MUTATIONAL SIGNATURES AND HETEROGENEOUS HOST RESPONSE REVEALED VIA LARGE-SCALE CHARACTERIZATION OF SARS-CoV-2 GENOMIC DIVERSITY

Alex Graudenzi[1,+,*], **Davide Maspero**[2,1,+], **Fabrizio Angaroni**[2,+],
Rocco Piazza[3,†,*], **Daniele Ramazzotti**[3,†,*]
[1] Inst. of Molecular Bioimaging and Physiology,
Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[2] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy
[3] Dept. of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy
[+] Equal contributors † Co-senior authors
[*] Correspondence: alex.graudenzi@ibfm.cnr.it
rocco.piazza@unimib.it | daniele.ramazzotti@unimib.it

ABSTRACT

To dissect the mechanisms underlying the observed inflation of variants in SARS-CoV-2 genome, we present the largest up-to-date analysis of intra-host genomic diversity, which reveals that the majority of samples present a complex sublineage architecture, due to the interplay between host-related mutational processes and transmission dynamics.

Strikingly, the deconvolution of the entire set of intra-host variants reveals the existence of mutually exclusive viral mutational signatures, which prove that distinct hosts differently respond to SARS-CoV-2 infections. In particular, two signatures are likely ruled by APOBEC and Reactive Oxygen Species (ROS), which induce hypermutation in a significant number of samples, and appear to be affected by severe purifying selection.

Conversely, several mutations linked to low-rate mutational processes appear to transit to clonality in the population, eventually leading to the definition of new viral genotypes and to an increase of overall genomic diversity. Finally, we demonstrate that a high number of variants are observed in samples associated to independent lineages, likely due to signature-related mutational hotspots or to positive selection.

Introduction

The COVID-19 pandemic has currently affected 188 countries worldwide with $9,129,146$ people being infected, while the number of casualties has reached the impressive number of $473,797$ [1] (update 24 June 2020). The origin and the main features of SARS-CoV-2 evolution have been investigated [2, 3, 4, 5, 6], also due to the impressive amount of consensus viral sequences included in public databases, such as GISAID [7]. However, only a few currently available datasets include raw sequencing data, which are necessary to quantify intra-host genomic variability.

Due to the combination of high error and replication rates of viral polymerase, subpopulations of viruses with distinct genotypes, also known as viral quasispecies [8], usually coexist within single hosts. Such heterogeneous mixtures are supposed to underlie the adaptive potential of RNA viruses to internal and external selection phenomena, which are related, e.g., to the interaction with the host's immune system or to the response to antiviral agents. For instance, it was hypothesized that intra-host heterogeneity may be correlated with prognosis and clinical outcome [9, 10]. Furthermore, even if the modes of transmission of intra-host variants in the population are still elusive, one may hypothesize that, in certain circumstances, infections allow such variants to spread, sometimes inducing significant changes in their frequency [11].

In particular, several studies on SARS-CoV-2 support the presence of intra-host genomic diversity in clinical samples and primary isolates [12, 13, 14, 15, 16, 17, 11, 18], whereas similar results were obtained on SARS-CoV [19], MERS [20], EBOLA [21] and H1N1 influenza [22]. We here present the largest up-to-date study on intra-host genomic diversity of SARS-CoV-2, based on 1003 high-quality samples for which raw sequencing data are available.

Our analysis shows that $\approx 31\%$ of the SARS-CoV-2 genome has already mutated in at least one sample, with most mutations ($\approx 69\%$) being detected at a low frequency in a very limited number of hypermutated samples ($\approx 13\%$ of the cohort displays more than 50 minor variants). The large majority of samples shows highly heterogeneous sublineage architectures, in which patterns of co-occurence of minor variants seem to suggest the existence of uncovered infection paths and of homoplasies, either due to functional convergent evolution or to mutational hotspots. Importantly, several variants are observed as clonal in certain samples and at a relatively low frequency in others, demonstrating that transition to clonality might be due not only to functional selection shifts, but also to complex transmission dynamics involving stochastic and founder effects.

Strikingly, our analysis allowed to identify three non-overlapping mutational signatures, i.e., specific distributions of nucleotide substitutions, which are observed in distinct clusters of samples in a mutually exclusive fashion, suggesting the presence of host-related mutational processes. One might hypothesize that such different processes are related to the interaction of the virus with the host's immune system and/or with antiviral therapies, and might pave the way for a better understanding of the molecular mechanisms underlying different clinical outcomes. In particular, the first signature is characterized by C>T:G>A substitution and is likely related to APOBEC activity, the second signature is characterized by G>T:C>A substitution and might be associated to ROS-related processes, while a third signature is mostly associated to A>G:T>C substitution, which is usually imputed to ADAR activity. The two first signatures, in particular, appear to cause hypermutation of the samples, up to extremely high values (up to 5255 variants detected in a single host), and the dN/dS analysis would suggest that they are both affected by purifying selection in the population.

We finally provide a high-resolution model of viral evolution via VERSO [12], which allows to identify a robust phylogenomic lineage tree, as well as to quantify the intra- and inter-host genomic diversity even among samples harboring the same viral lineages. Interestingly, 2 clonal variants are detected as convergent in two distinct clades, and it is sound to hypothesize they might have been related to transmission-related founder effects, following random mutation acquisition. Importantly, a significantly high number of minor variants is observed in independent lineages and is unlikely related to infection events, suggesting that in such cases, the same mutation can emerge in independent samples due to the presence of mutational hotspots, especially in hypermutated samples affected by mutational signatures.

Results

Mutational landscape of SARS-CoV-2 from variant frequency profiles of 1003 samples. We performed variant calling from Amplicon raw sequencing data of 1171 samples from the NCBI BioProject `PRJNA613958` and by aligning sequences to reference genome SARS-CoV-2-ANC, which is the likely ancestral SARS-CoV-2 genome [12]. We first filtered-out samples showing more than 10% of the genomic positions covered by less than 50 reads. We also verified the absence of any possible bias in the detection of minor variants due to sequencing artifacts. As one can see in Suppl. Fig. 1, no correlation between the number of SNVs and both total coverage and the mean sequencing depth is observed ($R^2 = 0.0013$ in both cases), which proves the good quality of the calls. As a result, we extensively analyzed the mutational profiles of 1003 high-quality samples; 9899 distinct single-nucleotide variants (SNVs, identified by genome location and nucleotide substitution) were detected in the cohort, for a total of $62,945$ mutations (see Methods for further details; the variant frequency profiles of all samples are included in Suppl. Table 1; see Fig. 1I for a graphical representation of an example dataset). In particular, in our analysis we consider any SNV detected in any given sample as *clonal*, if its variant frequency (VF) is $> 90\%$ and as *minor* if its variant frequency is $> 5\%$ and $\leq 90\%$.

The distribution of the number of minor and clonal variants observed in each sample (Fig. 1A) unveils a bimodal distribution of clonal variants (1st mode = 4, 2nd mode = 10, median = 10, max = 18) likely due to the partitioning of samples in distinct phylogenomic clades (see below). Minor variants are detected in $\approx 74\%$ of the samples and show a long-tail distribution (median = 2, max = 5255). 131 samples ($\approx 13\%$ of the cohort) display a significantly large number of minor variants (> 50) and hint at the existence of hypermutation processes, which is confirmed in successive analyses (see below). Accordingly, we label such samples as *hypermutated*.

Interestingly, we observe a statistically significant increase of genomic diversity on clonal variants with respect to collection week (Mann-Kendall test for trend on median number of clonal variants $p = 0.02$, Fig. 1B), due to the accumulation of clonal variants in the population, and which confirms recent findings [23, 13, 12], whereas, as expected, this phenomenon is less evident for minor variants (Fig. 1C). This aspect is further investigated in the following and hints at the interplay involving the evolutionary dynamics within hosts and the transmission among hosts, which differently affects clonal and minor variants [24].

