medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Variant-driven multi-wave pattern of COVID-19 via Machine Learning clustering of spike protein mutations

Adele de Hoffer[1,+], Shahram Vatani[2,3,+], Corentin Cot[2,3,+], Giacomo Cacciapaglia[2,3,*], Francesco Conventi[4,5], Antonio Giannini[4,6], Stefan Hohenegger[2,3], and Francesco Sannino[4,6,7,8,*]

[1]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2]Institut de Physique des 2 Infinis (IP2I), CNRS/IN2P3, UMR5822, 69622 Villeurbanne, France
[3]Université de Lyon, Université Claude Bernard Lyon 1, 69001 Lyon, France
[4]INFN sezione di Napoli, Complesso Universitario di Monte S. Angelo Edificio 6, via Cintia, 80126 Napoli, Italy
[5]Universitá di Napoli Parthenope, Napoli NA, Italy
[6]Dipartimento di Fisica E. Pancini, Università di Napoli Federico II, Complesso Universitario di Monte S. Angelo Edificio 6, via Cintia, 80126 Napoli, Italy
[7]Scuola Superiore Meridionale, Largo S. Marcellino 10, 80138 Napoli NA, Italy
[8]CP3-Origins & the Danish Institute for Advanced Study, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark
[*]g.cacciapaglia@ipnl.in2p3.fr, sannino@cp3.sdu.dk
[+]these authors contributed equally to this work

ABSTRACT

Never before such a vast amount of data has been collected for any viral pandemic than for the current case of COVID-19. This offers the possibility to answer a number of highly relevant questions, regarding the evolution of the virus and the role mutations play in its spread among the population. We focus on spike proteins, as they bear the main responsibility for the effectiveness of the virus diffusion by controlling the interactions with the host cells. Using the available temporal structure of the sequencing data for the SARS-CoV-2 spike protein in the UK, we demonstrate that every wave of the pandemic is dominated by a different variant. Consequently, the time evolution of each variant follows a temporal structure encoded in the epidemiological Renormalisation Group approach to compartmental models. Machine learning is the tool of choice to determine the variants at play, independent of (but complementary to) the virological classification. Our Machine Learning algorithm on spike protein sequencing provides a simple and unbiased way to identify, classify and track relevant virus variants without any prior knowledge of their characteristics. Hence, we propose a new tool that can help preventing and forecasting the emergence of new waves, and that can be used by decision makers to define short and long term strategies to curb the current COVID-19 pandemic or future ones.

Highlights

- **Objectives** To study the relation between mutations of SARS-CoV-2, the emergence of relevant variants and the multi-wave pattern of the COVID-19 pandemic.
- **Setting** Genomic sequencing of the SARS-CoV-2 spike proteins in the UK nations (England, Scotland, Wales). Epidemiological data for the number of infections in the UK nations, South Africa, California and India.
- **Methodology** We designed a simple Machine Learning algorithm based on the Levenshtein distance on the spike protein sequences to cluster the available dataset and define variants. We set up a time-sensitive procedure that allows to define a variant as a chain of subsequent clusters. The Mutation epidemiological Renormalisation Group (MeRG) framework is used to describe the epidemiological data.
- **Results** Our analysis of the sequencing data from England, Wales and Scotland shows that:

    1. A Machine Learning analysis based only on the spike proteins allows to efficiently identify the variants of concern and of interest, as well as other variants relevant for the diffusion of the virus.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2. We identify a branching relation between variants, thus reconstructing the phylogeny of the main variants.

3. Comparison with the epidemiological data demonstrates that each new wave is dominated by a new emerging variant, thus confirming the hypothesis that there is a strong correlation between the emergence of variants and the multi-wave pattern.

4. The number of infected by each variant can be modelled via an independent logistic function (sigmoid), thus confirming the MeRG approach. Analyses of the epidemiological data for South Africa, California and India further corroborate this result.

- **Conclusions** Using a simple Machine Learning algorithm, we are able to identify, classify and track relevant virus variants without any prior knowledge of their characteristics. While our analysis is only based on spike protein sequencing and is unbiased, the results are validated by other informed methods based on the complete genome. By correlating the variant definition to epidemiological data, we discover that each new wave of the COVID-19 pandemic is driven and dominated by a new emerging variant, as identified by our Machine Learning analysis. The results are seminal to the development of a new strategy to study how SARS-CoV-2 variants emerge and to predict the characteristics of future mutations of the spike proteins. Furthermore, the same methodology can be applied to other viral diseases, like influenza, if sufficient sequencing data is available. Hence, we provide an effective and unbiased method to identify new emerging variants that can be responsible for the onset of a new epidemiological wave. Our Machine Learning strategy is, in fact, a new tool that can help preventing and forecasting the emergence of new waves, and it can be used by decision makers to define short and long term strategies to curb the current COVID-19 pandemic or future ones.

Introduction

It is of primary importance to understand the diffusion of a virus and its variants, especially in view of an efficient vaccination campaign. This task has been difficult in the past, mainly due to the scarce data available for extended pandemics caused by infectious diseases, like for instance the "Spanish" Influenza of 1918-19[1]. COVID-19 is revolutionising our understanding of pandemics because we have now access to real time data about, for example, the genome sequencing of the virus and its proteins. Among the latter, spike proteins play a special role, as they are responsible for the interaction between the virus and the host cells, and for the effectiveness of the virus in spreading and multiplying. Like other coronaviruses, the SARS-CoV-2 virus has relatively low mutation rates[2], nevertheless the current COVID-19 pandemic has seen the emergence of epidemiologically relevant variants. Genomic sequencing has allowed to track the mutations of the spike proteins, and to identify potentially dangerous variants[3,4] that may have an increased infectivity compared to the initial form. Since the second half of 2020, variants of concern (VoC) and of interest (VoI) have been identified in various regions of the world: for instance, the Alpha VoC (B.1.1.7, GRY), first identified in September 2020 in the UK[5,6]; the Beta VoC (B.1.351, GH/501Y.V2) first found in South Africa in May 2020[7]; the Gamma VoC (P.1, GR/501Y.V3) first detected in Brazil in November 2020[8], which has been spreading in Manaus notwithstanding the high rate of previous infections; the Delta VoC (B.1.617.2, G/478K.V1) identified in India in October 2020; and the Epsilon VoI (B.1.427+429, GH/452R.V1) found in California in March 2020[9]. An exhaustive list can be found on the WHO website (www.who.int/en/activities/tracking-SARS-CoV-2-variants/). Note that we follow the WHO naming scheme[10], while indicating in parenthesis the classification from the Pango lineage[11] and GISAID[12,13]. Considering the Alpha VoC as an example, it has been possible to study in the lab its infectious power, finding a higher rate of transmission by $67 \div 75\%$, compared to the previous ones[6]. The transmission advantage has been confirmed by epidemiological data in the UK[14]. Most analyses of the epidemiological data are done applying the time-honoured compartmental models of the SIR type[15–17], appropriately extended by including more compartments[18]. The main drawback in this approach is the large number of parameters, which need to be fixed by hand or extracted from the data. In this work, we want to bypass this difficulty by using a simplified and effective approach based on theoretical physics methods, the epidemic Renormalisation Group (eRG) framework[19–21], combined with information directly extracted from the spike protein sequencing via a simple Machine Learning approach. This novel method allowed us to analyse, at the same time, the variant structure of SARS-CoV-2 in multiple countries and regions of the world, and thus provide a direct comparison of their epidemiological impact. A theoretical analysis of the variants within the eRG framework is presented in a companion publication[22].

We first collect the spike protein sequencing data for the UK nations from the GISAID repository[12,13]. For each nation we analyse the data via a simple Machine Learning (ML) algorithm based on the Levenshtein measure (LM)[23,24], which quantifies the difference between two strings of text. This ML approach has been long used, greatly refined and optimised, in biology[25],

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

while more recently deep learning approaches[26–28] are becoming more effective. In the present work, we decided to rely on the original definition of LM in order to cluster protein sequences based on the number of mutations needed to connect them. Following this approach, we introduce a neat mathematical definition of a virus variant in terms of the LM across the various spike protein sequences. A variant is, therefore, defined as an ensemble of sequences that are relatively close to each other. In contrast, a mutation is a single change in the amino-acid sequence of the protein. Our approach is, therefore, complementary to that of the many packages used in computational biology[29–31], which focus on the alignment and similarity between the mRNA sequences and aim at reconstructing the phylogenic relation between them.

The procedure above is independently repeated for each geographical region in our study. We start by considering first the whole available dataset and then dividing the sequencing per month. This allows us to follow the evolution of the variants and construct chains connecting the clusters we identify by month. We validate our results by showing that our approach identifies the Alpha VoC in all the distinct UK regions we study. Once the dominant variants are identified we analyse their temporal spreading within the affected population. Given that only a small fraction of the infected individuals have their viral charge sampled and sequenced, we estimate the number of people infected by each variant by multiplying the number of positive tests by the rate of occurrence of each variant in the sequencing data. This rough approximation allows us to reliably extract the temporal evolution of each variant in the population. Note that each infected individual is, in practice, associated to the variant that is dominant in their viral charge, following the practice of the sequence reporting. Thus, the data we use track the time development of the dominance of each variant at the individual level.

To analyse the time evolution of the individuals infected by each variant, we employ the economical eRG approach[19] that allows to organise the pandemic waves according to temporal symmetry principles stemming from high energy physics[32,33]. The approach has been extensively tested[21,34], confronted with traditional SIR compartmental models[35], and, last but not least, summarised in a comprehensive review alongside other approaches[36]. The economy of the model rests in the fact that, once fixed the overall number of infected, the diffusion rate of the virus is encoded in a single parameter γ. The latter measures the speed at which the virus spreads in the population. This value can be extracted by fitting the new daily infected (or equivalently the cumulated number of infections). We remark that this approach can be put in correspondence to a SIR model with time-dependent reproduction number $R_0(t)$[35], which fits the data better than traditional compartmental models with constant parameters. We apply the eRG to each variant, thus yielding a classification of their aggressiveness via a single quantifier: their individual γ. A visual summary of the methodology followed by our analysis is shown in Fig. 1, while more details are reported in the supplementary material.

The main goal of this work is to understand the viral dynamics that characterises wave patterns stemming from infectious diseases like COVID-19. The eRG approach additionally offers a natural mathematical understanding in terms of dynamical flows of the system[37,38]. Importantly, by employing ML analysis to genomic data, we discover that each pandemic wave is driven by a single dominant variant. The findings demonstrate that the variant dynamics is one of the main engines behind the emergence of wave patterns for COVID-19. This result can be used as a template for similar infectious diseases. As direct consequence of our studies we arrive at a novel evolutionary model for the interpretation of the virus diffusion that is mutation driven.

