CHARACTERIZATION OF INTRA-HOST SARS-COV-2 VARIANTS IMPROVES PHYLOGENOMIC RECONSTRUCTION AND MAY REVEAL FUNCTIONALLY CONVERGENT MUTATIONS

Daniele Ramazzotti[1], **Fabrizio Angaroni**[2], **Davide Maspero**[2,3],
Carlo Gambacorti-Passerini[1], **Marco Antoniotti**[2,4], **Alex Graudenzi**[3,†,*], **Rocco Piazza**[1,†,*]
[1] School of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy
[2] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy
[3] Inst. of Molecular Bioimaging and Physiology,
Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[4] Bicocca Bioinformatics Biostatistics and Bioimaging Centre – B4, Milan, Italy
† Co-senior authors
* Corresponding authors: rocco.piazza@unimib.it | alex.graudenzi@ibfm.cnr.it

ABSTRACT

A global cross-discipline effort is ongoing to characterize the evolution of SARS-CoV-2 virus and generate reliable epidemiological models of its diffusion. To this end, phylogenomic approaches leverage accumulating genomic mutations as barcodes to track the evolutionary history of the virus and can benefit from the surge of sequences deposited in public databases. Yet, such methods typically rely on consensus sequences representing the dominant virus lineage, whereas a complex sublineage architecture is often observed within single hosts. Furthermore, most approaches do not account for variants accumulation processes and might produce inaccurate results in condition of limited sampling, as witnessed in most countries currently affected by the epidemics.

We here introduce a new framework for the characterization of viral (sub)lineage evolution and transmission of SARS-CoV-2, which considers both clonal and intra-host minor variants and exploits the achievements of cancer evolution research to account for mutation accumulation and uncertainty in the data.

The application of our approach to 18 SARS-CoV-2 samples for which raw sequencing data are available reveals a high-resolution phylogenomic model, which confirms and improves recent findings on viral types and highlights the existence of patterns of co-occurrence of minor variants, uncovering likely infection paths among hosts harboring the same viral lineage. Our findings confirm a significant increase of genomic diversity of SARS-CoV-2 in time, which is reflected in minor variants, and show that standard methods may struggle when handling datasets with important sampling limitations.

Importantly, our framework allows to pinpoint minor variants that might be positively selected across distinct lineages and regions of the viral genome under purifying selection, thus driving the design of treatments and vaccines. In particular, minor variant g.29039A>U, detected in multiple viral lineages and validated on an independent dataset, shows that SARS-CoV-2 can lose its main Nucleocapsid immunogenic epitopes, raising concerns about the effectiveness of vaccines targeting the C-terminus of this protein.

To conclude, we advocate the use of our framework in combination with data-driven epidemiological models, to deliver a high-precision platform for pathogen detection, surveillance and analysis.

Introduction

The outbreak of novel pneumonia COVID-19, which started in late 2019 in Wuhan (China) [1, 2] and was recently declared pandemic by the World Health Organization, is fueling the publication of an increasing number of studies aimed at exploiting the information contained in the viral genome of SARS-CoV-2 virus to identify its proximal origin, characterize the mode and timing of its evolution, as well as to define descriptive and predictive models of geographical spread and evaluate the related clinical impact [3]. In fact, the mutations that rapidly accumulate in the viral genome [4] can be used as natural *barcodes* to track the evolution of the virus and, accordingly, unravel the viral infection network [5, 6].

At the time of this writing, numerous independent laboratories around the world are isolating and sequencing SARS-CoV-2 samples and depositing them on public databases, e.g., GISAID [7], whose data are accessible via the Nextstrain portal [8]. Such data can be employed to estimate models from genomic epidemiology and may serve, for instance, to estimate the proportion of undetected infected people by uncovering cryptic transmissions, as well as to predict likely trends in the number of infected, hospitalized, dead and recovered people [9, 10].

Most studies typically employ standard phylogenomics approaches that, roughly, compare *consensus sequences* representing the dominant virus lineage within each infected host and rely on some measure of *genetic distance* among samples [11, 12]. However, while such analyses are undoubtedly effective in unraveling the main patterns of evolution of the viral genome, at least two issues may suggest that they are suboptimal when applied to the characterization of viral variations under the current circumstances.

First, most methods are missing key information on *intra-host minor variants*, which can be retrieved from whole-genome deep sequencing raw data [13] and might be essential to produce high-resolution models of the SARS-CoV-2 evolution and, accordingly, of the likely transmission chain. Second, most of the approaches currently used do not explicitly account for the processes of *accumulation* of genomic mutations in the population resulting from complex infection chains. For these reasons, they can produce unreliable results when processing data collected on a short timescale and with significant sampling limitations, as it is currently occurring for most countries affected by the epidemics [14]. We here introduce a novel framework to overcome these issues.

Intra-host minor variant analysis enhances the characterization of viral evolution. Due to the combination of high replication and mutation rates, populations of viruses with distinct sublineages can coexist within the same host, as shown by studies on numerous diseases [15, 16, 6]. Most mutations have no functional effect and follow a neutral evolutionary dynamics, driven by replication and degradation, and frequently involve genome positions that are highly prone to errors [17]. Certain variants, however, can be positively selected as a result of the strong immunologic pressure within human hosts [18]. Consequently, a complex architecture of viral (sub)lineages is often observed in infected individuals. In particular, it was recently shown that an unexpected high number of intra-hosts variants is observed in SARS-CoV-2 genomes, a proportion of which are observed in distinct hosts [19, 20]. Yet, the mechanisms of emergence and transmission of such variants in the population are still elusive.

In this respect, the characterization of intra-host minor variants detected within and across distinct viral lineages may allow to unveil two main scenarios (see Fig. 1A–D).

(1) Minor variants detected in hosts infected by the same viral lineage (i.e., displaying the same clonal variants) might indicate the possible presence of a (direct or indirect) *transmission path*, in which at least a portion of viral sublineages is transferred from an individual to another *during the infection* (see below). Accordingly, hosts sharing a significant proportion of identical minor variants might be at smaller evolutionary distance than hosts sharing clonal variants only (see Fig. 1A–C).

(2) *Homoplasies*, i.e., minor variants observed in hosts infected by different viral lineages (i.e., showing distinct clonal variants), might be emerged *after the infection*, because either positively selected in a parallel/convergent evolution scenario, or neutral, but falling in error-prone regions of the genome (see Fig. 1D).

