, X_2)}{H(X_1, X_2)}. \quad (8)$$

The Jensen-Shannon divergence is the symmetric version of the Kullback-Leibler divergence $d_{KL}(X_1 \parallel X_{12})$ and is given by:

$$d_{JS}(X_1, X_2) = \frac{1}{2}[d_{KL}(X_1 \parallel X_{12}) + d_{KL}(X_2 \parallel X_{12})], \quad (9)$$

where $X_{12} = \frac{1}{2}(X_1 + X_2)$.

The expression of $d_{JS}(X_1, X_2)$ can be rewritten as:

$$\begin{aligned} & d_{JS}(X_1, X_2) \\ & = \frac{1}{2}\sum_i P(X_1 = x_i) \ln(P(X_1 = x_i)) \\ & + \frac{1}{2}\sum_i P(X_2 = x_i) \ln(P(X_2 = x_i)) \\ & - \sum_i P(X_{12} = x_i) \ln(P(X_{12} = x_i)). \end{aligned} \quad (10)$$

In the Shannon's entropy family we include also the Jeffreys (d_{Je}) and Topsøe (d_{To}) expressions given by:

$$\begin{aligned} & d_{Je}(X_1, X_2) \\ & = \sum_i (P(X_1 = x_i) - P(X_2 = x_i)) \ln \\ & \left(\frac{P(X_1 = x_i)}{P(X_2 = x_i)}\right), \end{aligned} \quad (11)$$

$$\begin{aligned} & d_{To}(X_1, X_2) \\ & = \sum_i [P(X_1 = x_i) \ln \\ & \left(\frac{2P(X_1 = x_i)}{P(X_1 = x_i) + P(X_2 = x_i)}\right) + \\ & P(X_2 = x_i) \ln\left(\frac{2P(X_2 = x_i)}{P(X_1 = x_i) + P(X_2 = x_i)}\right)]. \end{aligned} \quad (12)$$

Hamming distance

We adopt also a distance based on standard concepts for comparing the symbols forming the genetic sequences (GS). In information processing, the Hamming metric between two strings of equal length is given by the number of positions at which the corresponding symbols are different [26,69]. Therefore, we start by considering triplets of k_s consecutive symbols in each string. This means that we form sub-strings with k_s successive symbols in the DNA strand, so that for a genetic sequence x with length L_x we obtain $n_x = \left\lfloor \frac{L_x}{3} \right\rfloor$ of k_s-tuples. The assessment of similarity between the k-th character, $k = 1, \ldots, k_s$, in the i-th pair of sub-strings (i.e., the x_k and y_k symbols in the two sub-strings i extracted from the x and y sequences), is then based on the concept of Hamming metric:

$$\delta_i(x_k, y_k) = \begin{cases} 1, \text{ if } x_k = y_k \\ 0, \text{ if } x_k \neq y_k \end{cases}, \quad (13)$$

where the indices k and i stand for the k-th symbol in the i-th sub-string ($k = 1, \ldots, k_s$, and $i = 1, \ldots, n_x$), respectively. In general the strings x and y have slightly different lengths and we simply adopt the length $n_{xy} = \min(n_x, n_y)$. For calculating the total distance between two strings we test the 'normalized' version of the Hamming distance:

$$d_{Ha}(x, y) = \frac{1}{n_{xy}} \sum_{i=1}^{n_{xy}} \sum_{k=1}^{k_s} \overline{\delta_i(x_k, y_k)}, \quad (14)$$

where $\overline{\delta_i(x_k, y_k)} = 1 - \delta_i(x_k, y_k)$ and the term 'normalization' means that the final values are given in relation to the total number of compared sub-strings, n_{xy}.