Results

The spike protein sequences are extracted, country by country, from the GISAID repository: the data contains the full amino-acid structure of the protein and a date-stamp from the laboratory where the sampling was done. The latter allows us to study the time evolution of each variant. Note that each genome corresponds to the dominant mutation occurring in a single infected individual. The dataset is pruned to remove potentially incomplete sequences. We then apply our ML approach to cluster the various sequences in variants. First, we computed the LM between each pair of sequences, thus counting without any weight the minimal number of mutations (substitutions, deletions and insertions) that are needed to transition from one sequence to the other. Secondly, the algorithm constructs a tree of proximity by pairing sequences that are the closest to each other into a higher branch. To combine branches that contain more than one sequence, we use the Ward's method, after having checked that other choices do not significantly affect the results (more details in the supplementary material). The tree is completed when all sequences are grouped into a single cluster. To define the variants, we consider a cut in the distance so that branches whose Ward distance is larger than the cut are considered as separate variants. To keep the results simple, and free from biases, we base our analysis on the traditional Levenshtein distance between sequences, without introducing weights which are commonly adopted in computational biology. Furthermore, we apply the same cut to all branches. As England has the largest available sequencing sample, with 436.073 sequences as of the end of June, we will mainly focus on this dataset. This minimises statistical and sampling bias errors. After pruning, i.e. removing all sequences with less than 1.250 amino-acids and missing reported ones, 329.384 sequences are retained, out of which we identify 9.480 different ones.

As a first step, we apply the ML algorithm to all sequences. The clustering is done on the dataset that contains only different sequences for the spike proteins to reduce the number of data points. After defining the clustering, the number of total sequences

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

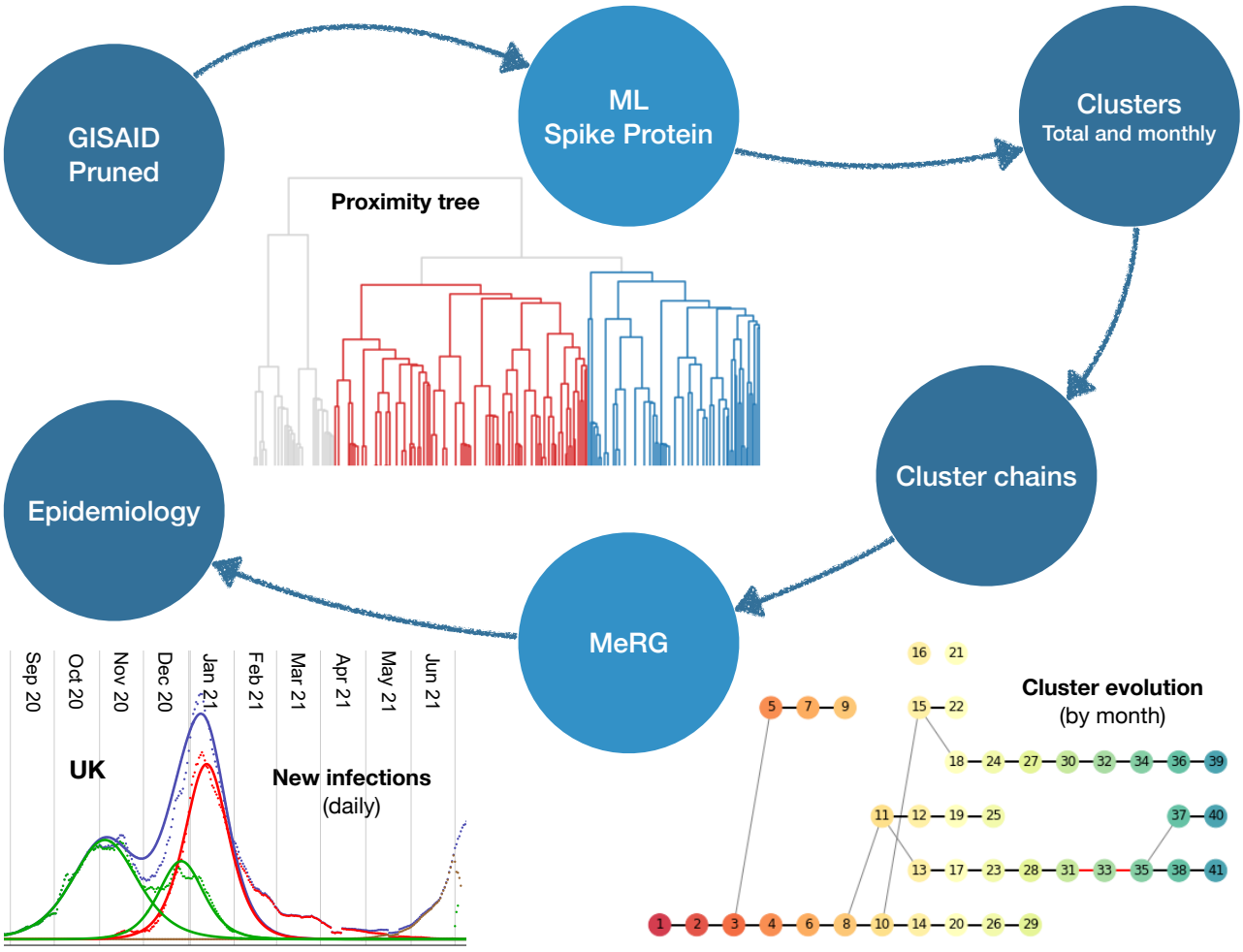

Figure 1. Methodology. Schematic representation of the methodology we follow.

in each cluster is counted, after pruning, to obtain the frequency for the occurrence of each cluster over time. The results are shown in Fig. 2 for two choices of the cut in the Ward distance r_W: $r_W = 50$ in the left plots and $r_W = 100$ in the right plots. We consider as relevant clusters only the ones that contain at least 1% of the sequences within the full dataset. We identify, therefore, 5 clusters for the lower cut (v0, v1a, v1b, v2, v3), and 4 for the upper cut (v0, v1, v2, v3). Increasing the cut allows to merge the clusters v1a and v1b into v1. Interestingly, the cluster v2 corresponds to the Alpha VoC: we have verified that the frequency of occurrence matches the one of the VoC tagging in the GISAID dataset, and furthermore checked that the most common sequence in the cluster v2 features the mutations associated with the Alpha VoC. Furthermore, the variant v3 corresponds to the Delta VoC, which is currently spreading in the UK. We also checked that by increasing the r_W cut, it is v0 and v1 that merge into a single variant, while v3 remains separated. In the middle plots of Fig. 2 we report our estimate of the number of daily new infections in each clustered variant, computed by multiplying the measured frequency in the sequencing dataset with the reported number of new infections. We observe that each wave can be associated with a different dominant variant: v0 for the first wave occurring in March-May 2020; v1 for the second wave in October-November 2020; v2 (the Alpha VoC) for the third wave in December-February 2021; and v3 for the last wave starting in May-June 2021. This feature supports the hypothesis that the occurrence of a new wave is related to the emergence of a new variant.

Before further analysing the data, we validated our method by comparing the variant definition, associated to the clusters in our analysis, to more standard methods used in computational biology. For this purpose, we have chosen the clade classification[40], as defined by the Nextstrain initiative[39] and embedded in the data from the GISAID repository. The Nextstrain clade definitions are informed by the statistical distribution of genome distances in phylogenetic clusters[41], followed by the merging of smaller lineages into major clades. The latter is based on shared marker variants. The main difference compared to our analysis is that the comparison is based on the whole genome sequence, while we only analyse spike proteins. Note that the

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

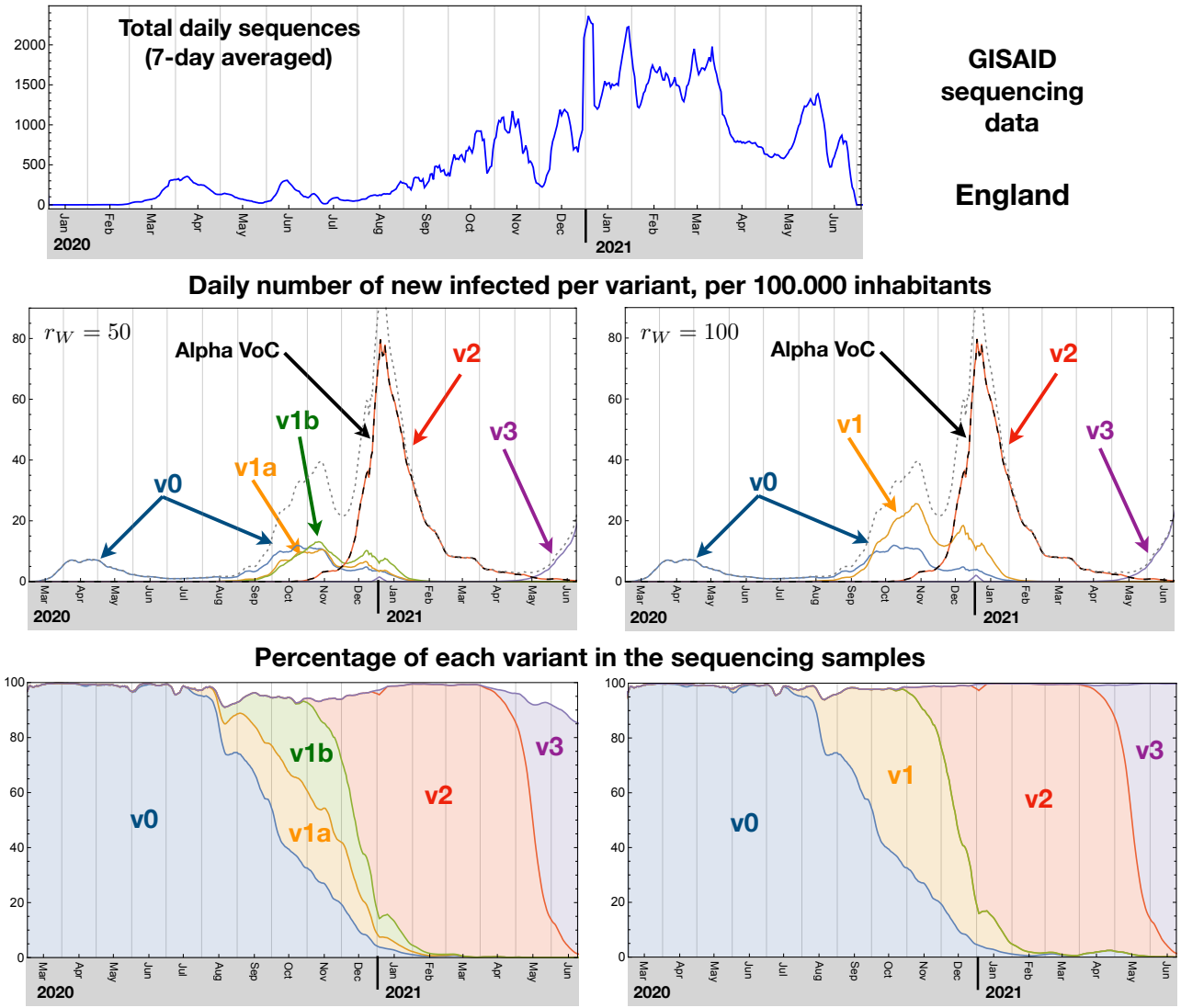

Figure 2. Machine Learning results. For England, the top panel shows the time-distribution of the sequences used in this analysis. Below, we present the results of the clustering for two choices of cut: at a Ward distance $r_W = 50$ (left) and $r_W = 100$ (right). The middle plots show an estimate of the daily number of new infections per cluster, compared to the ones attributed to the Alpha VoC in the GISAID dataset. In the bottom plots, we show the percentage of sequences in each cluster.