Two related questions arise: *can the characterization of intra-host minor variants* (i) *improve the resolution of transmission chain maps, and* (ii) *be used to identify possibly functionally (positively) selected or recurrent neutral variants?*

Lessons from cancer evolution to reduce the impact of sampling limitations. The process of diffusion of variants in the population is driven by the complex interplay between genomic evolution of the virus *within* hosts and transmission *among* hosts. Even if the discussion on the topology of viral evolution is ongoing [21], it is reasonable to assume that all clonal mutations are typically transferred from a host to another during an infection event, while the extent of transmission of intra-host minor variants is still uncertain and depends on the combination of sublineage clonality, virulence and contact dynamics.

However, some of the most widely-used phylogenomic methods, such as MrBayes [22], Beast [23] and Nextstrain-Augur [8], rely on distance-based or parsimony algorithms and do not include any explicit constraint on the process of variants accumulation. Even though, in optimal sampling conditions and in a timescale adequate to cover the whole evolutionary history of a virus, such methods are expected to produce reliable results, current circumstances have highlighted extremely partial and inhomogeneous samplings, both in geographical and temporal terms, in several countries worldwide [14]. In such cases, the genetic distance among hosts might not reflect the temporal ordering among mutational events and, accordingly, may induce erroneous inferences on the infection chain (see Fig. 1E). Notice also that some of the aforementioned approaches integrate genomic and clinical data, such as collection date,

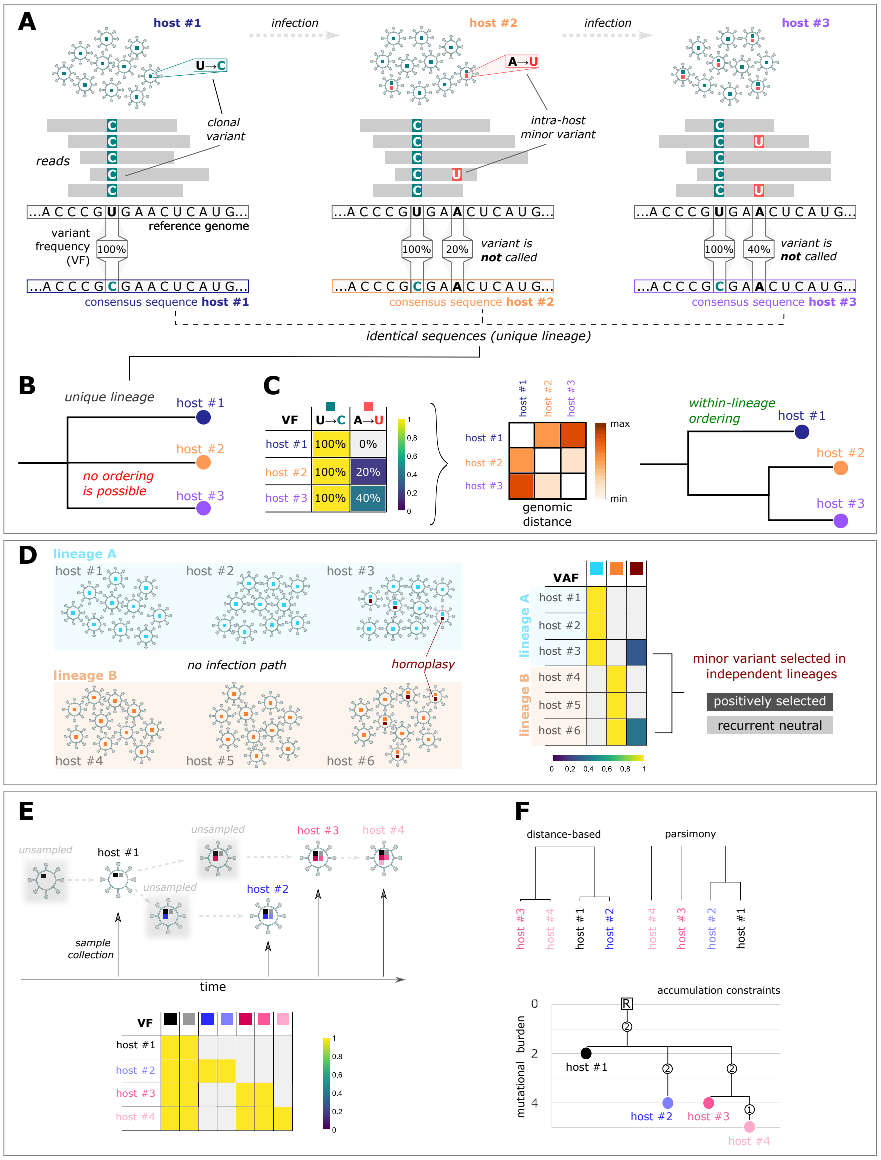

Figure 1: **(Sub)lineage evolution and transmission of viral genomes. (A)** In this toy example, three hosts infected by the same viral lineage are sequenced. In particular, all hosts share the same clonal mutation (U/C, green), but two of them (#2 and #3) are characterized by a distinct minor mutation (A/U, red), which randomly emerged in host #2 and was transferred to host #3 during the infection. Standard sequencing experiments return an identical consensus RNA sequence for all samples, by employing a threshold on variant frequency (VF) and by selecting mutations characterizing the dominant lineage. **(B)** By analyzing identical sequences, standard phylogenetic algorithms cannot disentangle any ordering or evolutionary relation among hosts infected by the same viral lineage. **(C)** By considering the VF distribution, it is possible to compute a refined genomic distance among hosts, as well as to identify a higher-resolution ordering within hosts infected by the same viral lineage, which may indicate possible transmission paths (in the example, we show a distance-based dendrogram). **(D)** In this second toy example, 6 infected hosts infected by two independent viral lineages are shown. An individual of each lineage (#3 and #6) display the same minor variant (dark red), which might indicate homoplasy. By analyzing the VF profile of all samples, it is possible to pinpoint such variant, which either is positively selected or recurrent neutral. **(E)** In this example, the branched evolution of 7 viral lineages is displayed (for simplicity all shown mutations are clonal and no sublineages are considered). 4 infected hosts harboring distinct clonal variants are tested and sequenced during the epidemics, revealing a typical scenario affected by sampling limitations. **(F)** From sequencing data of such hosts, distance-based and parsimony phylogenetic methods might return partial or incorrect evolutionary trees. By employing methods that account for mutation accumulation, the correct evolutionary model is inferred. In this representation, each sample is a leaf of the tree, positioned at a level corresponding to its mutational burden, whereas edges starting from the root R are labeled with the number of accumulating mutations. In the example, host #1 is parent of host #2 and #3, and the latter is a parent of host #4.

to better estimate temporal orderings among samples [23]. Unfortunately, however, collection date is often scarcely correlated with the date of contagion or with the onset of the disease. This aspect is intensified in COVID-19 for which the incubation period spans over a significantly large window (median 5.1 days, 95% CI, 4.5 to 5.8 days) [24] and for which the ratio of asymptomatic infected individuals is extremely high [25]. All these aspects may impact the reliability of downstream contact-tracing algorithms [26].