References

1. Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y., Mills, M.C.: Demographic science aids in understanding the spread and fatality rates of COVID-19. Proc. Natl. Acad. Sci. **117**(18), 9696–9698 (2020). https://doi.org/10.1073/pnas.2004911117
2. Mercatelli, D., Giorgi, F.M.: Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. (2020). https://doi.org/10.3389/fmicb.2020.01800
3. Ceraolo, C., Giorgi, F.M.: Genomic variance of the 2019-nCoV coronavirus. J. Med. Virol. **92**(5), 522–528 (2020). https://doi.org/10.1002/jmv.25700
4. Mallapaty, S.: COVID mink analysis shows mutations are not dangerous - yet. Nature **587**(7834), 340–341 (2020). https://doi.org/10.1038/d41586-020-03218-z
5. Hamed, S.M., Elkhatib, W.F., Khairalla, A.S., Noreddin, A.M.: Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Sci. Rep. (2021). https://doi.org/10.1038/s41598-021-87713-x
6. Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D.L., (on behalf of COVID-19 Genomics Consortium UK (CoG-UK)9, E.V.: Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characteri

sation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020). [Online; posted 18-December-2020]
7. Frampton, D., Rampling, T., Cross, A., Bailey, H., Heaney, J., Byott, M., Scott, R., Sconza, R., Price, J., Margaritis, M., Bergstrom, M., Spyer, M.J., Miralhes, P.B., Grant, P., Kirk, S., Valerio, C., Mangera, Z., Prabhahar, T., Moreno-Cuesta, J., Arulkumaran, N., Singer, M., Shin, G.Y., Sanchez, E., Paraskevopoulou, S.M., Pillay, D., McKendry, R.A., Mirfenderesky, M., Houlihan, C.F., Nastouli, E.: Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. The Lancet Infectious Diseases (2021). https://doi.org/10.1016/s1473-3099(21)00170-5
8. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, M., Korsman, S., Davies, M.A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Pond, S.L.K., Weaver, S., Martin, D., Lessells, R.J., Bhiman, J.N., Williamson, C., de Oliveira, T.: Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv (2020). https://doi.org/10.1101/2020.12.21.20248640. https://www.medrxiv.org/content/early/2020/12/22/2020.12.21.20248640
9. Long, S.W., Olsen, R.J., Christensen, P.A., Subedi, S., Olson, R., Davis, J.J., Saavedra, M.O., Yerramilli, P., Pruitt, L., Reppond, K., Shyer, M.N., Cambric, J., Finkelstein, I.J., Gollihar, J., Musser, J.M.: Sequence analysis of 20, 453 severe acute respiratory syndrome coronavirus 2 genomes from the Houston metropolitan area identifies the emergence and widespread distribution of multiple isolates of all major variants of concern. Am. J. Pathol. (2021). https://doi.org/10.1016/j.ajpath.2021.03.004
10. Davies, N., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A., Russell, T.W., Tully, D.C., Abbott, S., Gimma, A., Waites, W., Wong, K.L., van Zandvoort, K., working group, C., Eggo, R.M., Funk, S., Jit, M., Atkins, K.E., Edmunds, W.J.: Estimated transmissibility and severity of novel SARS-CoV-2 variant of concern 2020/12/01 in England. https://cmmid.github.io/topics/covid19/uk-novel-variant.html (2020). [First online: 23-12-2020, Last update: 03-03-2021]
11. Brown, C.M., Vostok, J., Johnson, H., Burns, M., Gharpure, R., Sami, S., Sabo, R.T., Hall, N., Foreman, A., Schubert, P.L., Gallagher, G.R., Fink, T., Madoff, L.C., Gabriel, S.B., MacInnis, B., Park, D.J., Siddle, K.J., Harik, V., Arvidson, D., Brock-Fisher, T., Dunn, M., Kearns, A., Laney, A.S.: Outbreak of SARS-CoV-2 infections, including COVID-19 vaccine breakthrough infections, associated with large public gatherings – Barnstable county, Massachusetts, July 2021. MMWR. Morbidity and Mortality Weekly Report **70**(31) (2021). https://doi.org/10.15585/mmwr.mm7031e2
