Analysis of co-occurring and mutually exclusive amino acid changes and detection of convergent and divergent evolution events in SARS-CoV-2

Ruba Al Khalaf[a], Anna Bernasconi[a,*], Pietro Pinoli[a] and Stefano Ceri[a]

[a]*Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy*

ARTICLE INFO

Keywords:
SARS-CoV-2
Co-occurring mutations
Mutually exclusive mutations
Convergent evolution
Divergent evolution
Statistical testing

ABSTRACT

The inflation of SARS-CoV-2 lineages with a high number of accumulated mutations (such as the recent case of Omicron) has risen concerns about the evolutionary capacity of this virus. Here, we propose a computational study to examine non-synonymous mutations gathered within genomes of SARS-CoV-2 from the beginning of the pandemic until February 2022. We provide both qualitative and quantitative descriptions of such corpus, focusing on statistically significant co-occurring and mutually exclusive mutations within single genomes. Then, we examine in depth the distributions of mutations over defined lineages and compare those of frequently co-occurring mutation pairs. Based on this comparison, we study mutations' convergence/divergence on the phylogenetic tree. As a result, we identify 1,818 co-occurring pairs of non-synonymous mutations showing at least one event of convergent evolution and 6,625 co-occurring pairs with at least one event of divergent evolution. Notable examples of both types are shown by means of a tree-based representation of lineages, visually capturing mutations' behaviors. Our method confirms several well-known cases; moreover, the provided evidence suggests that our workflow can explain aspects of the future mutational evolution of SARS-CoV-2.

1. Introduction

All viruses, including SARS-CoV-2, change over time. Many organizations, such as the Global Initiative on Sharing all Influenza Data (GISAID) [42], Nextstrain [18] and Pangolin [40], are studying the phylodynamics of SARS-CoV-2 genomes to track and define new variants. At the end of May 2021, the World Health Organization (WHO) announced the usage of a novel nomenclature system for naming and tracking SARS-CoV-2 genetic lineages using letters of the Greek alphabet, now offering a reference to refer to viral variants across the world. In the same announcement, the WHO also introduced a multi-level categorization of variants based on the levels of attention they should raise; namely, variants under monitoring (VUM), variants of interest (VOI), and variants of concern (VOCs). Such levels were defined by evaluating specific measures that define the virus properties and virulence. On a different level, several initiatives such as CoVariants [20] and outbreak.info [15] are working on determining the defining (i.e., characterizing) mutations for each variant.

The evolutionary dynamics of the virus were predominantly characterized by a mutational pattern of slow and selectively neutral random genetic drift. Past pandemics and long-term evolutionary dynamics of RNA viruses attest to the fact that such an evolutionary "lull" rarely lasts [28]. Indeed, in late 2020, three relatively divergent SARS-CoV-2 lineages emerged in rapid succession: B.1.1.7 (Alpha), B.1.351 (Beta), and P.1 (Gamma). Those three lineages were considered as VOCs according to the WHO. Due to the continuation of the pandemic, other lineages emerged and, at the beginning of 2022 we faced the emergence of BA.1 (Omicron) and of its descendant lineages. From the point of view of single mutations accumulated since the beginning of the pandemic, most of them had little to no impact on the virus' properties (i.e., not epidemiologically significant). However, some changes have arisen that affect the virus to gain specific advances, such as a spread advantage [7], the associated disease severity [16], or the resistance to vaccines [17], antiviral medicines [44], diagnostic tools, and other public health / social measures [34] (as we analyzed in [1, 2]).

Thanks to the continuous spreading of SARS-CoV-2 and the contextual deposition of its viral sequences to public repositories (even considering possible delays [24]), the viral evolution can be monitored and studied to understand SARS-CoV-2 variants and the risks that they pose. Methods that study the virus characteristics have been developed independently from the phylogenetic techniques traditionally employed in this field. For instance, considerable efforts have been dedicated to building surveillance systems that employ temporal analysis of SARS-CoV-2 mutations to assist in the identification of candidate variants of clinical importance. A number of studies have described typical SARS-CoV-2 mutational profiles across different countries and regions [31], proposing statistical indicators for location-based mutation evolution [45] and observing changes that become recurrently prevalent in different locations, thus suggesting selective advantages [26]. Time-series analyses have been considered for clustering of prevalent SARS-CoV-2 mutations over time [50, 8, 5, 22, 11, 21], trend detection in

*Corresponding author

ruba.al@polimi.it (R. Al Khalaf); anna.bernasconi@polimi.it (A. Bernasconi); pietro.pinoli@polimi.it (P. Pinoli); stefano.ceri@polimi.it (S. Ceri)

ORCID(s): 0000-0002-5645-5886 (R. Al Khalaf); 0000-0001-8016-5750 (A. Bernasconi); 0000-0001-9786-2851 (P. Pinoli); 0000-0003-0671-2415 (S. Ceri)

SARS-CoV-2 short nucleotide sequences [46], and single amino acid changes [41]. Such works consider functions that describe the prevalence of different mutations and, when a number of these are behaving similarly, they recognize a possible distinct variant. These approaches usually have an epidemiological angle and are focused on highly present forms of the virus. A different work [12] focuses on patterns of mutations located in relevant domains of the virus that are found in variants of concern but also in emerging variants, suggesting they can be used as a guidance for next evolution moves. At the same time, several studies have analyzed the evolution of mutational patterns that are typical of a specific geographical area [9, 10, 47, 6]. Our approach does not focus on the most spread variants or on groups of numerous mutations, nor it restricts to specific locations. We take an interest in a more micro-level phenomenon, i.e., involving the relationship between single mutations that appear together or separated in different fragments of the phylogenetic story of SARS-CoV-2.

Co-occurrences are analysed in a number of works [39, 52, 43]. Specifically, Qin et al. [39] used small groups of co-occurring mutations as drivers to define groups of sequences (some of which are location-specific) that are then validated on the phylogenetic tree. This analysis is conducted on less then a million genomes up to the beginning of 2021. For co-occurrences we next use a definition that is close to the co-mutations of Zhang et al. [52], which is however focused only on B.1.1.7, digging deep in the evolution and transmission chains of this variant. For this purpose, they have studied the co-occurrence of SARS-CoV-2 mutations by using only high frequency mutations. Mutation rates are compared with the spatial information: mutations found in places with highly similar mutation rates are considered as potential co-mutation patterns. Finally, Singh et al. [43] proposed an original approach to analyse occurrence/co-occurrence of genomic mutations with NLP techniques by exploiting their conceptual equivalence to the occurrence of words in a textual document – mutational signatures could be understood as topics of a document; the approach is still preliminary.

In this research, we employ a previously unexplored perspective: we consider 8 million sequences and we perform a systematic co-occurrence and mutual exclusion analysis of non-synonymous mutations' pairs. These become our entry point to the SARS-CoV-2 evolution aspects. In the following, we provide a workflow to determine 1) co-occurring pairs of mutations (as preliminarily tested in [1]) and 2) mutually exclusive pairs of mutations (as inspired by research on mutually exclusive human gene pairs [38]). Then, we employ the results from the first phase (co-occurring pairs) to study convergent and divergent evolution events of mutation pairs, also analyzing them from the point of view of single mutations participating to the event. An intuitive visual representation is employed to explain such events and point the attention to a number of interesting cases, which have been validated in the literature. As future work, our workflow encourages the design of a light-weight prediction procedure for the mutational evolution of SARS-CoV-2.

2. Methods

The framework of our study is divided into four parts, discussed in the remainder of this section and overviewed (as dotted-framed areas) in Figure 1:

1) *Data preparation* (described in Section 2.1), where we analyze the initial dataset of viral sequences and prepare aggregated intermediate tables.
2) *Data analysis* (Section 2.2), composed of three parts that use specific statistical tests to, respectively i) identify the co-occurring and mutually exclusive pairs of mutations, ii) compare the distributions over lineages of mutations that appear in previously identified co-occurring pairs, and iii) test the evolution events happened between co-occurring pairs.
3) *Lineages distribution-dependent analysis* (Section 2.3), i.e., an analysis on mutation pairs that takes into consideration how mutations are spread across the lineages; it includes the visualization of convergence/divergence events involving mutation pairs along tree-based structures representing the hierarchy of Pango lineages.
4) *Lineages distribution-independent analysis* (Section 2.4), i.e., an analysis of co-occurring and mutual exclusive mutation pairs that ignores how mutations are spread across the lineages.

The complete analysis has been performed in Python (Version 3.7.6), using classical data science libraries, i.e., Pandas (Version 1.3.5) for data extraction and aggregation, Scipy (Version 1.4.1) for statistical analysis, Seaborn (Version 0.11.2) and Matplotlib (Version 3.1.3) for data visualization.

2.1. Data preparation

A set of over 8.2 million sequences of SARS-CoV-2 genome was collected from GISAID since the beginning of the pandemic until mid of February 2022 to be used in this large scale analysis. Each entry of the initial dataset represents one viral sequence with their assigned Pango lineage [35] and a list of its non-synonymous mutations; these belong to all the proteins of SARS-CoV-2, expressed considering the reference SARS-CoV-2 genome hCoV-19/Wuhan/WIV04/2019 [53].

Data aggregation. From the initial dataset, we prepared three intermediate tables that were used in following analyses as shown in the *Data preparation* box of Figure 1:

1) triplets of the form $\langle m, l, S_{m,l} \rangle$, where $S_{m,l}$ represents the number of sequences from lineage l having a mutation m;
2) for the lineages included in 1), triplets of the form $\langle l, \{seq_1, \ldots, seq_i, \ldots, seq_{S_l}\}, S_l \rangle$, where for each lineage l we have the set of sequences assigned to that Pango lineage and its cardinality S_l;
3) for the mutations included in 1), triplets of the form $\langle m, \{seq_1, \ldots, seq_i, \ldots, seq_{S_m}\}, S_m \rangle$, where for each mutation m we have the set of sequences exhibiting that mutation and its cardinality S_m.

