QUANTIFICATION OF INTRA-HOST GENOMIC DIVERSITY OF SARS-CoV-2 ALLOWS A HIGH-RESOLUTION CHARACTERIZATION OF VIRAL EVOLUTION AND REVEALS FUNCTIONALLY CONVERGENT VARIANTS

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] Dept. 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 affected by the epidemics.

We introduce VERSO (Viral Evolution ReconStructiOn), a new comprehensive framework for the characterization of viral evolution and transmission from sequencing data of viral genomes. Our approach accounts for accumulation of clonal mutations and uncertainty in the data, by taking advantage of the achievements of research in cancer evolution, to deliver robust phylogenomic lineage models, and exploits intra-host variant frequency profiles to characterize the sublineage similarity among samples, which may derive from uncovered infection events.

The application of our approach to RNA-sequencing data of 162 SARS-CoV-2 samples generates a high-resolution model of evolution and spread, which improves recent findings on viral types and highlights the existence of patterns of co-occurrence of minor variants, revealing likely infection paths among hosts harboring the same viral lineage. The in-depth analysis of the mutational landscape of SARS-CoV-2 confirms a statistically significant increase of genomic diversity in time and identifies a number of variants that are transiting from minor to clonal state in the population. We also show that standard phylogenetic methods can produce unreliable results when handling datasets with noise and sampling limitations, as proven by the further application of VERSO to 12419 consensus sequences included in GISAID database.

Notably, VERSO allows to pinpoint minor variants that might be positively or negatively selected across distinct lineages, thus driving the design of treatments and vaccines. In particular, minor variant g.29039A>U, detected in multiple viral lineages and validated on independent samples, 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. Finally, we here release the likely SARS-CoV-2 ancestral genome, obtained by resolving ambiguous SNPs that distinguish two widely-used reference genomes from human samples, by employing the Pangolin-CoV and the Bat-CoV-RaTG13 genomes.
Our results show that the joint application of our framework and data-driven epidemiological models might allow to deliver a high-resolution platform for pathogen surveillance and analysis.

Introduction

The outbreak of coronavirus disease 2019 (COVID-19), which started in late 2019 in Wuhan (China) [1, 2] and was declared pandemic by the World Health Organization, is fueling the publication of an increasing number of studies aimed at exploiting the information provided by 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, 4, 5]. As a matter of fact, the mutations that rapidly accumulate in the viral genome [6] can be used as natural *barcodes* to track the evolution of the virus and, accordingly, unravel the viral infection network [7, 8].

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 [9], whose data are accessible via the Nextstrain portal [10]. 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 [11, 12].

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 [13, 14]. However, while such analyses are undoubtedly effective in unraveling the main patterns of evolution of the viral genome, at least two issues can be raised.

First, most methods are missing key information on *intra-host minor variants*, which can be retrieved from whole-genome deep sequencing raw data [15]. In our opinion, this information might be essential to produce high-resolution models of the SARS-CoV-2 evolution and, accordingly, of the likely transmission chain.

Second, most approaches do not explicitly model the processes of *accumulation* of clonal variants in the population and do not account for *uncertainty* in the data due to, e.g., sequencing artifacts [16]. For these reasons, they can produce unreliable results when processing data collected on a short timescale and with significant sampling limitations, as it has occurred for most countries during the epidemics [17, 18]. We here introduce a novel framework to overcome these issues.

The analysis of intra-host genomic diversity enhances the characterization of viral evolution. Due to the high replication and mutation rates of RNA viruses, populations of distinct viral sublineages (also known as viral *quasispecies* [19]) can often coexist within the same host, as shown by studies on numerous diseases [20, 21, 22, 15, 8, 23].

In particular, most intra-host minor variants have no functional effect, follow a neutral evolutionary dynamics and often involve *mutational hotspots* [24]. Certain variants, instead, can be positively selected as a result of the strong immunologic pressure within human hosts [25]. Furthermore, the mechanisms of transmission of minor variants in the population are still elusive and depend on the combination of sublineage clonality, virulence and contact dynamics, which results in the presence of *bottlenecks*, *founder effects* and *stochastic fluctuations* [26, 22]. The complex architecture of viral sublineages observed in infected individuals is the result of this complex and undeciphered interplay. 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 detected in distinct hosts [27, 28].

In this regard, the analysis of variant frequency profiles retrieved from raw sequencing data allows to overcome the limitations of approaches based on consensus sequences, which, on the one hand, do not consider intra-host minor variants and, on the other hand, might erroneously call certain minor variants as clonal, according to arbitrary frequency thresholds. Moreover, such analysis allows to examine three important phenomena (see Fig. 1A and Suppl. Fig. 1).

(1) Patterns of co-occurrence of minor variants in hosts infected by the same viral lineage (i.e., displaying the same clonal variants) might indicate the possible presence of a *transmission path*, in which at least a portion of viral sublineages has been transferred from an individual to another *during the infection*. Accordingly, one can hypothesize that hosts sharing a significant proportion of identical minor variants might be at smaller evolutionary distance than hosts sharing clonal variants only (see Fig. 1A).

(2) *Homoplasies*, i.e., minor variants observed in hosts infected by different viral lineages (i.e., displaying distinct clonal variants), might have 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 Suppl. Fig. 1A).

(3) Variants that are detected with low frequency in certain individuals and as clonal in others might indicate the existence of *selection shifts* and/or of *founder effects* involving specific sublineages during infection events.