12. Li, B., Deng, A., Li, K., Hu, Y., Li, Z., Xiong, Q., Liu, Z., Guo, Q., Zou, L., Zhang, H., Zhang, M., Ouyang, F., Su, J., Su, W., Xu, J., Lin, H., Sun, J., Peng, J., Jiang, H., Zhou, P., Hu, T., Luo, M., Zhang, Y., Zheng, H., Xiao, J., Liu, T., Che, R., Zeng, H., Zheng, Z., Huang, Y., Yu, J., Yi, L., Wu, J., Chen, J., Zhong, H., Deng, X., Kang, M., Pybus, O.G., Hall, M., Lythgoe, K.A., Li, Y., Yuan, J., He, J., Lu, J.: Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 delta variant (2021). https://doi.org/10.1101/2021.07.07.21260122
13. Allen, H., Vusirikala, A., Flannagan, J., Twohig, K.A., Zaidi, A., Groves, N., Lopez-Bernal, J., Harris, R., Charlett, A., Dabrera, G., Kall, M.: Increased household transmission of COVID-19 cases associated with SARS-CoV-2 variant of concern B.1.617.2: a national case-control study (2021). https://www.gov.uk/government/collections/new-sars-cov-2-variant
14. Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J.L., Marc, G.P., Moreira, E.D., Zerbini, C., Bailey, R., Swanson, K.A., Roychoudhury, S., Koury, K., Li, P., Kalina, W.V., Cooper, D., Frenck, R.W., Hammitt, L.L., Türeci, Özlem., Nell, H., Schaefer, A., Ünal, S., Tresnan, D.B., Mather, S., Dormitzer, P.R., Şahin, U., Jansen, K.U., Gruber, W.C.: Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. New Eng. J. Med. (2020). https://doi.org/10.1056/nejmoa2034577
15. Voysey, M., Clemens, S.A.C., Madhi, S.A., Weckx, L.Y., Folegatti, P.M., Aley, P.K., Angus, B., Baillie, V.L., Barnabas, S.L., Bhorat, Q.E., Bibi, S., Briner, C., Cicconi, P., Collins, A.M., Colin-Jones, R., Cutland, C.L., Darton, T.C., Dheda, K., Duncan, C.J.A., Emary, K.R.W., Ewer, K.J., Fairlie, L., Faust, S.N., Feng, S., Ferreira, D.M., Finn, A., Goodman, A.L., Green, C.M., Green, C.A., Heath, P.T., Hill, C., Hill, H., Hirsch, I., Hodgson, S.H.C., Izu, A., Jackson, S., Jenkin, D., Joe, C.C.D., Kerridge, S., Koen, A., Kwatra, G., Lazarus, R., Lawrie, A.M., Lelliott, A., Libri, V., Lillie, P.J., Mallory, R., Mendes, A.V.A., Milan, E.P., Minassian, A.M., McGregor, A., Morrison, H., Mujadidi, Y.F., Nana, A., O'Reilly, P.J., Padayachee, S.D., Pittella, A., Plested, E., Pollock, K.M., Ramasamy, M.N., Rhead, S., Schwarzbold, A.V., Singh, N., Smith, A., Song, R., Snape, M.D., Sprinz, E., Sutherland, R.K., Tarrant, R., Thomson, E.C., Török, M.E., Toshner, M., Turner, D.P.J., Vekemans, J., Villafana, T.L., Watson, M.E.E., Williams, C.J., Douglas, A.D., Hill, A.V.S., Lambe, T., Gilbert, S.C., Pollard, A.J., Aban, M., Abayomi, F., Abeyskera, K., Aboagye, J., Adam, M., Adams, K., Adamson, J., Adelaja, Y.A., Adlou, S., Ahmed, K., Akhalwaya, Y., Akhalwaya, S., Alcock, A., Ali, A., Allen, E.R., Allen, L., Almeida, T.C.D.S.C., Alves, M.P., Amorim, F., Andritsou, F., Anslow, R., Appleby, M., Arbe-Barnes, E.H., Ariaans, M.P., Arns, B., Arruda, L., Awedetan, G., Azi, P., Azi, L., Babbage, G., Bailey, C., Baker, K.F., Baker, M., Baker, N., Baker, P., Baldwin, L., Baleanu, I., Bandeira, D., Bara, A., Barbosa, M.A., Barker, D., Barlow, G.D., Barnes, E., Barr, A.S., Barrett, J.R., Barrett, J., Bates, L., Batten, A., Beadon, K., Beales, E., Beckley, R., Belij-Rammerstorfer, S., Bell, J., Bellamy, D., Bellei, N., Belton, S., Berg, A., Bermejo, L., Berrie, E., Berry, L., Berzenyi, D., Beveridge, A., Bewley, K.R., Bexhell, H., Bhikha, S., Bhorat, A.E., Bhorat, Z.E., Bijker, E., Birch, G., Birch, S., Bird, A., Bird, O., Bisnauthsing, K., Bittaye, M., Blackstone, K.,