In order to reduce the impact of sampling limitations on the accuracy of predictions, it may be effective to employ methods that account for accumulation processes and for the possible existence of heterogeneous subpopulations within samples. In this regard, many computational methods have been developed for the analysis of somatic cancer evolution and intra-tumor heterogeneity from sequencing data and may be relevant alternatives in the context of viral evolution, due to the many analogies and to the similar data types [27]. Some methods, in particular, were proven to outperform standard distance-based phylogenetic models when a complex polyclonal architecture and conditions of limited samplings are observed [28], as it is expected in viral evolution and transmission scenarios.

A further question arises: *can methods accounting for mutation accumulation and uncertainty in the data be used to reduce the impact of sampling limitations and produce more reliable phylogenomic models?*

Results

A novel framework to investigate viral (sub)lineage evolution and transmission. In order to answer to the questions above, we here propose a new framework that: (1) considers both clonal and intra-host minor variants, (2) explicitly account for the accumulation of mutations following infection events and for the uncertainty in the data due to, e.g., partial transmission of minor variants or sequencing artifacts. Three major results have been achieved in this regard.

(A) By explicitly considering both clonal and intra-host minor variants and by employing probabilistic approaches that account for mutation accumulation processes and uncertainty in the data, it is possible to estimate highly refined phylogenomic models, as compared to standard approaches (see Methods and Fig. 1). On the one hand, this allows to reduce the impact of sampling limitations on the reconstruction of lineage ancestral relations. On the other hand, our approach allows to identify likely transmission paths (evolutionary relations) among hosts infected by the same viral lineage, while this is prevented to methods that process consensus sequences.

Such higher-resolution phylogenomic models can be then employed in downstream analyses, for instance by improving the estimation of molecular clocks and, accordingly, the predictive accuracy of epidemiological models, which typically rely on limited and inhomogeneous data [9, 14].

(B) Recent evidences point at an increasing genetic diversity of SARS-CoV-2 in human hosts [29, 19]. In this respect, the characterization of the viral (sub)lineage composition allows to improve the quantification of the genetic diversity in a given host or in a given cluster of samples, for instance by highlighting possible temporal trends.

(C) The analysis of homoplasies, i.e., of minor variants detected in independent viral lineages [21], may allow to identify positively selected genomic variants, as a result of functional *convergent evolution* and, similarly, to isolate specific regions of the viral genome under *purifying selection* (see Fig. 1D). This information might be then used to drive the design of opportune treatments/vaccines, for instance by blacklisting positively selected genomic regions and prioritizing target loci exposed to purifying selection.

We here present the (sub)lineage characterization of 18 SARS-CoV-2^+ patients from multiple studies, for which, at the time of writing, raw Illumina sequencing data are available in public databases (see Methods). By processing variant frequency profiles of both clonal and minor variants and by employing methods originally designed for the inference of cancer phylogenies from multiple samples [30, 31], we first provide a high-resolution phylogenomic model, which allows to identify a likely transmission network also among individuals infected by the same viral lineage. We then provide a quantitative evaluation of viral diversity in each individual, at the level of both clonal and intra-host minor variants. Finally, we identify a number of minor variants shared across independent viral lineages and which might likely be positively selected. In particular, mutation g.29039A>U, responsible for the Nucleocapsid variant p.Lys256*, is of particular interest because it abrogates the vast majority of the B and T bona fide antigenic epitopes of the target protein.

Phylogenomic analysis of 18 SARS-CoV-2^+ samples. In Figure 2A one can find the variant frequency profiles of 18 SARS-CoV-2^+ samples on 44 selected single-nucleotide variants, 21 of which are detected with high frequency (VF > 0.50) in at least one sample (samples are annotated with collection date and location; complete data are provided in Suppl. Table 1). Even though, as expected, the VF profiles of minor variants are noisy, patterns of co-occurrence are evident across samples and highlight a complex sublineage architecture.