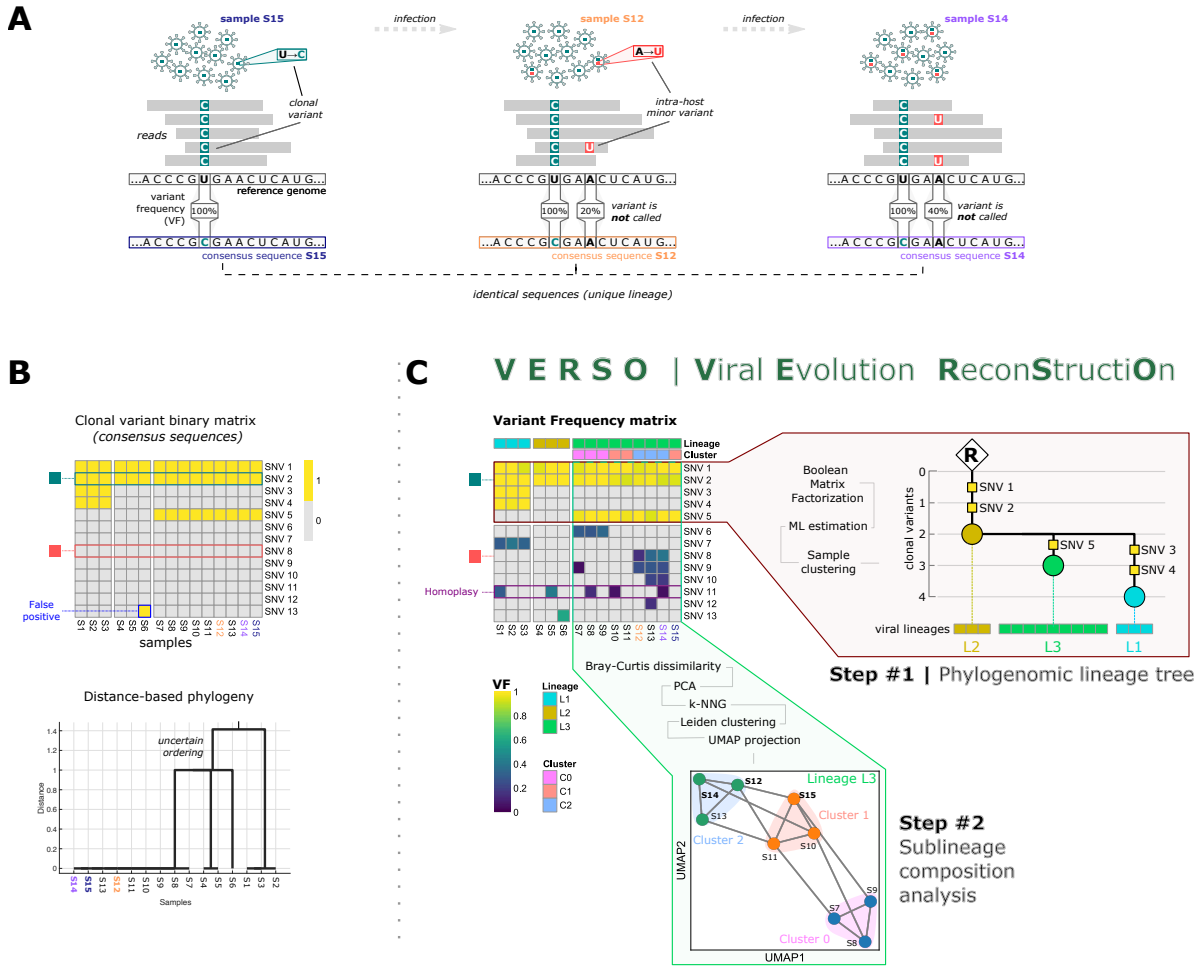

Figure 1: **VERSO framework for viral evolution and sublineage characterization. (A)** In this toy example, three hosts infected by the same viral lineage are sequenced. In particular, all hosts share the same clonal mutation (U/C, green), but two of them (#2 and #3) are characterized by a distinct minor 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 and not considering any constraint on variant accumulation, standard phylogenetic algorithms cannot disentangle any ordering or evolutionary relation among hosts infected by the same viral lineage and can return uncertain ordering among lineages. Furthermore, false positives can confound the analysis. **(C)** VERSO is a new framework for the characterization of viral evolution, which processes VF profiles from raw sequencing data. In step #1 VERSO returns a phylogenomic lineage tree on binarized profiles of clonal variants (VF $> 90\%$), by relying on a probabilistic framework and assuming variant accumulation (see Methods). The output tree model describes viral genotypes or lineages, associated with clusters of samples, whereas edges represent mutation accumulating from reference genome. In step #2 VERSO computes a refined genomic distance among hosts on VF profiles, which is employed with clustering and dimensionality reduction techniques to project and visualize samples on a 2D space, representing the sublineage composition and the distance among hosts, also allowing to identify possible intra-lineage transmission paths. Homoplasies can be investigated to pinpoint positively selected variants or mutational hotspots.

Two related questions arise: *can the characterization of intra-host genomic diversity* (i) *improve the resolution of transmission chain maps, and* (ii) *be used to identify positively and negatively selected variants and variants undergoing transition to clonality?*

Lessons from cancer evolution to reduce the impact of noise and 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 [29], it is reasonable to assume that clonal mutations (i.e., with a variant frequency close to 100%) are likely transferred from a host to another during an infection event, while, as specified above, the extent of transmission of intra-host minor variants is

still uncertain. We notice also that, while multiple infections of the same host from distinct viral lineages might in principle affect variant clonality, their occurrence is extremely rare.

However, some of the most widely-used phylogenomic methods, such as MrBayes [30], Beast [31] and Augur [10] (used for the analyses in Nextstrain), rely on distance-based or parsimony algorithms and do not include any explicit constraint on the process of variants accumulation. Furthermore, most techniques do not typically account for noisy observations, which can derive, e.g., from sequencing issues or from biological variability.

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 [18]. In such cases, the simple 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. 1B and Suppl. Fig. 1B-C). Notice also that some of the aforementioned approaches integrate genomic and clinical data, such as collection dates, to better estimate temporal orderings among samples [31]. Unfortunately, however, collection dates may be 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 is typically long (symptoms onset occurs within 11.5 days for 97.5% of infected persons, CI, 8.2 to 15.6 days [32]) and for which the ratio of asymptomatic infected individuals is extremely high [33]. All these aspects may impact the reliability of downstream contact-tracing algorithms [34].

In order to reduce the impact of noise and sampling limitations on the accuracy of predictions, it may be effective to employ methods that account for accumulation processes and uncertainty in the data. 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 [35, 36, 37, 38]. Some early work on the topic [39], in particular, proved that methods accounting for mutation accumulation outperform standard distance-based phylogenetic models when a complex polyclonal architecture and conditions of limited samplings are observed, 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

To answer the questions above, in the following we first introduce the VERSO (Viral Evolution ReconStructiOn) framework for viral evolution and sublineage characterization (see Fig. 1C), which is further detailed in the Methods Section. We then present the detailed analysis of 162 SARS-CoV-2 samples from multiple studies, for which, at the time of writing, raw Illumina sequencing data are available in public databases, also discussing the definition and release of the likely SARS-CoV-2 ancestral genome. In detail, we provide both the phylogenomic lineage tree and the sublineage characterization of samples, which allows to identify likely transmission paths also among individuals infected by the same viral lineage. We then provide an in-depth analysis of the mutational landscape of both clonal and minor variants, also identifying variants undergoing transition to clonality, as well as a number of mutations shared across distinct lineages and which might be positively selected. In particular, we identify mutation g.29039A>U, which is responsible for the Nucleocapsid variant p.Lys256* and might be of particular interest because it abrogates the vast majority of the B and T bona fide antigenic epitopes of the target protein. We finally show the phylogenomic lineage tree reconstructed by our method on a large cohort comprising 12419 GISAID consensus sequences.

VERSO: a new framework for the characterization of viral evolution and transmission. We introduce VERSO (Viral Evolution ReconStructiOn), a new framework for the inference of high-resolution models of viral evolution from raw sequencing data of viral genomes (see Fig. 1C). Our framework introduces several major novelties.

(A) VERSO employs a probabilistic approach that models the process of *accumulation of clonal variants* following infection events and explicitly accounts for *uncertainty in the data*, due to technical noise and biological variability.

Accordingly, it delivers robust phylogenomic models of viral evolution, also in conditions of severe sampling limitations [17]. Such 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 [11, 18].

(B) VERSO exploits the information on intra-host *variant frequencies* to characterize the *(sub)lineage composition* of samples and to define a highly refined *genomic distance* among hosts.

More in detail, by employing methods for *dimensionality reduction* and *clustering* originally employed in single-cell analyses [40], VERSO first allows to reconstruct and visualize likely infection chains among hosts harboring the same viral lineage, while this is prevented to methods that process consensus sequences.

Moreover, VERSO improves the quantification of the *genomic diversity* in a given host or in a given group of samples, based on sublineage composition and heterogeneity, also allowing to identify variants undergoing selection shifts or being possibly affected by founder effects during transmission.

Finally, our framework allows to investigate *homoplasies*, i.e., variants detected in independent viral lineages [29] and which may result from functional *convergent evolution*, as well as to isolate specific regions of the viral genome under *purifying selection* (see Suppl. Fig. 1A). 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.

Reference genome. 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, ref. $\#1$ in the following), used, e.g., in [43] and sequence `EPI_ISL_402125`, ref. $\#2$ in the following, employed in several studies on SARS-CoV-2, e.g., [3]. The sequence comparison highlighted the presence of 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).