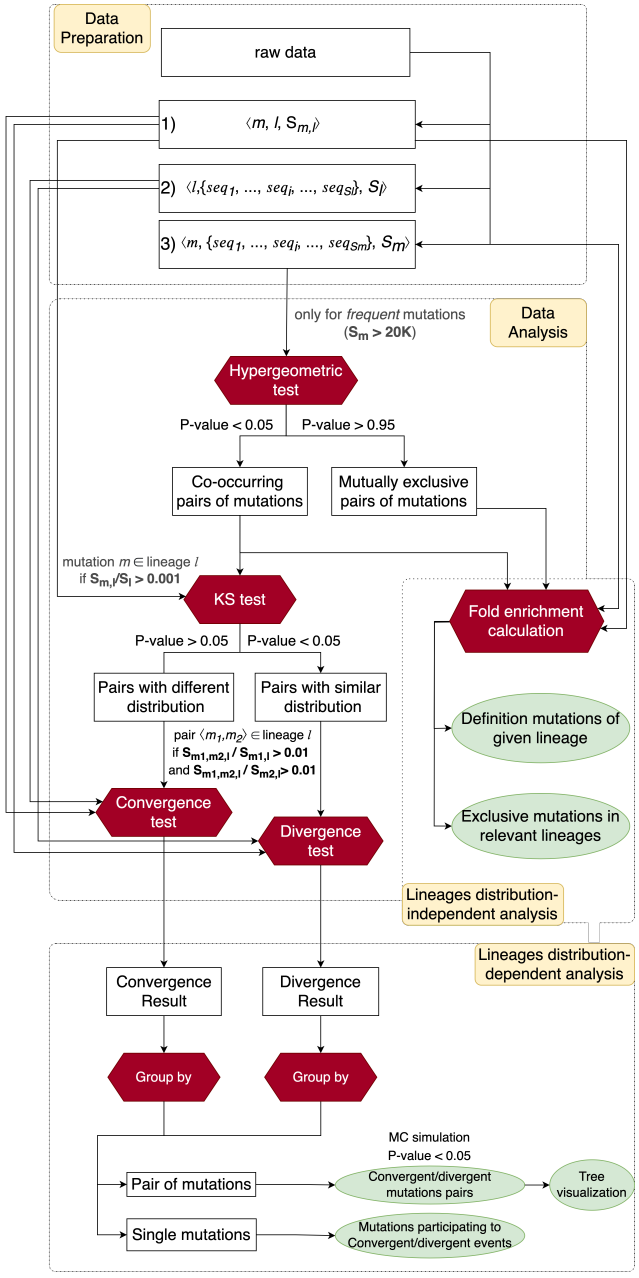

Figure 1: Methodological workflow of the study. The schema is composed of four main parts enclosed in dotted-framed areas: 1) Data preparation; 2) Data Analysis; 3) Lineages distribution-dependent analysis; and 4) Lineages distribution-independent analysis. Legend: S_l: number of sequences assigned to lineage l; S_m: number of sequences holding a mutation m; $S_{m,l}$: number of sequences from lineage l holding mutation m; $S_{m_1,m_2,l}$: number of sequences assigned to lineage l holding the pair of mutations m_1 and m_2; KS test: Kolmogorov-Smirnov test; MC simulation: Monte Carlo simulation.

To reduce the size of the initial dataset, we only considered the most *frequent mutations*, i.e., those found in at least 20K viral sequences as extracted from the intermediate table 3); they amount to 421. In the following, we refer to these as frequent mutations: we only performed the analysis over this list of mutations.

2.2. Data analysis

2.2.1. Detection of co-occurring and mutually exclusive mutations pairs

In a given population (e.g., the sequences that are associated to a lineage), we define:

- *co-occurring pairs of mutations*: pairs of mutations that are observed in the same sequences of the reference population a number of times that is significantly *higher* than the expected one when the two mutations are independent from each other (i.e., frequency of first mutation × frequency of second mutation × size of the population);
- *mutually exclusive pairs of mutations*: pairs of mutations that are observed in the same sequences of the reference population a number of times that is significantly *lower* than the expected one when the two mutations are independent from each other.

Both co-occurring and mutually exclusive pairs are found at the two tails of the hypergeometric distribution of pairs of mutations found within the population. We employ as p-value for our selection the cumulative distribution function (cdf) of the hypergeometric distribution:

$$P(S_{m_1,m_2} = n) = \frac{\binom{S_{m_1}}{n}\binom{P-S_{m_1}}{S_{m_2}-n}}{\binom{P}{S_{m_2}}}$$

where S_{m_1}, S_{m_2} and S_{m_1,m_2} are the numbers of sequences harbouring mutations m_1, m_2, and both m_1 and m_2, respectively, while P is the size of the reference population. The p-value of observing N sequences with both mutation m_1 and mutation m_2 can be computed as:

$$P(S_{m_1,m_2} > N) = 1 - P(S_{m_1,m_2} \leq N) = 1 - \sum_{n=0}^{N} P(S_{m_1,m_2} = n)$$

We employ the Python `cdf` (Cumulative Distribution Function) function of the `scipy.stats.hypergeom` that, for a pair $\langle m_1, m_2 \rangle$ takes as input the number of sequences with both mutations (S_{m_1,m_2}, the total count of available sequences P, and the number of sequences with m_1 and with m_2 (S_{m_1} and S_{m_2}) in order to compute the probability $P(S_{m_1,m_2} \leq N)$. We compute p-values for all the possible pairs of frequent mutations (FM), that are $FM*(FM-1)/2$. Results are filtered by using suitable p-values (lower than 0.05 correspond to co-occurring pairs and higher than 0.95 correspond to mutually exclusive pairs) as shown in the *Data analysis* box of Figure 1.

2.2.2. Distribution of frequent mutations over lineages

We employed lineages as they are assigned by Pangolin [35]. For each mutation m of our dataset and each available Pangolin lineage l, we computed the count of sequences assigned to l that hold m, obtaining the triplet $\langle m, l, S_{m,l} \rangle$. Such data is collected within a 421 (frequent mutations) × 1,587 (all lineages) matrix. We refer to the *mutation m's distribution over lineages* as the row of the matrix

corresponding to m. To focus only on lineages that well-represent the mutation, for each row we set to zero the sequence counts that are less than 0.001 of the total lineage sequences. For each pair of co-occurring mutations extracted during the previous step, we compared the distributions over lineages of each of its two mutations; the comparison was performed by employing the Kolmogorov-Smirnov (KS) test [23] using the `scipy.stats.ks_2samp` Python method, which is efficient in determining if two samples are significantly different from each other. For a pair $\langle m_1, m_2 \rangle$ it takes as input the two arrays of m_1 and m_2's distributions over lineages and the 'two-sided' option. A reasonable p-value < 0.05 is used as a threshold of significance (see *Data analysis* box of Figure 1).

To support the understanding of these methods' steps, we exemplify them on a minimal example data structure. Assume we prepare a 3×3 matrix (Table 1), which collects the counts of sequences holding mutations m_0, m_1, and m_2 taken from sequences assigned to lineages L_1, L_2, and L_3. The total number of sequences of each lineages are given in parenthesis in the header.

	L_1 (10000)	L_2 (100)	L_3 (300)
m_0	4	10	100
m_1	4	10	110
m_2	300	10	100

Table 1: Minimal example matrix of counts of sequences assigned to lineages and holding mutations

As L_1 sequences with m_0 and with m_1 are less than 0.001*10000, then the corresponding counts are set to zero (see Table 2).

	L_1 (10000)	L_2 (100)	L_3 (300)
m_0	0	10	100
m_1	0	10	110
m_2	300	10	100

Table 2: Minimal example matrix from Table 1 where non-representative counts are set to zero

The first row of the matrix is the m_0's distribution over lineages. All possible mutation pairs are tested for co-occurrence with the hypergeometric test; let the following pairs be selected: $\langle m_0, m_1 \rangle$ and $\langle m_0, m_2 \rangle$. For each such pair, we compare their distributions over lineages. The KS test is employed to conclude that line m_0 is not significantly different from line m_1 and that line m_0 is significantly different from line m_2, yielding to the corresponding alternative steps in the workflow. Depending on the results of the KS test, each pair of co-occurring frequent mutations can be classified in one of the two categories:

- pairs with different distributions over lineages;
- pairs with similar distributions over lineages.

2.2.3. Convergent and divergent evolution testing method

We developed a method to identify convergence and divergence events involving two mutations at a time. The method was run for each pair of mutations extracted from the previous step, analyzing their presence within couples of lineages in 1-step relationship on the hierarchical lineages tree (i.e., each node with its direct ancestor). To support the explanation, we refer to Figure 2, showing a co-occurrence graph (panel A) where mutations (i.e., nodes) are exhibited by the same sequences (i.e., edges that connect co-occurring mutations). This graph is connected with a lineages tree (panel B), where lineages are nodes and edges explain their hierarchical relationships according to the phylogenesis. The mutation graph and the tree are connected when a mutation occurs above the threshold 0.001 of the sequences of a specific lineage.

When considering two mutations *co-occurring* across the dataset and having *different distributions over lineages*, we observe a *convergence event* when both the following conditions occur:

- for each lineage l, its sequences holding both mutations of a pair $\langle m_1, m_2 \rangle$ represent a large-enough fraction (> 0.01) of the sequences holding exclusively one mutation of the pair (i.e., $S_{m_1,m_2,l}/S_{m_1,l} > 0.01$ and $S_{m_1,m_2,l}/S_{m_2,l} > 0.01$);
- only m_1 or m_2 is found in the direct ancestor of l.

In panel B of Figure 2, the blue rectangle surrounds an event of convergence between mutations, in which only mutation B was found in the parent lineage `lin.1` whereas both mutations B and C were found together on the same sequences of the sub-lineage `lin.1.1`.

Several kinds of divergence events can occur, but here we focus on a particular definition. When considering two mutations *co-occurring* together across the dataset and having *similar distributions over lineages*, we observe a *divergence event* when both the following conditions occur:

- $S_{m_1,m_2,l}/S_{m_1,l} > 0.01$ and $S_{m_1,m_2,l}/S_{m_2,l} > 0.01$ (same condition as for the convergence event);
- only one of the two mutations is found in the direct descendant lineage of l.