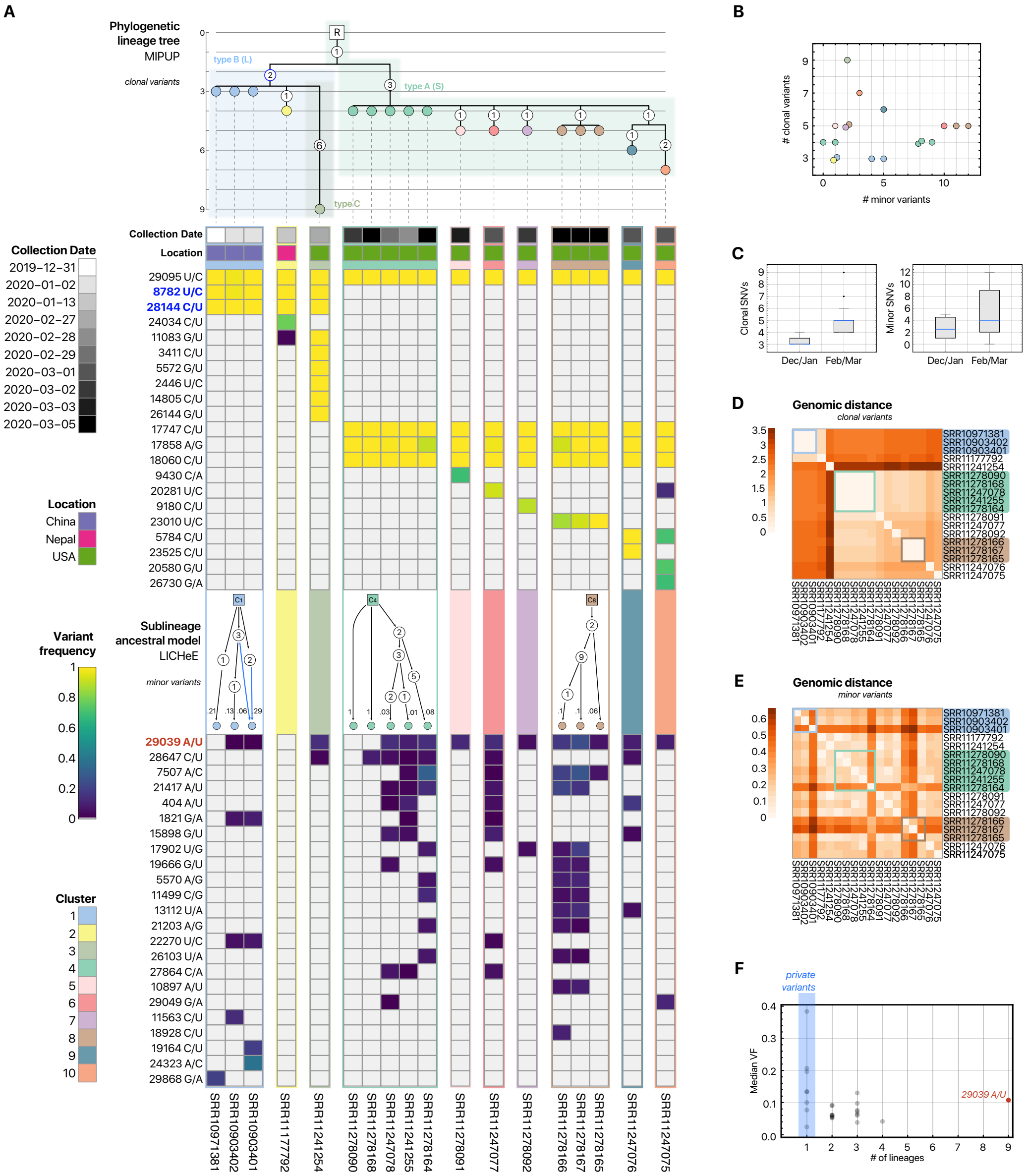

Figure 2: **(Sub)lineage characterization of** 18 **SARS-CoV-2$^+$ patients. (A)** Heatmap returning the variant frequency (VF) profiles of 18 SARS-CoV-2$^+$ patients on 44 selected single-nucleotide variants. Samples are annotated with collection date and location. The phylogenetic lineage tree returned by MIPUP [22] by considering clonal variants (VF > 0.50) is displayed at the top of the heatmap. The squared node labeled with capital R indicates the reference genome REF-ANC, whereas the circled nodes indicate the number of variants characterizing each ancestral relation; colored shades indicate SARS-CoV-2 types from [12, 32]. In the middle panel, the model returned by LICHeE on each non-singleton cluster, by considering minor variants only is shown. Leafs represent samples and the internal nodes indicate the number of minor variants accumulating along that path. Confluences (marked in blue) represent uncertain ancestral relations for samples with noisy VF profiles. Variants colored in blue characterize the B (L) type [12, 32] and the left branch of the model, whereas variant g.29039A>U is colored in red. **(B)** Scatterplot displaying for each sample the number of clonal and minor variants (colors represent the clusters of panel A). **(C)** Boxplots returning the distribution of the number of clonal and minor variants, obtained by grouping samples according to collection date (Dec/Jan vs. Feb/Mar). **(D–E)** Heatmamps returning the genomic distance computed on either clonal or minor variants. **(F)** Scatterplot returning, for each minor variant, the number of clusters/lineages in which the variant is found (x axis) and the median variant frequency (y axis). The blue shade indicates variants that are private in clusters/lineages. Mutation g.29039A>U, which is discussed in the text, is colored in red.

A phylogenetic lineage tree was reconstructed by applying MIPUP [31] to variants with VF $>$ 0.50, which would be included in consensus sequences and which we here consider as clonal (see Methods). The analysis allowed to identify 10 clusters of samples, corresponding to distinct viral lineages (i.e., displaying the same set of clonal mutations) and the evolutionary relations among them, which might represent major transmission events. Three viral lineages in particular are found in a significant number of samples: cluster C_1 (light blue, China, 3 samples), cluster C_4 (light green, Washington, 5 samples), cluster C_8 (brown, Washington, 3 samples). By construction, the model cannot disentangle any ordering within samples infected by the same lineage.

We then refined the analysis, by applying LICHeE to each (non-singleton) cluster and considering intra-host minor variants only (see Methods). This allowed to identify a high-resolution evolutionary model, which describes likely transmission paths among hosts characterized by the same viral lineage, and which would not be identifiable by processing consensus sequences only (see Fig. 2A). For example, samples SRR11278166 and SRR11278167 share 11 minor mutations, which might have been transferred from a host to another during infection. Clearly, once detailed contact tracing of considered individuals would be available, this will allow to validate the hypothesized transmission paths.

The whole phylogenomic model reveals the presence of two major branches, corresponding to previously identified SARS-CoV-2 types [12, 32]. In detail, the right branch of our model is characterized by the absence of SNVs g.8782U>C and g.28144C>U and corresponds to type A [12] (also type S [32]), which was hypothesized to be the ancestral SARS-CoV-2 type, according to the bat outgroup coronavirus and the similarity with highly related viruses. Following this classification, reference genome REF-ANC (see Methods) would belong to type A (S) and, in particular, to the T-subcluster identified in [12] as ancestral genome.

More in detail, variant g.29095U>C, which is present in all 18 samples of the dataset, is the likely first evolutionary event from reference genome REF-ANC and precedes the emergence of the two major branches, whereas SNVs g.17747C>U, g.17858A>G and g.18060C>U indicate subsequent events characterizing the evolution of the right branch. Notice that all samples of the right branch are from Washington and, from the analysis of the lineage phylogenetic tree, the transmission chain likely started by first involving the samples of cluster C_4. In particular, according to the sublineage analysis, an infection event might have likely involved samples SRR11241255 and SRR11247078, which share 5 minor variants, and possibly sample SRR11278164.

The left branch of our model is characterized by the presence of both SNVs g.8782U>C and g.28144C>U and corresponds to type B [12] (also type L [32]). This branch includes all Asian samples plus a sample from Washington (SRR112412), who displays a significant number of additional clonal variants, including mutation g.26144G>U, which identifies type C in [12]. Notice that, from our analysis of minor variants, the most likely ancestral sample of this viral lineage in this dataset is SRR10971381 (Wuhan; also employed in other studies as reference genome [3], GenBank: MN908947.3), a result that is confirmed by the collection date.

Interestingly, the B (L) type appear to be more abundant in the population ($5247/6187 \approx 85\%$ of the samples on GISAID database [7]; update April, 15th, 2020) and was diffused especially in the early stages of the epidemic in Wuhan [32], whereas the ancestral A (S) type appears to be particularly spread outside China. We also note that considering the actual shortage of clinical data on SARS-CoV-2$^+$ patients, there are insufficient elements to support any epidemiological claim on virulence and pathogenicity of the distinct SARS-CoV-2 types.