In order to disambiguate the conflicting SNPs, we here employed the Bat-CoV-RaTG13 genome [1] (sequence `EPI_ISL_402131`) and the Pangolin-CoV genome [4] (sequence `EPI_ISL_410721`), which both display haplotype UCUCU at locations 8782, 9561, 15607, 28144 and 29095. Accordingly, we generated the reference genome SARS-CoV-2-ANC, which most likely represents the ancestral SARS-CoV-2 genome, preceding both ref. $\#1$ and ref. $\#2$, and which is released in FASTA format as Suppl. File 1. SARS-CoV-2-ANC was used for variant calling in downstream analyses.

Viral evolution and sublineage characterization of 162 SARS-CoV-2 samples via VERSO (Dataset $\#1$). We applied VERSO to the mutational profiles of 162 SARS-CoV-2 samples of Dataset $\#1$ (see Methods), after executing variant calling from raw RNA-sequencing data (see Methods). 684 distinct high-quality single-nucleotide variants are observed in the cohort, 180 of which are always detected as clonal (variant frequency, VF $> 90\%$). For clarity reasons, in Fig. 2A we show the variant frequency profiles of 58 SNVs detected in in at least 2% of the samples of the cohort, 19 of which are clonal (samples are annotated with collection date and location; complete data are provided in Suppl. Table 1, whereas in Suppl. Fig. 2 the geo-temporal localization of samples via Microreact [44] is displayed).

Even though, as expected, the VF profiles of minor variants are noisy, a complex sublineage architecture is observed in several individuals. Moreover, patterns of co-occurrence of minor variants across samples support the hypothesis of transmission of at least a portion of sublineages from a host to another, as well as the presence of homoplasies.

The phylogenomic lineage tree was reconstructed by applying step $\#1$ of VERSO to the 19 variants detected as clonal (VF $> 90\%$) in at least 2% of the samples. The model describes the presence of 11 viral genotypes or lineages L1-11, which display the same set of clonal mutations and are associated to different clusters of samples (Fig. 2A; see Methods). Notably, data are extremely consistent with the perfect phylogeny assumptions (only 1 clonal variant in 1 sample shows a possible violation; see Methods).

We then applied step $\#2$ of VERSO to the complete VF profiles of the samples associated with the lineages with the largest size (≥ 10 samples), and projected the sublineage composition and the genomic distance among samples on the UMAP low-dimensional space (see Methods). Such maps describe the likely transmission paths among hosts characterized by the same viral lineage, and which would not be identifiable by processing consensus sequences only (see Fig. 2B).

More in detail, variant g.29095U>C (mapped on ORF *N*, synonymous) is the earliest evolutionary event from reference genome SARS-CoV-2-ANC and is present in all 162 samples of the dataset. The related viral lineage L1, which displays no further mutations, identifies a cluster including 3 Chinese and 1 American samples.

Two principal clades originate from lineage L1 and correspond to previously identified SARS-CoV-2 types in [14, 41]. In particular, both lineage L1 and the left clade of our model (lineage L2) are characterized by the absence of SNVs g.8782U>C and g.28144C>U and correspond to type A [14] (also type S [41]), which was hypothesized to be a early SARS-CoV-2 type. Lineage L2 is characterized by three private SNVs: g.17747C>U (*ORF1ab*, p.5828P>L), g.17858A>G (*ORF1ab*, p.5865Y>C) and g.18060C>U (*ORF1ab*, synonymous), and includes 41 samples from USA.

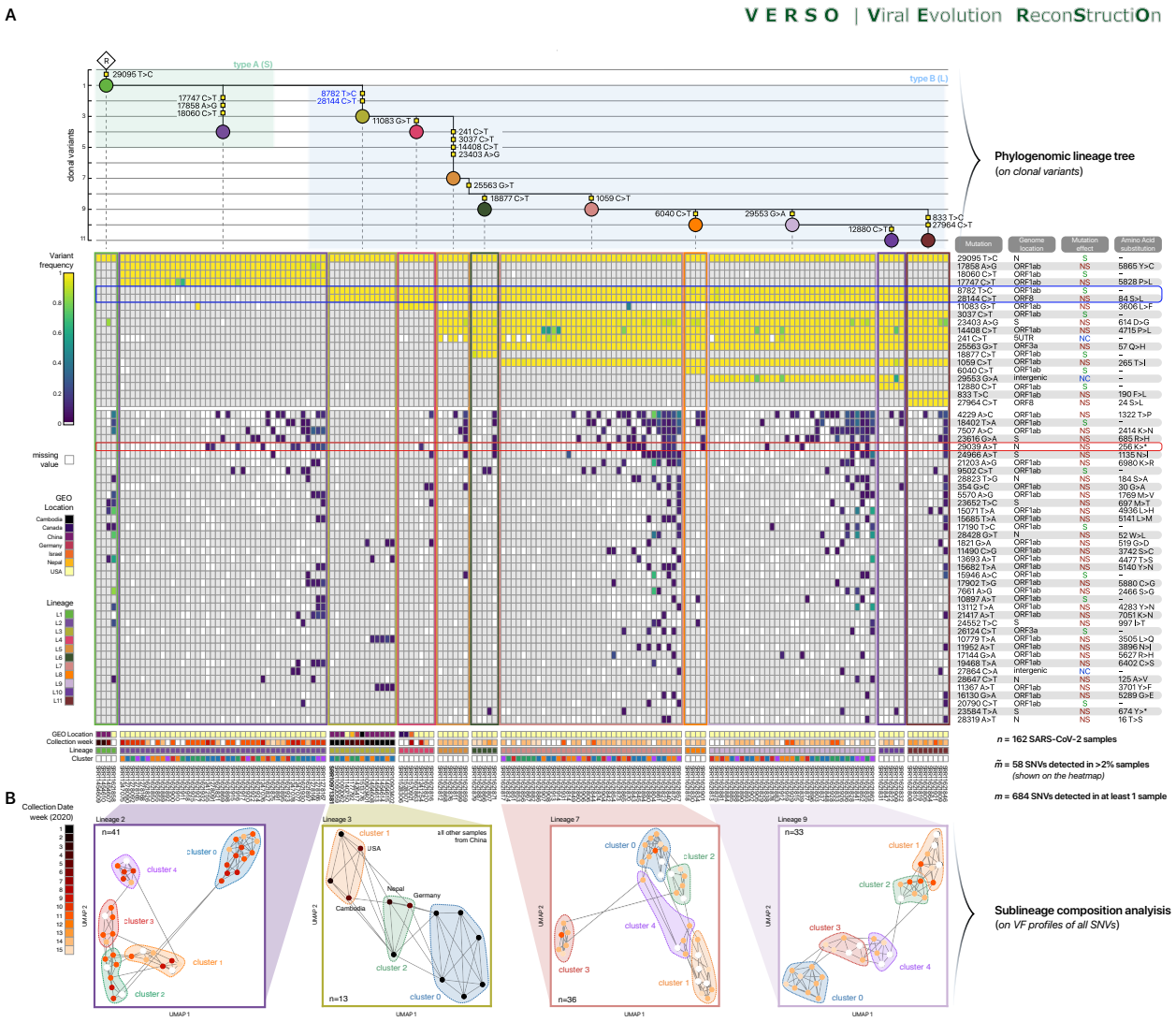

Figure 2: **Viral evolution and sublineage characterization of** 162 **SARS-CoV-2 samples of via VERSO (Dataset** #1**). (A)** Heatmap returning the variant frequency (VF) profiles of 162 SARS-CoV-2 samples on 58 high-quality single-nucleotide variants (SNVs), detected in at least 2% of the samples. Samples are annotated with collection date and location. SNVs are annotated with mapping on ORFs, synonymous (S), nonsynonymous (NS) and non-coding (NC) states, and related amino acid substitutions (asterisks label nonsense mutations). Variants colored in blue characterize the B (L) type [14, 41], whereas variant g.29039A>U is colored in red. The phylogenomic lineage tree returned by VERSO considering only clonal variants (VF > 0.90) is displayed at the top of the heatmap. The rhombus labeled with capital R indicates the reference genome SARS-CoV-2-ANC, yellow squared nodes indicate the variants characterizing each ancestral relation, whereas circle nodes represent viral lineages L1–11 and associated to clusters of samples. Colored shades indicate SARS-CoV-2 types A(S) and B(L) from [14, 41]. **(B)** The 4 panels display the projection of the viral sublineage composition and of the genomic distance among all samples on the UMAP low-dimensional space, as computed by VERSO from VF profiles, for the lineages including ≥ 10 samples. Samples are clustered via Leiden algorithm on the k-nearest neighbor graph (k-NNG), computed on the Bray–Curtis dissimilarity on VF profiles [42]. The color gradient indicates collection date (in week), whereas solid lines represent the edges of the k-NNG.