Blackwell, L., Bletchly, H., Blundell, C.L., Blundell, S.R., Bodalia, P., Boettger, B.C., Bolam, E., Boland, E., Bormans, D., Borthwick, N., Bowring, F., Boyd, A., Bradley, P., Brenner, T., Brown, P., Brown, C., Brown-O-Sullivan, C., Bruce, S., Brunt, E., Buchan, R., Budd, W., Bulbulia, Y.A., Bull, M., Burbage, J., Burhan, H., Burn, A., Buttigieg, K.R., Byard, N., Puig, I.C., Calderon, G., Calvert, A., Camara, S., Cao, M., Cappuccini, F., Cardoso, J.R., Carr, M., Carroll, M.W., Carson-Stevens, A., de M. Carvalho, Y., Carvalho, J.A., Casey, H.R., Cashen, P., Castro, T., Castro, L.C., Cathie, K., Cavey, A., Cerbino-Neto, J., Chadwick, J., Chapman, D., Charlton, S., Chelysheva, I., Chester, O., Chita, S., Cho, J.S., Cifuentes, L., Clark, E., Clark, M., Clarke, A., Clutterbuck, E.A., Collins, S.L., Conlon, C.P., Connarty, S., Coombes, N., Cooper, C., Cooper, R., Cornelissen, L., Corrah, T., Cosgrove, C., Cox, T., Crocker, W.E., Crosbie, S., Cullen, L., Cullen, D., Cunha, D.R., Cunningham, C., Cuthbertson, F.C., Guarda, S.N.F.D., da Silva, L.P., Damratoski, B.E., Danos, Z., Dantas, M.T., Darroch, P., Datoo, M.S., Datta, C., Davids, M., Davies, S.L., Davies, H., Davis, E., Davis, J., Davis, J., Nobrega, M.M.D., Kalid, L.M.D.O., Dearlove, D., Demissie, T., Desai, A., Marco, S.D., Maso, C.D., Dinelli, M.I., Dinesh, T., Docksey, C., Dold, C., Dong, T., Donnellan, F.R., Santos, T.D., dos Santos, T.G., Santos, E.P.D., Douglas, N., Downing, C., Drake, J., Drake-Brockman, R., Driver, K., Drury, R., Dunachie, S.J., Durham, B.S., Dutra, L., Easom, N.J., van Eck, S., Edwards, M., Edwards, N.J., Muhanna, O.M.E., Elias, S.C., Elmore, M., English, M., Esmail, A., Essack, Y.M., Farmer, E., Farooq, M., Farrar, M., Farrugia, L., Faulkner, B., Fedosyuk, S., Felle, S., Feng, S., Silva, C.F.D., Field, S., Fisher, R., Flaxman, A., Fletcher, J., Fofie, H., Fok, H., Ford, K.J., Fowler, J., Fraiman, P.H., Francis, E., Franco, M.M., Frater, J., Freire, M.S., Fry, S.H., Fudge, S., Furze, J., Fuskova, M., Galian-Rubio, P., Galiza, E., Garlant, H., Gavrila, M., Geddes, A., Gibbons, K.A., Gilbride, C., Gill, H., Glynn, S., Godwin, K., Gokani, K., Goldoni, U.C., Goncalves, M., Gonzalez, I.G., Goodwin, J., Goondiwala, A., Gordon-Quayle, K., Gorini, G., Grab, J., Gracie, L., Greenland, M., Greenwood, N., Greffrath, J., Groenewald, M.M., Grossi, L., Gupta, G., Hackett, M., Hallis, B., Hamaluba, M., Hamilton, E., Hammersley, D., Hanrath, A.T., Hanumunthadu, B., Harris, S.A., Harris, C., Harris, T., Harrison, T.D., Harrison, D., Hart, T.C., Hartnell, B., Hassan, S., Haughney, J., Hawkins, S., Hay, J., Head, I., Henry, J., Herrera, M.H., Hettle, D.B., Hill, J., Hodges, G., Horne, E., Hou, M.M., Houlihan, C., Howe, E., Howell, N., Humphreys, J., Humphries, H.E., Hurley, K., Huson, C., Hyder-Wright, A., Hyamns, C., Ikram, S., Ishwarbhai, A., Ivan, M., Iveson, P., Iyer, V., Jackson, F., Jager, J.D., Jaumdally, S., Jeffers, H., Jesudason, N., Jones, B., Jones, K., Jones, E., Jones, C., Jorge, M.R., Jose, A., Joshi, A., Júnior, E.A., Kadziola, J., Kailath, R., Kana, F., Karampatsas, K., Kasanyinga, M., Keen, J., Kelly, E.J., Kelly, D.M., Kelly, D., Kelly, S., Kerr, D., de Ávila