Standard methods for phylogenomic inference appear to struggle in the analysis of this dataset. In Suppl. Fig. 1 we show the phylogenetic model returned by MrBayes on clonal variants. First, the model returns a polytomy in which all the elements of the left branch of our model (type B (L)) and a clade including all samples of the right branch (type A (S)) are positioned at the same level of the tree. Second, for most samples characterized by distinct viral lineages the method cannot return any temporal ordering, despite different distances with the ancestors are returned. For instance, from the MrBayes model it is not possible to determine whether sample SRR11241254 is a child of either sample SRR11177792 or of cluster C_1, while it is evident from both our model and the heatmap (Fig. 2A) that two independent branches originate from cluster C_1 toward both samples. Finally, by construction, the method cannot identify any distance-based ordering among individuals infected by the same viral lineage. Also the Nextstrain-Augur model (Suppl. Fig. 2) appears to show some limits. For instance, the model is able to distinguish the two main branches, but the ordering among clades appear to be quite arbitrary, especially with respect to the elements of type A (S) (right branch of our model). The results of both approaches likely derive from the sampling limitations of currently available datasets and show that, in such cases, methods that explicitly account for mutation accumulation might be more appropriate to reconstruct reliable models of viral evolution and transmission.

By comparing the number of clonal and of minor variants in each host (Figure 2B), it is possible to notice that even individuals characterized by the same viral lineage may display a significantly different number of minor variants, with different distributions observed across lineages. In this respect, the quantitative analysis of genomic distance among hosts (Fig. 2D–E, see Methods) shows that minor variants can be used to better assess within-lineage genomic diversity.

Importantly, the comparison of the distribution of the number variants obtained by grouping samples with respect to collection dates (December/January vs. February/March; see Fig. 2C) allows to highlight a noteworthy increase for both clonal and minor variants (median number of variants $+100\%$, one-sided Mann–Whitney U test on all variants $p = 0.039$). This result would support the hypothesis that the overall genomic diversity of SARS-CoV-2 is progressively increasing [29, 19], also proving that such phenomenon is partially reflected on minor variants. g

Interestingly, several intra-host minor variants are found in a relatively large number of samples, as well as in different clusters/lineages and might indicate homoplasies. In this respect, in Fig. 2F we display for each minor variant the number of clusters/lineages in which it is observed, as compared to its median variant frequency. One could suppose that variants that are privately detected in single clusters (left region of the scatterplot) might result from infection events, whereas those found in a considerable number of lineages (right region of the scatterplot) are either positively selected or recurrent neutral mutations, although the presence of sequencing artifacts cannot be formally excluded, despite the high quality of all the reported variant calls. Moreover, higher values of median variant frequency indicate increased confidence on the variant, as well as possible ongoing selection shifts.

A variant in particular (g.29039A>U) is found in significant abundance in $14/18$ samples and in $9/10$ lineages (median VF ≈ 0.11, max VF ≈ 0.25) and deserves further investigation, as it might suggest the presence of functional convergence.

SARS-CoV-2 mutation g.29039A>U causes a stop gain event occurring in the linker region of the Nucleocapsid protein. The g.29039A>U is a nonsense mutation occurring in the linker region of the Nucleocapsid protein, close to the $3'$ end of the viral genome and causing the substitution of Lysine 256 with a stop codon (p.Lys256*; see Fig. 3). This mutation is present, at subclonal level, in $14/18$ samples of our study cohort, with a median intra-host variant frequency of 0.1076 and is also confirmed in dataset PRJNA6079. The locus of the g.29039A>U mutation is particularly complex, with the mutated nucleotide being the center of symmetry of a perfect 19bp palindromic hairpin sequence. Notably, in many samples the presence of a large number of supplementary reads suggests that this hairpin region may be also a hotspot for more complex rearrangements, likely occurring through RdRp mediated template switching (Suppl. Fig. 3). The functional effect of these rearrangements is the generation of a stop codon in the linker region, similarly to what observed for g.29039A>U. Studies done on SARS-CoV Nucleocapsid (N) protein allowed to identify two non-interacting structural domains, one N-terminal, known as the RNA-Binding Domain (RBD) and one C-terminal, known as the dimerization domain (DD) [33]. RBD and DD domains are separated by a disordered region playing a putative role as a flexible linker. The RBD directly interacts with the viral single-stranded RNA, generating the ribonucleoprotein core, while the C-terminal DD domain is responsible for protein dimerization, although its functional role is yet unknown. A comparison between the N protein of SARS-CoV and SARS-CoV-2 reveals a percent identity and similarity of 90.5% and 94.1%, respectively (Suppl. Fig. 4), indicating that the two N proteins share a very similar structural organization. Recently, a set of putative SARS-CoV-2 B- and T-immunogenic peptides were identified by homology to those experimentally defined in SARS-CoV [34, 35]. Interestingly, the majority of the immunogenic epitopes of the Nucleocapsid protein (13 out of 18) are located in the C-terminus of the protein, in the linker region and DD domain and right after Lysine 256 (see Fig. 3; Suppl. Table 2).

In this context the identification of a subloclonal nonsense variant such as the p.Lys256* occurring before the C-terminus of the SARS-CoV-2 Nucleocapsid is particularly intriguing because it suggests that the mutation may have been selected in order to mask one of the most immunogenic viral regions [34, 35] under the selective pressure of the host immune response. Indeed, several studies demonstrated that antibodies raised against the N protein upon SARS-CoV infection are abundantly expressed [36], albeit short-lived [37]. In contrast, T cell-mediated immune responses can provide long-term protection after SARS-CoV infection [38, 39, 37]. In line with these data, sequence homology and bioinformatic approaches indicate that the Nucleocapsid, together with the Spike protein, are among the most promising targets for DNA vaccine development [40]. Not surprisingly, vaccines for both proteins have been already developed for SARS [41]. In this context, the identification of a subclonal nonsense variant, occurring in patients' samples and suppressing the expression of the immunogenic Nucleocapsid DD domain, may suggest a potentially limited effectiveness of veccines targeting this region for at least two reasons: *i)* because resistant subclones, such as the g.29039A>U variant, are already present in the wild, probably owing to the natural selective pressure exerted by the host immune system and *ii)* because the DD domain could be non-essential for the survival and infectiousness of the virus, as opposite to the RBD and therefore it could be easily lost upon exposure to the selective pressure raised by the vaccine.