marker variants used in the lineage merging contain specific information on proteins, including the spike. It has been noted that this way of defining clades provides similar results to the Pango lineage classification[11], and other variations. For the UK, we show in Fig. 3 the frequency and number of infections assigned to each clade, where we note that GRY corresponds broadly to the Alpha VoC (or variant v2 in our results). These plots confirm that each wave is dominated by a single group of mutations. By comparing the frequencies, we also observe that our variant v1 matches the clade GV, while v0 groups all the other ones that mainly occurred during the first wave. We also note that our method allows to clearly identify the Delta VoC (v3), while in the Nextstrain clades it is associated with the clade G. This validation demonstrates that the clustering based on the spike protein sequences alone is able to identify relevant variants for the SARS-CoV-2.

Variants as time-ordered cluster chains

Having validated our method to define variants in terms of ML clusters, we now turn our attention to study the time evolution of the mutations and how new relevant variants emerge from the old ones. To do so, we have divided the sequencing dataset

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

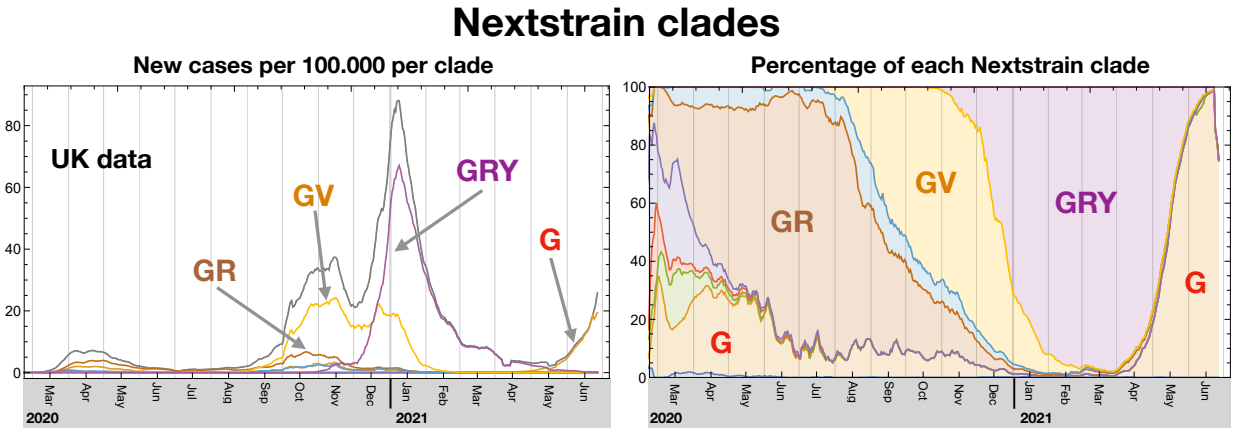

Figure 3. Nextstrain clades. Representation of the "clades" as defined by the Nextstrain initiative[39]. In the two panels we show the estimate of the number of infected by each clade (left) and the percentage in the sequencing data (right) for the UK. We indicate the names of only the major clades. Note that GRY coincides with the Alpha VoC, while the Delta VoC is behind the late dominance of the clade G.

for England into months, following the date tag in the GISAID repository. For each month, we run the ML algorithm on the pruned data to define clusters, retaining only the ones comprising at least 1% of the dataset. Then, we compare the clusters in consecutive months to link the ones that have high degree of "similarity": we establish strong links if the most common spike sequence in the two clusters is exactly the same, and weak links if the strong link fails but the "average distance" between the two clusters is below a given threshold. More details on this procedure and its validation can be found in the supplementary material. The linkage algorithm we employ allows to define unique chains of clusters, that we associate to variants. The results are shown in Fig. 4 for two choices of the Ward distance: $r_W = 100$ in the left and $r_W = 200$ in the right plots. Interestingly, as in Fig. 2, we identify 5 and 4 clusters for the two choices. By comparing the plots of the frequencies and infection numbers, we see that the cluster chain variants defined here coincide with the ones found in the global analysis in Fig. 2, thus further validating our method.

The chain analysis, however, allows us to better probe the time evolution and emergence of the variants. To do so, for the clusters at the beginning of each chain, we define a branching link with the cluster in the previous month that is the closest in terms of the Ward distance. These connections, which do not qualify as weak links, are shown in grey in the top plots in Fig. 4. From the case $r_W = 200$, we clearly see that v1, which is responsible for the second wave, branched off from v0 in October. Similarly, v2, which corresponds to the Alpha VoC, also branched off from v0 a month later. The Delta VoC v3, instead, branched off from v1 in April, at the time when the links defining the chain turn from strong to weak. The three-node chain appearing in June corresponds to a short lived variant that appeared just after the end of the first wave, but was not able to ignite a new wave. By lowering the cut that defined clusters, see left plots in Fig. 4, one can see how v1 splits in two distinct, but closely related, chains. The Delta VoC v3 is now seen as branching off from v1b.

These results firstly show that the phylogenetic relation between variants can be recovered by our simple ML algorithm applied to the spike protein sequences alone. Furthermore, we see a distinctive pattern relating the emergence of a relevant variant and the exponential increase in infections that ignites a new pandemic wave. A new wave only emerges when a new variant is generated, which has the virological strength to overcome the old ones. This is seen very clearly with v2 (or Alpha VoC) which spins off from v0 closely to v1 and takes over by generating a third wave. We also see the emergence of short lived variants that do not have the power to start a new wave and therefore die off without infecting a sizeable number of individuals.

Epidemiological data and MeRG

The results of our ML analysis firmly suggest that there is a strong relation between the genesis of a new relevant variant and the emergence of a new wave, with exponential increase in the number of infections, in the epidemiological data. In a companion article[22] we developed a framework that can be used to describe the evolution of each variant. The model is based on the eRG approach by including mutations (MeRG).

The MeRG framework models the time evolution of the cumulated number of infected by each variant in terms of a logistic

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

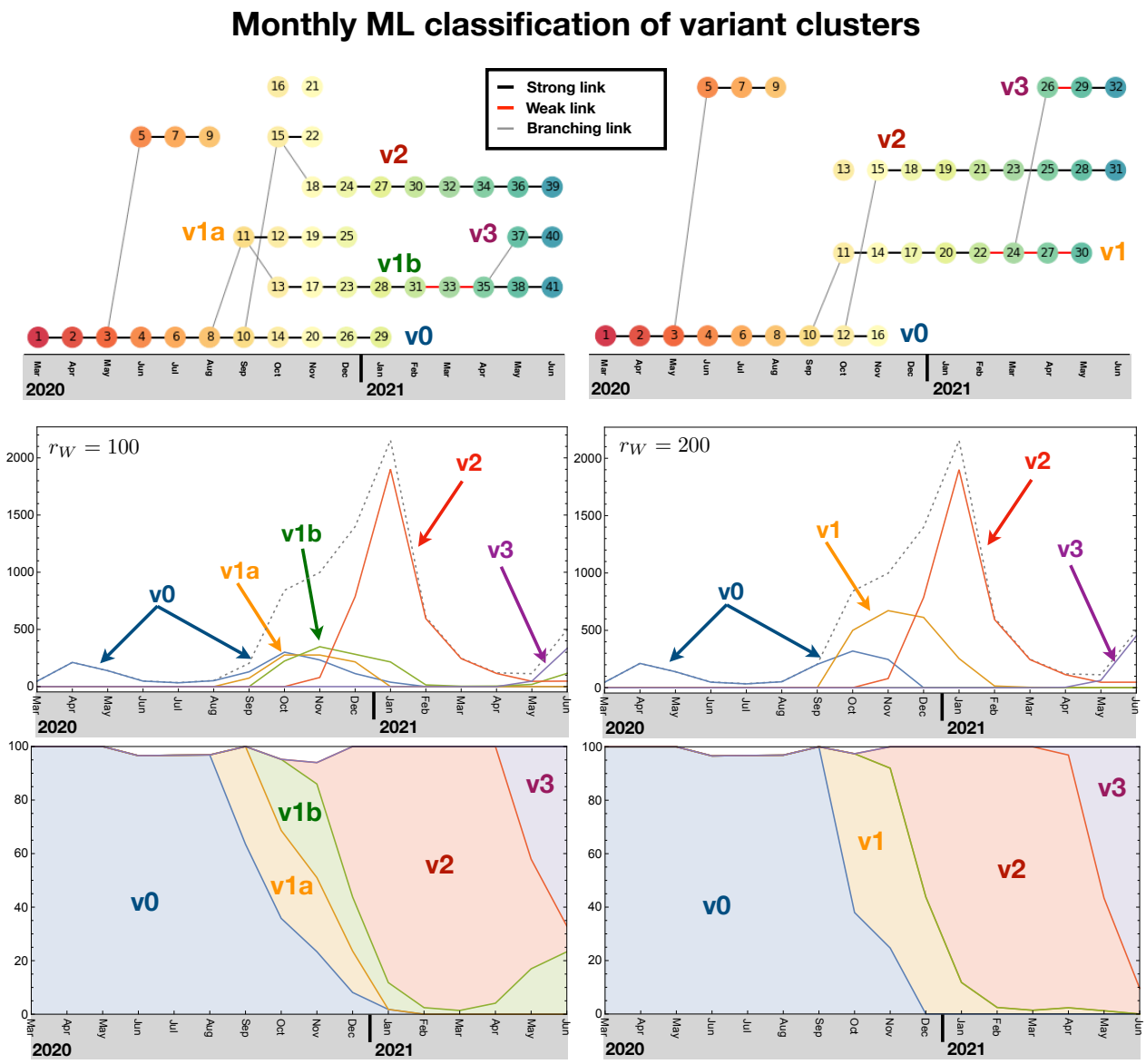

Figure 4. Monthly ML analysis and chain variants. The clusters are linked to form chains, which are then identified with relevant variants, as shown in the top plots. In the middle and bottom plots we show the number of daily infected and percentage of occurrence for each variant, similarly to Fig. 2. The left plots correspond to $r_W = 100$ while the right ones to $r_W = 200$. Note that the chains v2 and v3 can be associated to the Alpha and Delta VoC, respectively.