Kfouri, R., Khan, L., Khozoee, B., Kidd, S., Killen, A., Kinch, J., Kinch, P., King, L.D., King, T.B., Kingham, L., Klenerman, P., Knapper, F., Knight, J.C., Knott, D., Koleva, S., Lang, M., Lang, G., Larkworthy, C.W., Larwood, J.P., Law, R., Lazarus, E.M., Leach, A., Lees, E.A., Lemm, N.M., Lessa, A., Leung, S., Li, Y., Lias, A.M., Liatsikos, K., Linder, A., Lipworth, S., Liu, S., Liu, X., Lloyd, A., Lloyd, S., Loew, L., Ramon, R.L., Lora, L., Lowthorpe, V., Luz, K., MacDonald, J.C., MacGregor, G., Madhavan, M., Mainwaring, D.O., Makambwa, E., Makinson, R., Malahleha, M., Malamatsho, R., Mallett, G., Mansatta, K., Maoko, T., Mapetla, K., Marchevsky, N.G., Marinou, S., Marlow, E., Marques, G.N., Marriott, P., Marshall, R.P., Marshall, J.L., Martins, F.J., Masenya, M., Masilela, M., Masters, S.K., Mathew, M., Matlebjane, H., Matshidiso, K., Mazur, O., Mazzella, A., McCaughan, H., McEwan, J., McGlashan, J., McInroy, L., McIntyre, Z., McLenaghan, D., McRobert, N., McSwiggan, S., Megson, C., Mehdipour, S., Meijs, W., Mendonça, R.N., Mentzer, A.J., Mirtorabi, N., Mitton, C., Mnyakeni, S., Moghaddas, F., Molapo, K., Moloi, M., Moore, M., Moraes-Pinto, M.I., Moran, M., Morey, E., Morgans, R., Morris, S., Morris, S., Morris, H.C., Morselli, F., Morshead, G., Morter, R., Mottal, L., Moultrie, A., Moya, N., Mpelembue, M., Msomi, S., Mugodi, Y., Mukhopadhyay, E., Muller, J., Munro, A., Munro, C., Murphy, S., Mweu, P., Myasaki, C.H., Naik, G., Naker, K., Nastouli, E., Nazir, A., Ndlovu, B., Neffa, F., Njenga, C., Noal, H., Noé, A., Novaes, G., Nugent, F.L., Nunes, G., O-Brien, K., O-Connor, D., Odam, M., Oelofse, S., Oguti, B., Olchawski, V., Oldfield, N.J., Oliveira, M.G., Oliveira, C., Oosthuizen, A., O-Reilly, P., Osborne, P., Owen, D.R., Owen, L., Owens, D., Owino, N., Pacurar, M., Paiva, B.V., Palhares, E.M., Palmer, S., Parkinson, S., Parracho, H.M., Parsons, K., Patel, D., Patel, B., Patel, F., Patel, K., Patrick-Smith, M., Payne, R.O., Peng, Y., Penn, E.J., Pennington, A., Alvarez, M.P.P., Perring, J., Perry, N., Perumal, R., Petkar, S., Philip, T., Phillips, D.J., Phillips, J., Phohu, M.K., Pickup, L., Pieterse, S., Piper, J., Pipini, D., Plank, M., Plessis, J.D., Pollard, S., Pooley, J., Pooran, A., Poulton, I., Powers, C., Presa, F.B., Price, D.A., Price, V., Primeira, M., Proud, P.C., Provstgaard-Morys, S., Pueschel, S., Pulido, D., Quaid, S., Rabara, R., Radford, A., Radia, K., Rajapaska, D., Rajeswaran, T., Ramos, A.S.F., Lopez, F.R., Rampling, T., Rand, J., Ratcliffe, H., Rawlinson, T., Rea, D., Rees, B., Reiné, J., Resuello-Dauti, M., Pabon, E.R., Ribiero, C.M., Ricamara, M., Richter, A., Ritchie, N., Ritchie, A.J., Robbins, A.J., Roberts, H., Robinson, R.E., Robinson, H., Rocchetti, T.T., Rocha, B.P., Roche, S., Rollier, C., Rose, L., Russell, A.L.R., Rossouw, L., Royal, S., Rudiansyah, I., Ruiz, S., Saich, S., Sala, C., Sale, J., Salman, A.M., Salvador, N., Salvador, S., Sampaio, M., Samson, A.D., Sanchez-Gonzalez, A., Sanders, H., Sanders, K., Santos, E., Guerra, M.F.S., Satti, I., Saunders, J.E., Saunders, C., Sayed, A., van der Loeff, I.S., Schmid, A.B., Schofield, E., Screaton, G., Seddiqi, S., Segireddy, R.R., Senger, R., Serrano, S., Shah, R., Shaik, I., Sharpe, H.E., Sharrocks, K., Shaw, R., Shea, A., Shepherd, A., Shepherd, J.G., Shiham, F., Sidhom, E., Silk, S.E., da Silva Moraes, A.C., Silva-Junior, G., Silva-Reyes, L., Silveira, A.D., Silveira, M.B., Sinha, J., Skelly, D.T., Smith, D.C., Smith, N., Smith, H.E., Smith, D.J., Smith, C.C., Soares, A., Soares, T., Solórzano, C., Sorio, G.L., Sorley, K., Sosa-Rodriguez, T., Souza, C.M., Souza, B.S., Souza, A.R., Spencer, A.J., Spina, F., Spoors, L., Stafford, L., Stamford, I., Starinskij, I., Stein, R., Steven, J., Stockdale, L., Stockwell, L.V., Strickland, L.H., Stuart, A.C., Sturdy, A., Sutton, N., Szigeti, A., Tahiri-Alaoui, A., Tanner, R., Taoushanis, C., Tarr, A.W., Taylor, K., Taylor, U., Taylor, I.J., Taylor, J., te Water Naude, R., Themistocleous, Y., Themistocleous, A., Thomas, M.,