The right major clade - which includes lineages L3–11 – is characterized by the presence of both SNVs g.8782U>C (*ORF1ab*, synonymous) and g.28144C>U (*ORF1ab*, p.84S>L) and corresponds to type B [14] (also type L [41]). L3 includes all remaining Chinese samples, among which also sample SRR10971381 (EPI_ISL_402125), which was employed as reference genome in several studies [3], plus several samples from other countries (USA, Cambodia, Germany and Nepal). Lineage L4 includes samples from USA, Canada and Israel, whereas lineages L5-11 are associated with American samples only.

Consistently with our model, the B (L) type appear to be more abundant in the population ($10453/12229 \approx 85\%$ of the samples on GISAID database [9]; update May, 7th, 2020) and was diffused especially in the early stages of the epidemic in Wuhan [41]. We also note that, considering the actual shortage of clinical data on SARS-CoV-2$^+$ patients associated to genomic sequences, there are insufficient elements to support any epidemiological claim on virulence and pathogenicity of the distinct SARS-CoV-2 types, even if very recent evidences would suggest the existence of a low correlation [45].

The analysis of the sublineage composition of the lineages characterized by the largest sample size, i.e., L2, L3, L7 and L9 (Fig. 2B), allows to uncover distinct intra-lineage clusters, which identify samples with similar VF profiles on all variants, as well as the possible links among clusters, as determined by the k-nearest neighbors graph (see Methods). This result would suggest that the inherently complex sublineage architecture of samples, as well as possible infection paths, may be obfuscated in a consensus sequences analysis or in an analysis that only considers clonal variants.

For instance, with respect to lineage L3, a cluster including 6 very early samples from China is identified, with links connecting to a distinct cluster with subsequent samples from China, Germany and Cambodia. As a further example, one can notice that within lineage L7 two distinct clusters (clusters 0 and 3) include single samples with a significantly antecedent collection date, namely samples SRR11621844 and SRR11621931, and which one might suppose to be early spreaders for that cluster. This high-resolution representation of the sublineage composition of individuals and of the likely within-lineage transmission paths is a major novelty introduced by VERSO and might pave the way for improved genomic-based contact tracing algorithms.

Standard methods for phylogenomic inference appear to struggle in the analysis of this dataset. In Suppl. Fig. 3 we show the phylogenetic model returned by Nextstrain-Augur on 133 (of 162) samples of Dataset $\#1$ for which high-quality consensus sequences are currently available in GISAID database, whereas in Suppl. Fig. 4 we display a heatmap showing the distance among samples returned by the method, as compared with that returned by VERSO. Since there is no ground-truth to be used as benchmark in the evaluation of the phylogenetic tree inference, we here compared the capability of the methods in grouping samples in homogeneous and stable clusters. In particular, we employed an heuristic based on the amount of the variance explained by each principal component, as computed on the distance matrix returned by Nextstrain-Augur, to determine the optimal number of clusters, which in this case is 8, and compared the results with the clustering executed via VERSO.

Interestingly, the normalized mutual information among the two partitionings shows a value of 0.59, proving that the two approaches produce significantly different results. In addition, the average Silhouette coefficient for Nextstrain-Augur clustering is 0.41 whereas for VERSO is 0.73, thus demonstrating a significantly higher stability for the latter method.

All in all, these results would suggest that VERSO is able to partition samples in highly stable clusters, as opposed to standard distance-based methods such as Nextstrain-Augur, which provide more uncertain assignments to samples, often violating the accumulation hypothesis. This might 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.

Mutational landscape of clonal and minor variants. The comparison of the number of clonal (VF $> 90\%$) and minor variants detected in each host (Fig. 3A) reveals a bimodal distribution of clonal variants (1st mode $= 4$, 2nd mode $= 10$, median 9, max 15), apparently related to the two major clades of the phylogenomic lineage tree (see above), whereas minor variants display a more dispersed long-tailed distribution (median 3, max 44). From the plot, it is also clear that individuals characterized by the same viral lineage may display a significantly different number of minor variants, with distinct distributions observed across lineages.

Importantly, the comparison of the distribution of the number of variants obtained by grouping the samples with respect to collection week (Fig. 3B-C) allows to highlight a statistically significant increasing trend for clonal variants (Mann-Kendall test for trend on median number of clonal variants $p = 0.008$). This result would strongly support both the hypotheses of accumulation of clonal variants in the population and that of a concurrent increase of overall genomic diversity of SARS-CoV-2 [46, 27], whereas the relevance of this phenomenon on minor variants is more uncertain.

We then focused on the properties of the SNVs detected in the cohort. Surprisingly, the distribution of the median VF for each detected variants (Fig. 3D) reveals a bimodal distribution, with the large majority of variants showing either a very low or a very high VF, with only a small proportion of variants showing a median VF within the range $10 - 90\%$. In this respect, we further categorized the variants as *always clonal*, (i.e., SNVs detected with VF $> 90\%$ in all samples), *always minor* (i.e., SNVs detected with VF $> 0.05\%$ and $\leq 90\%$ in all samples) and *transiting* (i.e., SNVs detected as clonal in at least one sample and as minor in at least another sample). As one can see in Fig. 3E, 26.3%, 71.5%, and 2.2% of all SNVs are respectively detected as always clonal, always minor or transiting in our dataset. Moreover, 56.1%, 73.8% and 60% of always clonal, always minor and transiting variants, respectively, are