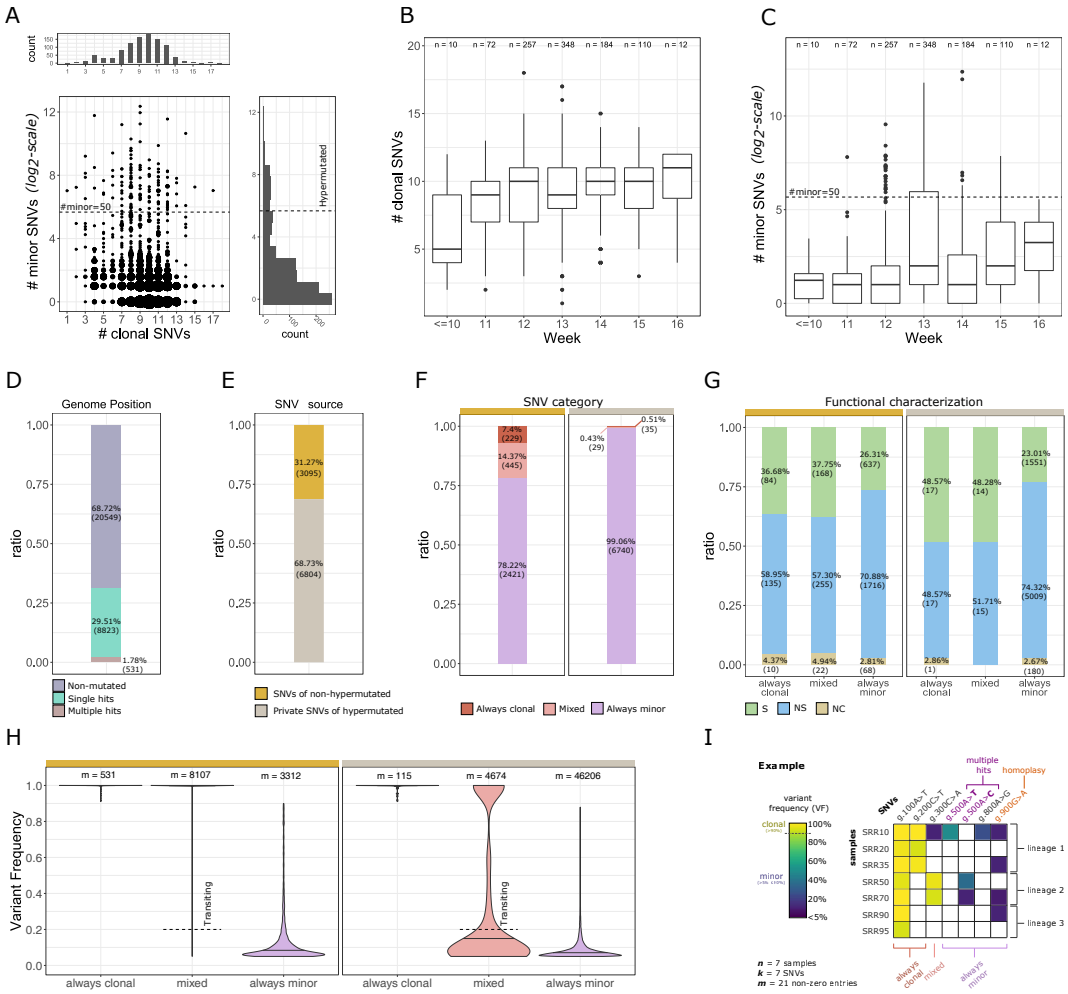

Figure 1: **Mutational landscape of** 1003 **SARS-CoV-2 samples.** **(A)** Scatter-plot displaying the number of clonal (VF $> 90\%$) and minor (VF $> 5\%$ and $\leq 90\%$) variants for 1003 samples (node size proportional to the number of samples; y-axis in $\log_2$ scale). **(B)** Box-plots returning the distribution of the number of clonal and **(B)** minor variants, obtained by grouping samples according to collection date (weeks, 2020; y-axis of panel C is in $\log_2$ scale). **(D)** Bar-plot returning the proportion of sites of SARS-CoV-2 genome that are either non-mutated, mutated with a unique SNV, or mutated with multiple SNVs. **(E)** The proportion of SNVs either detected in non-hypermutated samples or privately detected in hypermutated samples. **(F)** Stacked bar-plots returning, for each source, the proportion of SNVs detected as always clonal, mixed or always minor. **(G)** The ratio of synonymous (S), non-synonymous (NS) and non-coding (NC) mutations, for each category. **(H)** Violin plots returning the distribution of VF for the SNVs either detected in non-hypermuted samples or privately detected in hypermutated samples (m returns the number of non-zero entries of the VF matrix). **(I)** Graphical representation of an example dataset with $n = 7$ samples and the VF profiles of $k = 7$ SNVs.

Evidence of transition to clonality. We further categorize each detected SNV as: (i) *always clonal*, if clonal in all samples in which it is detected, (ii) *always minor*, if minor in all samples in which it is detected, (iii) *mixed*, if observed as clonal in at least one sample and as minor in at least another sample.

9899 single-nucleotide variants were detected on 9354 distinct genome sites (31.28% of SARS-CoV-2 genome), of which 531 sites (1.78% of the genome) display multiple nucleotide substitutions (see Fig. 1D). This suggests that the proportion of mutated genomic sites might be considerably higher in the overall population, especially if considering minor variants. Remarkably, 68.73% of all SNVs (6804) are privately detected in hypermutated samples, which represent the $\approx 13\%$ of the cohort (Fig. 1E). 7.4%, 14.37% and 78.22% SNVs are detected as always clonal, mixed and always minor, respectively, in non hyper-mutated samples, whereas hyper-mutated samples are characterized by a large majority of always minor variants (99.06%) (Fig. 1F; notice that, in this categorization, SNVs detected in non-hypermutated samples can be present also in hypermutated samples, yet not privately). In most cases, most variants are non-synonymous (see Fig. 1G).

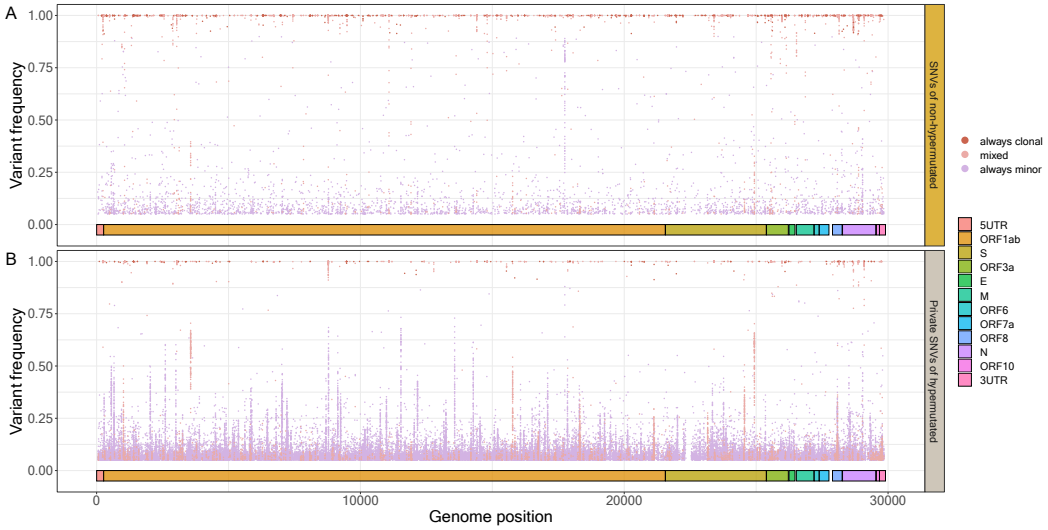

Figure 2: **Distribution of SNVs detected on SARS-CoV-2 genome. (A)** Scatter-plot returning the genome location and the VF of all SNVs detected in non-hypermutated samples and **(B)** privately detected in hypermutated samples, colored according to category.

The analysis of the VF distribution (Fig. 1H) unveils an impressive scarcity of variants showing VF in the middle-range, i.e., between 30% and 80%, for all categories. This phenomenon is likely due to transmission *bottlenecks*, which tend to purify low-frequency variants in the population. Nonetheless, both mixed and always minor variants display broad VF spectra, an aspect that is particularly relevant for the former category, for both hypermutated and non-hypermutated samples. In this respect, 6.96% of all mixed variants (33 on 474) never display a VF $\leq 20\%$: one may hypothesize that such variants are indeed *transiting to clonality* in the population, because either positively selected, as a result of the strong immunologic pressure within human hosts [25], or because affected by transmission phenomena involving *founder effects* and *stochastic fluctuations* [26, 10]. The former hypothesis is supported by the ratio of non-synonymous and synonymous substitutions $= 1.38$ ($18/13$, 2 variants are non-coding), which is slightly superior to the theoretical NS/S ratio of the SARS-CoV-2-ANC reference genome, which is $= 1.20$ (resulting ratio $= 1.15$), suggesting a mild tendency toward positive selection for transiting variants.