Thomas, K., Thomas, T.M., Thombrayil, A., Thompson, F., Thompson, A., Thompson, K., Thompson, A., Thomson, J., Thornton-Jones, V., Tighe, P.J., Tinoco, L.A., Tiongson, G., Tladinyane, B., Tomasicchio, M., Tomic, A., Tonks, S., Tran, N., Tree, J., Trillana, G., Trinham, C., Trivett, R., Truby, A., Tsheko, B.L., Turabi, A., Turner, R., Turner, C., Ulaszewska, M., Underwood, B.R., Varughese, R., Verbart, D., Verheul, M., Vichos, I., Vieira, T., Waddington, C.S., Walker, L., Wallis, E., Wand, M., Warbick, D., Wardell, T., Warimwe, G., Warren, S.C., Watkins, B., Watson, E., Webb, S., Webb-Bridges, A., Webster, A., Welch, J., Wells, J., West, A., White, C., White, R., Williams, P., Williams, R.L., Winslow, R., Woodyer, M., Worth, A.T., Wright, D., Wroblewska, M., Yao, A., Zimmer, R., Zizi, D., Zuidewind, P.: Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. The Lancet (2020). https://doi.org/10.1016/s0140-6736(20)32661-1

16. Knoll, M.D., Wonodi, C.: Oxford-AstraZeneca COVID-19 vaccine efficacy. Lancet (2020). https://doi.org/10.1016/s0140-6736(20)32623-4
17. Machado, J.A.T., Lopes, A.M.: Rare and extreme events: the case of COVID-19 pandemic. Nonlinear Dyn. **100**(3), 2953–2972 (2020). https://doi.org/10.1007/s11071-020-05680-w
18. Lopes, A.M., Andrade, J.P., Machado, J.T.: Multidimensional scaling analysis of virus diseases. Comput. Methods Progr. Biomed. **131**, 97–110 (2016). https://doi.org/10.1016/j.cmpb.2016.03.029
19. Machado, J.A.T., Rocha-Neves, J.M., Andrade, J.P.: Computational analysis of the SARS-CoV-2 and other viruses based on the Kolmogorov's complexity and Shannon's information theories. Nonlinear Dyn. **101**(3), 1731–1750 (2020). https://doi.org/10.1007/s11071-020-05771-8
20. Machado, J.T., Lopes, A.M.: A computational perspective of the periodic table of elements. Commun. Nonlinear Sci. Num. Simul. **78**, 104883 (2019). https://doi.org/10.1016/j.cnsns.2019.104883
21. Machado, J.T., Lopes, A.M.: Multidimensional scaling and visualization of patterns in prime numbers. Commun. Nonlinear Sci. Num. Simul. **83**, 105128 (2020). https://doi.org/10.1016/j.cnsns.2019.105128
22. Bennett, C.H., Gács, P., Li, M., Vitányi, P., Zurek, W.H.: Information distance. IEEE Trans. Inf. Theory **44**(4), 1407–1423 (1998)
23. Fortnow, L., Lee, T., Vereshchagin, N.: Kolmogorov complexity with error. In: Durand, B., Thomas, W. (eds.) STACS 2006–23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, France, February 23–25, 2006. Lecture Notes in Computer Science, pp. 137–148. Springer, Berlin, Heidelberg (2006)
24. Cha, S.: Taxonomy of nominal type histogram distance measures. In: Proceedings of the American Conference on Applied Mathematics, pp. 325–330. Harvard, Massachusetts, USA (2008)
25. Deza, M.M., Deza, E.: Encyclopedia of distances. Springer-Verlag, Berlin, Heidelberg (2009)
26. Hamming, R.W.: Error detecting and error correcting codes. Bell Syst. Tech. J. **29**(2), 147–160 (1950). https://doi.org/10.1002/j.1538-7305.1950.tb00463.x