function (sigmoid), solution of the eRG equation, and given by:

$$\mathscr{I}_c(t) = \frac{Ae^{\gamma t}}{B + e^{\gamma t}}, \tag{1}$$

where $\mathscr{I}_c$ is the cumulative number of infected, γ is the infection rate (in inverse days) and A is the total affected individuals after the wave (per 100.000 inhabitants). The parameter B controls the timing of the wave, and is of no concern in this study. We recall that the parameter γ encodes the effective diffusion speed of the variant, including not only its intrinsic viral power but also the effect of pharmaceutical measures (like vaccinations) and social distancing measures. Nevertheless, it is possible to compare the value of these parameters between different variants. If the diffusion occurs under similar social conditions, this represents a measure of the ability of the new variant to spread and infect new individuals.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

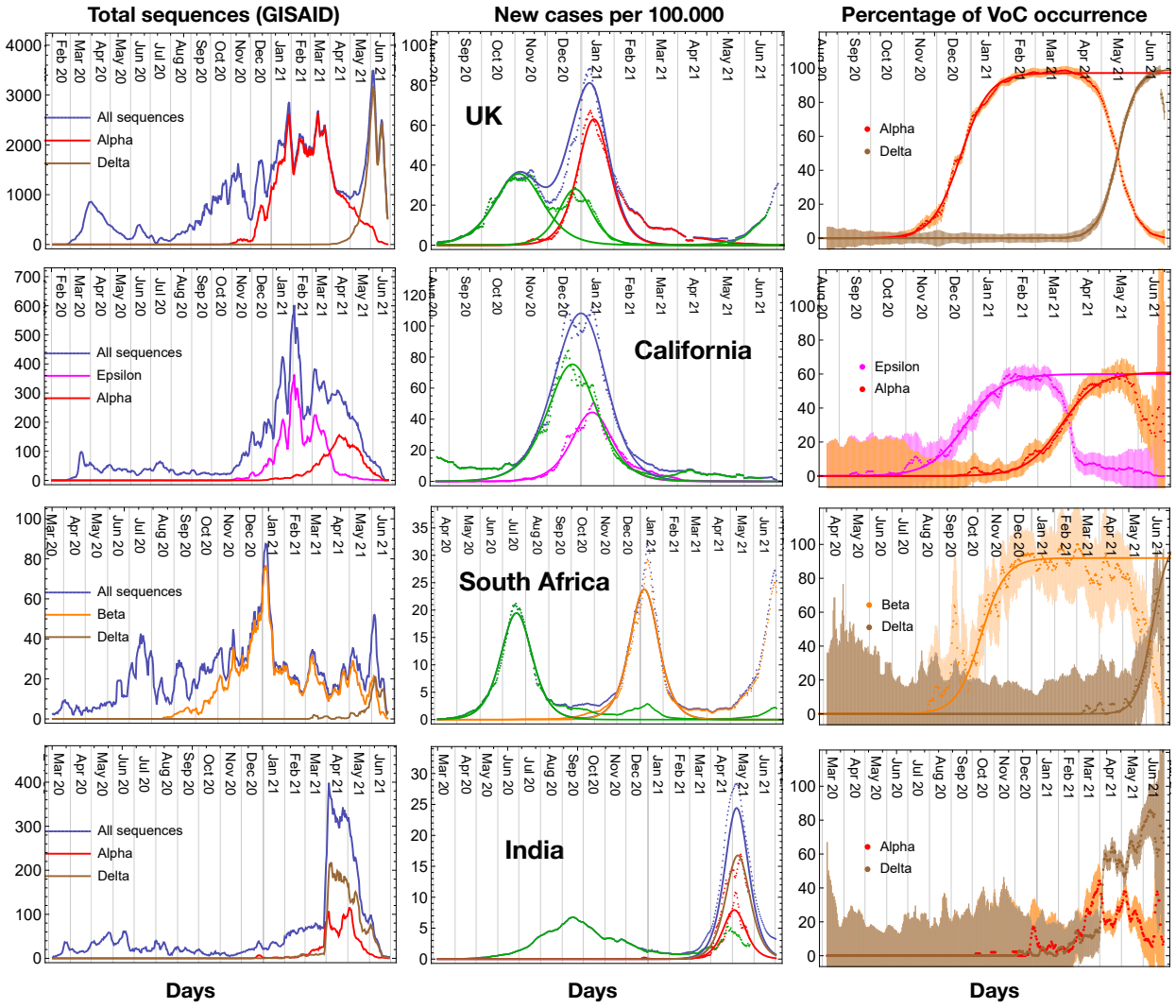

Figure 5. MeRG model for epidemiological data of variants. Results of the MeRG fitting of the number of infected associated to each relevant variant. Each row corresponds to a geographical region. In the left column we show the total number of sequencing available on GISAID (in colour the ones associated to the relevant VoC or VoI); the middle column shows the number of new daily infected (per 100.000 inhabitants); the right column shows the percentage of each VoC or VoI in the sequencing data. All plots show daily rates, with data smoothened over a period of 7 days. In the middle plots, the data are shown by dots, where blue corresponds to the total and the colours show the number of infected associated to each variant. The solid lines show the result of the fits to the MeRG model (note that only for the UK we fit the "standard variant" - in green - with two logistic functions). In the left plots, the error derives from the expected statistical variation on the number of daily sequences (after smoothening). For all the plots, the classification in variants derived from the GISAID data.

Thus, we used the logistic function above to fit the epidemiological data, after distributing the new daily infected to each variant proportionally to the variant frequency observed in the sequencing data. This procedure yields a reliable estimate of the diffusion of each variant. For this purpose, we use the full dataset from GISAID for the whole UK, using the VoC classification embedded in the data to define the variants. As shown before, this classification is equivalent to the result of our ML approach. The result is shown in the top row of Fig. 5, where we show the number of sequences (left plot), the new number of infections per variant and the result of the MeRG fit (middle) and the frequency of the VoCs (right). Note that the total numbers are plotted in blue, while the VoCs in colours. We considered the epidemiological data from the most recent waves, which developed between September 2020 and February 2021. The green curve in the middle plot shows that, after the first peak at the beginning

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Region	standard variant		variant of concern			transmissibility	VoC percentage	
	A	γ	A_{VoC}	γ_{VoC}	VoC/VoI	increase	$A_{\%}$	$\gamma_{\%}$
UK	2140(12)	0.0668(5)	2530(10)	0.0994(7)	Alpha	49%	97.3(3)%	0.076(1)
			–	–	Delta	–	99(1)%	0.115(2)
South Africa	1104(2)	0.0705(4)	1161(2)	0.0904(5)	Beta	28%	91.9(8)%	0.061(4)
			–	–	Delta	–	96(6)%	0.090(7)
India	717(3)	0.0358(4)	497.8(8)	0.0858(3)	Alpha	140%	–	–
			908(5)	0.0747(6)	Delta	109%	–	–
California	4773(7)	0.0620(3)	2250(5)	0.0758(5)	Epsilon	22%	59.9(6)%	0.059(4)
			–	–	Alpha	–	61.0(6)%	0.0610(2)

Table 1. MeRG fit parameters. Parameters from the fit of the VoC/VoI for the UK, South Africa, California and India, also shown in Fig. 5. The fit follows the MeRG model, according to which each variant can be fitted by an independent logistic function. For the UK, the "standard variant" fit corresponds to the first peak, in October-November 2020. The transmissibility increase is computed by comparing the gamma of the VoC with that of the standard variant in the same country. For the new variants that have not reached the peak of diffusion, it is not possible to extract reliable values for the eRG parameters.

of November, a second smaller peak developed. We describe the two with two independent sigmoids. The second sigmoid is subtracted from the data when fitting for the Alpha VoC data. The parameters from the fit are reported in Table1. As the social conditions during this period did not change substantially, it is meaningful to compare the γ parameters for the Alpha and Delta VoC with the other ones (in green). We observed a marked increase in the infectivity, by 49% for Alpha, which is compatible with laboratory tests. Interestingly, the frequency percentage for the VoCs, show in the left plot, can also be fitted very accurately with a logistic function in Eq. (1) as long as only one VoC dominates. The results are also reported in Table 1. The fit parameter $\gamma_{\%}$ is a measure of how more infectious is the new VoC with respect to the previously dominant one. This plot also shows very effectively the switch between the two variants, occurring in May 2021.

We repeated the same analysis for South Africa, California and India, which show very good fits notwithstanding the more limited sequencing statistics available on GISAID. This is clearly shown in the left plots, where we report the statistical uncertainly at 65% confidence level, due to the available sequencing. The results, shown in Fig. 5 and Table 1, demonstrate that the MeRG framework provides an excellent modelling of the data.

Discussion

We present a simple Machine Learning algorithm based on the Levenshtein distance that allows us to identify, classify and track relevant virus variants without any prior knowledge of their characteristics. While our procedure is based on spike protein sequencing only and is not biased by any knowledge of the probability and relevance of each mutation, the results are validated by comparison to other informed methods based on the complete genome. We applied the algorithm to the sequencing data for England, which offers the largest dataset on the GISAID open-source genome repository. The results show that the relevant VoCs (Alpha and Delta, in the case under study) can be clearly identified and isolated. The effectiveness of the algorithm is also confirmed by the data from Wales and Scotland, which have more limited numbers of available sequences. Our results prove that the method, based on spike protein sequences alone, is as effective as and complementary to other methods used in the literature. Furthermore, we designed a procedure to classify variants in terms of time-ordered chains of clusters. As a practical application, we applied the algorithm to the England data binned in months, and then defined links between the clusters independently determined for each month. This technique also allows us to define a branching link, which reconstructs the proximity of the new variants with previous ones. Hence, we establish that the Alpha VoC is closely related to the variant that first spread in Europe, while the Delta one stems from the variant that dominated during the second wave in Europe, in September-October 2020.