Conversely, one might hypothesize that most remaining mixed variants may result from random mutations hitting positions of SNVs that are already present as clonal in the population; this is especially relevant for hypermutated samples, which show an extremely high number of low-frequency variants.

Furthermore, the distribution of SNVs with respect to each region of the genome in Fig. 2A–B demonstrate that mutations are uniformly distributed across the genome, with noteworthy predominance of minor SNVs in hyper-mutated samples (see also Suppl. Fig 2). Overall, this analysis provides the first large-scale quantification of transition to clonality in SARS-CoV-2 and might serve to intercept variants possibly involved in functional modifications, bottlenecks or founder effects.

De novo decomposition of SARS-CoV-2 mutational signatures. In order to investigate the existence of mutational processes related to the interaction between the host and the SARS-CoV-2 virus, we analyzed the distribution of nucleotide substitutions for all SNVs detected in the cohort. In Fig. 3A one can see the proportion of SNVs for each of the 12 nucleotide substitution types (e.g., number of C>T's) over the total number of nucleotides present in the reference genome for each substitution type (e.g., number of C's); SNVs are then grouped in those detected in non hyper-mutated samples or privately detected in hyper-mutated samples.

Certain substitutions present a significantly higher normalized abundance, confirming recent findings on distinct cohorts [27, 28]. In particular, C>T substitutions are observed in 19.3% and 78.13% of all C nucleotides in SARS-CoV-2 genome, in either non-hypermutated or (privately) in hypermutated samples, respectively (97.43% by considering all SNVs from all samples); G>T's in 11.19% and 21.73% of all G's, T>C's in 3.9% and 3.73% of all T's, and A>G's in 3.07% and 1.65% of all A's. Although, traditionally, a 12-substitution pattern has been used in order to report mutations occurring in single-stranded genomes, we reasoned that, owing to the intrinsically double-stranded nature of the viral life-cycle (i.e., a mutation occurring on a *plus* strand can be transferred on the *minus* strand by RdRP and vice versa) it is sound to consider a total of 6 substitution classes (obtained by merging equivalent substitutions in complementary strands) to investigate the possible presence of viral mutational signatures [29]. Clonal variants were not considered in

Figure 3: **Distribution of nucleotide substitutions in the cohort.** **(A)** Stacked bar-plots returning the normalized substitution proportion of all SNVs detected in at least one sample of the cohort, with respect to all 12 possible nucleotide substitutions, grouped by variant type and source. **(B)** (Average) proportion of substitution classes of always minor variants for all samples of the cohort, grouped and sorted by the number of minor SNVs (e.g., at position 10 of the x-axis one can find the average proportion of substitution classes for all samples with 10 minor SNVs). The right plot is drawn with a different scale to depict highly-hypermutated samples; colored shades were added for clarity sake, to highlight the variation of the relative proportion of substitution classes.

the analysis, to focus on SNVs likely related to host-specific mutational processes and by excluding variants presumably transmitted during infection events.

In Fig. 3B we return the (average) proportion of substitution classes obtained by grouping samples with respect to the number of minor SNVs. One can notice that samples with less than 50 minor variants display a relatively even proportion of the substitution classes C>T:G>A, C>A:G>T and A>G:T>C, plus some residuals from other classes. Remarkably, samples with more than 50 and less than 500 minor variants show a clear prevalence of C>A:G>T class, whereas highly hypermutated samples ($>$ 500 minor SNVs) are predominantly characterized by C>T:G>A class. This result strongly supports the existence of focal processes causing mutations at different rates in distinct samples, and which we here investigate via *mutational signature* analysis (see Methods).

In detail, to identify the mutational processes responsible for the SARS-CoV-2 variants with a statistically grounded approach, we applied a Non-Negative Matrix Factorization (NMF) [30] and standard metrics to determine the optimal rank (see Methods). In particular, we analyzed the mutational profiles of 218 samples with at least 6 minor variants (on 1003 total samples), to ensure a sufficient sampling of the distributions. Strikingly, 3 distinct and non-overlapping mutational signatures are found and explain 99.99% of the variance in the data (Fig. 4A and Suppl. Fig. 3; Cophenetic correlation coefficient $= 1.00$, Pearson average correlation of predictions 0.98). In particular, signature $S\#1$ is predominantly related to substitution C>T:G>A (99.95%), signature $S\#2$ to substitution C>A:G>T (97.14%), while signature $S\#3$ is composed by a mixture of all remaining substitutions with different prevalence, namely T>C:A>G (52.17%), T>A:A>T (20.03%), C>G:G>C (14.88%) and T>G:A>C (12.91%).

Characterization of mutational signatures of SARS-CoV-2. Signature S#1 is related to C>T:G>A substitution, which was often associated to APOBEC (*Apolipoprotein B mRNA Editing Catalytic Polypeptide-like*), i.e., a cytidine deaminases involved in the inhibition of several viruses and retrotransposons [31]. An insurgence of APOBEC-related

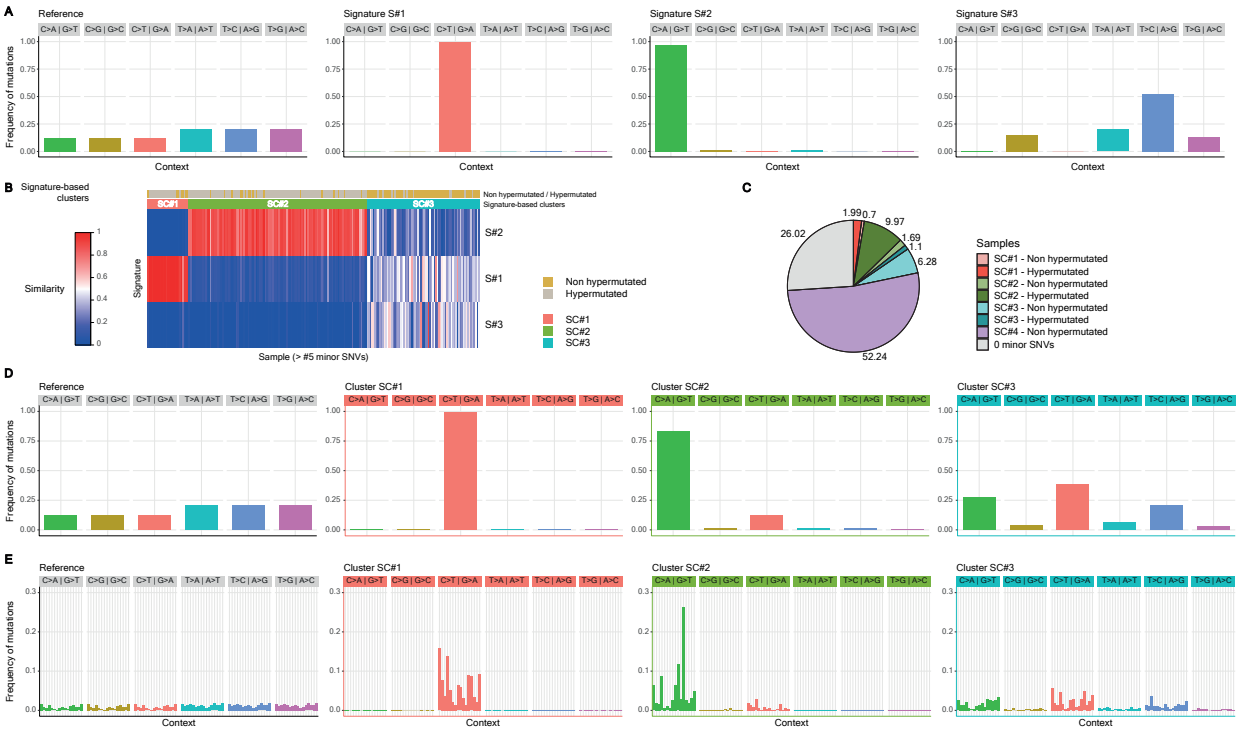

Figure 4: **Mutational signatures of SARS-CoV-2**. **(A)** The nucleotide class distribution in SARS-CoV-2-ANC reference genome [12] and for the 3 SARS-CoV-2 mutational signatures retrieved via NMF on 6 substitution classes are shown. **(B)** Heatmap returning the clustering of 218 samples with ≥ 6 minor variants (21.73% of the cohort), computed via k-means on the low-rank latent NMF matrix (see Methods). **(C)** Pie-chart returning the proportion of samples in the three signature-based clusters, plus a fourth cluster SC#4 including all samples with ≥ 1 and < 6 minor variants, and the group of samples with 0 minor SNVs. **(D)** Categorical normalized cumulative VF distribution of all SNVs detected in each signature-based cluster, with respect to 6 substitution classes and to **(E)** 96 trinucleotide contexts, as compared to the theoretical distribution in SARS-CoV-2-ANC reference genome (left).

mutations was observed in other coronaviruses shortly after spillover [32] and it was recently hypothesized that APOBEC-like editing processes might have a role in the response of the host to SARS-CoV-2 [27].