27. Cilibrasi, R., Vitany, P.M.B.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005). https://doi.org/10.1109/TIT.2005.844059
28. Yin, C., Chen, Y., Yau, S.S.T.: A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering complexity for DNA sequences. J. Theor. Biol. **359**, 18–28 (2014). https://doi.org/10.1016/j.jtbi.2014.05.043
29. Kubicova, V., Provaznik, I.: Relationship of bacteria using comparison of whole genome sequences in frequency domain. Inf. Technol. Biomed. **3**, 397–408 (2014). https://doi.org/10.1007/978-3-319-06593-9_35
30. Glunčić, M., Paar, V.: Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res. (2013). https://doi.org/10.1093/nar/gks721
31. Hautamaki, V., Pollanen, A., Kinnunen, T., Aik, K., Haizhou, L., Franti, L.: A comparison of categorical attribute data clustering methods, pp. 53–62. Berlin, Springer (2014). https://doi.org/10.1007/978-3-662-44415-3_6
32. Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.: The distance function effect on k-nearest neighbor classification for medical datasets. Springer Plus **5**,(2016). https://doi.org/10.1186/s40064-016-2941-7
33. Aziz, M., Alhadidi, D., Mohammed, N.: Secure approximation of edit distance on genomic data. BMC Med Genom. (2017). https://doi.org/10.1186/s12920-017-0279-9
34. Yianilos, P.N.: Normalized forms of two common metrics. Tech. Rep. Report 91–082-9027-1, NEC Research Institute (1991)
35. Yu, J., Amores, J., Sebe, N., Tian, Q.: A new study on distance metrics as similarity measurement. In: IEEE International Conference on Multimedia and Expo, pp. 533–536 (2006). https://doi.org/10.1109/ICME.2006.262443
36. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.): Feature extraction: foundations and applications. Springer, Berlin (2008)
37. Russel, R., Sinha, P.: Perceptually based comparison of image similarity metrics. Perception **40**, 1269–1281 (2011). https://doi.org/10.1068/p7063
38. Burrows, M., Wheeler, D.J.: A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation (1994)
39. Welch, T.: A technique for high-performance data compression. Computer **17**(6), 8–19 (1984). https://doi.org/10.1109/mc.1984.1659158
40. Kodituwakku, S.: Comparison of lossless data compression algorithms for text data. Indian J. Comput. Sci. Eng. **1**(4), 416–425 (2010)
41. Saeed, N., Haewoon, Imtiaz, Saqib, M.: A survey on multidimensional scaling. ACM Comput. Surv. (CSUR) **51**(3), 47 (2018). https://doi.org/10.1145/3178155
42. Hartigan, J.A.: Clustering algorithms. Wiley, London (1975)
43. Tenreiro Machado, J.A., Galhano, A.M.: Multidimensional scaling visualization using parametric similarity indices. Entropy **17**(4), 1775–1794 (2015). https://doi.org/10.3390/e17041775
44. Machado, J.A.T.: Relativistic time effects in financial dynamics. Nonlinear Dyn. **75**(4), 735–744 (2014). https://doi.org/10.1007/s11071-013-1100-8