As shown in panel B of Figure 2, the orange rectangle surrounds an event of divergence of mutations, in which both co-occurring mutations C and D were found together on the same sequences of the parent lineage `lin.2.2.2`, whereas its sub-lineages have exclusively mutation C or D.

Based on these definitions, we generate two tables containing: 1) the converging pairs of mutations ('Convergence Result' in the *Lineages distribution-dependent analysis* box of Figure 1); 2) the diverging pairs of mutations ('Divergence Result' in Figure 1).

For each convergence event (sketched in panel C of Figure 2), we find a corresponding row in the first table, showing: the *pair* of converging mutations, the Distinct Ancestor Lineage (*DAL*, i.e., the lineage with only one of the two pair mutations), the Common Descendant Lineage (*CDL*, i.e., the

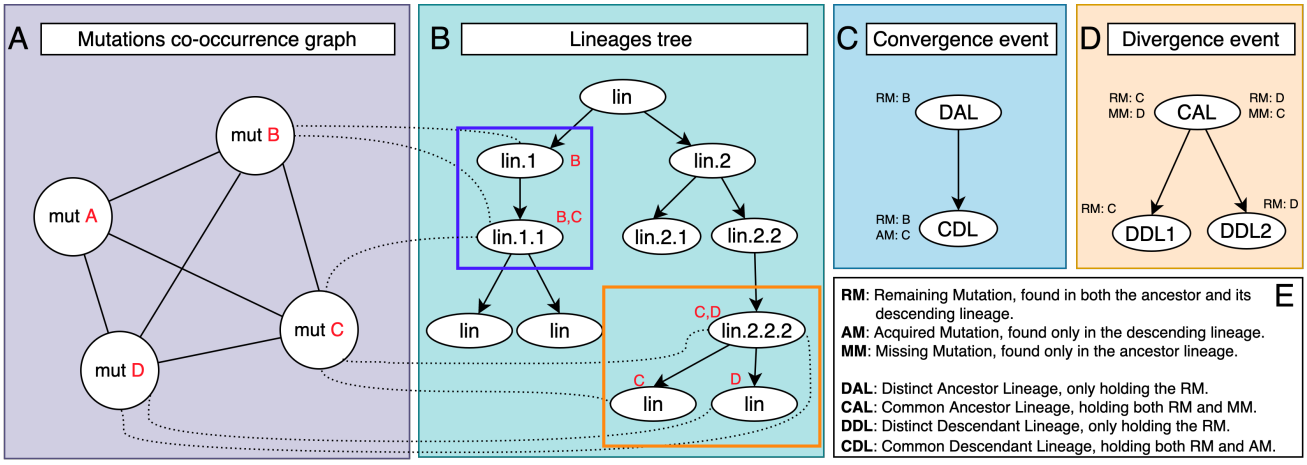

Figure 2: A. Frequent mutations graph whose nodes represent mutations and edges represent the co-occurrence of the two connected nodes within same sequences. B. Tree representing the phylogenetic hierarchy among lineages. C. Schematic representation of a convergence event. D. Schematic representation of a divergence event E. Terminology used in the methods.

sub-lineage with both pair mutations), the Remaining Mutation (*RM*, i.e., the mutation found both in the ancestor and its descendant lineage), the Acquired Mutation (*AM*, i.e., the mutation found only in the descendant lineage), the count of sequences assigned to *CDL* holding both *RM* and *AM*, the count of sequences assigned to *DAL* holding only *RM*, and the *depth* of *CDL* in the lineages tree.

For each divergence event (sketched in panel D of Figure 2), we find one row in the second table, showing: the *pair* of diverging mutations, the Common Ancestor Lineage (*CAL*, i.e. the lineage with both pair mutations), a Distinct Descendant Lineage (*DDL*, i.e., a sub-lineage with only one of the two pair mutations), the Remaining Mutation (*RM*, i.e., as before, the mutation found both in the ancestor and its descendant lineage), the Missing Mutation (*MM*, i.e., the mutation found only in the ancestor lineage), the count of sequences assigned to *CAL* holding both *RM* and *MM*, the count of sequences assigned to *DDL* holding only *RM*, and the *depth* of CAL in the lineages tree.

2.3. Lineages distribution-dependent analysis

To study the behavior of pairs of co-occurring mutations along the different lineages in the phylogenetic tree, two aggregated tables were generated from the convergence and divergence tables' results (shown in the *Lineages distribution-dependent analysis* box of Figure 1), by grouping according to the following fields:

- *Pairs of co-occurring mutations*. We grouped both 'Convergence Result' and 'Divergence Result' tables by the $\langle m_1, m_2 \rangle$ pair. Since a stronger evidence for converging mutations pairs (resp. diverging) is available when a pair shows more convergence (resp. divergence) events) in the lineages tree – also at different depths – the resulting aggregated tables were ranked by the count of *CDL* (convergent evolution) or by the count of *CAL* (divergent evolution).
- *Remaining mutation*. We grouped both 'Convergence Result' and 'Divergence Result' tables by *RM* and then ranked the aggregated tables by descending count of *DAL* (convergent evolution) and by descending count of *DDL* (divergent evolution).

Monte Carlo simulation. We used a Monte Carlo (MC) simulation approach to simulate the convergence and divergence tests and calculate the significance of our findings. By employing the Python method `sample` of the `random` library, we performed a simulation of 10,000 rounds, each composed as follows: 1) randomly select as many pairs of mutations as the ones obtained by the hypergeometric test out of the primary pool of unique pairs of mutations; 2) count how many distinct pairs of mutations pass the convergence test; 3) count how many distinct pairs of mutations pass the divergence test; 4) calculate the p-values by comparing the random distributions generated according to 2) and 3) and the real data.

Sub-trees visualization. Graphviz [14] and NetworkX [19] were used to draw lineages trees. Tree-shaped graphs are generated from the aggregated tables representing pairs of co-occurring mutations. One graph is produced for each converging or diverging pair of mutations, following the relationships between lineages as indicated by the Pangolin nomenclature. Each node represents a lineage and lineages are connected by hierarchical relationships (arrows). We start from the original SARS-CoV-2 haplotypes (A or B) and descend the tree up to the node where an evolution event is identified. Since several pairs have a high number of evolution events, we omit the visualization of branches without such events. A color code is used to highlight where convergent/divergent evolution events are identified and which mutations from the considered pair are observed.

2.4. Lineages distribution-independent analysis

Following the hypergeometric test that identifies co-occurring and mutually exclusive mutation pairs, we also perform a lineage-independent analysis, which is not related to the analysis reported in Section 2.3, as it does not consider how mutations are spread across the lineages. More precisely, we employ co-occurring pairs to investigate the defining mutations of lineages. Instead, we employ mutually exclusive pairs to gain insights on the mutations that are preferred during the virus evolution. To further study these two happenings, we incorporate the calculation of fold enrichment as shown in the *Lineages distribution-independent analysis* box of Figure 1.

Fold enrichment calculation. The Fold Enrichment (*FE*) is a general statistical term that indicates how many folds a phenomenon happened more (or less) than expected by random chance. For instance, *FE=3* means than the event happened three times the random expectation for that event. We use the following formula to calculate it:

$$logFE_{m,l} = log_2(S_{m,l} * (L/S_m))$$

where we have that $S_{m,l}$ is the total number of sequences having a mutation m in a given lineage l; L is the total number of lineages found in the population; S_m is the total number of sequences having mutation m in the population. Here we considered $S_{m,l}$ as the count of observed events and S_m/L as the count of expected events. By using $logFE_{m,l}$, the values of enrichment range from $-\infty$ to $+\infty$, where negative $logFE_{m,l}$ values indicate that the mutation m is less enriched in lineage l and positive values indicate that the mutation m is more enriched in lineage l, while zero means the enrichment is as expected by random chance. Details on how this measure is employed are given in the results (Section 3.2.1), where noteworthy variants are analyzed.

3. Results

3.1. Dataset description

By analyzing the considered dataset (see Section 2.1 in the Methods), we observed the distributions of the sequences over continents and lineages, provided respectively in Figure 3A and Figure 3B. Half of the viral sequences of the analysed dataset were collected in the European continent. Sequences were assigned to 1,587 distinct Pango lineages, the most represented lineages being B.1.1.7 (Alpha), AY.4, and BA.1 (Omicron), respectively represening the 13.8%, 10.1%, and 9.2% of the total population. According to the WHO, the current variants of concern (VOCs) are B.1.617.2 (Delta) and Omicron (including all BA.1, BA.2, BA.3, and BA.1.1), while previously circulating VOCs were B.1.1.7 (Alpha), B.1.351 (Beta), and P.1 (Gamma). Figure 3C presents the counts of sequences from the dataset assigned to distinct lineages that are currently or were previously considered as VOCs; these are distributed by collection date for the whole course of the pandemic.

Almost all the sequences – except for five – exhibit at least one non-synonymous mutation with a total number of 156,951 non-synonymous mutations (substitutions, deletions, or insertions) found in the population with an average of 32.6 mutations per sequence. The most dominant non-synonymous mutations are the substitution D614G in the spike protein, the substitution P323L in the non-structural protein 12, and the substitution T478K in the spike protein, found in 97%, 96.5%, and 63.8% of the population, respectively.

3.2. Co-occurring and mutually exclusive mutation pairs

A total of 88,410 unique pairs were generated from the list of 421 frequent mutations. Using the hypergeometric test (see Methods, Section 2.2.1), we extracted 16,692 co-occurring pairs by using p-value < 0.05; note that, using a smaller p-value < 0.01 produces 16,415 co-occurring pairs, thus a small decrease. Then, we extracted 69,903 mutually exclusive pairs by using a p-value > 0.95; note that, out of these, 62,491 have p-value $= 1$, with 374 pairs never appearing together in the same sequence in the entire dataset. Differently from Zhang et al. [52] (where mutations are considered as "high frequency" if found in at least 1% of sequences), we have used a much lower frequency threshold (found in more than 20K sequences, i.e. 0.25%). Consequently, when performing the hypergeometric test on pairs of mutations, this choice has allowed to widen the possibility of co-occurrence/mutual exclusion detection.