As specified above, a mutational process occurring on single-stranded RNA with a given pattern, e.g., C>T, could occur as a C>T mutation on the *plus* reference strand, but could similarly occur on the *minus* strand, again as a C>T substitution. However, C>T events originally occurring in the *minus* strand would be recorded as G>A owing to the mapping of the mutational event as a reverse-complement on the *plus* reference genome. Starting from these considerations and hypothesizing that the C>T:G>A substitution is mediated by APOBEC, which operates on single-stranded RNA and is similarly active on both strands, the analysis of the C>T / G>A ratio (or, more generally, of a *plus/minus* substitution ratio) should give an accurate measurement of the molar ratio between the two viral strands inside the infected cells.

In our case, by looking at the relative proportion of substitutions one can notice that there is a large disproportion between C>T and G>A, with a ≈ 17-fold substitution ratio in favor of the former (97.43%/ 5.75%). This result allows us to hypothesize that *plus* and *minus* viral strands of SARS-CoV-2 genome are present in infected cells with a molar ratio of approximately $17:1$ in favor of the *plus* strand and are consistent with the expected activity of APOBEC on single-stranded RNA. Further experimental analyses will be required to confirm this hypothesis.

The second signature S#2 is predominantly characterized by substitution C>A:G>T, whose origin is however still obscure. To gain insight into the mechanisms responsible for its onset, also in this case we analyzed the C>A and G>T substitution frequency, which revealed a strong disproportion in favor of the latter (1.95% vs 32.92%). Strikingly, the G>T to C>A ratio is $17:1$, virtually identical to the C>T to G>A ratio shown for the single-stranded APOBEC process, therefore suggesting that: (i) the C>T:G>A mutational process is able to operate on single-stranded RNA, (ii) the G>T substitution is the active mutational process, (iii) the ratio value of $17:1$ represents a constant, likely reflecting the molar ratio between the two viral strands.

In this respect, one might hypothesize a role for *Reactive Oxygen Species* (ROS) as mutagenic agent underlying this signature, as observed for instance in clonal cancer evolution [33]. ROS are extremely reactive species formed by the partial reduction of oxygen. A large number of ROS-mediated DNA modifications have already been identified; in particular however, guanine is extremely vulnerable to ROS, because of its low redox potential [34]. ROS activity on guanine causes its oxidation to 7,8-dihydro-8-oxo-2′-deoxyguanine (oxoguanine). Notably: (i) oxoguanine can pair with adenine, ultimately causing G>T transversions, and (ii) ROS are able to operate on single-stranded RNA, therefore their mutational process closely resembles the C>A:G>T pattern we see in signature S#2. Thus, it is sound to hypothesize that the C>A:G>T substitution is generated by ROS, whose production is triggered upon infection, in line with several reports indicating that a strong ROS burst is often triggered during the early phases of several viral infections [35, 36].

Finally, signature S#3 is primarily characterized by A>G:T>C substitution, which is typically imputed to the ADAR deaminase mutational process [37]. ADAR targets adenosine nucleotides, causing deamination of the adenine to inosine, which is structurally similar to guanine, ultimately leading to an A>G substitution. Unlike APOBEC, ADAR targets double-stranded RNA, hence it is active only on *plus/minus* RNA dimers. In line with this mechanism and in sharp contrast with APOBEC, A>G's and the equivalent T>C's show a similar prevalence (4.72% and 7.63%,respectively; ratio ≈ 1.62), therefore supporting the notion that the A>G:T>C mutational process is exquisitely selective for double-stranded RNA, where it can similarly targets adenines present on both strands.

Identification of signature-based clusters. We then clustered the 218 samples with at least 6 mutations (on 1003), by applying k-means on the low-rank latent NMF matrix and employing standard heuristics to determine the optimal number of clusters (see Methods). As a result, 3 signature-based clusters (SC#1, SC#2 and SC#3) are retrieved, including 27, 117 and 74 samples respectively (see Fig. 4B).

Remarkably, cluster SC#1 and SC#2 are characterized by distinctive signatures, S#1 (dominated by substitution C>T:G>A) and S#2 (C>A:G>T), respectively, whereas cluster SC#3 is characterized by a mixtures of all three signatures. The Silhouette width coefficients demonstrate that clusters are robust and well-separated (0.90, 0.86, 0.52 for the three clusters, respectively; average Silhouette $= 0.76$) and, in particular, the samples of clusters SC#1 and SC#2 display non-overlapping categorical VF distributions (see Fig. 3D; Pearson correlation coefficient between the exposure matrix $= -0.83$). This impressive result points at the existence of two mutually exclusive host-related mutational processes involving different groups of samples.

We here recall that samples with a number of minor variants between 1 and 5 (52.24% of the cohort) cannot be reliably associated to signature-based clusters, due to the low number of SNVs. For this reason, such samples were considered separately in the analysis and were labeled as cluster SC#4 from now on (Fig. 4C).

Importantly, by computing the categorical VF distribution of all minor SNVs with respect to all 96 trinucleotide contexts (i.e., by considering flanking bases), one can notice that clusters SC#1 and SC#3 display profiles that closely resemble that of the theoretical substitution distribution of the reference genome, thus suggesting that, in such cases, the host-related mutational processes are likely independent from flanking bases. Conversely, SC#2 displays a distribution of C>A:G>T substitutions with prevalent peaks in certain contexts and, especially, on GCT>GAT context. We finally note that, due to the possible transmission of minor variants among hosts during infections (see above), signature-based clusters might include both samples with host-related mutational processes and samples with minor variants herited from infecting hosts.

Characterization of signature-based clusters. We analyzed in-depth the sublineage composition of the 4 different signature-based clusters. As a first noteworthy result, while the distributions of the number of clonal variants are significantly alike across clusters (Kolmogorov-Smirnov, KS test $p > 0.10$ for $5/6$ pairwise comparisons, KS $p = 0.003$ for SC#3 vs. SC#4; see Fig. 5A and Suppl. Table 2), the distributions of the number of minor variants is significantly different among clusters (KS $p < 0.01$ for all pairwise comparisons; see Fig. 5B and Suppl. Table 2).

In particular, clusters SC#1 and SC#2 are characterized by a significantly high number of minor variants (1st quartile $= 59$ and $= 67$ minor variants , median $= 590$ and $= 141$, 3rd quartile $= 1428$ and $= 241$, max $= 5255$ and $= 754$, for SC#1 and SC#2, respectively). Accordingly, both clusters include a large majority of hypermutated samples (with > 50 minor variants), 20 on 27 and 100 on 117 for SC#1 and SC#2, respectively. This result supports the existence of highly active mutational processes and is consistent with the hypothesis of APOBEC- and ROS-related mechanisms for cluster SC#1 and SC#2, respectively. Conversely, cluster SC#3 displays a much lower number of minor variants and most samples are non-hypermutated (1st quartile $= 18$ minor variants, median $= 29$, 3rd quartile $= 42.5$, max $= 232$; hypermutated samples 11 on 63). This finding hints at the existence of mild spontaneous mutational processes affecting this signature-based cluster.