45. Liébecq, C. (ed.): IUPAC-IUBMB Joint Commission on Biochemical Nomenclature and Nomenclature Commission of IUBMB. In: Biochemical Nomenclature and Related Documents. Portland Press (1992)
46. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2009). https://doi.org/10.1007/s11192-009-0146-3
47. Waltman, L., van Eck, N.J., Noyons, E.C.: A unified approach to mapping and clustering of bibliometric networks. J. Inf. **4**(4), 629–635 (2010). https://doi.org/10.1016/j.joi.2010.07.002
48. van Eck, N.J., Waltman, L.: Visualizing bibliometric networks. In: Measuring Scholarly Impact, pp. 285–320. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10377-8_13
49. Perianes-Rodriguez, A., Waltman, L., van Eck, N.J.: Constructing bibliometric networks: a comparison between full and fractional counting. J. Inf. **10**(4), 1178–1195 (2016). https://doi.org/10.1016/j.joi.2016.10.006
50. van Eck, N.J., Waltman, L.: Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics **111**(2), 1053–1070 (2017). https://doi.org/10.1007/s11192-017-2300-7
51. Machado, J.A.T.: Shannon information and power law analysis of the chromosome code. Abstr. Appl. Anal. **2012**, 1–13 (2012). https://doi.org/10.1155/2012/439089
52. Machado, J.A.T., Costa, A.C., Quelhas, M.D.: Can power laws help us understand gene and proteome information? Adv. Math. Phys. **2013**, 1–10 (2013). https://doi.org/10.1155/2013/917153
53. Machado, J.T.: Fractional order description of DNA. Appl. Math. Model. **39**(14), 4095–4102 (2015). https://doi.org/10.1016/j.apm.2014.12.037
54. Sculley, D., Brodley, C.: Compression and machine learning: a new perspective on feature space vectors, p. 332. IEEE (2006). https://doi.org/10.1109/dcc.2006.13
55. Felsenstein, J.: PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein (1993)
56. Tuimala, J.: A primer to phylogenetic analysis using the PHYLIP package. CSC - Scientific Computing Ltd., Finland (2006)
57. Kolmogorov, A.: Three approaches to the quantitative definition of information. Int. J. Comput. Math. **2**(1–4), 157–168 (1968)
58. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004). https://doi.org/10.1109/tit.2004.838101
59. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 863–872 (2004)
60. Chen, X., Kwong, S., Li, M.: A compression algorithm for DNA sequences and its applications in genome comparison. In: Genome Informatics: Proceedings of the 10th Workshop on Genome Informatics, pp. 51–61 (1999)
61. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics **17**(2), 149–154 (2001). https://doi.org/10.1093/bioinformatics/17.2.149
62. Chen, X., Francia, B., Li, M., McKinnon, B., Seker, A.: Shared information and program plagiarism detection. IEEE Trans. Inf. Theory **50**(7), 1545–1551 (2004). https://doi.org/10.1109/tit.2004.830793
63. Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Towards parameter-free data mining. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 206–215. ACM Press (2004). https://doi.org/10.1145/1014052.1014077
64. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 623–656 (1948)
65. Gray, R.M.: Entropy and information theory. Springer-Verlag, New York (2011)
66. Beck, C.: Generalised information and entropy measures in physics. Contemp. Phys. **50**(4), 495–510 (2009). https://doi.org/10.1080/00107510902823517
67. Khinchin, A.I.: Mathematical foundations of information theory. Dover, New York (1957)
68. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**(6), 620–630 (1957)
69. Pilcher, C.D., Wong, J.K., Pillai, S.K.: Inferring HIV transmission dynamics from phylogenetic sequence relationships. PLoS Med. **5**(3), e69 (2008). https://doi.org/10.1371/journal.pmed.0050069