3.2.1. Lineage distribution-independent results

Before proceeding with our main analysis thread, which evaluates mutations w.r.t. their distributions over lineages, we derive a series of interesting observations that can be made just on the basis of the results of the hypergeometric test. Such results are derived using the methods described in Section 2.4. First, by exploiting the extracted co-occurring pairs, we show how they could be used to complement lists of variants' defining mutations. Second, by exploiting the extracted mutually exclusive pairs, we show how the pairs could be used as insights of the natural evolution of the virus. Results are explained using notable examples.

Defining mutations of a given lineage: the case of the Delta variant. The B.1.617.2 variant (Delta) has been considered for a long time as one of the VOCs by WHO. Almost all (19) of its *defining* mutations (according to CoVariants [20]) are found in our list of frequent mutations (specifically, only 2 out of 21 mutations are missing, but they are also absent in the full dataset). Table 3 presents the p-values derived from running a hypergeometric test on each pair formed by Delta defining mutations in the Spike protein. All pairs have a significant p-value, except for the pair composed by Spike_D614G and Spike_E156- (p-value = 0.54) (likely due to the huge difference in the populations of these two mutations). In general, these p-values suggest that the pairs of mutations tend to co-occur together. Out of all the pairs $\langle m_1, m_2 \rangle$ resulting from the hypergeometric test, we con-

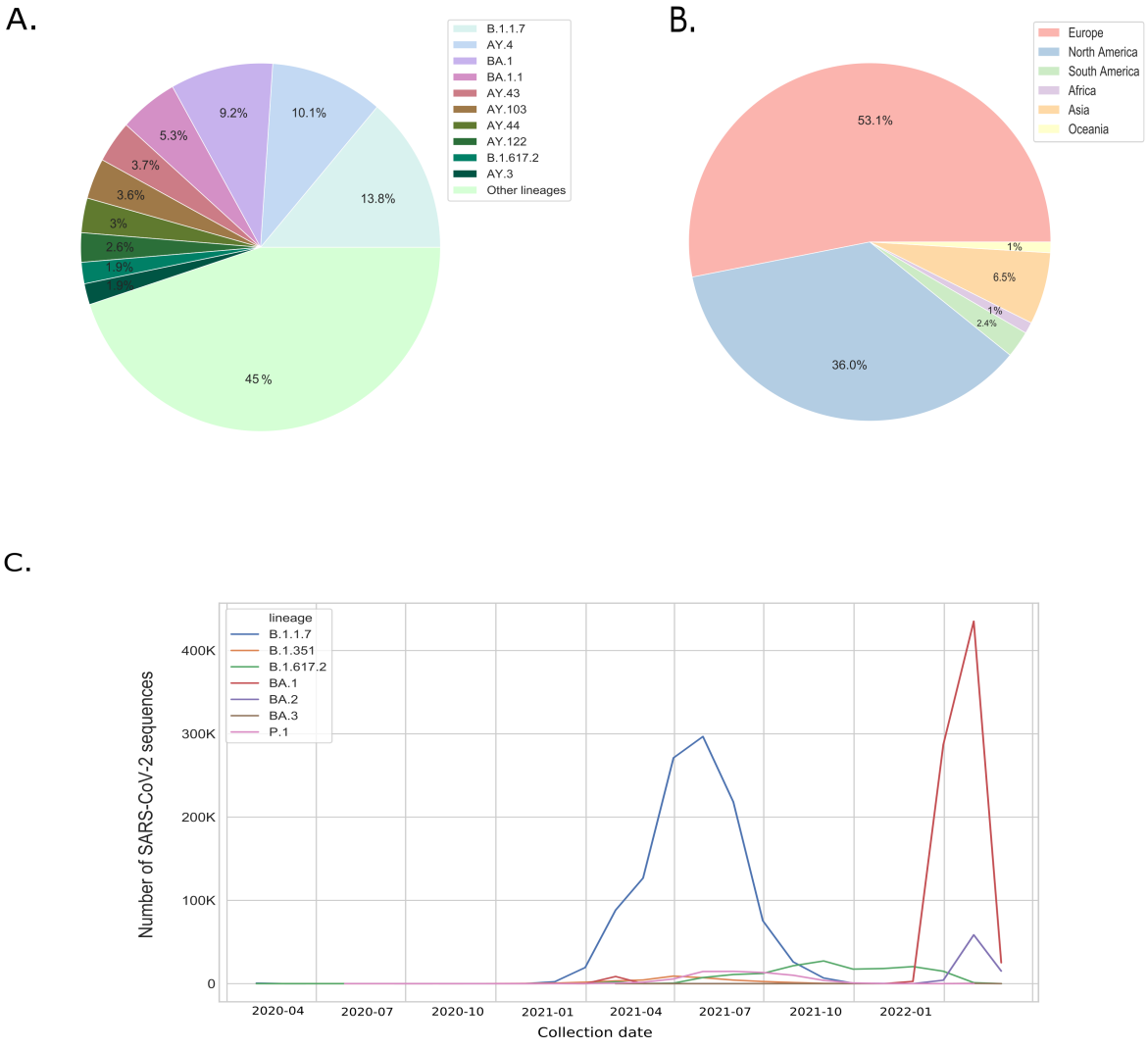

Figure 3: A. Sequences' distribution across lineages, detailing the ten most representative lineages. B. Sequences' distribution across all the continents. C. Number of sequences assigned to the lineages currently or previously considered as VOCs according to the WHO with their collection dates.

sidered the ones where m_1 belongs to Delta defining mutations whereas m_2 does not. Only mutations m_2 with a *logFE* above a threshold of 1 (i.e., mutations that have been found in Delta's population at least twice more than expected) were further considered. We identified other 68 mutations that co-occur with all Delta defining mutations and that *logFE* values above our threshold. Three of these are found in the Spike protein: Spike_V1104L (*FE*: 44.87, *logFE*: 5.48), Spike_G142D (*FE*: 41.06, *logFE*: 5.35), and Spike_S112L (*FE*: 13.75, *logFE*: 3.78); the complete set is found in Supplementary Table S1. Our findings suggest that the identified mutations could be considered as additional defining mutations of the Delta variant, complementing the list provided by CoVariants [20]. Properties of significant mutations co-occurring with Delta defining mutations are shown in the scatter plots drawn in the panels of Figure 4.

Mutually exclusive mutations: the case of the Alpha and Omicron lineages. The variants B.1.1.7 (Alpha) and BA.1 (Omicron) have a very high number of common defining mutations. Among their defining mutations lists, we identified possible mutually exclusive pairs; see Table 4, where pairs with p-value = 1 and mutations from these two variants are reported. We focus on the first two rows where the number of sequences having both mutation equals to zero: ⟨Spike_S371L, Spike_A570D⟩ and ⟨Spike_S371L, NSP3_A890D⟩. Spike_S371L is one of the defining mutations of BA.1, while both Spike_A570D and NSP3_A890D are from the defining mutations of the lineage B.1.1.7 (Alpha). Indeed, Spike_S371L has a *logFE* of 9.91 in Omicron and $-\infty$ in Alpha, whereas Spike_A570D and NSP3_A890D both have *logFE* of $-\infty$ in Omicron and of 10.58 in Alpha.