We used the relative percentage of each variant in the sequencing dataset to estimate the number of individuals infected by each variant. Hence, by correlating the variant definition to epidemiological data, we discover that each new wave of the COVID-19 pandemic is driven and dominated by a new emerging variant, as identified by our ML analysis. This observation corroborates the hypothesis that there exists a strong and direct causal relation between the emergence of a new variant and of a new epidemic wave. We model the number of infected by use of the eRG framework. Each variant can be modelled by an independent eRG function, in agreement with an evolutionary theory we proposed in a companion manuscript (MeRG)[22]. We also use epidemiological data from the whole UK, California, India and South Africa to confirm the validity of the model.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Limitations

The main limitation of this study is the fact that it is applied only to the sequencing data from a single region, England. This is justified by the fact that sequencing dataset associated to England on the GISAID open-source genome repository is by far the largest compared to other countries/regions. Thus, it is the only dataset that allows for reliable classification of the variants. To validate the results, we have also analysed the data for Wales and Scotland, as presented in the supplementary material. We chose the other two nations of the Great Britain island because they have a very similar epidemiological history compared to England, thus we would expect the same results. As such, by comparing the results we would test the reliability of the ML procedure alone. In fact, the results for Wales and Scotland, while less significant with respect to statistics, show the same patterns we obtained for England.

Conclusions

The results of our ML analysis are seminal to the development of a new strategy to study how SARS-CoV-2 variants emerge and to predict the characteristic of future mutations of the spike proteins. Furthermore, the same methodology can be applied to other viral diseases, like influenza, if sufficient sequencing data is available. Hence, we provided an effective and unbiased procedure to identify new emerging variants that can be responsible for the onset of a new epidemiological wave. The main novelty of our approach is that it is based on information about the spike protein alone, thus making the analysis computationally less intensive than more standard approaches based on the whole genome. Furthermore, no prior knowledge of the variant characteristics is necessary to obtain reliable results.

The results we present here constitute a milestone for the development of a new exploratory strategy of the genesis of variants for coronaviruses. The time sequence definition of a variant, in fact, allows us to study the branching off of new relevant variants, and track the changes in the spike protein associated to the same variant over time or to stemming new variants. Further studies are necessary to fully exploit this information. Furthermore, the same procedure can be applied to other countries for the COVID-19 pandemic, if sufficiently extensive sequencing datasets are available. We also plan to apply the same analysis to other viral infectious diseases.

Our ML strategy is a new tool that can help preventing and forecasting the emergence of new epidemic waves by offering a simple and computationally light procedure to identify new relevant variants. Hence, it can be used by decision makers to define short and long term strategies to curb the current COVID-19 pandemic or future ones.

Acknowledgements

We acknowledge with gratitude the authors, originating and submitting laboratories of the genetic sequence and metadata made available through GISAID. A full listing of all authors and laboratories is available on the GISAID website.

Author contributions statement

This work has been designed and performed conjointly and equally by all the authors. In particular: AdH, SV, AG and FC have developed the Machine Learning algorithm and analysed the spike protein sequencing data; CC has collected and analysed the clade and variant of concern data from GISAID; GC, CC, SH and FS have developed the theoretical framework. All authors have equally contributed to the writing of the text.

Additional information

The authors declare no competing interests.

Data and code availability

All raw data used in this work are obtained from open-source repositories: GISAID for the sequencing and Ourworldindata.org for the epidemiological data. The Machine Learning code is available at https://github.com/AdeledeHoffer/ML-Covid

References

1. Taubenberger, J. K. & Morens, D. M. 1918 influenza: The mother of all pandemics. *Rev Biomed* **17(1)**, 69–79 (2006).

2. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**, 9733–9748, DOI: 10.1128/JVI.00694-10 (2010). https://jvi.asm.org/content/84/19/9733.full.pdf.

3. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116 – 121, DOI: https://doi.org/10.1038/s41586-020-2895-3 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

4. Korber, B. *et al.* Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19, DOI: https://doi.org/10.1016/j.cell.2020.06.043 (2020).

5. Rambaud, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *COVID-19 Genomics Consortium UK (CoG-UK) Rep.* (2020).

6. Mahase, E. COVID-19: What have we learnt about the new variant in the UK? *BMJ* **371**, DOI: 10.1136/bmj.m4944 (2020). https://www.bmj.com/content/371/bmj.m4944.full.pdf.

7. Tegally, H. *et al.* 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* DOI: 10.1101/2020.12.21.20248640 (2020). https://www.medrxiv.org/content/early/2020/12/22/2020.12.21.20248640.full.pdf.

8. Sabino, E. C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet* **397**, 452–455, DOI: https://doi.org/10.1016/S0140-6736(21)00183-5 (2021).

9. Pater, A. A. *et al.* Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. *bioRxiv* DOI: 10.1101/2021.01.11.426287 (2021). https://www.biorxiv.org/content/early/2021/01/19/2021.01.11.426287.full.pdf.

10. Konings, F. *et al.* SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiology* **6**, 821–823, DOI: https://doi.org/10.1038/s41564-021-00932-w (2021).

11. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407, DOI: 10.1093/nsr/nwaa036ttps://doi.org/10.1038/s41564-020-0770-5 (2020).

12. Elbe, S. & Buckland-Merret, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46, DOI: https://doi.org/10.1002/gch2.1018 (2017).

13. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance* **22 (13)**, DOI: https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).

14. Volz, E. *et al.* Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv* DOI: 10.1101/2020.12.30.20249034 (2021). https://www.medrxiv.org/content/early/2021/01/04/2020.12.30.20249034.1.full.pdf.

15. Kermack, W. O., McKendrick, A. & Walker, G. T. A contribution to the mathematical theory of epidemics. *Proc. Royal Soc. A* **115**, 700–721, DOI: https://doi.org/10.1098/rspa.1927.0118 (1927).

16. Perc, M. *et al.* Statistical physics of human cooperation. *Phys. Reports* **687**, 1 – 51, DOI: https://doi.org/10.1016/j.physrep.2017.05.004 (2017).

17. Wang, Z., Andrews, M. A., Wu, Z.-X., Wang, L. & Bauch, C. T. Coupled disease–behavior dynamics on complex networks: A review. *Phys. Life Rev.* **15**, 1 – 29, DOI: https://doi.org/10.1016/j.plrev.2015.07.006 (2015).

18. Giordano, G. *et al.* Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat. Medicine* DOI: https://doi.org/10.1038/s41591-021-01334-5 (2021).

19. Della Morte, M., Orlando, D. & Sannino, F. Renormalization Group Approach to Pandemics: The COVID-19 Case. *Front. Phys.* **8**, 144, DOI: https://doi.org/10.3389/fphy.2020.00144 (2020).

20. Cacciapaglia, G. & Sannino, F. Interplay of social distancing and border restrictions for pandemics (COVID-19) via the epidemic Renormalisation Group framework. *Sci Rep* **10**, 15828, DOI: https://doi.org/10.1038/s41598-020-72175-4 (2020). 2005.04956.

21. Cacciapaglia, G., Cot, C. & Sannino, F. Second wave COVID-19 pandemics in Europe: A temporal playbook. *Sci Rep* **10**, 15514, DOI: https://doi.org/10.1038/s41598-020-72611-5 (2020). 2007.13100.

22. Cacciapaglia, G. *et al.* Epidemiological theory of virus variants (2021). 2106.14982.

23. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk* **163**, 845–848 (1965).

24. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Cybern. Control. Theory* **10**, 707–710 (1966).

25. Berger, B., Waterman, M. S. & Yu, Y. W. Levenshtein distance, sequence comparison and biological database search. *IEEE Transactions on Inf. Theory* 1–1, DOI: https://doi.org/10.1109/TIT.2020.2996543 (2020).

26. Koumakis, L. Deep learning models in genomics; are we there yet? *Comput. Struct. Biotechnol. J.* **18**, 1466–1473, DOI: https://doi.org/10.1016/j.csbj.2020.06.017 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

337 **27.** Kopp, W., Monti, R., Tamburini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nat. Commun.* **11**,
338 3488, DOI: https://doi.org/10.1038/s41467-020-17155-y (2020).

339 **28.** Yang, A. *et al.* Review on the application of machine learning algorithms in the sequence data mining of DNA. *Front.*
340 *Bioeng. Biotechnol.* **8**, 1032, DOI: https://doi.org/10.3389/fbioe.2020.01032 (2020).

341 **29.** Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
342 (1990).

343 **30.** Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595
344 (2010).

345 **31.** Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).

346 **32.** Wilson, K. G. Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture.
347 *Phys. Rev. B* **4**, 3174–3183, DOI: https://doi.org/10.1103/PhysRevB.4.3174 (1971).

348 **33.** Wilson, K. G. Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior. *Phys. Rev.*
349 *B* **4**, 3184–3205, DOI: https://doi.org/10.1103/PhysRevB.4.3184 (1971).

350 **34.** Cacciapaglia, G., Cot, C., Islind, A. S., Óskarsdóttir, M. & Sannino, F. Impact of US vaccination strategy on COVID-19
351 wave dynamics. *Sci. Reports* (2021). 2012.12004.

352 **35.** Della Morte, M. & Sannino, F. Renormalization group approach to pandemics as a time-dependent SIR model. *Front.*
353 *Phys.* **8**, DOI: 10.3389/fphy.2020.591876 (2021).

354 **36.** Cacciapaglia, G. *et al.* The field theoretical ABC of epidemic dynamics (2021). 2101.11399.

355 **37.** Cacciapaglia, G., Cot, C. & Sannino, F. Multiwave pandemic dynamics explained: How to tame the next wave of infectious
356 diseases. *Sci. Reports* **11**, 6638 (2021). 2011.12846.

357 **38.** Cacciapaglia, G. & Sannino, F. Evidence for complex fixed points in pandemic data. *Front. Appl. Math. Stat.* **7**, 659580,
358 DOI: https://doi.org/10.3389/fams.2021.659580 (2021). 2009.08861.

359 **39.** Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123, DOI: 10.1093/
360 bioinformatics/bty407 (2018). https://academic.oup.com/bioinformatics/article-pdf/34/23/4121/26676762/bty407.pdf.

361 **40.** Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **7**, 1012–1023, DOI: https:
362 //doi.org/10.1093/nsr/nwaa036 (2020). https://academic.oup.com/nsr/article-pdf/7/6/1012/33408507/nwaa036.pdf.

363 **41.** Han, A. X., Parker, E., Scholer, F., Maurer-Stroh, S. & Russell, C. A. Phylogenetic Clustering by Linear Integer
364 Programming (PhyCLIP). *Mol. Biol. Evol.* **36**, 1580–1595, DOI: https://doi.org/10.1093/molbev/msz053 (2019). https:
365 //academic.oup.com/mbe/article-pdf/36/7/1580/28833695/msz053.pdf.

366 **42.** Cacciapaglia, G., Cot, C. & Sannino, F. Mining Google and Apple mobility data: Temporal anatomy for COVID-19 social
367 distancing. *Sci Rep* **11**, 4150, DOI: https://doi.org/10.1038/s41598-021-83441-4 (2020). 2008.02117.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary material

S1 Theoretical modelling of variant diffusion

The current paper deals with the time evolution and spread of different variants of SARS-CoV-2 in a given population. A theoretical study of the underlying processes in the framework of the epidemic Renormalisation Group (eRG) approach[19] has recently been presented in a companion paper[22]. In this section, we shall briefly review the relevant formalism.