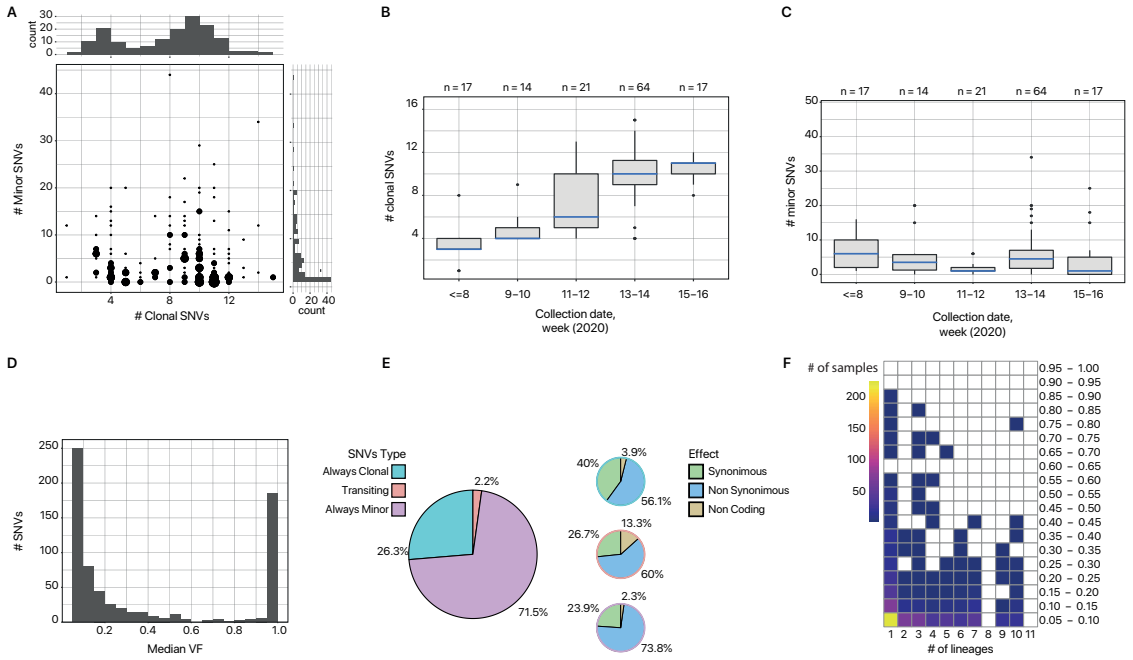

Figure 3: **Mutational landscape of** 162 **SARS-CoV-2 samples (Dataset** #1**). (A)** Scatterplot displaying, for each sample, the number of clonal (VF $> 90\%$) and minor variants (node size proportional to the number of samples). **(B-C)** Boxplots returning the distribution of the number of clonal and minor variants, obtained by grouping samples according to collection date (weeks, 2020). **(D)** Distribution of the median VF for all SNVs detected in the cohort. **(E)** Pie-charts returning: (left) the proportion of SNVs detected as always clonal, always minor or transiting; (right), for each category, the proportion of synonymous, non-synonymous and non coding variants. **(F)** Heatmap returning the distribution of the VF of all SNVs detected in the cohort, grouped by the number of distinct lineages of the phylogenomic model in Fig. 2 in which such variants are observed.

nonsynonimous, whereas the large majority of remaining variants are synonimous (the mapping on SARS-CoV-2 ORF of variants detected in $> 2\%$ of the samples is shown in Fig. 2A).

These results would suggest that, in most cases, randomly emerging SARS-CoV-2 minor variants tend to remain at a low frequency in the population, whereas, in some circumstances, certain variants can undergo frequency increases and even become clonal, due to uncovered mixed transmission events or to selection shifts, as it was observed in [8] for the case of H3N2 and H1N1/2009 influenzas. This result is further confirmed by analyzing the data included in GISAID database [9]. In fact, 115 (of 489) always minor variants detected in our dataset are found as clonal in at least one sample included in GISAID database [9] (update May, 7th, 2020).

Interestingly, several variants classified as always minor are found in a relatively high number of distinct lineages identified by the phylogenomic analysis (see above), and might indicate the presence of either mutational hotspots or homoplasies. In this respect, the heatmap in Fig. 3F returns the distribution of SNVs in VF bins, grouped according to the number of lineages in which they are found.

On the one hand, one might suppose that variants that are detected in single lineages (left region of the scatterplot) either are randomly emerged private mutations or the result of intra-cluster infection events (see above). On the other hand, SNVs found in multiple lineages (right region of the scatterplot) may either be positively selected, in a parallel/convergent evolution scenario, or recurrent neutral mutations due, e.g., to mutational hotspots.

As expected, lineage-specific variants are the most abundant and also display a larger dispersion and higher values of median VF; variants detected in multiple lineages, instead, display significantly lower VF values, and can be found in a maximum of $10/11$ lineages. A variant in particular, g.29039A>U (N, p.256K>*, nonsense) is found in significant abundance in $26/162$ samples and in $7/11$ lineages (median VF ≈ 0.099, max VF ≈ 0.25) and was further investigated, 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. 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.256K>*; see Fig. 4).

This mutation is present, at subclonal level, in $26/162$ samples of our study cohort, with a median intra-host variant frequency of 0.099 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. 5). 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) [47]. 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. 6), 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 [48, 49]. 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. 4; Suppl. Table 2).

In this context the identification of a subloclonal nonsense variant such as the p.256K>* 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 [48, 49] 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 [50], albeit short-lived [51]. In contrast, T cell-mediated immune responses can provide long-term protection after SARS-CoV infection [52, 53, 51]. 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 [54]. Not surprisingly, vaccines for both proteins have been already developed for SARS [55]. 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, one will be able to quantitatively assess the possible increase of frequency of minor variants in the population, as this might be used to evaluate the presence of selection-driven shifts.

VERSO phylogenomic lineage tree of 12419 GISAID samples (Dataset $\#2$). We finally applied VERSO to the 12419 high-quality samples available on GISAID database on May 7th, 2020, by aligning consensus sequences to SARS-CoV-2-ANC reference genome with Augur pipeline, and calling clonal variants in unambiguous locations (ambiguous locations are set as NA; details on Dataset $\#2$ are provided in Methods Section). The phylogenomic lineage tree returned by VERSO is displayed in Fig. 5A and reveals the presence of 17 viral lineages, and the corresponding clusters of samples (see Suppl. Fig. 7 for the complete mutational profile of all samples). As raw sequencing data are not available in GISAID database, step $\#2$ of the VERSO pipeline could not be performed.

Remarkably, this model is consistent with the analysis made on 162 samples of Dataset $\#1$ and discussed above (Fig. 1), despite minor differences due to the only partially overlapping composition of the cohorts. Specifically, all lineages present in both models display the same variants accumulation ordering. Moreover, as previously shown, the two major clades that characterize both models correspond to A(S) and B(L) SARS-CoV-2 major types [14, 41]. Noteworthy, a specific lineage is inferred as hidden (i.e., with no attached samples) in the analysis of Dataset $\#1$, whereas it includes a significant number of samples in the analysis of Dataset $\#2$ (lineage L14 in Fig. 5, with g.25563G>T as characterizing variant). This proves that VERSO is effective in handling sampling limitation issues.

Furthermore, error rates computed comparing observed data and theoretical genotypes derived from VERSO model (i.e., implying accumulation of variants), show significantly low values for most SNVs (see Methods). Specifically, all 27 variants show false positive/negative rates lower then 2%, whereas only 3 SNVs display false positive rates greater than 2%. This findings demonstrate the existence of a consistent process of accumulation of variants in the SARS-CoV-2 genome, as observed in the population. However, a variant in particular, g.14805C>U, shows a ratio