This observation suggests that the more recent circulat-

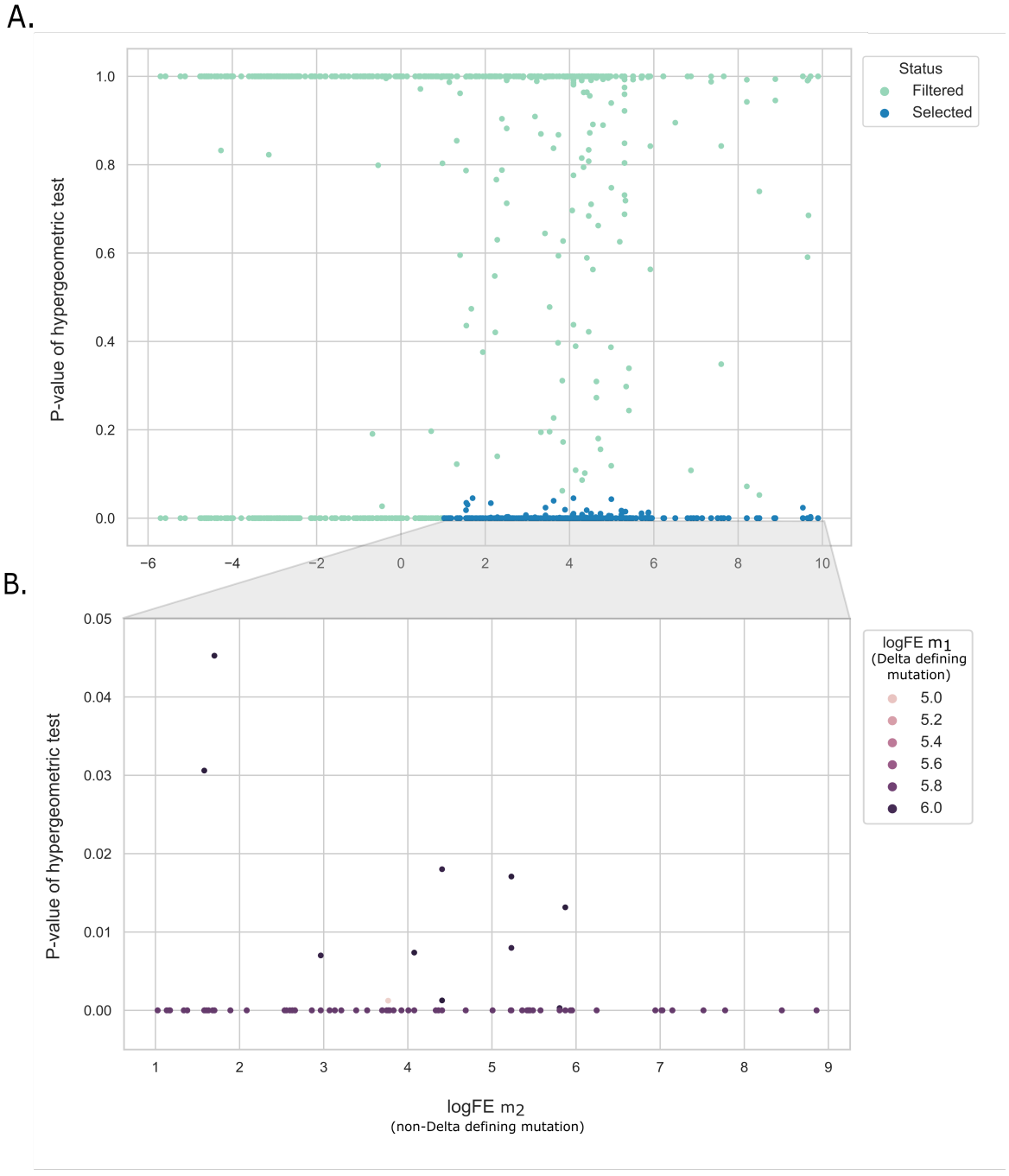

Figure 4: A. 2D Scatter plot of the p-values of the hypergeometric tests on all the pairs of mutations in which only one of the two is a defining mutation of Delta variant, mapped on the values of *logFE* of the mutation of the pair that is not a Delta-defining mutation (outside of the list). A total of 7,638 pairs of mutations were considered in the figure after removing mutations with FE=0. Blue dots indicate the p-values < 0.05, whose corresponding non-Delta mutations have a *logFE*>1. B. Zoomed version of Panel A scatter plot; it includes 68 mutations extracted from the 'blue' pairs selected in Panel A (co-occurring with all 19 Delta-defining mutations); here, a color scale is used to indicate the *logFE* of the pair mutation that belongs to the Delta variant: the darker the color, the higher the value. The corresponding 3D scatter plot is provided in Supplementary Figure S1.

Table 3
Examples of co-occurring pairs of mutations in the Spike protein, extracted from the defining mutations list of the Delta variant [20]. The provided number of sequences is evaluated on the complete dataset.

m_1	m_2	#Seq. with m_1	#Seq. with m_2	#Seq. with m_1, m_2	P-value
T19R	E156-	4145223	21528	20285	4.72e-09
T19R	F157-	4145223	3867174	3833412	0.0
T19R	R158G	4145223	24435	23862	0.0
T19R	L452R	4145223	4199166	4034866	0.0
T19R	T478K	4145223	5252904	4046410	3.74e-09
T19R	D614G	4145223	7981457	4134564	1.04e-08
T19R	P681R	4145223	4208095	4123831	3.28e-09
T19R	D950N	4145223	4033822	3960828	0.0
E156-	F157-	21528	3867174	20395	0.0
E156-	R158G	21528	24435	18717	0.0
E156-	L452R	21528	4199166	17429	1.01e-09
E156-	T478K	21528	5252904	17870	7.69e-09
E156-	D614G	21528	7981457	21339	0.5496
E156-	P681R	21528	4208095	20315	5.83e-9
E156-	D950N	21528	4033822	19943	2.06e-09
F157-	R158G	3867174	24435	17664	0.0
F157-	L452R	3867174	4199166	3785418	0.0
F157-	T478K	3867174	5252904	3795616	0.0
F157-	D614G	3867174	7981457	3857745	7.53e-09
F157-	P681R	3867174	4208095	3849316	5.77e-10
F157-	D950N	3867174	4033822	3724976	0.0
R158G	L452R	24435	4199166	20816	0.0
R158G	T478K	24435	5252904	21209	0.0
R158G	D614G	24435	7981457	24247	0.0418
R158G	P681R	24435	4208095	23830	0.0
R158G	D950N	24435	4033822	22455	0.0
L452R	T478K	4199166	5252904	4080758	1.02e-09
L452R	D614G	4199166	7981457	4189517	4.32e-09
L452R	P681R	4199166	4208095	4078817	0.0
L452R	D950N	4199166	4033822	3907159	9.18e-10
T478K	D614G	5252904	7981457	5241411	8.97e-09
T478K	P681R	5252904	4208095	4082650	4.22e-09
T478K	D950N	5252904	4033822	3919358	2.71e-10
D614G	P681R	7981457	4208095	4197588	4.37e-09
D614G	D950N	7981457	4033822	4025217	5.00e-09
P681R	D950N	4208095	4033822	4001531	5.86e-09

Table 4
Examples of mutually exclusive pair of mutations. m_1 are extracted from the defining mutations' list of the Omicron variant; m_2 are defining mutations of the Alpha variant.

m_1	m_2	#Seq. with m_1	#Seq. with m_2	#Seq. with m_1, m_2	P-value
Spike_S371L	Spike_A570D	1012982	1161124	0	1.0
Spike_S371L	NSP3_A890D	1012982	1159088	0	1.0
Spike_S371L	Spike_D1118H	1012982	1153130	20	1.0
Spike_Y505H	Spike_A570D	1115034	1161124	109	1.0
Spike_Y505H	NSP3_A890D	1115034	1159088	110	1.0
Spike_Y505H	Spike_D1118H	1115034	1153130	131	1.0
Spike_T547K	Spike_A570D	1184888	1161124	116	1.0
Spike_T547K	NSP3_A890D	1184888	1159088	116	1.0
Spike_T547K	Spike_D1118H	1184888	1153130	140	1.0

ing lineage, i.e. Omicron, tends to favor Spike_S371L over Spike_A570D or NSP3_A890D. Note that BA.1 (Omicron) and B.1.1.7 (Alpha) are sharing many defining mutations (e.g., H69-, V70-, Y144-, D614G, N501Y, and P681H in Spike_protein and other mutations in other proteins) – almost half of the defining mutations of Alpha are considered defining mutations also of Omicron – and that they are not closely related lineages. This may suggest that the virus could be evolving into the direction of collecting new mutations that might enhance its features, e.g., Omicron is now the most complete 'escapee' from neutralization by currently available antibodies in comparison to other SARS-CoV-2 variants, including the Alpha variant [27].

3.3. Lineages distribution analysis for co-occurring mutations pairs

Following the methods described in Section 2.2.2, we were able to study the quantitative behavior of pair mutations over lineages. We extracted distributions over lineages for each frequent mutation. Figure 5 presents the number of sequences and lineages of each frequent mutation considered for the Spike protein. As expected, the most dominant mutation is Spike_D614G, found in 1,534 lineages out of 1,587 total lineages. Then, we compared each such distributions in pairs, using the KS test. As a result, we obtained 4,000 pairs with different distributions over lineages and 12,692 pairs with similar distributions over lineages. A 3D scatter plot of the KS test results for all the 421 frequent mutations is provided in Supplementary Figure S2.

Based on the procedure described in Section 2.2.3, from the group of pairs with different lineages distributions, we identified 4,489 distinct events of convergent evolution, whereas from the group of pairs with similar lineages distributions, we identified 415,892 distinct events of divergent evolution.

Along the Methods described in Section 2.3, we then produced the tables aggregated by mutation pairs and remaining mutations, which are analyzed in the next section.

3.4. Convergent and divergent evolution of pairs of mutations

3.4.1. Convergence

By using pairs of mutations as a grouping factor in the 'Convergence Result' table, we generated a table with 1,818 unique pairs of co-occurring mutations having at least one event of convergent evolution. Table 5 presents the pairs of mutations with the highest count of convergence events; the complete result is provided in Supplementary Table S2.

Example 1. The first line of the table shows that the co-occurring pair composed by NSP1_V84- and NSP1_V86- is found separately in 8 Distinct Ancestor Lineages (#DAL) and appears in 24 of their Descendant Lineages (#CDL), occurring at various depths of the lineages tree (between the third and the sixth level). In all such cases, the Remaining Mutation is, alternatively, the first or the second mutation of the pair. The lineages tree related to this pair is shown in Supplementary Figure S3.

Example 2. In addition to several co-occurring (and closely positioned) deletions, in the seventh row of Table 5 we can observe a pair composed by NSP4_V167L and Spike_P681R, which are present separately in 4 lineages (#DAL) and appear together in 17 of their descending lineages (#CDL), spotted at both the third and fifth level of the tree. Figure 6 shows the representative lineages tree of this pair. Note that having found a pattern of converging mutations in such a high number of lineages suggests the existence of selection advantages for such mutations.

We deepen our analysis by using remaining mutations (*RM*) as a grouping factor in the 'Convergence Result' table. We generate a table of 308 converging mutations. Table 6