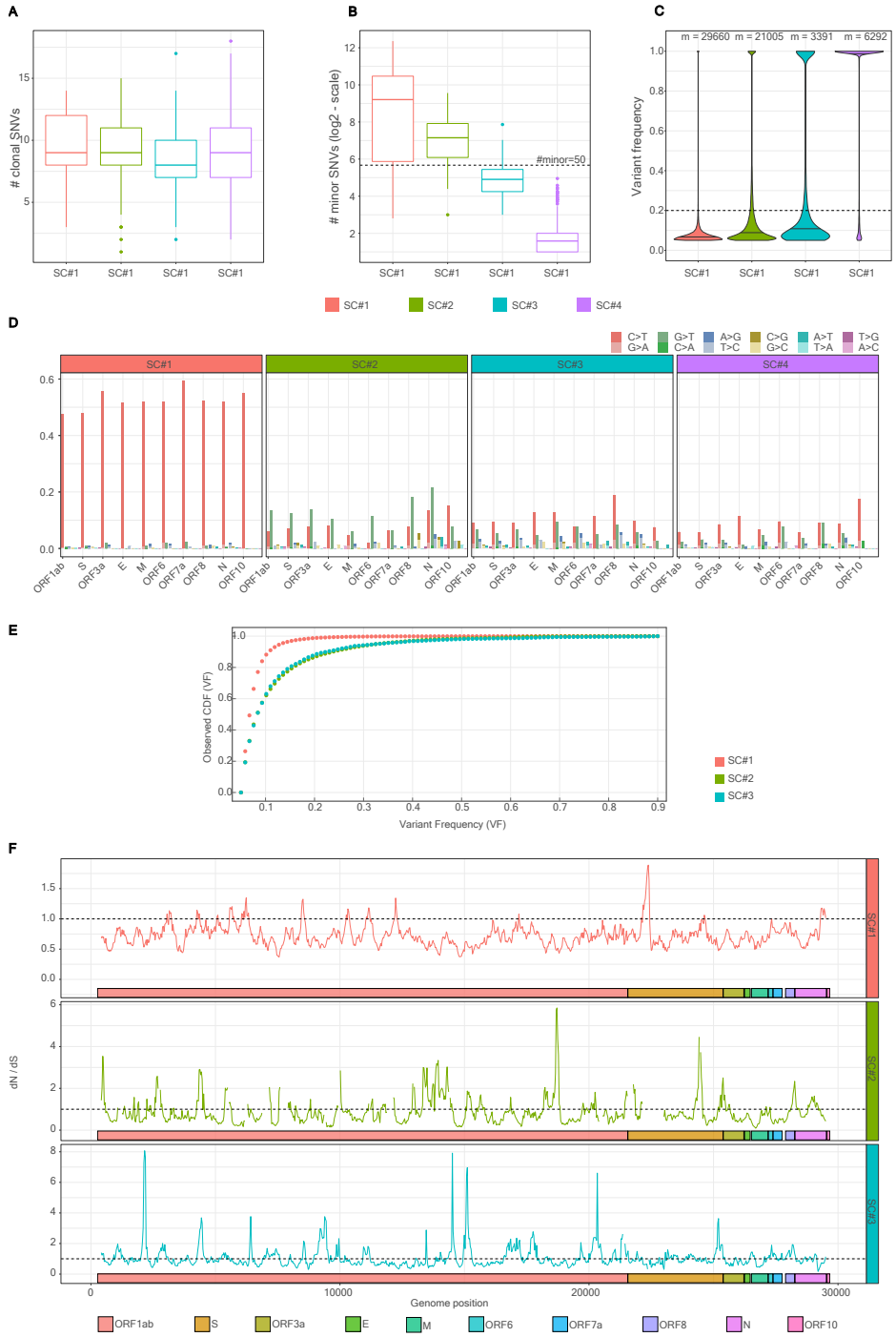

Figure 5: **Characterization of signature-based clusters of SARS-CoV-2 samples**. **(A)** Distribution of the number of clonal variants with respect to the 4 signature-based clusters described in the text. **(B)** Distribution of the number of minor variants for the 4 signature-based clusters ($\log_2$ scale). **(C)** Violin plot returning the VF distribution with respect to signature-based clusters (m returns the number of non-zero entries of the VF matrix). **(D)** Barplot displaying the categorical distribution of all SNVs detected in all signature-based clusters, with respect SARS-CoV-2 ORFs (normalized by reference allele count in each ORF). **(E)** Cumulative distribution function computed on the VF of all minor SNVs, grouped by three signature-based clusters. **(F)** Corrected dN/dS ratio plot, as computed by normalizing the ratio on cluster substitution distribution, on a 300-base sliding window, with respect to signature-based clusters (see Methods).

In this regard, the accumulation of a large number of mutations is well-known to be detrimental for the overall fitness of the virus. For this reason, one might hypothesize that in presence of hypermutation, e.g., when the APOBEC or ROS responses are triggered, the ability of the virus to proceed through its vital cycle, therefore accumulating further mutations, is greatly impaired or completely abolished. As opposite, in absence of major mutational processes, the virus could undergo a slower accumulation of variants, a process that is likely generated by a mixture of different spontaneous processes and which might be consistent with the mutational profile observed in cluster SC$\#3$.

To further investigate this issue, we analyzed the distribution of always clonal variants with regard to the whole cohort (see Methods). As one can see in Suppl. Figure 4, the categorical VF distribution on all substitution classes closely resembles that of cluster SC$\#3$ on minor variants (Pearson correlation coefficient $= 0.95$). The fact that the cluster SC$\#3$ is highly correlated with the clonal variants profile and the relative scarcity of C>T:G>A and C>A:G>T substitutions in the clonal population would support the hypothesis of purifying selection against (hyper)mutational processes characterizing signatures S$\#1$ and S$\#2$. This aspect is additionally discussed in the dN/dS analysis presented in the next subsection.

The VF distribution for all SNVs highlights additional differences among signature-based clusters. In particular, cluster SC$\#1$ displays higher abundance of low-frequency variants (in the range $5 - 10\%$), with respect to the other clusters (Fig. 5C). Moreover, the categorical distributions of substitutions with respect to SARS-CoV-2 Open Reading Frames (ORFs) seem to be significantly different among signature-based clusters and are consistent with the composition of the corresponding mutational signatures (Fig. 5D). In fact, while in most cases variants appear to be randomly distributed across ORFs, one can notice that signature-cluster SC$\#2$ displays a certain bias toward ORF N, i.e., the SARS-CoV-2 nucleocapsid protein. Overall, these results reinforce the hypothesis of distinct mutational processes active in different patients. When clinical data would be available in combination to sequencing data, this will allow to assess the correlation with clinical outcomes.

Evidence of purifying selection against signature-related (hyper)mutational processes. To investigate the evolutionary dynamics of SARS-CoV-2, we first analyzed the cumulative distribution function (CDF) of the VF of all minor variants detected in the cohort, grouped by signature-based cluster (we did not include cluster SC$\#4$ in the analysis, due to the low number of minor variants). Also in this case, we excluded clonal mutations from the analysis, as mostly related to transmission and accumulation evolutionary events [12].

Remarkably, the CDF of cluster SC$\#1$ is significantly different from that of the other clusters and presents a steeper slope, suggesting that the mutational processes related to signature S$\#1$ (i.e., APOBEC) produce a considerably larger number of low-frequency variants, likely due to a higher mutational rate, and in spite of possible selection and transmission phenomena (Fig. 5E).

To further investigate this issue, we performed a corrected version of the dN/dS ratio analysis, i.e., obtained by normalizing the S/NS rate with respect to the theoretical distribution of substitutions detected in each cluster, as suggested in a different context in [38]. In Fig. 5F, one can notice that the dN/dS ratio trend is significantly different for the three signature-based clusters.

In particular, SC$\#1$ show a ratio that is generally below 1, suggesting the existence of significant purifying selection, a result that is consistent with our previous conclusions. The dN/dS ratio is slightly higher for SC$\#2$, yet mostly lower than 1, hinting at the presence of milder purifying selection processes. Finally, as expected, signature-based cluster SC$\#3$ displays a dN/dS ratio close to 1 on the whole genome, suggesting that the mutational processes underlying this cluster might be spontaneous and neutral in the population, as also confirmed by the similarity with the clonal variant substitution profile shown above.

High-resolution model of SARS-CoV-2 evolution via VERSO reveals convergent variants and allows to quantify minor variants transmission. We employed VERSO, a framework introduced in [12] for the reconstruction of high-resolution models of viral evolution from raw sequencing data of viral samples. In particular, we first applied VERSO to the binarized VF profiles of 24 clonal variants (VF $> 90\%$) detected in at least 5% of the cohort, in order to obtain a phylogenomic lineage tree, which describes the existence 13 different viral genotypes (or lineages) and the ancestral relations among them (Fig. 6A).