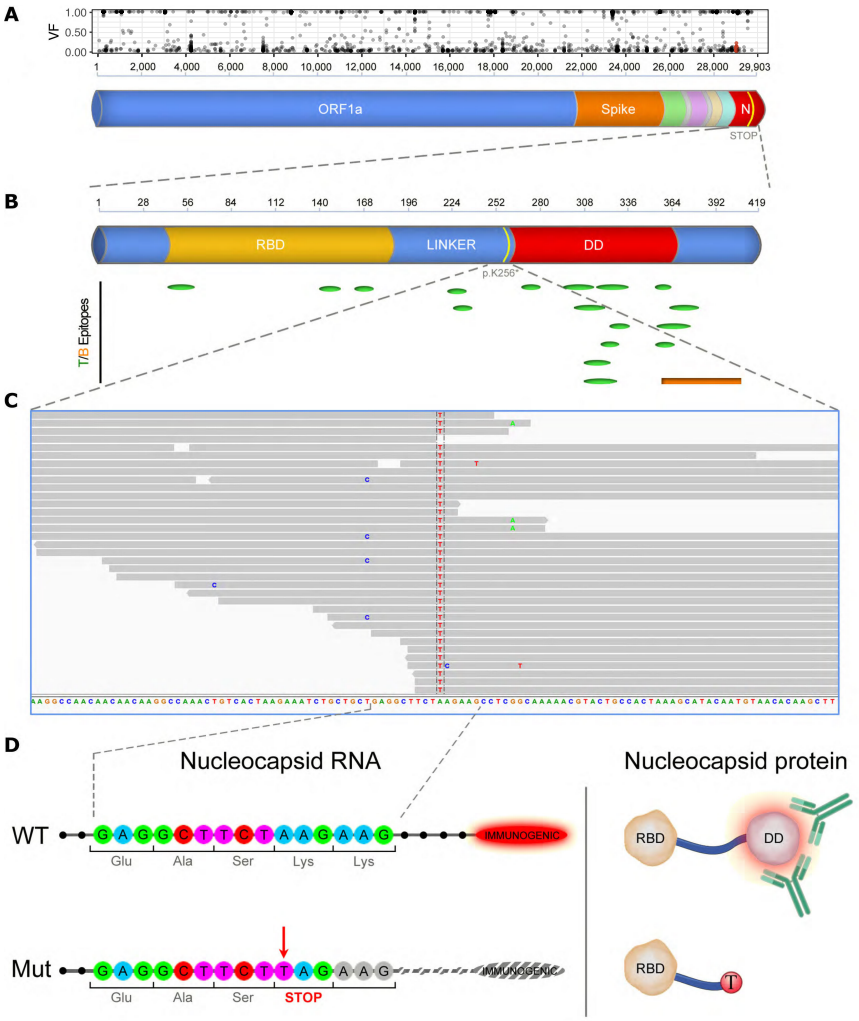

Figure 4: **Functional effect of g.29039A>U mutation.** **(A)** Schematic representation of the entire SARS-CoV2 genome structure. The *N* protein is shown in red (C-terminus). The yellow band indicates the position of the p.256K>* mutation. In the top panel, the VF profiles of all 684 SNVs detected in the cohort 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.256K>* 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.256K>*) 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.

of false positives equal to 7.3% and which significantly deviates from the average value of the dataset (= 0.73%). One could speculate that such SNV may be involved in a violation of the perfect phylogenetic constraints due to, e.g., convergent evolution.

By looking at the geo-temporal localization of samples obtained via Microreact [44] (Fig. 5B-C), one can see that the different lineages are distributed across the world in distinct complex patterns, suggesting that most countries have suffered from multiple infections, especially in the early phases of the outbreak, as confirmed also in [14]. The country showing most cases is USA, which currently includes 3899 samples, distributed in all 17 lineages, but with lineage L2 and L16 accounting together for 65% of the samples. 381 samples from China are included in the dataset and are distributed in 7 distinct lineages, with L1 and L3 the most abundant (33% and 55% of the samples respectively). These two lineages represent the early samples differentiating types A(S) and B(L) and are consistently most observed in China, with lineage L3 being the more abundant.

We finally note that the average Silhouette coefficient as computed on sample clusters (see above) is 0.98 for VERSO and 0.35 for the model returned by Nexstrain-Augur on the same dataset, confirming also in this case the significantly greater robustness of clustering assignments of our approach.

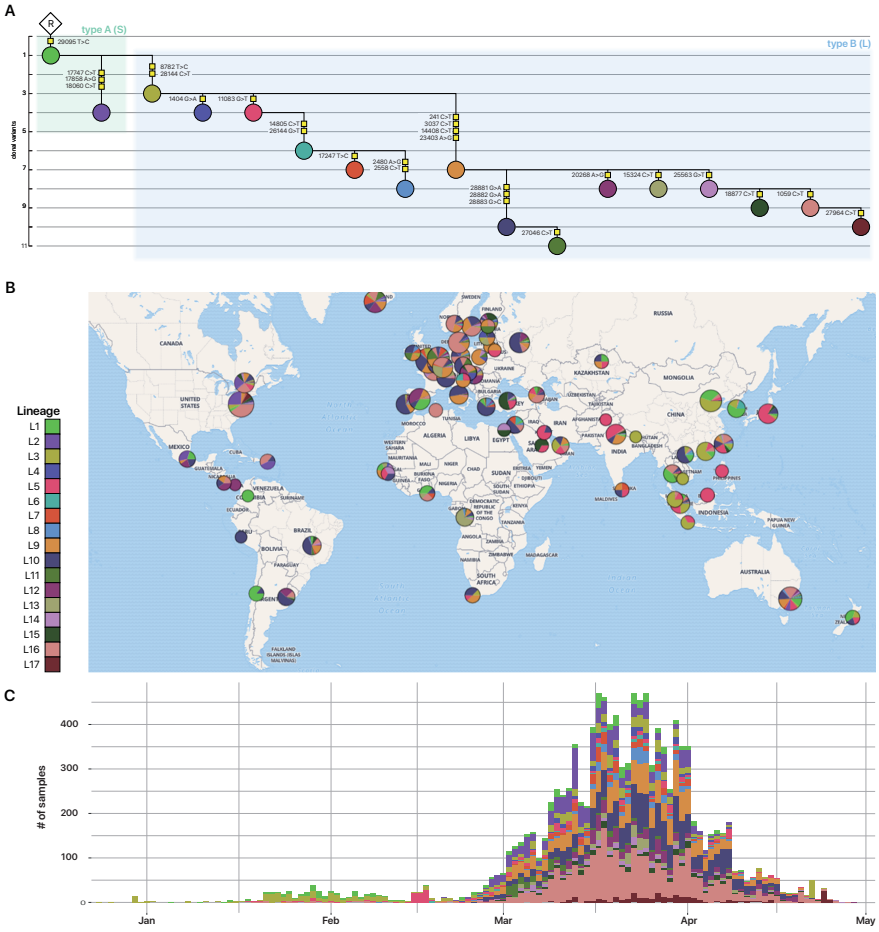

Figure 5: **VERSO phylogenomic lineage tree of** 12419 **GISAID samples (Dataset** #2**). (A)** The phylogenomic lineage tree reconstructed via VERSO on 12419 sample currently present on GISAID database [9] is displayed. **(B)** The geo-temporal localization of samples with respect to the 17 lineages of the phylogenomic model is displayed via Microreact [44] **(C)** Stacked barplots returning the distribution of samples with respect to lineage and collection date.

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 [12, 56].

This might be particularly relevant for countries which suffered outbreaks of exceptional proportions, such as USA, Brazil, Spain and Italy, 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 hypothesized that the rapid diffusion of COVID-19 might be likely due to the extremely high number of untested asymptomatic hosts [33]. 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.

We finally recall that we release as Suppl. File 1 the sequence SARS-CoV-2-ANC, which likely represents the ancestral SARS-CoV-2 genome after spillover and which might allow to improve the resolution and accuracy of phylogenomic analyses, thus enhancing the understanding of SARS-CoV-2 evolution.

Methods

VERSO computational framework. VERSO is a new framework for the reconstruction of viral evolution models from raw sequencing data of viral genomes. VERSO includes a two-step procedure, which takes advantage of the many achievements of research in cancer evolution and in sequencing data analysis.