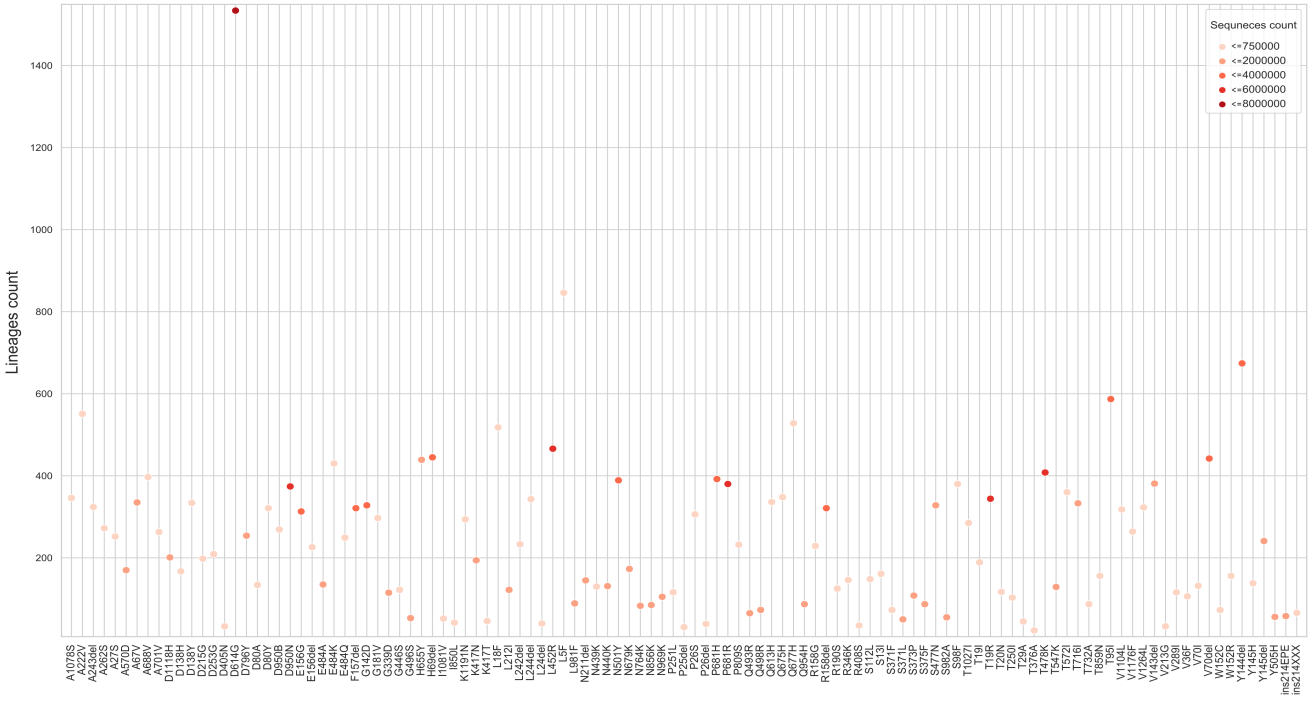

Figure 5: Scatter plot indicating the number of lineages containing the frequent mutations of the Spike protein; the higher the number, the more spread a mutation is over the lineages. A color scale is used to express the number of sequence exhibiting each mutation.

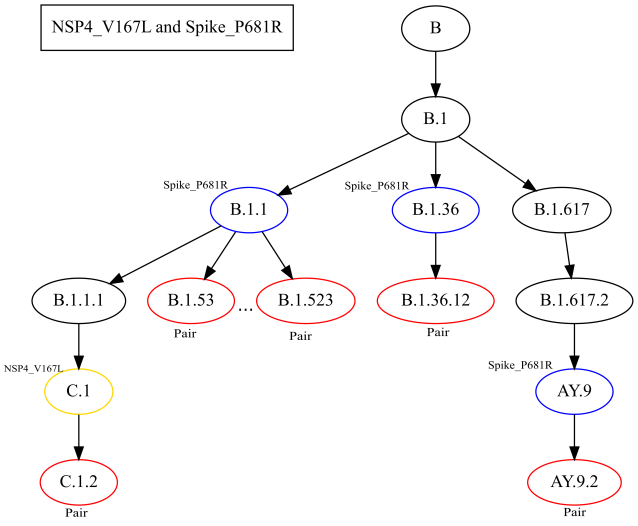

Figure 6: Tree-based representation of lineages involved in the evolution of NSP4_V167L and Spike_P681R. Each node represents one lineage and the arrow between two lineages draws the phylogenetic relation between an ancestor lineage and its descendant lineage. For ease of visualization, we show only part of the tree (5 convergence events out of 17 events detected by the pair). Colors are used to indicate which mutation is present in the indicated lineage: *black* when both mutations are present, *blue* when only Spike_P681R is present, and *yellow* when only NSP4_V167L is present.

shows the converging mutations participating to the highest number of convergent evolution events across the whole lineages tree; the complete list is provided in Supplementary Table S4. In Table 6 we find the three consecutive deletions occurring on the non-structural protein 6 (NSP6) at positions 106–108. These three mutations are co-occurring and converging in 28 different lineages, suggesting that the deletion of 9 nucleotides in ORF1ab gene that generates the 'SGF

Table 5
Top 10 converging pairs of mutations ranked by descending #CDL.

$\langle m_1, m_2 \rangle$	#CDL	CDL depth	#DAL	#RM
NSP1_V84-+NSP1_V86-	24	3,4,5,6	8	NSP1_V84-,NSP1_V86-
NSP1_H83-+NSP1_V86-	23	3,4,5	7	NSP1_H83-,NSP1_V86-
NSP1_G82-+NSP1_V86-	23	3,4,5	7	NSP1_G82-,NSP1_V86-
NSP6_F108-+NSP6_S106-	18	3,4	3	NSP6_S106-
NS8_P93S+NSP3_V932A	18	4	1	NSP3_V932A
NSP6_F108-+NSP6_G107-	18	3,4	3	NSP6_G107-
NSP4_V167L+Spike_P681R	17	3,5	4	NSP4_V167L,Spike_P681R
NSP6_L260F+NSP6_S106-	15	2,3,4,5	4	NSP6_S106-
N_G204R+N_R203K	15	1,3	4	N_R203K,N_G204R
NSP6_G107-+NSP6_L260F	14	2,3,4,5	4	NSP6_G107-

Table 6
Top 10 remaining mutations ranked by descending #DAL (counting converging events).

RM	#CDL	#DAL	#AM
NSP6_G107-	53	17	24
NSP6_S106-	54	17	24
Spike_N501Y	24	14	14
NS8_R52I	30	12	15
Spike_L452R	20	12	9
Spike_P681H	17	12	17
NSP12_F694Y	15	11	12
NS8_Q27stop	28	11	15
NSP6_F108-	35	11	28
N_T205I	23	10	18

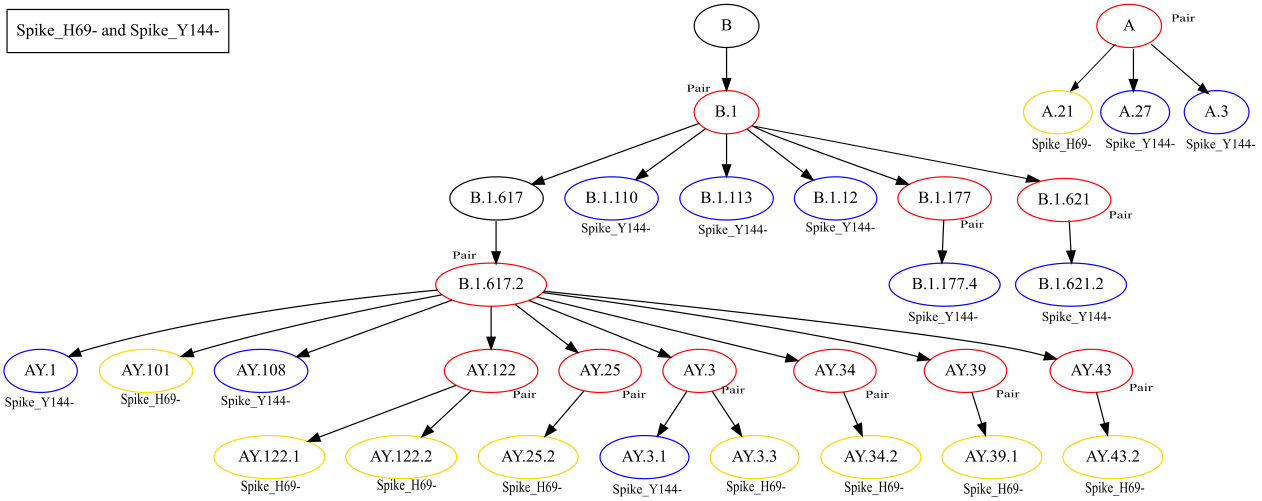

Figure 7: Tree-based representation of lineages involved in the evolution of Spike_H69- and Spike_Y144-. For ease of visualization, here we present a portion of the original tree, with 19 out of 307 total divergent events detected for the pair of deletions.

Table 7
Top 10 diverging pairs of mutations ranked by descending #CAL.

$\langle m_1, m_2 \rangle$	#CAL	CAL depth	#DDL	#RM
Spike_D614G+Spike_L5F	22	1,2,3,4	584	Spike_D614G,Spike_L5F
NSP1_H83-+NSP1_M85-	21	1,2,3,4,5	236	NSP1_H83-,NSP1_M85-
NSP1_G82-+NSP1_M85-	21	1,2,3,4,5	235	NSP1_G82-,NSP1_M85-
NSP12_P323L+Spike_L5F	21	1,2,3,4	587	NSP12_P323L,Spike_L5F
NSP1_M85-+NSP1_V84-	20	1,2,3,4,5	202	NSP1_M85-,NSP1_V84-
NSP1_M85-+NSP1_V86-	19	1,2,3,4,5	124	NSP1_M85-,NSP1_V86-
Spike_H69-+Spike_Y144-	17	1,2,3,4	307	Spike_H69-,Spike_Y144-
Spike_V70-+Spike_Y144-	17	1,2,3,4	307	Spike_V70-,Spike_Y144-
NSP5_K90R+Spike_D614G	17	1,2,3,4,5	406	Spike_D614G,NSP5_K90R
NSP16_K160R+Spike_D614G	15	2,3,4	63	Spike_D614G

deletion' is among the most prevalent remaining mutations in convergent evolution events in the population. Note that this triple amino acid deletions is included in the defining mutations lists of three previous VOCs, i.e., Alpha, Beta, and Gamma. NSP6 is a multi-pass transmembrane protein that is thought to be involved in autophagy and antagonism of innate immune responses, but it remains unclear what influence this deletions has on virus phenotype [49, 37]. Other mutations in Table 6 are N501Y, L452R, and P681H in Spike. These mutations started to converge from the second wave of COVID-19 and have been reported in many globally circulating lineages. They are considered as defining mutations of different lineages considered as VOCs, namely N501Y is a defining mutation of Alpha, Beta, Gamma, and Omicron; L452R is a defining mutation of Delta; and P681H is a defining mutation of Alpha and Omicron. Moreover, N501Y may increase the binding affinity to ACE2 [3] and affect the immune response to possible vaccines and treatments [4, 48]; L452R is one of the RBD mutations that possibly enhance the binding affinity to ACE2 receptor and reduce the binding affinity of many antibodies [51]; finally, P681H enhances the furin binding and viral infectivity [33].

Table 8
Top 10 remaining mutations ranked by descending #DDL (counting diverging events).