We finally specify that currently available data do not allow to assess the VF distribution of g.29039A>U in the population, because for the large majority of samples in public databases only consensus sequences are available. To this extent, we stress that once a larger amount of raw sequencing data on SARS-CoV-2 would be available, it will be possible to analyze which and how minor variants undergo an increase in their frequency in the population, as this might be used to assess the presence of selection-driven shifts.

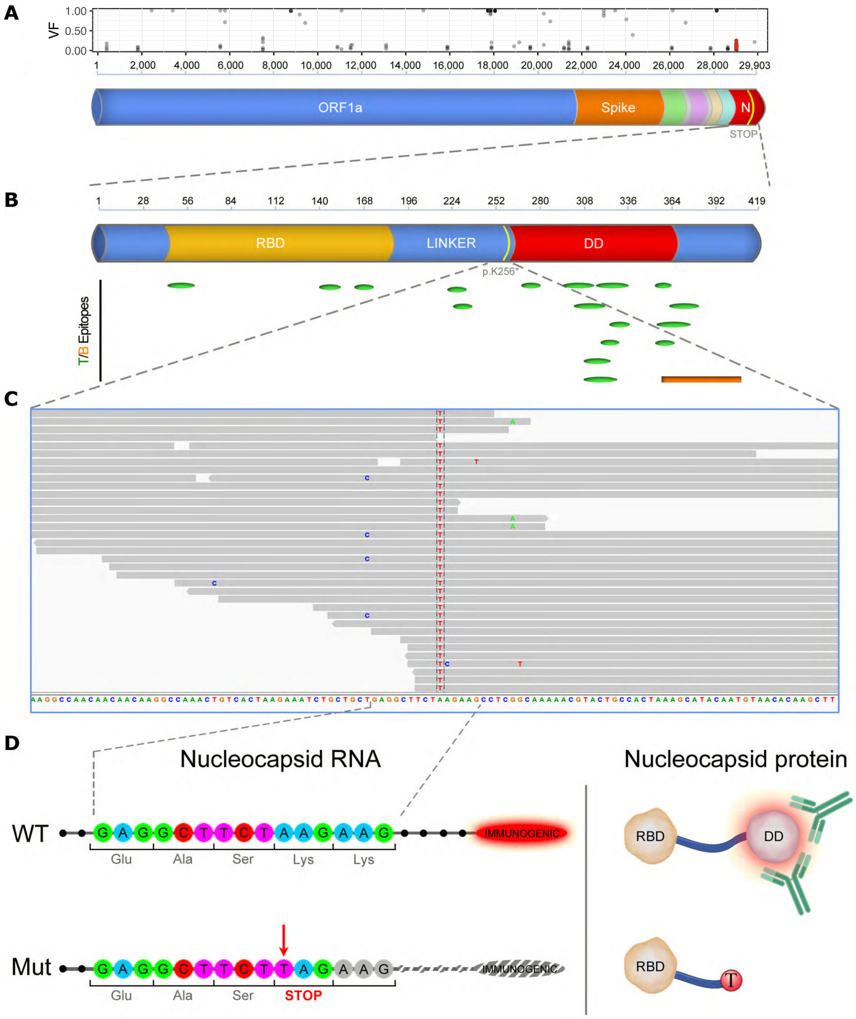

Figure 3: **Functional effect of g.29039A>U mutation.** **(A)** Schematic representation of the entire SARS-CoV2 genome structure. The N protein is shown in red (C-terminus). The yellow band indicates the position of the p.Lys256* mutation. In the top panel, the VF profiles of all 44 SNVs used in the phylogenomic analysis are shown in correspondence to genome position. **(B)** Nucleocapsid protein. The yellow region represents the RNA Binding Domain (RBD), the central blue region is the linker and the red one represents the Dimerization Domain (DD). The yellow band indicates the position of the p.Lys256* mutation within the linker region. T and B epitopes are represented as green ovals and orange rectangle, respectively. **(C)** Representative set of reads from validation sample SRR11140744 showing the presence of the g.29039A>U (p.Lys256*) variant. The mutated base is highlighted by the presence of the red 'T'. **(D)** Schematic cartoon showing the effect of the g.29039A>U variant at RNA (left) and protein (right) level in comparison with the wild-type sequence.

Discussion

We have shown that, by considering both clonal and intra-host minor variants and by explicitly accounting for the mutation accumulation process, it is possible to highly improve the resolution of the phylogenomic analysis of the SARS-CoV-2 evolution and reduce the impact of sampling limitations. This may represent a paradigm shift in the analysis of viral genomes and should be quickly implemented in combination to data-driven epidemiological models, to deliver a high-precision platform for pathogen detection and surveillance [10, 42].

This might be particularly relevant for countries which recently suffered outbreaks of exceptional proportions, such as Italy, Spain and USA, and for which the limitations and inhomogeneity of diagnostic tests have proved insufficient to define reliable descriptive/predictive models of disease evolution and spread. For instance it was recently hypothesized that the rapid diffusion of COVID-19 might be likely due to the extremely high number of untested asymptomatic hosts [25]. A better estimation of undetected infected hosts can be derived from high-resolution phylogenomic analyses of tested individuals, also in conditions of severe sampling limitations.

We have also shown that the analysis of intra-host minor variants can produce experimental hypotheses with translational relevance and which might drive the design of treatment or vaccines. For instance, variant g.29039A>U was found to suppress the vast majority of the B and T epitopes of the highly immunogenic Nucleocapsid protein, while leaving its entire RNA Binding Domain intact. It is therefore tempting to speculate that this subclonal event may be the result of the selection pressure exerted by the host immune system against the virus and may raise an alert on the development of vaccines specifically targeting the highly immunogenic C-terminal region of the Nucleocapsid protein.

Methods

Dataset description. We studied a cohort of 18 samples from distinct individuals obtained from 4 NCBI BioProjects, which, at the time of writing, are the only publicly available datasets including raw Illumina sequencing data. In detail, we selected the following project: (1) PRJNA601736, 2 individuals geographically located in Wuhan (China) for which we considered RNA-seq (Illumina MiSeq) from bronchoalveolar lavage fluid (BALF); (2) PRJNA603194, 1 individual from Wuhan (China) with RNA-seq (Illumina MiSeq, BALF) from [2]; (3) PRJNA608651, 1 individual from Nepal with RNA-seq (Illumina MiSeq) isolate from oro-pharyngeal swab of [43]; (4) PRJNA610428, 14 individuals geographically located in the State of Washington (USA) for which we considered RNA-seq data (6 Illumina MiSeq and 8 Illumina NextSeq 500). Furthermore, we considered 4 additional samples from NCBI BioProject PRJNA607948 all obtained from one unique individual from Wisconsin, USA (1 swab and 3 independent passage isolates; Illumina MiSeq), which were used to validate the discovered minor variant g.29039A>U.