Step $\#1$**.** In the first step, VERSO employs an evolved version of the algorithm introduced in [57] for the reconstruction of models of cancer evolution from single-cell mutational profiles, in order to identify the phylogenomic lineage tree of viral evolution. Our probabilistic framework solves a Boolean matrix factorization problem with perfect phylogenetic constraints, thus assuming a consistent process of accumulation of clonal variants in the population, and accounts for uncertainty in the data by employing a maximum likelihood approach (via MCMC search) that allows for the presence of false positives, false negatives and missing data points (see the SI for further details).

In particular, the method takes as input a binary mutational profile matrix, as defined on clonal mutations only: an entry in a given sample is equal to 1 (present) if the VF is larger than a certain threshold (in our case, equal to 90%), it is equal to 0 if lower than a distinct threshold (in this case, equal to 5%), and is considered as missing (NA) in the other cases, thus modeling possible uncertainty in sequencing data. We recall that the accumulation hypothesis holds only when considering clonal mutations, which are most likely transmitted from a host to another during the infection, whereas this might not be the case with variants with lower frequency, due to bottlenecks, founder effects and stochasticity (see below).

The inference returns a set of maximum likelihood phylogenetic trees (minimum 1) depicting the variants as sampled during the MCMC search. Accordingly, such trees provide a set of maximum likelihood attachments of samples to variants, which in turn can be used to estimate a similarity measure between samples, namely by counting how many times pairs of samples are attached to the same variant. VERSO then performs hierarchical clustering of samples on such similarity measure, by selecting the optimal number of clusters via standard heuristics (see Parameter settings). We notice that, in case only 1 unique maximum likelihood phylogenetic tree is found during the inference, the unique maximum likelihood attachment of samples to variants can directly be used for clustering.

In the VERSO output model, each node correspond to a viral genotype (or lineage), whereas edges model parental relations, which stand for major evolutionary events, and are characterized by one or more accumulating variants. Each viral lineage is associated to a cluster of samples, via the procedure described above. As shown in [57] in a different experimental context, our algorithmic framework ensures robustness and scalability also in case of high rates of errors and missing data, due for instance to sampling limitations. Finally, VERSO allows to visualize the geo-temporal localization of clusters via Microreact [44].

We note that violations to perfect phylogeny constraints are possible, due, e.g., to the emergence of manifold mutations at single locations or to multiple infections of the same host, yet they are unlikely when considering clonal variants only. In any case, VERSO allows to identify clonal variants that display significant deviations from the expected error rates of the dataset, thus indicating likely violations of the accumulation hypothesis. In particular, given the maximum likelihood phylogenetic tree and the related attachments of samples to variants, VERSO can estimate the variants theoretically expected in each sample and subsuming the accumulation hypothesis. Hence, by comparing the theoretical observations with input data, VERSO can estimate the rate of false positives, i.e., variants that are observed in the data but are not predicted by VERSO, and false negatives, i.e., variants that are not observed, but predicted. Variants that show a particularly high level of estimated error rates represent possible violations of perfect phylogeny constraints, e.g., due to convergent evolution.

Since, by construction, no ordering is possible among individual infected by the same viral lineage when considering clonal variants only, VERSO aims at overcoming this limitation with step $\#2$.

Step $\#2$**.** In the second step, VERSO takes into account the variant frequency profiles of intra-host variants, by exploiting widely-used methods for the analysis of single-cell transcriptomic data and included in the SCANPY suite [40]. We remark that, while it is sound to binarize clonal variant profiles to reconstruct a phylogenetic lineage tree, the variant frequency profile has to be considered when analyzing intra-host variants, for several reasons.

First, variant frequency profiles describe the sublineage composition of any given host, and this information would be lost during binarization.

Second, minor variant profiles might be noisy, due to the relatively low abundance and to the technical limitations 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. However, binarization with arbitrary thresholds might increase the false positive rate, compromising the accuracy of any downstream analysis.

Third, as specified above, the extent of transmission of minor variants among individuals is still partially obscure. Due to the low frequency, for instance, minor variants may be affected by stochastic fluctuations, bottlenecks and founder effects, according to which a certain sublineage might significantly change its variant frequency, not be transmitted or become clonal in the infected host [26].

The latter issue also suggests that the hypothesis of accumulation of minor variants during infection may not hold and should be relaxed. For this reason, VERSO defines a pairwise genomic distance, as computed on the variant frequency profiles, to be used in downstream analyses. The intuition is that samples displaying similar patterns of co-occurrence of minor variants might have a similar sublineage architecture, thus being at a small evolutionary distance. Accordingly, this might indicate a direct or indirect infection event. In particular, in this work we employed the Bray-Curtis dissimilarity, which is defined as follows: given the ordered VF vectors of two samples, i.e. $v_i = \{VF_1^i, \ldots, VF_r^i, \}$ and $v_j = \{VF_1^j, \ldots, VF_r^j, \}$, the pairwise Bray-Curtis dissimilarity $d(i, j)$ is given by:

$$d(v_i, v_j) = \frac{\sum_{l=1}^{r} |VF_l^i - VF_l^j|}{\sum_{l=1}^{r} |VF_l^i + VF_l^j|}. \tag{1}$$

Since this measure weights the pairwise VF dissimilarity on each variant with respect to the sum of the VF of all variants detected in both samples, it can be effectively used to compare the sublineage composition of samples, as proposed for instance in [42]. However, VERSO allows to employ different distance metrics, such as correlation or Euclidean distance.

As a design choice, in VERSO the genomic distance is computed among all samples associated to any given viral lineage, as inferred in step $\#1$. The rationale is that, in a statistical inference framework modeling a complex interplay involving heterogeneous dynamical processes, it is crucial to stratify samples into homogeneous groups, to reduce the impact of possible confounding effects [58].

In the context of viral evolution and transmission, noisy observations affecting low-frequency variants, e.g., due to sequencing limitations, as well as homoplasies, such as mutational hotspots or convergent mutations observed across lineages, might distort the relative distance among samples, when considering unpartitioned cohorts. Conversely, the identification of a phylogenomic lineage tree, which describes the process of accumulation of clonal variants, allows to stratify samples in robust groups that are consistent with major evolutionary events, therefore improving the reliability of downstream analyses. Accordingly, due to the distinct properties of clonal and minor variants during transmission, it is reasonable to assume that the event in which certain minor variants and no clonal variants are transmitted from a host to another during the infection is extremely unlikely.

The genomic distance is then employed in a standard SCANPY [40] workflow for dimensionality reduction and clustering, and which includes three steps: (i) the computation of the *k-nearest neighbour* graph (k-NNG), which can be executed on the original variant frequency matrix, or after applying *principal component analysis* (PCA), to possibly reduce the effect of noisy observations (when the number of samples and variants is sufficiently high); (ii) the clustering of samples via either *Louvain* or *Leiden* algorithms for community detection [59]; (iii) the projection of samples on a low-dimensional space via standard tSNE [60] or UMAP [61] plots.

As output, VERSO delivers both the partitioning of samples in homogeneous clusters and the visualization in a low-dimensional space, also allowing to label samples according to other covariates, such as, e.g., collection date or geographical location. In the map in Fig. 2B, for instance, the sublineage composition of each sample and the genomic distance among samples are projected on the first two UMAP components, whereas samples that are connected by k-NNG edges display similar patterns of co-occurrence of variants. Accordingly, the map show clusters of samples likely affected by infection events, in which (a portion of) sublineages might have been transmitted from a host to another. This represent a major novelty introduced by VERSO and also allows to visualize for the first time the space of sublineages (also known as quasispecies [22]).