Interestingly, SNV g.29095T>C (mapped on ORF *N*, synonymous) appear to be the earliest evolutionary event from reference genome SARS-CoV-2-ANC [12]. Two major clades characterize the model and correspond to known SARS-CoV-2 types [39, 40], as determined by presence/absence of mutations g.8782T>C (*ORF1ab*, synonymous) and g.28144C>T (*ORF1ab*, p.84S>L). Furthermore, the model unveils the presence of *homoplasies*, as 2 clonal variants are found in separate clades in a convergent evolution scenario, namely g.14805C>T (*ORF1ab*, synonymous) and g.28311C>T (*N*, p.13P>L) (Fig. 6B). One might hypothesize that such SNVs have spontaneously emerged in unrelated samples and were selected either due to some functional advantage or, more likely, to the combination

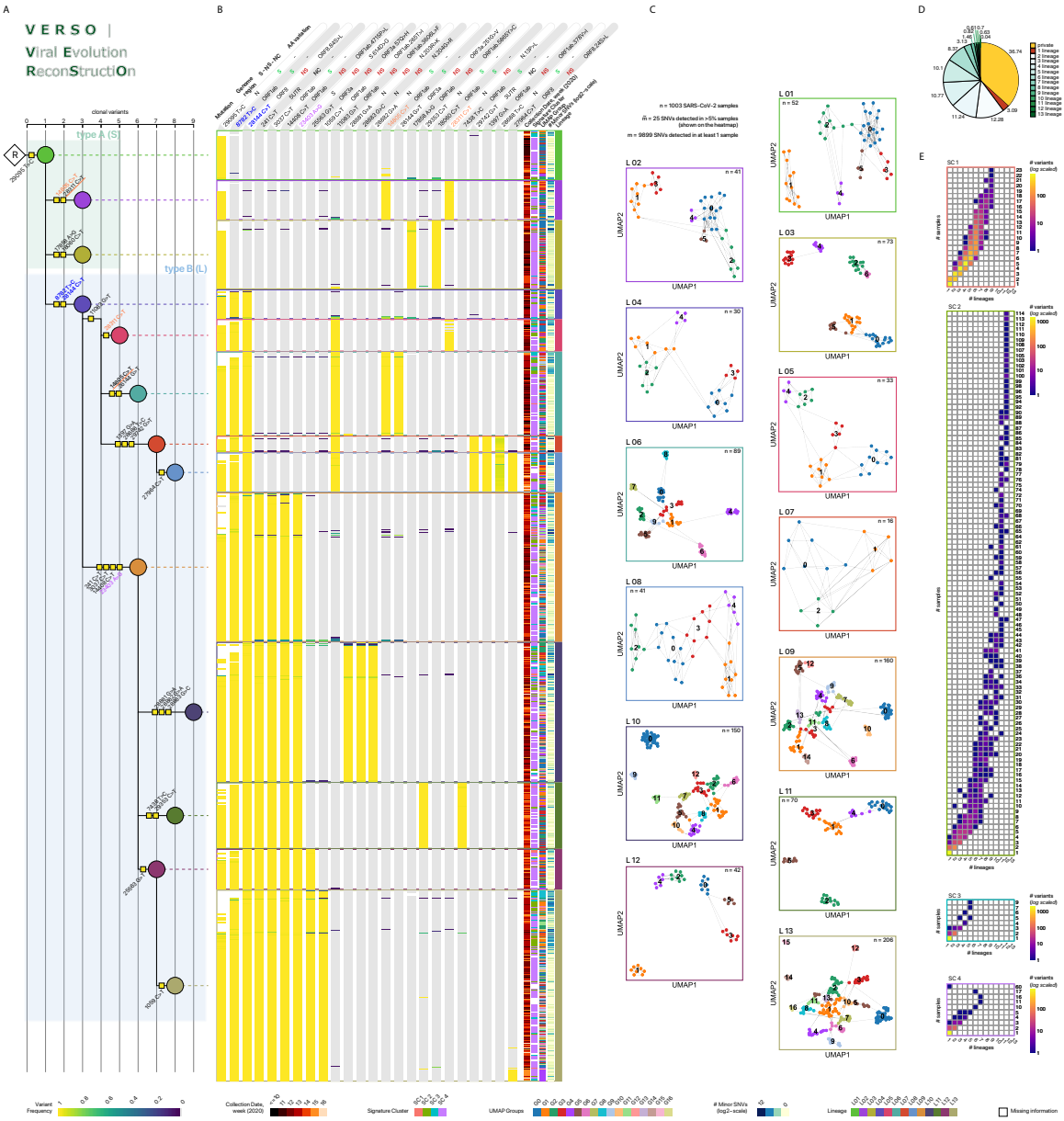

Figure 6: **Viral evolution and sublineage characterization of** 1003 **SARS-CoV-2 samples of via VERSO [12].** **(A)**The phylogenomic lineage tree returned by VERSO by considering 24 clonal variants (VF > 0.90) detected in at least 5% of the 1003 samples of the cohort is displayed. The R-labeled rhombus indicates the reference genome SARS-CoV-2-ANC, yellow squared nodes mark the variants characterizing the ancestral relation, while circle nodes represent viral genotypes or lineages L1–13 and are associated to samples partitions. **(B)** Heatmap returning the variant frequency (VF) profiles of 1003 SARS-CoV-2 samples on selected clonal variants. Samples are annotated with collection date, location and signature-based cluster. SNVs are annotated with mapping on ORFs, synonymous (S), nonsynonymous (NS) and non-coding (NC) states, and amino acid substitutions (asterisks label nonsense mutations). Variants in blue characterize the B (L) type [39, 40], variant in purple was recently correlated with clincal outcome [41, 42, 43], variants in orange indicate homoplasies. **(C)** The 13 panels display the projection of the viral sublineage composition and of the genomic distance among the samples on the UMAP low-dimensional space, as computed by VERSO from VF profiles. Samples are clustered by employing the Leiden algorithm on the k-nearest neighbor graph (k-NNG), by using the Bray–Curtis dissimilarity on VF profiles [44]. Solid lines are the edges of the k-NNG. **(D)** Pie-chart returning the proportion of SNVs privately detected in single samples, detected in multiple samples of the same lineage and in multiple samples and multiple lineages. **(E)** Heatmaps displaying the count of minor variants with respect to the number of lineages and samples in which they are found, grouped by signature-based cluster (e.g., at row 3 and column 5, the color represents the number of SNVs found in 3 lineages and 5 samples).

of founder and stochastic effects involved in variant transmission during infections, which might lead certain minor SNVs transiting to clonality in the population (see above). The latter hypothesis is further confirmed by the fact that g.14805C>T is a silent mutation, and points at the complex interplay between evolution within hosts and transmission among hosts. Furthermore, we note that variant g.23403A>G (*S*, p.614D>G), whose correlation with clinical outcome was recently hypothesized [41, 42, 43], is found in 5 viral lineages, which include 628 samples of the cohort.

A second step of VERSO allows to process the complete VF profiles of the samples in each lineage, to project the sublineage composition and the genomic distance among samples on a low-dimensional space (see Methods). Such maps return the genomic similarity among hosts harboring the same viral lineage (i.e., same clonal mutation), and which would not be identifiable by considering consensus sequences. In Fig. 6C one can see that many distinct intra-lineage clusters are found, which identify samples with similar VF profiles on all variants. Also, a complex inter-cluster network is revealed in each case, which connect clusters via samples presenting a relatively small evolutionary distance. This genomic distance measures the impact of both transmission processes of minor variants during infections and of shared mutational hotspots hitting independent samples; in particular, the latter occurrence is more likely when considering samples affected by hypermutation processes, as observed for signature clusters SC#1 and SC#2.

We finally quantified the number of homoplasies involving minor variants (i.e., identical minor SNVs observed in samples of independent lineages). As extensively discussed in [12], while all the clonal variants of a host are most likely transmitted during an infection, the extent of transmission of minor variants is still baffling and is highly influenced by bottlenecks, founder effects and stochasticity [26, 12]. Multiple simultaneous infections of the same host from individuals harboring distinct viral lineages, also named *superinfections*, might in principle affect variant clonality, yet their occurrence is extremely rare [11]. For such reasons, homoplasies involving minor variants are most likely due to: (i) positive selection of the variants due to some functional advantage, in a scenario of parallel/convergent evolution, (ii) mutational hotspots, i.e., SVNs falling in mutation-prone sites or regions of the viral genome, (iii) complex transmission dynamics involving founder effects and stochasticity, which may allow certain minor variants to transit to clonality, eventually leading to a lineage transmutation (see above). We also note complex combinations of such phenomena are possible, as proven by the complex sublineage architecture often observed in individual samples.