Pipeline for variant calling. We downloaded SRA files from 5 NCBI BioProjects with the following accession numbers: PRJNA601736, PRJNA603194, PRJNA607948, PRJNA608651 and PRJNA610428; then we converted the samples to FASTQ files using SRA toolkit. Following [43], we used Trimmomatic (version 0.39) to remove the nucleotides with low quality score from the RNA sequences with the following settings: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:40.

Since different reference genomes have been employed in the analysis of SARS-CoV-2, we here selected as candidate references two high-quality genome sequences from human samples: sequence EPI_ISL_405839 (downloaded from GISAID, GenBank: MN975262.1, ref. #1 in the following), used, e.g., in [43] and sequence EPI_ISL_402125 (GenBank: MN908947.3, ref. #2 in the following), employed in several studies on SARS-CoV-2, e.g., [3]. The sequence comparison highlighted the presence of only 5 SNPs, at locations: 8782 (ref. #1 U, ref. #2 C), 9561 (ref. #1 U, ref. #2 C, respectively), 15607 (ref. #1 C, ref. #2 U), 28144 (ref. #1 C, ref. #2 U) and 29095 (ref. #1 U, ref. #2 C). Hence, in order to define a unique ancestral reference genome to be used in downstream analyses, we here employed the bat coronavirus RaTG13 genome from [1] (GenBank: MN996532.1) to disambiguate the SNPs at those locations. Accordingly, we here define the reference genome REF-ANC with haplotype UCUCU at the 5 locations listed above. Reference genome REF-ANC represents an extremely likely ancestral genome, preceding both ref. #1 and ref. #2, and was used for variant calling (REF-ANC is released in FASTA format as Suppl. File 1).

We then used bwa mem (version 0.7.17) to map reads to REF-ANC. We generated sorted BAM files from bwa mem results using SAMtools (version 1.10) and removed duplicates with Picard (version 2.22.2). Variant calling was performed generating mpileup files using SAMtools and then running VarScan (min-var-freq parameter set to 0.01) [44].

Finally notice that, one should be extremely careful when considering low-frequency variants, which might possibly result from sequencing artifacts, even in case of high-coverage experiments. For this reason, we here employed further significance filter on variants. In particular, we kept only the mutations: (1) showing a VarScan significance p-value < 0.05 (Fisher's Exact Test on the read counts supporting reference and variant alleles) in at least 50% of the samples, (2) displaying a variant frequency $VF > 5\%$ in 2 or more samples or $VF > 10\%$ in a single sample. As a result, we selected a list of 44 highly-confident SNVs to be included in the analysis.

Phylogenomic analysis. In order to reconstruct a high-resolution phylogenomic model of viral evolution, also in condition of sampling limitations, we designed a two-step procedure that employs two methods originally designed for the inference of cancer phylogenies from sequencing data of multiple samples, namely MIPUP [31] and LICHeE [30].

In the first step, MIPUP is employed to explicitly model the accumulation process of clonal variants in the population and identify the phylogenetic lineage tree. In detail, MIPUP searches minimum perfect unmixed phylogenies on binarized VF profiles and without the employment of any further data-specific hypothesis. The method first binarizes the VF profiles according to a user-defined threshold, then it solves a minimum conflict-free row split problem via Integer Linear Programming and finally returns all the orderings of mutations that do not violate the accumulation hypothesis. In the output tree (graph), samples are the leafs and every edge is marked by the set of mutations that occurred along that path.

In particular, we applied MIPUP to clonal variants, i.e., by selecting a VF threshold equal to 0.50, since such variants would be included in consensus sequences. The phylogenetic lineage model so obtained describes clusters of samples as lineages and highlights the ancestral relations among them (see Fig. 2A). Notice that, by construction, no ordering is possible among individual infected by the same viral lineage when considering clonal variants only. Standard phylogenomic analyses were also performed by applying MrBayes [22] to binarized VF profiles (default parameters, VF binarization threshold $= 0.50$) and Nexstrain-Augur [8] to consensus sequences of the 18 samples retrieved by GISAID (default parameters; the models are shown in the Suppl. Fig. 3 and 4.)

A second step is then defined to improve the resolution of the phylogenomic analysis and identify the likely ancestral relations within each cluster of samples (lineage) retrieved during the first step, by considering minor variant profiles. While it is safe to binarize clonal variant profiles to reconstruct a phylogenetic lineage tree, the analysis of minor variants requires certain precautions. First, minor variant profiles might be noisy, due to the relatively low abundance and to the technical features of sequencing experiments. Accordingly, such data may possibly include artifacts, which can be partially mitigated during the quality-check phase and by including in the analysis only highly-confident variants.

Second, the extent of transmission of minor variants among individual is still uncertain and depends on sublineage clonality, virulence and contact dynamics [6]. Due to the low frequency, for instance, minor variants may be also affected by both bottlenecks and founder effects, according to which a certain sublineage might either be not transmitted or become clonal in the infected host [45]. Even if an in-depth investigation of inter-host transmission of minor variants is beyond the scope of the current work, a preliminary analysis of our dataset showed evident patterns of co-occurrence of minor variants among individuals infected by the same viral lineage, which would support the hypothesis of transmission of at least of portion of sublineages from a host to another.

For these reasons and at a first approximation, we can here reasonably assume that, in a significant number of cases, sublineages are transmitted from an individual to another in proportion to their abundance in the infecting host. In such cases, the variants accumulation hypothesis hold and can be used to reconstruct a likely within-lineage transmission chain. In particular, in order to manage uncertainty in the data and to handle variant transmission and accumulation processes in samples composed by heterogeneous sublineage mixtures, we here employed LICHeE [30] a method originally designed for the inference of cancer phylogenies from multi-sample sequencing data.

More in detail, we applied LICHeE to each non-singleton cluster of samples retrieved in the previous analysis. LICHeE deconvolves the VF profiles of samples possibly composed by heterogeneous subpopulations, under the hypothesis of a process of variants accumulation (notice the no further data-specific hypotheses are employed). The method first partitions variants into groups according to the occurrence in each sample, then it clusters those showing similar VF profiles across samples, and finally returns all the orderings of such clusters that do not violate the accumulation hypothesis. The result is a Directed Acyclic Graph where samples are the leafs and can be reached from single or multiple directed paths. In this context, the latter case (marked in blue in the output model) would suggest that the sublineage architecture is insufficient to allow the identification of a unique ancestral path.