Datasets description. *Dataset* $\#1$. We analyzed a cohort of 162 samples from distinct individuals obtained from 11 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 projects: (1) `PRJNA601736`, 2 individuals geographically located

in Wuhan (China) for which we considered RNA-seq from bronchoalveolar lavage fluid (BALF); (2) PRJNA603194, 1 individual from Wuhan (China) with RNA-seq from [2]; (3) PRJNA605983, 2 individuals from China with Illumina HiSeq RNA-seq data; (4) PRJNA607948, 1 individual from Wisconsin (USA); (5) PRJNA608651, 1 individual from Nepal with RNA-seq isolate from oro-pharyngeal swab of [43]; (6) PRJNA610428, 143 individuals geographically located in the State of Washington (USA); (7) PRJNA615319, 1 individual geographically located in Germany; (8) PRJNA616446, 7 individuals from China; (9) PRJNA623895, 1 individual from Cambodia; (10) PRJNA624792, 2 individuals from Canada and (11) PRJNA626526, 1 individual from Israel. Furthermore, we considered 3 additional samples from NCBI BioProject PRJNA607948, all obtained from the same individual from Wisconsin already included in the dataset, representing 3 independent passage isolates, which were used to validate the discovered minor variant g.29039A>U.

Dataset #2. We considered the whole set of complete, high-coverage and human-derived consensus sequences available on GISAID on May 7th, 2020. The dataset includes a total of 12419 samples.

Variant calling. *Dataset* #1. With respect to dataset #1, we downloaded SRA files from 11 NCBI BioProjects with the following accession numbers: PRJNA601736, PRJNA603194, PRJNA605983, PRJNA607948, PRJNA608651, PRJNA610428, PRJNA615319, PRJNA616446, PRJNA623895, PRJNA624792 and PRJNA626526; 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.

We then used bwa mem (version 0.7.17) to map reads to SARS-CoV-2-ANC reference genome (Suppl. File 1; see Results). 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) [62]. Variants were mapped on SARS-CoV-2 coding sequences (CDSs) via a custom R script, also by highlighting synonymous/nonsynonymous states and amino acid substitutions for the related Open Reading Frame (ORF) product. In particular, we translated reference and mutated CDSs with the seqinr R package to obtain the relative amino acid sequences, which we compared to assess the effect of each nucleotide variation in terms of amino acid substitution.

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 filters 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) and more than 8 reads of support in at least 70% of the samples, (2) displaying a variant frequency VF $> 5\%$. As a result, we selected a list of 684 highly-confident SNVs to be included in the analysis.

Dataset #2. With respect to dataset #2, we aligned the 12419 consensus sequences present on GISAID on May 7th, 2020 to SARS-CoV-2-ANC reference genome via Augur pipeline. 7001 distinct variants were detected in at least 1 sample. We called as clonal the variants detected in unambiguous locations, whereas ambiguous locations were set to NA. We then removed regions highly affected by sequencing artifacts by selecting the 6713 variants showing less than 10% of ambiguous calls. Finally, we selected the list of variants to be considered as input to VERSO by selecting the 27 mutations observed in at least 2% of samples.

A table including for each sample, the Run ID, the BioProject ID, the BioSample ID, the instrument type, the sample name, the GISAID ID and origin country is returned for Dataset #1 as Suppl. Table 3, whereas name, accession date, region and host type are provided for each sample of Dataset #2 as Suppl. Table 4.

Parameter settings. The phylogenomic analysis via VERSO was performed on dataset #1 and #2 by considering only clonal variants (VF $> 90\%$ for dataset #1) detected in at least 2% of the samples of the cohort. A grid search comprising 64 different error rates was employed. For Dataset #1, hierarchical clustering was performed on the similarity matrix returned by VERSO (see Methods). For Dataset #2, instead, as VERSO returned only 1 unique maximum likelihood phylogenetic tree, hierarchical clustering was performed directly on the maximum likelihood attachment of samples to variants; in this latter case, we first removed samples with ambiguous attachment (229 out of 12419 samples) and then we performed PCA analysis (with a maximum of 27 PCs) on the attachment matrix and used the results as inputs for hierarchical clustering. The optimal number of clusters was determined with an heuristic based on the amount of the variance explained by each principal component.

The analysis of the sublineage composition of Dataset #1 was performed by considering the VF profiles of all 684 SNVs detected in the cohort, with respect to the clusters with ≥ 10 samples (L2, L3, L7 and L9, with 41, 13, 36 and 33 samples, respectively). Missing values (NA) are present in less than 10% entries in each variant and were imputed to 0 for downstream analysis. A number of PCs equals to the number of samples was employed in PCA analysis, prior

to the computation of the k-nearest neighbour graph ($k = 5$) on the Bray-Curtis dissimilarity of VF profiles. Leiden algorithm was applied with resolution $= 1$. The UMAP plot was finally obtained by setting parameters spread $= 2$ and min_dist $= 0.5$ (see Suppl. Table 5 for the parameter settings of VERSO employed in the case studies).

Standard phylogenomic analyses were also performed by applying Nexstrain-Augur [10] to 133 on 162 consensus sequences of Dataset #1 present in GISAID database and on the 12419 samples of Dataset #2 (default parameters; the model for Dataset #1 is shown in Suppl. Fig. 3).

Software availability

The source code to replicate all the analyses is available at: `https://github.com/BIMIB-DISCo/SARS-CoV-2-IHMV`. The Web based tool for the geo-temporal visualization of samples, Microreact, is available at this link: `https://microreact.org/showcase`. SCANPY is available at this link: `https://scanpy.readthedocs.io/en/stable/`.

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. Partial 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. We also thank David Posada for interesting suggestions on the preliminary version of the manuscript.

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] Xiao, K. *et al.* Isolation of sars-cov-2-related coronavirus from malayan pangolins. *Nature* (2020). URL `https://doi.org/10.1038/s41586-020-2313-x`.

[5] Deng, X. *et al.* Genomic surveillance reveals multiple introductions of sars-cov-2 into northern california. *Science* (2020). URL `https://science.sciencemag.org/content/early/2020/06/05/science.abb9263`.

[6] 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).

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

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

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

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

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

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

[13] 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).

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

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

[16] Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

[17] Villabona-Arenas, C. J., Hanage, W. P. & Tully, D. C. Phylogenetic interpretation during outbreaks requires caution. *Nature Microbiology* 1–2 (2020).

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

[19] Domingo, E. *et al.* The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevance—a review. *Gene* **40**, 1–8 (1985).

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

[21] 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).

[22] Domingo, E., Sheldon, J. & Perales, C. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.* **76**, 159–216 (2012).

[23] Ni, M. *et al.* Intra-host dynamics of ebola virus during 2014. *Nature microbiology* **1**, 1–9 (2016).

[24] 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).

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

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

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

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

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

[30] 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).

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

[32] 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).

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

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

[35] Ramazzotti, D. *et al.* Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31**, 3016–3026 (2015).

[36] Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. *Systematic biology* **64**, e1–e25 (2015).

[37] Caravagna, G. *et al.* Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences* **113**, E4025–E4034 (2016).

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

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

[40] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

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

[42] Srinivas, G. *et al.* Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nature communications* **4**, 2462 (2013).

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

[44] Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial genomics* **2** (2016).

[45] Zhang, X. *et al.* Viral and host factors related to the clinical outcome of covid-19. *Nature* 1–7 (2020).

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

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

[48] 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).

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

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

[51] 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).

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

[53] 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).

[54] 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).

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

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

[57] Ramazzotti, D. *et al.* Longitudinal cancer evolution from single cells. *bioRxiv* (2020).

[58] Pearl, J. *Causality* (Cambridge university press, 2009).

[59] Traag, V. A., Waltman, L. & van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).

[60] Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605 (2008).

[61] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[62] 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).