RM	#CAL	#DDL	#MM
Spike_D614G	66	1416	244
NSP12_P323L	62	1389	221
NSP6_L37F	47	873	88
Spike_L5F	28	685	58
NS3_Q57H	22	594	50
N_R203K	22	579	91
N_G204R	21	567	85
NSP1_M85-	36	551	115
Spike_Y144-	27	519	87
NSP16_K160R	24	452	47

3.4.2. Divergence

By using pairs of mutations as a grouping factor on the 'Divergence Result' table, we generated a table of 6,625 unique pairs of co-occurring mutations with at least one possible event of divergent evolution. Table 7 presents the pairs of mutations with the highest count of divergence events; the complete table is provided as Supplementary Table S3.

Example. Coronaviruses, including SARS-CoV-2, have lower substitution rates than other RNA viruses because of an RdRp with proofreading activity [13, 32]. However, proofreading cannot correct deletions. Even if deletions in the considered frequent mutation list are only 31 (7.36% of the total), we found that in 18.66% of the observed divergence events the remaining mutation (*RM*) was a deletion. This might be explained by the fact that, unlike substitutions, deletions cannot be corrected by proofreading activity [30]. Notable examples of pairs of deletions with divergent evolution through the lineages tree are ⟨Spike_H69-, Spike_Y144-⟩ and, similarly, ⟨Spike_V70-, Spike_Y144-⟩; see the seventh and eigth rows of Table 7. Spike_H69- and Spike_H70- are co-occurring mutations having very simi-

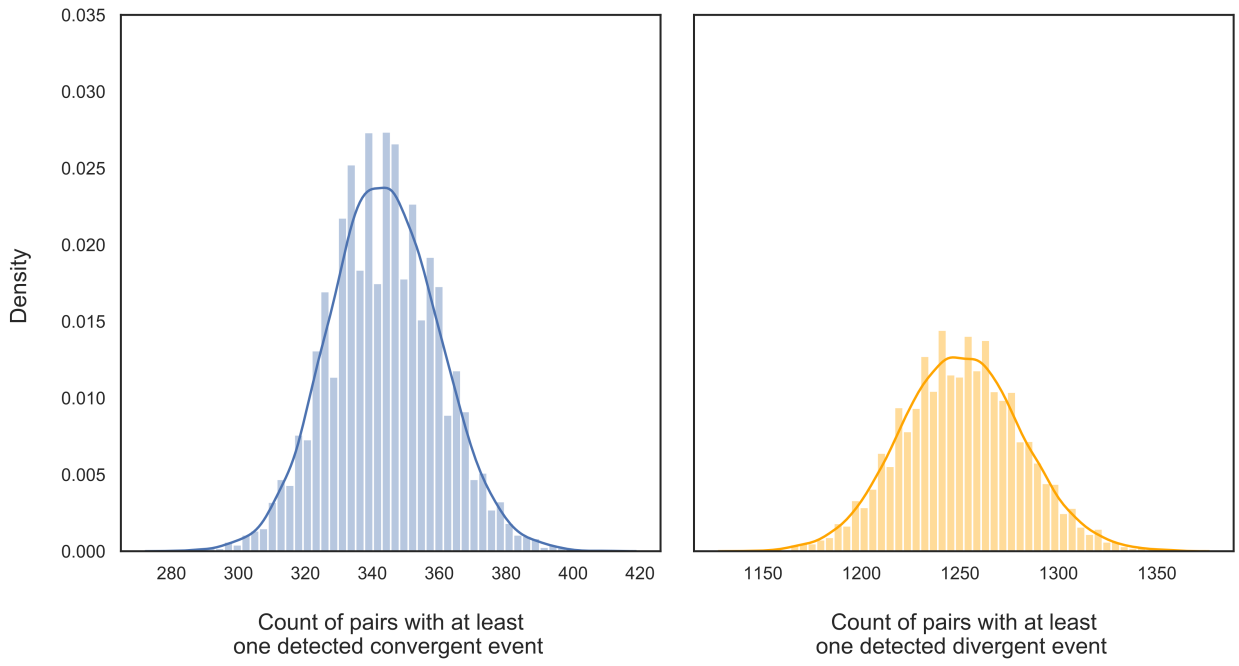

Figure 8: Distributions of the counts of pairs with randomly detected convergent (left) or divergent (right) events, extracted from a sample of 16,692 pairs of mutations repeated for 10,000 times.

lar lineages distribution according to the KS test (p-value = 1.0); they are spotted together in 439 different lineages (forming respectively 98% and 99% of the lineages they appear in). Therefore, we discuss about this double deletion as if it were a unique mutation and compare its pattern of distribution with the one of Spike_Y144-. We detected 304 divergent events, whose 86.18% includes Spike_Y144- as the diverging mutation (*RM*) passing to the new descendant lineage, while only 13.81% includes Spike_H69-/V70- as the remaining mutations. Both Spike_H69-/V70- and Spike_Y144- are deletions that lie in the NTD region of Spike and may modulate antigenicity [25, 29, 30]. Since the count of DDL is high (307), the generated tree is very large, thus Figure 7 shows only a portion of it.

We deepen our analysis by using remaining mutations (*RM*) as a grouping factor in the 'Divergence Result' table; we generate a table of 362 unique diverging mutations. Table 8 shows the remaining mutations participating to the highest number of divergent evolution events across the whole lineages tree; the complete list is provided in Supplementary Table S5. The table shows two mutations that were expected, i.e., Spike_D614G and NSP13_P323L, as they are the most dominant mutations across the population. With the exception of Spike_L5F, the other seven top mutations are well known mutations that were studied since almost the beginning of the pandemic, having a major role in forming the five well-known distinct clades [36]. This may explain why we detected a considerably high numbers of divergence events, compared to the convergence ones.

3.4.3. Validation

To assess the significance of our findings, we performed a Monte Carlo simulation to compare the resulting numbers of pairs showing convergence/divergence events with numbers of randomly selected pairs. A p-value $< 10e\text{-}4$ in both cases of convergent and divergent evolution was detected. Figure 8 shows the distribution of the counts of pairs of mutations with at least one convergent or divergence event detected following the Monte Carlo approach; both distributions are centered on mean values that are considerably distant from the observed values (1,818 for convergence events and 6,625 for divergence events).

4. Discussion

The SARS-CoV-2 pandemic is a major threat to the public health. In response to the continuous spreading of the virus the global community answered with an incredible effort to collect and deposit a huge amount of viral genomes to public repositories. Thanks to such availability, we were able to conduct a large-scale analysis that aimed at highlighting the role of non-synonymous mutations' pairs in identifying evolution events of SARS-CoV-2.

First, we analyzed the patterns of co-occurrence and mutual exclusion of pairs of mutations on sequences of the virus. We then focused on sets of significantly co-occurring pairs of mutations by analyzing how their distributions over lineages compare. Finally, we precisely described events of convergence (when two mutations are frequent in a lineage but only one of them is frequent in its ancestor) and of divergence (when two mutations are frequent in a lineage but

one of the two disappears in a sub-lineage). Based on this notion, we observed that co-occurring pairs with different distributions allow to identify convergence events, while co-occurring pairs with similar distributions allow to identify divergence events. The obtained results were grouped by 1) considering for each pair of mutations the number of lineages where it instantiates a converging/diverging behavior and 2) considering for each remaining mutation (i.e., maintained through two directly-related lineages) the number of lineages where it participated to a converting/diverging behavior.

The essence of this work can be summarized as follows. The lineage-independent analysis, which included a study of the most significant co-occurring mutation pairs, complemented by the fold enrichment calculation, suggested a way to find candidates for "variants' defining mutations" that are not currently listed by reference sources (e.g., CoVariants)—as shown for the Delta variant case, where we highlighted the mutations S112L, G142D, and V1104L in the spike protein. Then, considering the mutually exclusive mutation pairs assigned to two different variants, we regard them as candidates for characterizing different variant phenotypes. Among them, we spotted Spike_S371L in Omicron which never occurs together with Spike_A570D or NSP3_A890D in Alpha.

Convergent and divergent events can instead be used for anticipating lineage evolution. Specifically, if two mutations are co-occurring – one remaining (*RM*) and one acquired (*AM*) – whenever *RM* appears in a lineage then we expect that *AM* will be next acquired; in our analysis, we spotted several confirmations, e.g., Figure 6 represents 5 (out of 17) converging events having either NSP4_V167L or Spike_P681R as *RM* in a parent node and both mutations in one of its descendant nodes. In convergence events, we highlight the presence of remaining mutations that are widely spread and defining several VOCs (see Table 6).

Conversely, if two mutations are mutually exclusive – one remaining (*RM*) and one missing (*MM*) – whenever *RM* and *MM* appear in a lineage, then we expect that *MM* will not be present in some descendant node; in our analysis, we spotted several confirmations, e.g., Figure 7 represents 19 (out of 307) diverging events having both Spike_H69- or Spike_Y144- as *RM* in a parent not and only one of them in its descendant nodes. In divergence events, we highlight the presence of remaining mutations that are frequent in the dataset but not specifically defining notable variants (see Table 8).

The analysis elements proposed in this study pose the basis for anticipating the evolution of the SARS-CoV-2 virus, by observing mutation and lineages behaviors from a purely data-driven quantitative point of view.

CRediT authorship contribution statement. **Ruba Al Khalaf**: Data elaboration and analysis, Investigation, Visualization, Validation, Writing – original draft. **Anna Bernasconi**: Study conception, Design of methods workflow, Supervision, Writing – original draft, review & editing. **Pietro Pinoli**: Study conception, Design of statistical analysis, Supervision, Writing – review & editing. **Stefano Ceri**: Supervision, Project principal investigator, Writing – review & editing.

Declaration of Competing Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements. The authors of the manuscript gratefully acknowledge all data contributors, i.e. the Authors and their Originating Laboratories responsible for obtaining the specimens, and their Submitting Laboratories that generated the genetic sequence and metadata and shared via the GISAID Initiative the data on which this research is based.

Funding. This work was supported by the ERC Advanced Grant number 693174 GeCo (data-driven Genomic Computing).

Supplementary material. Supplementary materials associated with this article include Figures S1–S3 and Tables S1–S7 and attached to the manuscript.