The eRG approach is inspired by the running of fundamental couplings as a function of the energy scale in particle physics. Originally proposed[19] as an effective description of epidemic diffusion processes organised around time-scale invariances, it has been extended to account for geographic mobility across different countries[20], the multi-wave structure of the SARS-CoV-2 pandemic[38] as well as the impact of the US vaccination campaign[34]. The predictive power of this approach has been demonstrated by accurately describing the impact of non-pharmaceutical interventions[38,42] and predicting the starting date of the second wave in the fall of 2020 in Europe[21]. An interpretation of the eRG approach as a time-dependent SIR model has also been discussed[35], while the relation to other epidemiological approaches has been discussed at great depth in this review[36].

In the companion paper[22], the eRG framework has been further extended to describe the time evolution of two different variants of a disease. An epidemic coupling strength $\alpha_i(I_{c,i})$ is introduced for each variant (here $i = 1,\dots,n$ labels the n different variants). The latter is a function of the cumulative number of individuals $I_{c,i}$ that have been infected by this ith variant. The time evolution of the different variants is then described by a set of β-functions

$$-\beta_i(I_{c,i}) = \frac{d\alpha_i}{dt}\,, \qquad \forall i = 1,\dots,n\,, \tag{E1}$$

which constitute the core of the Mutation eRG (MeRG) approach. Inspired by the numerical study of a compartmental model and empirically validated by comparing with data from California, the United Kingdom and South Africa, the β-functions were written in the form of gradient equations[22] (in detail for the case $n = 2$)

$$-\beta_i(I_{c,i}) = \nabla_i \Phi(I_{c,j})\,, \qquad \text{with} \qquad \begin{array}{l} \nabla_i = \frac{\partial}{\partial I_{c,i}} \\ \Phi(I_{c,j}) = \sum_{k=1}^{n} I_{c,k}^2 \frac{\gamma_k}{2}\left(1 - \frac{2I_{c,k}}{3A_k}\right). \end{array} \tag{E2}$$

Here, γ_k is a measure for the infection rate and A_k is the asymptotic number of individuals infected by variant k. Solutions of the β-functions (E2) give cumulative numbers of infected individuals as functions of time for each variant that follow a logistics function

$$I_{c,i}(t) = \frac{A_i}{1 + e^{-\gamma_i(t-\kappa_i)}}\,, \tag{E3}$$

where κ_i is a parameter that governs the time of appearance of the variant. For given i, the function β_i in (E2) has two zeroes, namely $I_{c,i} = 0$ and $I_{c,i} = A_i$, corresponding to the complete absence of the variant i or the eradication of the latter, in the sense that there are no more infectious individuals left carrying it. The complete set of β-functions $(\beta_1,\dots,\beta_n)$ has 2^n fixed points and the epidemic is described by the flow equations connecting (some of) them.

S2 Machine Learning algorithm

We employ a Machine Learning algorithm based on the Levenshtein distance between spike protein sequencing. The procedure can be grossly divided into three steps, which we describe in detail below. All the python codes are provided (see link in the main publication), with reference to the main libraries provided in this material. While in this work we mainly focus on England, due to the larger dataset, and other nations of Great Britain, the ML codes can be run on any dataset, for different countries or regions of the world.

S2.1 Extraction and pruning of the raw data

The raw data are downloaded from the GISAID open-source repository (registration required) in the form of "fasta" files, which contain information on samples from COVID-19 infected cases. The files contain the full genome, including the spike protein sequences, but also the date when the sample was taken, the laboratory where it was analysed and the geographic information about the country or region of origin of the sample. This additional information allows to separate datasets based on a specific geographical origin, and sample them in time.

In this work, we focus only on the spike protein sequences. The data contains sequences with un-identified amino-acids (labelled with an X) and sequences with missing pieces (thus, with unusual lengths). The spike protein of the early SARS-CoV-2

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sequences have 1271 amino-acids. Hence, to remove data with missing pieces, we only keep sequences with at least 1250 amino-acids. Furthermore, we dismiss all sequences containing at least one X in the sequence. This pruning allows us to work only on a high purity dataset. The number of sequences before and after the pruning for England, Wales and Scotland are listed in Table T1. After the pruning, a significant number of sequences are in the dataset, many of which contain the same sequences for the spike protein. To accelerate the next step in the analysis, it is convenient to remove repetitions and thus only work on a dataset containing only different sequences (see right column in Table T1). In particular, this is necessary when we do a global analysis of the whole dataset, while the time-sampling automatically reduces the number of sequences in the dataset.

	Raw sequences	After pruning	Different sequences (after pruning)
England	436.073	329.384	9.480
Wales	36.423	24.761	1.101
Scotland	56.950	42.302	1.514

Table T1. Number of sequences in the datasets for England, Wales and Scotland before and after the pruning and selection.

The extraction and pruning of the data is done by the python program `extraction_country.py`, where the name of the country or region needs to be specified in the first lines of the program. The output is as follows:

- The list of the strictly different sequences as a text file `country_seq_ass.txt`.
- A csv table `country.csv` were lines correspond to selected sequences. The first column contains the date, the second the corresponding sequence in reference to the text file already saved and finally a column labelling the VoC or VoI the sequence belongs to, according to GISAID.
- A list of the different labs contributing to the sequencing.

Note, in passing, that the pruning could be by-passed by modifying the distance calculation in such a way that the presence of incomplete sequences is properly taken into account. This would require a more complex and optimised procedure. For our purposes, we wanted to remain as unbiased as possible, thus we opted for the pruning and decided to work with a dataset of the highest purity.

S2.2 Computation of the distance matrix

The output of the previous step is a list of strictly different spike protein sequences, with a record of their multiplicity in the dataset under study. At this stage, we need to compute the Levenshtein distances between them. This computation yields a symmetric matrix with zeros on the diagonal, such that only a triangular matrix needs to be computed on the data.

To efficiently compute the Levenshtein distances, we use the library polyleven. For this purpose, we created the python code `Launch_distance.py`, where a specific country or region has to be specified at the beginning of the file. The program calls `distance.sh` and `distance.py`, thus the process is multicore and very fast. The various lines of the upper triangular matrix are saved in the subfolder `create` as separate binary files. Launching now `concatenation.py` will load them all and save a single file containing the complete distance matrix, while the temporary files are deleted.

Note that to speed up the computation of the distances, we only considered strictly different spike protein sequences. This avoids repeated computations. Yet, the construction of the proximity tree, which will be the task of the next step, may depend on the multiplicity of each sequence in the starting dataset, as we will see. Hence, the program `concatenation.py` outputs two different distance matrices: `country.bin` where only strictly different sequences are included, and `country_complete.bin` where the multiplicity of the sequences are taken into account by copying multiple times the corresponding lines in the table. Thus, the latter contains a much bigger distance matrix than the former.

S2.3 The proximity tree

To create the proximity tree, various libraries are at our disposal. We chose to adopt the hierarchical clustering algorithm from the `scipy` library as it is easy to work with and adaptable. The only needed input is the distance matrix. For the algorithm, each initial sequence, i.e. a line or column of the distance matrix, is a *leaf*. The library then groups the leaves into *branches* based on the Levenshtein distance between them, as specified in the input file. To calculate the effective distance between branches, various methods are available, and they need to be selected. We anticipate that some methods are not sensitive to the multiplicity the identical leaves (which have distance zero among them), while others are.

The algorithm constructs the proximity tree starting from the leaves: hence, for a sample with n leaves (i.e. a $n \times n$ distance matrix as input), the code considers an initial state with n branches containing a single leaf each. Hence, the steps consist in regrouping 2 branches together forming a new branch. Naturally, there can be at most $n-1$ of those steps. At any step, the two

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

branches that are chosen to be regrouped (let us call them A and B) are the ones that have the smallest distance between them. Such distance $\mathrm{dis}(A,B)$ is computed in terms of the Levenshtein distance $d(x,y)$ between leaves on the two branches (x and y being a leaf from A and B respectively). Here is where different methods to compute the branch distance $\mathrm{dis}(A,B)$ can be employed. The most common choices are:

- Single Linkage Clustering, where

$$\mathrm{dis}(A,B) = \min_{x\in A\,,\,y\in B}\ \ d(x,y)\,. \tag{E4}$$

  This method is insensitive to the multiplicity of identical leaves, thus one can use the file `country.bin` as input.

- Complete Linkage Clustering, where

$$\mathrm{dis}(A,B) = \max_{x\in A\,,\,y\in B}\ \ d(x,y)\,. \tag{E5}$$

  As before, this is independent on the leaf multiplicity.

- Unweighted Average Linkage Clustering, where

$$\mathrm{dis}(A,B) = \frac{1}{|A||B|}\sum_{x\in A\,,\,y\in B}\ \ d(x,y)\,, \tag{E6}$$

  where $|X|$ is the number of leaves in the branch X. This method is sensitive to the multiplicity of identical leaves, thus is requires `country_complete.bin` as input.

- Ward's Method, an agglomerative clustering method based on the measure of the average squared distance of points in the cluster to its centre of gravity, or centroid. Hence, the effective distance between two branches is defined by the increase in the above measure in the merged cluster with respect to the two separate ones. In practice:

$$\mathrm{dis}(A,B) = \frac{|A||B|}{|A|+|B|}\left[\sum_{x\in A\,,\,y\in B}\frac{d(x,y)^2}{|A||B|} - \sum_{x,x'\in A}\frac{d(x,x')^2}{2|A|^2} - \sum_{y,y'\in B}\frac{d(y,y')^2}{2|B|^2}\right]. \tag{E7}$$

  This method is sensitive to the multiplicity of identical leaves, thus it requires `country_complete.bin` as input. Note that we also used this method on the distance matrix `country.bin` to speed up the computation.

At each step, the algorithm joins branches together, until all leaves are grouped together. This approach can be schematically represented by a dendrogram tree, as shown in Fig. F1. This step is executed by the program `linkage.py`. In order to define clusters and variants, we need to define a threshold in the effective distance, which will therefore be equivalent to a horizontal cut of the tree branches. In practice, all clusters whose effective distance is larger than the threshold are used to define the variants. The value of the threshold needs to be determined each time, as it crucially depends on the measure that is employed, and on the samples. Once the threshold is fixed, the clusters and their time evolution can be obtained. This is done by the executable called `time.py`.