In our case, we observe that 36.74% of minor variants are observed as private of single samples, 3.09% in multiple samples of the same lineage and 60.17% are detected in samples belonging to distinct lineages (Fig. 6D). Important conclusions can be drawn from these results. On the one hand, the quantity of minor variants most likely transmitted across hosts appears to be relatively limited, as proved by the ratio of mutations shared in different hosts harboring the same viral lineage. On the other hand, the number of minor variants observed in multiple independent lineages is surprisingly high; this is especially true for clusters related to (hyper)mutational processes (i.e., signature-based clusters SC#1 and SC#2), which can display an extremely high number of minor variants shared across a large number of lineages and samples (up to 9 and 12 lineages, and 23 and 114 distinct samples, respectively; Fig. 6E). This result would hint at the abundance of mutational hotspots inducing the occurrence of identical minor variants on likely unrelated samples, as well as the possible presence of positively selected variants.

Overall, these results suggest that the complex sublineage architecture of samples can be exploited to highly refine standard (phylo)genomic analyses and provide a quantitative measure of intra- and inter-host genomic diversity.

Validation

We employed two independent datasets comprising 179 and 204 samples, respectively (NCBI BioProjects: `PRJNA625551`, Amplicon, USA; `PRJNA633948`, Amplicon, Australia), to validate the presence of the mutational signatures described above.

In detail, we performed signature assignment with respect to the discovered signatures on 41 and 131 high-quality samples showing ≥ 6 minor variants. Three signature-based clusters are found for both datasets and explain more that 99% of the variance. Such clusters are related to combinations of signatures consistently to the analysis presented in the text and display alike distributions of minor SVNs (see Suppl. Figs. $5-8$).

Discussion

Standard (phylo)genomic analyses of viral consensus sequences might miss useful information to investigate the elusive mechanisms of viral evolution *within hosts* and of transmission *among hosts*. In this respect, raw sequencing data of viral samples can be effectively employed to deliver a high-resolution picture of *intra-host heterogeneity*, which might underlie different clinical outcomes and affect the efficacy of anti-viral therapies. This aspect is vital especially

during the critical phases of an outbreak, as experimental hypotheses are urgently needed to deliver effective prognostic, diagnostic and therapeutic strategies for infected patients.

We here presented the largest up-to-date quantitative analysis of intra-host genomic diversity of SARS-CoV-2, which revealed that the large majority of samples present a complex sublineage architecture, likely due to the interplay between host-related mutational processes and transmission dynamics.

In particular, we here proved the existence of mutually exclusive *viral mutational signatures*, i.e., nucleotide substitution patterns, which show that different hosts respond to SARS-CoV-2 infections in different ways, in many cases by suffering hypermutational processes ruled either by APOBEC or ROS-related activity. As a first consequence, $\approx 31\%$ of the SARS-CoV-2 viral genome is already mutated in at least one sample and, in most cases, such mutations are found at a low frequency in hypermutated individuals.

The dS/dN analysis shows that such numerous low-frequency variants tend to be purified in the population whereas, conversely, a significant number of variants linked to spontaneous low-rate mutational processes appear to consolidate. In particular, due to the still obscure combination of founder effects and selection phenomena, certain variants appear to transit to clonality in the population, eventually leading to the definition of new viral genotypes. Once become clonal, mutations tend to accumulate in the population, as proven by a statistically significant increase of genomic diversity, and might be used to reconstruct robust models of viral evolution via VERSO [12].

Finally, the analysis of homoplasies, i.e., (low-frequency) variants shared across distinct viral lineages and unlikely due to infection events, demonstrate that a high number of mutations can independently emerge in multiple samples, due to mutational hotspots often related to signatures or, possibly, to positive (functional) selection.

Methods

Dataset. We analyzed a cohort comprising 1171 samples from NCBI BioProject with accession number `PRJNA613958`. For all samples, Amplicon sequencing high-coverage raw data are provided; all patients were located in Australia. Within this cohort, we considered for our analyses, 1003 high-quality samples having coverage > 50 in more than 90% of the virus genome. We considered two additional datasets for validation, NCBI BioProject with accession numbers `PRJNA625551` (United States, 179 AMPLICON samples) and `PRJNA633948` (Australia, 204 AMPLICON samples). We applied the same quality filters to these datasets (coverage > 50 in $> 90\%$ of virus genome), to obtain two validation sets of 88 and 194 samples, respectively.

SNVs calling. We downloaded SRA files and converted them to FASTQ files using SRA toolkit. Following [12], we used Trimmomatic (version 0.39) to remove positions at low quality from the RNA sequences, using the following settings: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:40.

We used bwa mem (version 0.7.17) to map reads to the reference genome (SARS-CoV-2-ANC [12]). We then generated sorted BAM files from bwa mem results with SAMtools (version 1.10) and removed duplicates with Picard (version 2.22.2). Variant calling was performed generating mpileup files with SAMtools and then using VarScan (min-var-freq parameter set to 0.01) [45].

Signatures analysis. The analysis was performed on minor variants (VF $> 5\%$ and $\leq 90\%$), in order to ensure that the considered variants are not due to transmission, but are likely emerged in the host. In such way, we could associate to each discovered signature a mechanism causing variants in the viral genome related to the specific host.

Signatures decomposition was formulated as a Non-negative Matrix Factorization problem (NMF) [30]. Given n samples, r possible substitution classes (e.g., C>T:G>A) and s signatures, we can define the following objects:

- the input data matrix $\mathbf{D}$, a $n \times r$ dimensional matrix, where every element $d_{i,j}$ represents the cumulative VF of the SNVs with substitution class j in the ith sample. Note that $d_{i,j} \in \mathbb{R}^+$;
- the low-rank latent NMF matrix $\mathbf{A}$, a $n \times s$ dimensional matrix , where every element $a_{i,j}$ represents the linear combination coefficient of signature j in sample i (also *exposure* of the ith sample to signature j [29]). Note that $a_{i,j} \in \mathbb{R}^+$;
- the signature (or basis) matrix $\mathbf{B}$, a $s \times r$ dimensional matrix, where every row is a categorical distribution of all substitution classes in each signature. For this matrix we assume that every row must sum up to 1, then $b_{i,j} \in \{0, 1\}$ and $\sum_{i=1}^{r} b_{i,j} = 1$.

In particular, we here considered 6 substitution classes ($r = 6$) by merging equivalent substitution types, namely G>T:C>A, G>C:C>G, G>A:C>T, A>T:T>A, A>G:T>C and A>C:T>G.

Problem 1 *(Signature decomposition): Given the data matrix* $\mathbf{D}$*, we aim at finding the NMF latent matrix* $\mathbf{A}$ *and the signature matrix* $\mathbf{B}$*, such that* $||\mathbf{D} - \mathbf{A} \cdot \mathbf{B}||_2$ *is minimum.*

To solve the stated problem, we here performed a total of 1000 independent NMF runs with standard update [30], for solutions at ranks varying from 1 to 6, where initial solutions were randomly initialized; for each run a total of 20 iterations were performed where signatures and their assignments to samples were iteratively estimated by non-negative least squares [46]. The final solution was constructed as the consensus of the 1000 runs [30].

We then employed multiple state-of-the-art approaches to assess the optimal rank (optimal number of signatures s) for the NMF decomposition. We first assess the stability of NMF results over the 1000 runs, with the idea that stable solutions are preferable to unstable ones; to this extent, we computed Cophenetic correlation coefficient [30] and dispersion coefficient [47], which both showed a sharp drop at rank equal to 4, hinting at 3 signatures as a stable solution (see Suppl. Fig. 3A–B). Furthermore, we also evaluated the goodness of fit of NMF solutions at different ranks, with rank equals to 3 being able to explain 99.99% of variance in the data (see Suppl. Fig. 3A–C) and showing an average cross-validation error of 0.15 [48]; finally, we report as Suppl. Fig. 2D the average Pearson correlation between observations and predictions by NMF with rank equals to 3, showing a plateau with correlation equals 0.98 [48]. All of this supports 3 as the optimal rank for our decomposition problem and the presence of 3 distinct mutational signatures in out data.

Identification of signature-based clusters. In order to identify clusters of samples possibly affected in different proportions by the discovered mutational signatures, we considered the low-rank latent NMF matrix $\mathbf{A}$ defined above.