References

[1] Al Khalaf, R., Alfonsi, T., Ceri, S., Bernasconi, A., 2021. CoV2K: a knowledge base of SARS-CoV-2 variant impacts, in: International Conference on Research Challenges in Information Science, Springer. pp. 274–282.

[2] Alfonsi, T., Al Khalaf, R., Ceri, S., Bernasconi, A., 2022. CoV2K model, a comprehensive representation of SARS-CoV-2 knowledge and data interplay. Scientific Data 9, 1–12.

[3] Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. Medicine in Drug Discovery 10, 100086.

[4] Andreano, E., Rappuoli, R., 2021. SARS-CoV-2 escaped natural immunity, raising questions about vaccines and therapies. Nature medicine 27, 759–761.

[5] Bernasconi, A., Mari, L., Casagrandi, R., Ceri, S., 2021. Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. Scientific Reports 11, 1–10.

[6] Biswas, N., Mallick, P., Maity, S.K., Bhowmik, D., Mitra, A.G., Saha, S., Roy, A., Chakrabarti, P., Paul, S., Chakrabarti, S., 2021. Genomic Surveillance and Phylodynamic Analyses Reveal the Emergence of Novel Mutations and Co-mutation Patterns Within SARS-CoV-2 Variants Prevalent in India. Frontiers in Microbiology 12.

[7] Chen, J., Wang, R., Wang, M., Wei, G.W., 2020. Mutations strengthened SARS-CoV-2 infectivity. Journal of molecular biology 432, 5212–5226.

[8] Chiara, M., Horner, D.S., Gissi, C., Pesole, G., 2021. Comparative genomics reveals early emergence and biased spatiotemporal distribution of SARS-CoV-2. Molecular Biology and Evolution 38, 2547–2565.

[9] Bollen, N., Artesi, M., Durkin, K., Hong, S.L., Potter, B., Boujemla, B., Vanmechelen, B., Martí-Carreras, J., Wawina-Bokalanga, T., Meex, C., et al., 2021. Exploiting genomic surveillance to map the spatio-temporal dispersal of SARS-CoV-2 spike mutations in Belgium across 2020. Scientific reports 11, 1–8.

[10] Ko, K., Nagashima, S., E, B., Ouoba, S., Akita, T., Sugiyama, A., Ohisa, M., Sakaguchi, T., Tahara, H., Ohge, H., et al., 2021. Molecular characterization and the mutation pattern of SARS-CoV-2 during first and second wave outbreaks in Hiroshima, Japan. PLoS One 16, e0246383.

[11] Negi, S.S., Schein, C.H., Braun, W., 2022. Regional and temporal

coordinated mutation patterns in SARS-CoV-2 spike protein revealed by a clustering and network analysis. Scientific reports 12, 1–10.

[12] Ostrov, D.A., Knox, G.W., 2022. Emerging mutation patterns in SARS-CoV-2 variants. Biochemical and biophysical research communications 586, 87–92.

[13] Denison, M.R., Graham, R.L., Donaldson, E.F., Eckerle, L.D., Baric, R.S., 2011. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA biology 8, 270–279.

[14] Ellson, J., Gansner, E.R., Koutsofios, E., North, S.C., Woodhull, G., 2004. Graphviz and dynagraph—static and dynamic graph drawing tools, in: Graph drawing software. Springer, pp. 127–148.

[15] Gangavarapu, K., Latiff, A.A., Mullen, J.L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Zeller, M., Aceves, C.M., Zaiets, K., et al., 2022. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. medRxiv .

[16] Groves, D.C., Rowland-Jones, S.L., Angyal, A., 2021. The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design. Biochemical and biophysical research communications 538, 104–107.

[17] Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., et al., 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science 369, 1603–1607.

[18] Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123.

[19] Hagberg, A., Swart, P., S Chult, D., 2008. Exploring network structure, dynamics, and function using NetworkX. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[20] Hodcroft, E.B., . CoVariants: SARS-CoV-2 Mutations and Variants of Interest. https://covariants.org/. (2021). Last accessed: July 29th, 2022.

[21] de Hoffer, A., Vatani, S., Cot, C., Cacciapaglia, G., Chiusano, M.L., Cimarelli, A., Conventi, F., Giannini, A., Hohenegger, S., Sannino, F., 2022. Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19. Scientific Reports 12, 1–14.

[22] Huang, Q., Zhang, Q., Bible, P.W., Liang, Q., Zheng, F., Wang, Y., Hao, Y., Liu, Y., 2022. A New Way to Trace SARS-CoV-2 Variants Through Weighted Network Analysis of Frequency Trajectories of Mutations. Frontiers in Microbiology 13.

[23] Massey Jr, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association 46, 68–78.

[24] Kalia, K., Saberwal, G., Sharma, G., 2021. The lag in SARS-CoV-2 genome submissions to GISAID. Nature Biotechnology 39, 1058–1060.

[25] Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., et al., 2021. SARS-CoV-2 evolution during treatment of chronic infection. Nature 592, 277–282.

[26] Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al., 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182, 812–827.

[27] Liu, L., Iketani, S., Guo, Y., Chan, J.F.W., Wang, M., Liu, L., Luo, Y., Chu, H., Huang, Y., Nair, M.S., et al., 2022. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. Nature 602, 676–681.

[28] Martin, D.P., Weaver, S., Tegally, H., San, J.E., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., et al., 2021. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell 184, 5189–5200.

[29] McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello, M., Chen, A., Liu, Z., Zatta, F., et al., 2021. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. Cell 184, 2332–2347.

[30] McCarthy, K.R., Rennick, L.J., Nambulli, S., Robinson-McCarthy, L.R., Bain, W.G., Haidar, G., Duprex, W.P., 2021. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science 371, 1139–1142.

[31] Mercatelli, D., Giorgi, F.M., 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Frontiers in microbiology , 1800.

[32] Minskaia, E., Hertzig, T., Gorbalenya, A.E., Campanacci, V., Cambillau, C., Canard, B., Ziebuhr, J., 2006. Discovery of an RNA virus 3'→ 5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proceedings of the National Academy of Sciences 103, 5108–5113.

[33] Mohammad, A., Abubaker, J., Al-Mulla, F., 2021. Structural modelling of SARS-CoV-2 alpha variant (B.1.1.7) suggests enhanced furin binding and infectivity. Virus Research 303, 198522.

[34] Organization, W.H., . Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/. Accessed: July 29th, 2022.

[35] O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al., 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution 7, veab064.

[36] Patro, L.P.P., Sathyaseelan, C., Uttamrao, P.P., Rathinavelan, T., 2021. Global variation in SARS-CoV-2 proteome and its implication in pre-lockdown emergence and dissemination of 5 dominant SARS-CoV-2 clades. Infection, Genetics and Evolution 93, 104973.

[37] Peacock, T.P., Penrice-Randal, R., Hiscox, J.A., Barclay, W.S., 2021. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. The Journal of general virology 102.

[38] Pinoli, P., Srihari, S., Wong, L., Ceri, S., 2021. Identifying collateral and synthetic lethal vulnerabilities within the DNA-damage response. BMC bioinformatics 22, 1–17.

[39] Qin, L., Ding, X., Li, Y., Chen, Q., Meng, J., Jiang, T., 2021. Co-mutation modules capture the evolution and transmission patterns of SARS-CoV-2. Briefings in Bioinformatics 22, bbab222.

[40] Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology 5, 1403–1407.

[41] Showers, W.M., Leach, S.M., Kechris, K., Strong, M., 2022. Longitudinal analysis of SARS-CoV-2 spike and RNA-dependent RNA polymerase protein sequences reveals the emergence and geographic distribution of diverse mutations. Infection, Genetics and Evolution 97, 105153.

[42] Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22, 30494.

[43] Singh, R., Nagpal, S., Pinna, N.K., Mande, S.S., 2021. Tracking mutational semantics of SARS-CoV-2 genomes. medRxiv .

[44] Tchesnokov, E.P., Gordon, C.J., Woolner, E., Kocinkova, D., Perry, J.K., Feng, J.Y., Porter, D.P., Götte, M., 2020. Template-dependent inhibition of coronavirus RNA-dependent RNA polymerase by remdesivir reveals a second mechanism of action. Journal of Biological Chemistry 295, 16156–16165.

[45] Troyano-Hernáez, P., Reinosa, R., Holguín, Á., 2021. Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: a global and regional approach by epidemiological week. Viruses 13, 243.

[46] Wada, K., Wada, Y., Ikemura, T., 2020. Time-series analyses of directional sequence changes in SARS-CoV-2 genomes and an efficient search method for candidates for advantageous mutations for growth in human cells. Gene 763, 100038.

[47] Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., Wei, G.W., 2021a. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Communications biology 4, 1–14.

[48] Wang, R., Chen, J., Gao, K., Wei, G.W., 2021b. Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. Genomics 113, 2158–2170.

[49] Xia, H., Cao, Z., Xie, X., Zhang, X., Chen, J.Y.C., Wang, H., Menachery, V.D., Rajsbaum, R., Shi, P.Y., 2020. Evasion of type I interferon by SARS-CoV-2. Cell reports 33, 108234.

[50] Yang, H.C., Chen, C.h., Wang, J.H., Liao, H.C., Yang, C.T., Chen, C.W., Lin, Y.C., Kao, C.H., Lu, M.Y.J., Liao, J.C., 2020. Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. Proceedings of the National Academy of Sciences 117, 30679–30686.

[51] Yang, L., Li, J., Guo, S., Hou, C., Liao, C., Shi, L., Ma, X., Jiang, S., Zheng, B., Fang, Y., et al., 2021. SARS-CoV-2 Variants, RBD Mutations, Binding Affinity, and Antibody Escape. International journal of molecular sciences 22, 12114.

[52] Zhang, J., Zhang, Y., Kang, J.Y., Chen, S., He, Y., Han, B., Liu, M.F., Lu, L., Li, L., Yi, Z., et al., 2021. Potential transmission chains of variant B.1.1.7 and co-mutations of SARS-CoV-2. Cell Discovery 7, 1–10.

[53] Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.