S3 Linkage Analysis

Data from GISAID repository has been grouped on a monthly basis for the England, Wales and Scotland datasets, as shown in Table T2. Despite the rather large number of monthly recorded data, we found sequences with high replica rate, thus the percentage of different sequences over the whole dataset is also reported in the Table T2 and shown by the orange line in Fig. F2. It has been noted that the most frequent sequence always represents a remarkable amount of the whole dataset, as shown by the blue line in Fig. F2 for the England dataset. Furthermore, the Levenshtein distance between the most frequent sequences in two consecutive months is always equal to 0 (i.e., the same sequence is dominant) with the exception of months 9 and 11. This last month clearly represents the overtaking of the Alpha VoC B.1.1.7 over the previous ones. For each month the same hierarchical clustering algorithm described above based on the Levenshtein measure (LM) has been applied. Here the two main parameters to take in account are the clustering threshold (Ward distance) that mainly acts on the size of the clusters and the cut on the minimum amount of data that is needed to define a cluster. Detailed studies have been made to find the right trade off between the number of clusters and the coverage of the dataset.

We define the threshold based on the Ward distance, r_W, as defined in the previous section. The coverage cut-off is defined in terms of a minimum percentage of the whole sequence dataset (per month) that is covered by each branch above threshold.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

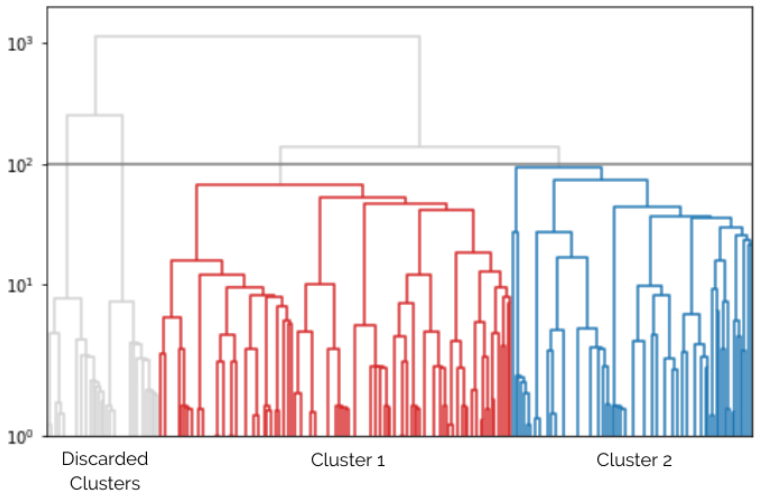

Figure F1. Dendrogram three for the England dataset, for month 9 (September 2020), built using the Ward's method. On the y-axis we show the effective distance between two clusters at the point when they merge. The branches in colour represent more than 1% of the total number of sequences for a cut at Ward distance 100.

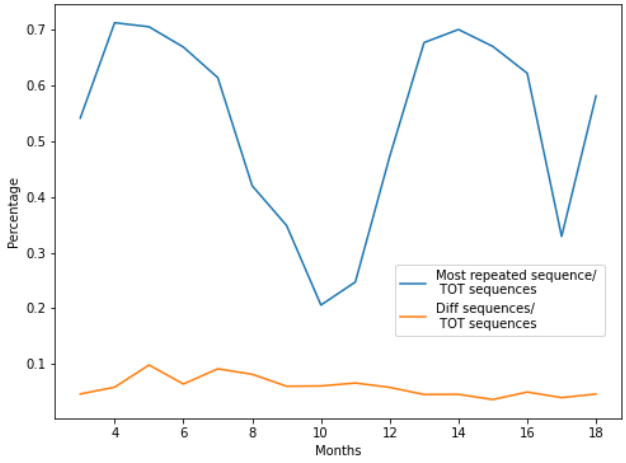

Figure F2. Dataset coverage of the most common sequence and different sequences in GISAID (England dataset).

Only branches above the cut-off are retained to define clusters. For each given threshold value r_W, the coverage of the dataset decreases while the cut-off value increases, hence increasing the coverage would require pushing the cut-off towards smaller values. However, the cut-off choice must take into account the large increase in the number of (small) clusters. The dataset coverage and number of clusters as functions of the threshold (Ward distance) are shown in Fig. F3. Quite independently of the month, the number of clusters is almost stable for any cut-off in the $[10^{-2}, 10^{-1}]$ range, while for smaller values it shows a linear increase. Thus we choose to set the cut-off value to 10^{-2} to maximise the mean coverage of the dataset (mean coverage $> 90\%$). Moreover it has been found that the mean coverage is almost stable for any threshold in the range $[50, 200]$ while the number of defined clusters decrease with the threshold value, as shown in Fig. F4. Thus the default working point has been set to threshold value of $r_W = 100$ and a cut-off of 10^{-2}.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Month	England	Scotland	Wales
1	2 (100%)	0	0
2	64 (9%)	0	1 (100%)
3	3944 (5%)	1109 (5%)	965 (5%)
4	7335 (6%)	1923 (6%)	2178 (5%)
5	2204 (10%)	306 (12%)	866 (7%)
6	5079 (6%)	58 (24%)	425 (7%)
7	2026 (9%)	67 (9%)	102 (11%)
8	4429 (8%)	937 (8%)	210 (15%)
9	9499 (6%)	1592 (6%)	1275 (6%)
10	19634 (6%)	1740 (8%)	2732 (5%)
11	23431 (7%)	602 (11%)	2497 (6%)
12	25339 (6%)	1161 (9%)	3989 (6%)
13	46685 (5%)	2196 (8%)	3508 (6%)
14	39326 (5%)	4600 (6%)	2769 (7%)
15	49545 (4%)	9655 (3%)	1958 (7%)
16	24057 (5%)	4800 (5%)	689 (12%)
17	24335 (4%)	6263 (3%)	301 (10%)
18	16133 (5%)	3638 (5%)	98 (12%)

Table T2. Number of GISAID recorded sequences from January 2020 (Month 1) to June 2021 (Month 18).

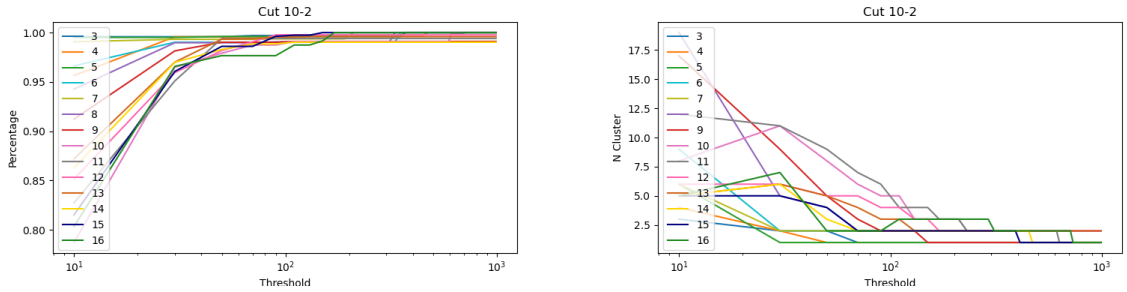

Figure F3. Dataset coverage (left) and number of clusters (right) with respect to threshold (Ward distance) value.

S3.1 Time-series sequences

To define and follow the time evolution of a given candidate variant, we build time-series of clusters starting from month 1 using the following algorithm:

- For each cluster in month i we selected the most frequent sequence and we tried to find a cluster with the same sequence in month $i+1$ (**strong link**). We iterate to build a path until the procedure fails or the last month is reached.

- If the strong link association fails but we still have clusters in consecutive months that are free from strong links, we connected them provided that the distance is less than a given threshold. If two clusters converge to the same node we keep the nearest one (**soft link**).

The threshold cut used for the soft link definition has been optimised looking at the distribution of the distances between clusters connected with strong links, as shown in Fig. F5. To preserve the topology based on the LM we choose the soft link threshold as the maximum distance found for a strong link.

We define a chain as a list of consecutive clusters connected by strong or weak links as best candidate to study the evolutionary paths of a candidate variant. For each well defined chain we also assign a branching link taking the first cluster of the chain and associate it with the cluster in the previous month that is the closest in terms of the Ward distance. In the following we assign as chain identification number the one of the first cluster in the chain.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

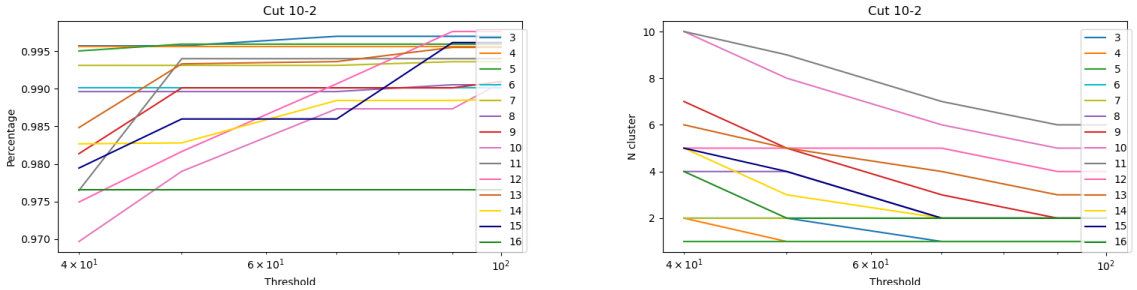

Figure F4. Dataset coverage (left) and number of clusters (right) with respect to cluster threshold value

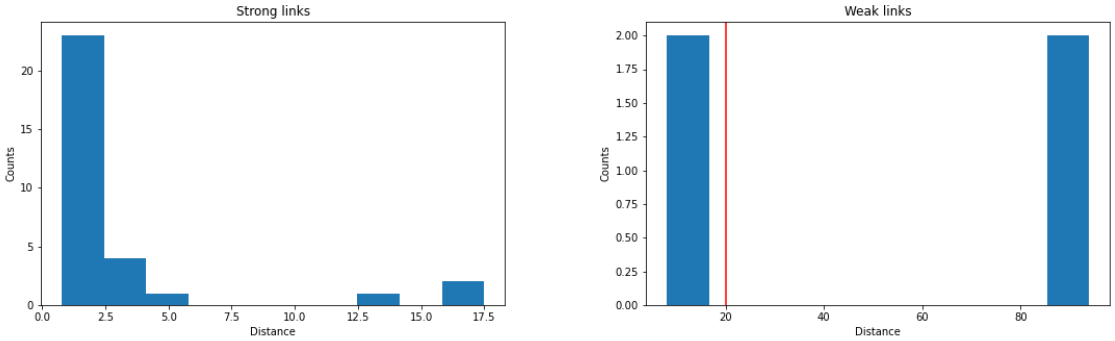

Figure F5. Strong link (left) and weak link (right) distances.