Specifically, we first normalized $\mathbf{A}$ such that each row of the matrix sums up to 1 and then computed the euclidean distance among each pair of samples. We next performed Principal Component Analysis (PCA) on the distance matrix to estimate the optimal number of clusters present in our data. In detail, the analysis of the eigenvalues of the distance matrix shows that 3 components explain $> 99\%$ of the variance, followed by a plateau. Accordingly, we performed k-means clustering with $k = 3$ on the first 3 Principal Components of the distance matrix to discover the signature-based clusters.

dN/dS analysis. In order to quantify the selection pressure in coding regions of SARS-CoV-2, we employed dN/dS analysis, which assesses and compares non-synonymous to synonymous substitution rates. In its standard version, this analysis assumes uniform nucleotide substitution probabilities across the genome; however, this hypothesis might not hold if different mutational processes are active with biases over a subset of substitutions (e.g., non-uniform distribution might be observed across signature-based clusters). If this bias is not taken in account, it may lead to erroneous estimation of the dN/dS ratio [38].

For this reason, since we discovered the existence of different host-related mutational processes (i.e., the mutational signatures) that are strongly biased toward specific substitutions, we developed a corrected dN/dS ratio analysis, as proposed in a different context in [38]. Specifically, given the ith sliding window of the coding region comprising l bases and considering the fth signature-based cluster, the corrected dN/dS ratio (for signature-cluster f and sliding window i) is given by:

$$\text{corrected} \quad \frac{dN}{dS}^{i,f} = \frac{N_{obs}^{i,f}/S_{obs}^{i,f}}{\sum_{k=1}^{l}\sum_{c} P_{f,c}N_{k,c}/\sum_{k=1}^{l}\sum_{c} P_{f,c}S_{k,c}} \qquad (1)$$
$$c \in \{\text{G>T:C>A}, \ldots, \text{A>C:T>G}\} \quad \text{(6 substitution classes)}$$

where, $N_{obs}^{i,f}$ and $S_{obs}^{i,f}$ are the numbers of non-synonymous and synonymous substitutions detected in the window i in at least one sample of signature-based cluster f, $P_{f,c}$ is the probability of substitution class c in signature-based cluster f (computed with respect to the categorical normalized cumulative VF distribution of all substitution classes in that cluster), and $N_{k,c}$ ($S_{k,c}$) is equal to 1 if class c identifies an admissible non-synonymous (synonymous) substitution in the kth position of the window i, 0 otherwise.

VERSO high-resolution model of viral evolution. We employed the VERSO framework introduced in [12] to reconstruct a high-resolution model of viral evolution. The framework processes variant frequency profiles generated from raw sequencing data and includes two subsequent algorithmic steps.

In the first step, VERSO takes as input the binarized mutational profiles of clonal mutations. In our analysis, we considered only clonal variants (VF $> 90\%$) detected in at least 5% of the samples of the cohort ($\tilde{m} = 24$ clonal variants on $n = 1003$ samples). VERSO employs a probabilistic approach to model the accumulation of clonal variants

and the presence of noise and uncertainty in the data, as proposed in the context of cancer evolution in [49]. As output, a phylogenomic lineage tree is returned, in which each node correspond to a viral genotype (or lineage), edges are parental relations, characterized by one or more accumulating variants, and any viral lineage is associated to a cluster of samples. Variants showing particularly high level of estimated error rates with respect to the theoretical model represent possible violations of perfect phylogenetic constraints, e.g., due to convergent evolution. In the analysis presented in the text, a grid search comprising 16 different error rates was employed. Hierarchical clustering was then performed on the similarity matrix returned by VERSO to associate samples to viral lineages [12].

In the second step, VERSO takes into account the VF profiles of all variants detected in the cohort and defines a pairwise genomic distance among samples, based on Bray-Curtis dissimilarity [44]. Samples showing similar patterns of co-occurrence of variants might have a similar sublineage architecture, therefore being at a small evolutionary distance and pinpointing possible uncovered infection events or the presence of shared of mutational hotspots. In this work the genomic distance is computed among all samples associated to any given viral lineage, as inferred in the first step, to reduce the impact of confounding effects [50]. The genomic distance is then used in a workflow for dimensionality reduction and clustering via SCANPY [51] including: (i) k-nearest neighbour graph (k-NNG) computation, executed after applying principal component analysis (PCA); (ii) clustering of samples via Leiden algorithms for community detection [52]; (iii) projection of samples on the UMAP low-dimensional space [53]. VERSO finally returns both the partitioning of samples in clusters and the visualization in a low-dimensional space.

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

A.G., D.M., F.A., R.P. and D.R. designed and developed the study. A.G, D.M., F.A. and D.R defined, implemented and executed the computational analyses. A.G., D.M., F.A., R.P. and D.R. analyzed the data and interpreted the results. A.G., D.R. and R.P. supervised the study. All authors wrote the manuscript, discussed the results, and commented on the manuscript.

Competing interests

The authors declare that they have no competing interests.

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. We thank Marco Antoniotti, Giulio Caravagna and Chiara Damiani for helpful discussions8.

References

[1] World Health Organization (WHO). Coronavirus disease 2019 (COVID-19): situation report, 156 (2020).

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

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

[4] 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).

[5] 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`.

[6] 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`.

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

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

[9] Novella, I. S., Domingo, E. & Holland, J. J. Rapid viral quasispecies evolution: implications for vaccine and drug strategies. *Molecular medicine today* **1**, 248–253 (1995).

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

[11] Lythgoe, K. A. *et al.* Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. *bioRxiv* (2020).

[12] Ramazzotti, D. *et al.* Quantification of intra-host genomic diversity of SARS-CoV-2 allows a high-resolution characterization of viral evolution and reveals functionally convergent variants. *bioRxiv* (2020).

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

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

[15] Capobianchi, M. R. *et al.* Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in italy. *Clinical Microbiology and Infection* (2020).

[16] Rose, R. *et al.* Intra-host site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. *medRxiv* (2020).

[17] Lu, J. *et al.* Genomic epidemiology of SARS-CoV-2 in guangdong province, china. *medRxiv* (2020).

[18] Seemann, T. *et al.* Tracking the covid-19 pandemic in australia using genomics. *medRxiv* (2020).

[19] Xu, D., Zhang, Z. & Wang, F.-S. Sars-associated coronavirus quasispecies in individual patients. *New England Journal of Medicine* **350**, 1366–1367 (2004).

[20] Park, D. *et al.* Analysis of intrapatient heterogeneity uncovers the microevolution of middle east respiratory syndrome coronavirus. *Molecular Case Studies* **2**, a001214 (2016).

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

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

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

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

[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] Simmonds, P. Rampant c-> u hypermutation in the genomes of SARS-CoV-2 and other coronaviruses–causes and consequences for their short and long evolutionary trajectories. *bioRxiv* (2020).

[28] Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent rna editing in the transcriptome of SARS-CoV-2. *Science Advances* eabb5813 (2020).

[29] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

[30] Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* **101**, 4164–4169 (2004).

[31] Sharma, S. *et al.* Apobec3a cytidine deaminase induces rna editing in monocytes and macrophages. *Nature communications* **6**, 1–15 (2015).

[32] Woo, P. C., Wong, B. H., Huang, Y., Lau, S. K. & Yuen, K.-Y. Cytosine deamination and selection of cpg suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. *Virology* **369**, 431–442 (2007).

[33] Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

[34] David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative dna damage. *Nature* **447**, 941–950 (2007).

[35] Molteni, C., Principi, N. & Esposito, S. Reactive oxygen and nitrogen species during viral infections. *Free radical research* **48**, 1163–1169 (2014).

[36] Reshi, M. L., Su, Y.-C. & Hong, J.-R. Rna viruses: Ros-mediated cell death. *International journal of cell biology* **2014** (2014).

[37] Nishikura, K. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry* **79**, 321–349 (2010).

[38] Van den Eynden, J. & Larsson, E. Mutational signatures are critical for proper estimation of purifying selection pressures in cancer somatic mutation data when using the dn/ds metric. *Frontiers in Genetics* **8**, 74 (2017).

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

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

[41] Lokman, S. M. *et al.* Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: a computational biology approach. *Infection, Genetics and Evolution* 104389 (2020).

[42] Daniloski, Z., Guo, X. & Sanjana, N. E. The d614g mutation in SARS-CoV-2 spike increases transduction of multiple human cell types. *bioRxiv* (2020).

[43] SKorber, B. *et al.* Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* (2020).

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

[45] 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).

[46] Chen, D. & Plemmons, R. J. Nonnegativity constraints in numerical analysis. In *The birth of numerical analysis*, 109–139 (World Scientific, 2010).

[47] Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495–1502 (2007).

[48] Ramazzotti, D., Lal, A., Liu, K., Tibshirani, R. & Sidow, A. De novo mutational signature discovery in tumor genomes using sparsesignatures. *bioRxiv* 384834 (2019).

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

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

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

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

[53] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018).