By applying LICHeE (with default parameters) to the VF profiles of minor variants of our dataset (VF < 0.50), we obtain a fine-grained phylogenetic model for each (non-singleton) cluster, which highlights the likely ancestral relations among hosts infected by the same viral lineage, i.e., indicating possible transmission paths in which a portion of minor variants (sublineages) is transferred from a host to another.

Similarly to [6], here we also define a genomic distance among samples, based on VF profiles, which can be computed on both clonal and intra-host minor variants. Let be $\mathbf{p}$ and $\mathbf{q}$ the n-dimensional vector representing the VF values of the n variants for two different samples; the L_2 distance $d(\mathbf{p}, \mathbf{q})$ between two samples is given by:

$$d(\mathbf{p}, \mathbf{q}) = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2} . \tag{1}$$

That is, we computed the pairwise L_2 distance among all samples, either considering clonal SNVs (VF > 0.50) or minor intra-hosts SNVs (VF < 0.50). Results are shown in Fig. 2D–E and demonstrate how the genomic distance computed on intra-host minor SNVs provides a fine-grained information.

Software availability

The source code used to replicate all the analyses is available at this link:
`https://github.com/BIMIB-DISCo/SARS-CoV-2-IHMV`.

Authors contributions

D.R., F.A., D.M., A.G. and R.P. designed the approach. D.R., F.A., D.M. and A.G. defined the computational methods, which were implemented and executed by D.R and D.M. R.P. analyzed and validated the variant g.29039A>U. D.R., F.A., D.M., C.G., M.A., A.G. and R.P. analyzed the data and interpreted the results. R.P. supervised the experimental data analysis. A.G. and D.R. supervised the computational analysis. A.G. and R.P. drafted the manuscript, which all authors discussed, reviewed and approved.

Acknowledgments

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures and by the AIRC-IG grant 22082. Support was also provided by the CRUK/AIRC Accelerator Award #22790, "Single-cell Cancer Evolution in the Clinic". We thank Giulio Caravagna and Chiara Damiani for helpful discussions.

References

[1] Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

[2] Wu, F. *et al.* A new coronavirus associated with human respiratory disease in china. *Nature* **579**, 265–269 (2020).

[3] Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* 1–3 (2020).

[4] Grubaugh, N. D., Petrone, M. E. & Holmes, E. C. We shouldn't worry when a virus mutates during disease outbreaks. *Nature Microbiology* 1–2 (2020).

[5] Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).

[6] Poon, L. L. *et al.* Quantifying influenza virus diversity and transmission in humans. *Nature genetics* **48**, 195 (2016).

[7] Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22** (2017).

[8] Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

[9] Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS computational biology* **9** (2013).

[10] Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in brazil and the americas. *Nature* **546**, 406–410 (2017).

[11] Lai, A., Bergna, A., Acciarri, C., Galli, M. & Zehender, G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *Journal of medical virology* (2020).

[12] Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* (2020).

[13] Wright, C. F. *et al.* Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology* **85**, 2266–2275 (2011).

[14] Mavian, C. *et al.* Regaining perspective on SARS-CoV-2 molecular tracing and its implications. *medRxiv* (2020).

[15] Miralles, R., Gerrish, P. J., Moya, A. & Elena, S. F. Clonal interference and the evolution of RNA viruses. *Science* **285**, 1745–1747 (1999).

[16] Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).

[17] Gojobori, T., Moriyama, E. N. & KimurA, M. Molecular clock of viral evolution, and the neutral theory. *Proceedings of the National Academy of Sciences* **87**, 10015–10018 (1990).

[18] Lucas, M., Karrer, U., Lucas, A. & Klenerman, P. Viral escape mechanisms–escapology taught by viruses. *International journal of experimental pathology* **82**, 269–286 (2001).

[19] Shen, Z. *et al.* Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. *Clinical Infectious Diseases* (2020).

[20] Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* 1–10 (2020).

[21] Chan, J. M., Carlsson, G. & Rabadan, R. Topology of viral evolution. *Proceedings of the National Academy of Sciences* **110**, 18566–18571 (2013).

[22] Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61**, 539–542 (2012).

[23] Bouckaert, R. *et al.* Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology* **15**, e1006650 (2019).

[24] Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine* (2020).

[25] Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (2020).

[26] Jombart, T., Eggo, R., Dodd, P. & Balloux, F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* **106**, 383–390 (2011).

[27] Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* (2017).

[28] Yuan, K., Sakoparnig, T., Markowetz, F. & Beerenwinkel, N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology* **16**, 36 (2015).

[29] Li, X. *et al.* Transmission dynamics and evolutionary history of 2019-nCoV. *Journal of Medical Virology* (2020).

[30] Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome biology* **16**, 91 (2015).

[31] Husić, E. *et al.* MIPUP: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP. *Bioinformatics* **35**, 769–777 (2019).

[32] Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* (2020).

[33] Chang, C.-k. *et al.* Modular organization of SARS coronavirus nucleocapsid protein. *Journal of biomedical science* **13**, 59–72 (2006).

[34] Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* **12**, 254 (2020).

[35] Fast, E. & Chen, B. Potential T-cell and B-cell epitopes of 2019-nCoV. *bioRxiv* (2020).

[36] Ying, L. *et al.* Identification of an epitope of SARS-coronavirus nucleocapsid protein. *Cell research* **13**, 141–145 (2003).

[37] Tang, F. *et al.* Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. *The Journal of Immunology* **186**, 7264–7268 (2011).

[38] Peng, H. *et al.* Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein in SARS-recovered patients. *Virology* **351**, 466–475 (2006).

[39] Fan, Y.-Y. *et al.* Characterization of SARS-CoV-specific memory T cells from recovered individuals 4 years after infection. *Archives of virology* **154**, 1093–1099 (2009).

[40] Grifoni, A. *et al.* A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host & Microbe* (2020).

[41] Ong, E., Wong, M. U., Huffman, A. & He, Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *BioRxiv* (2020).

[42] Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics* **19**, 9 (2018).

[43] Bastola, A. *et al.* The first 2019 novel coronavirus case in Nepal. *The Lancet Infectious Diseases* **20**, 279–280 (2020).

[44] Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568–576 (2012).

[45] Gutierrez, S. *et al.* Circulating virus load determines the size of bottlenecks in viral populations progressing within a host. *PLoS pathogens* **8** (2012).