S3.2 Results for England, Scotland and Wales

Chains (as candidate variants) for England, Wales and Scotland are shown in Fig. 4 (left plots) and F6, for the same threshold $r_W = 100$. For England, chains #1 (v0) and #18 (v2) clearly correspond to the original variant and to the Alpha VoC (B.1.1.7), while the chain #37 (v3) matches the Delta VoC.

We also found two other relevant chains, namely chain #11 (v1a) and chain #13 (v1b). Using the branching link it is possible to track also the evolutionary path of each single variant candidate. The Alpha VoC is clearly connected to the chain #1 via cluster #15 (indeed we found about 70 sequences of Alpha VoC in cluster #15). Similarly, the chains #11 and #13 are connected to the original strain and it is likely that they appeared before of the Alpha VoC. Similar results have been found using Scotland (and Wales) data in Fig. F6, where chain #1 corresponds to the original variant and chain #11 (#11 for Wales data) to the Alpha VoC B.1.1.7 and chain #18 (#20 for Wales data) to the Delta VoC B.1.617.2. Results for the variant frequency and daily cases are shown in Fig. F7 and F8 for Scotland and Wales data.

A comparison of the evolutionary path of the England chains suggests a competitive mechanism between variants. To further investigate this we study the spread of a chain using the distance between consecutive clusters, as shown in Fig. F9. The original variant and the Alpha VoC show more stable and lower distance values with respect to chains #11 and #13, thus suggesting that such variable can be used as an early discriminant between different evolutionary paths of variants. Namely, variants that show smaller distances between consecutive clusters appear to be more stable and able to ignite epidemiological episodes of exponential increase (waves). A comparison of the evolutionary path for the original variant and the Alpha VoC for England, Wales and Scotland is also shown in Fig. F10.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

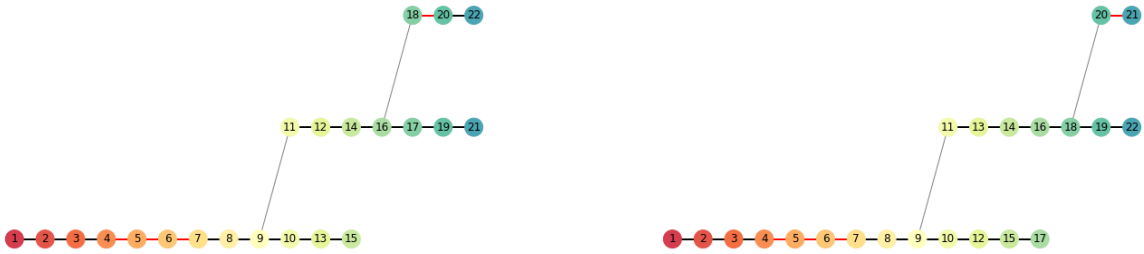

Figure F6. Chain results as candidate variants for the Scotland dataset (left) and the Wales one (right). Strong links (black line) and soft link (red line) are reported. Branching links (gray line) where defined are also shown.

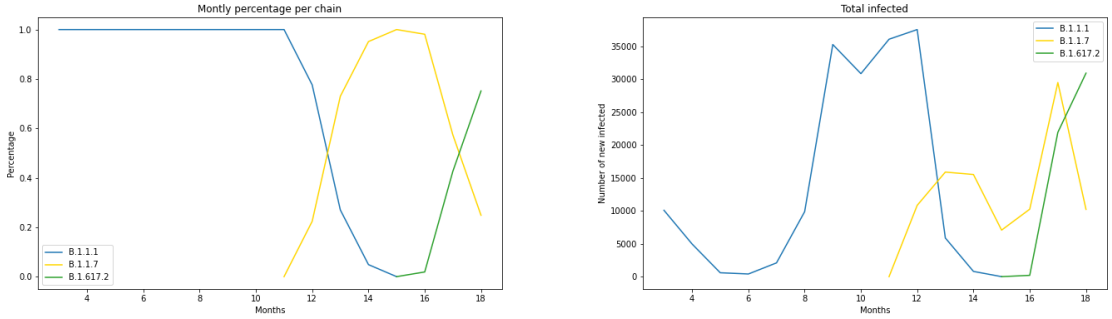

Figure F7. Frequency and daily cases for chains as candidate variants (Scotland)

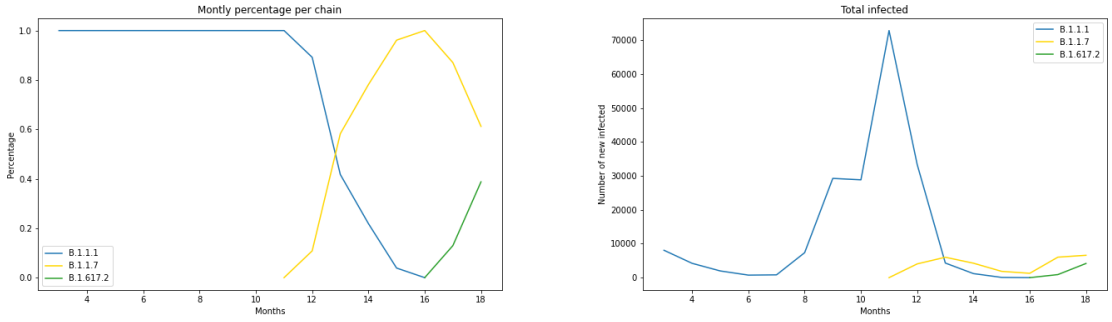

Figure F8. Frequency and daily cases for chains as candidate variants (Wales)

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

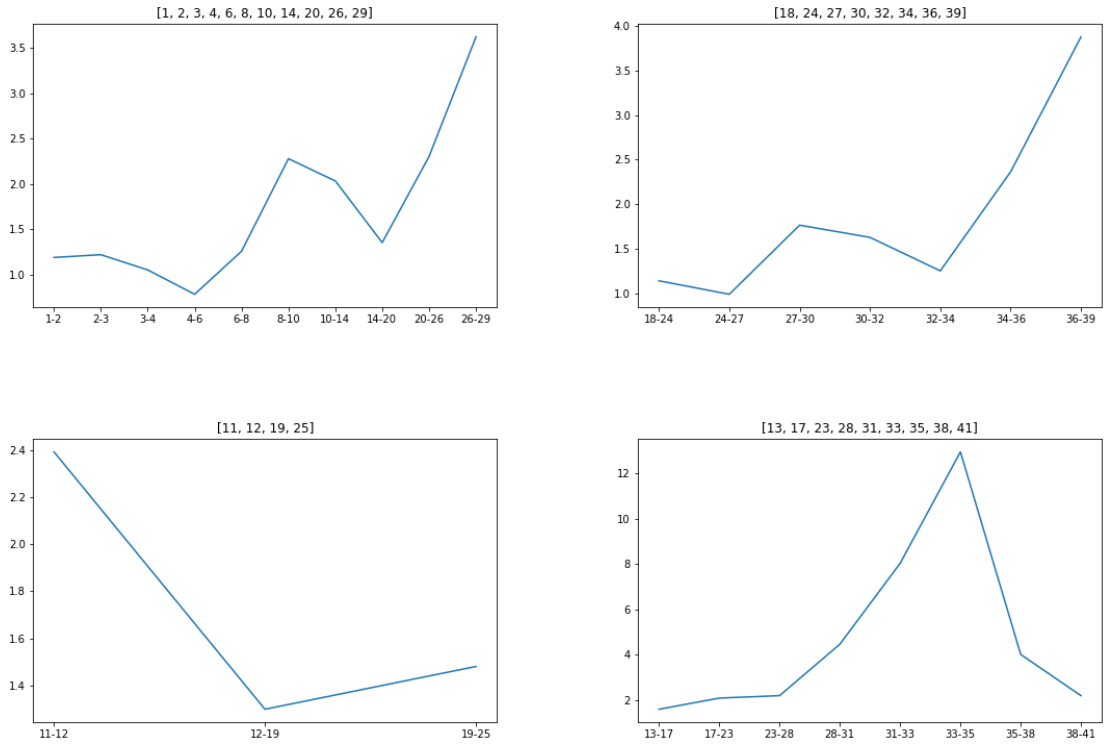

Figure F9. Distances between consecutive clusters for chain #1 (original strain, upper left), #18 (Alpha VoC B.1.1.7, upper right), #11 (bottom left) and #13 (bottom right)

medRxiv preprint doi: https://doi.org/10.1101/2021.07.22.21260952; this version posted July 24, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

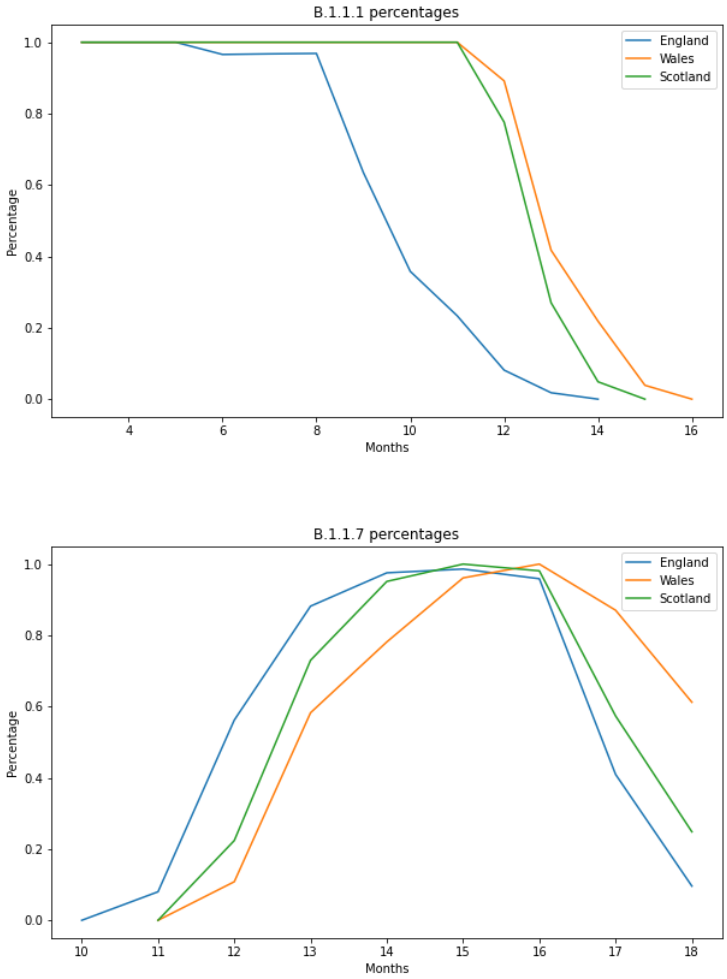

Figure F10. Time evolution comparisono for chains associated with the original strain (left) and Alpha VoC B.1.1.7 (right)