ConvMut: A Web tool to analyze viral convergent mutations along phylogenies.

Anna Bernasconi[1], Emma Fanfoni[1], Tommaso Alfonsi[1], Daniele Focosi[2]

[1]Department of Electronics, Information and Bioengineering. Politecnico di Milano, Milan, Italy.
anna.bernasconi@polimi.it. ORCID 0000-0001-8016-5750
[2]North-Western Tuscany Blood Bank. Pisa University Hospital, Pisa, Italy. daniele.focosi@gmail.com. ORCID 0000-0001-8811-195X

Author contributions: A.B. supervised the project and wrote the initial draft of the manuscript. E.F. developed the code. T.A. curated the web exposure. A.B., E.F., and T.A., jointly designed the computational methods. D.F. generated the project and revised the manuscript.

Abstract

Convergent evolution in protein antigens is common across pathogens and has also been documented in SARS-CoV-2; the most likely reason is the need to evade the selective pressure exerted by previous infection- or vaccine-elicited immunity. There is a pressing need for tools that allow automated analysis of convergent mutations.
In response to this need, we developed ConvMut, a tool for identifying patterns of recurrent mutations in SARS-CoV-2 evolution. To this end, we exploited the granular phylogenetic tree representation developed by PANGO, allowing us to observe what we call deltas, i.e., groups of mutations that are acquired on top of the immediately upstream tree nodes. Deltas comprise both amino acid substitutions and deletions. ConvMut can perform individual protein analysis to identify the most common single mutations acquired independently in a given subtree (starting from a user-selected root). Such mutations are represented in a barplot that can be sorted by frequency or position and filtered by region of interest. Lineages are then gathered into clusters according to their sets of shared mutations. Finally, an interactive graph orders the clusters, details the acquired amino acid changes for each sublineage, and allows us to the evolutive path until a selected lineage.
Other unique tools are paired with the main functionality of ConvMut to support a complete analysis, such as a frequency analysis for a given nucleotide or amino acid changes at a given residue across a selected phylogenetic subtree.
We expect that ConvMut will facilitate the design of antiviral anti-Spike monoclonal antibodies and Spike-based vaccines with longer-lasting efficacy, minimizing development and marketing failures.

Introduction

SARS-CoV-2 represents a panzootic of placental mammals [1], with multiple examples of reverse zoonosis documented so far [2,3,4] and the persistence of variants nearly extinct in humans [5,6]. Such massive circulation across different species creates huge opportunities for divergent evolution. The mutation rate of SARS-CoV-2 is estimated at 28.4 substitutions per year [7] and remains about 10-fold faster than for other RNA viruses established in humans since much more time (e.g. measles, mumps, RSV, dengue, Zika, WNV, enterovirus D68, or influenzavirus A(H_3N_2)), as documented by unprecedented genome sequencing efforts throughout the world. From December 2021 to December 2024, more than 3,500 sublineages have been designated by the PANGO group [8], all of them falling

under the WHO umbrella of the Omicron variant of concern (VoC) [9]. Such a granular phylogenetic tree represents a unique opportunity for studying convergent evolution. Globally, research has confirmed trends toward greater serological distance [10], translating into higher immune escape and finally into progressively higher reproductive numbers [11]. In proteomics, convergent evolution refers to the independent acquisition of genetic mutations leading to identical amino acid changes. The aim of those changes is adaptation to environmental variables, which for a pathogen implies immune evasion.
Convergent evolution in SARS-CoV-2 has been reviewed in detail elsewhere [12,13] and has led, e.g., to sudden failures of most of the therapeutic anti-Spike monoclonal antibodies [14]. To date, most insights on SARS-CoV-2 convergent evolution stem from manual work run by volunteers on social platforms, and no automated tool has been developed yet. We present here ConvMut, a comprehensive software for the analysis of convergent evolution in SARS-CoV-2 proteins.

Materials and Methods

In the following sections, we describe datasets that are used as reference and inputs for the developed tool. Dynamic inputs, which are processed and prepared by a bioinformatic pipeline before use, can typically be updated daily. To describe their main characteristics, we make reference to their versions available as of 3 December 2024.

Data references

We employ a series of data structures to store relevant information. Specifically, we use the reference genome of SARS-CoV-2 as encoded in the NC_045512.2 entry of the NCBI database [15], which we store in the 1-base coordinate system. For this reference sequence, we also store the positional annotations of all the proteins of the virus.
We employ an amino acid residue-codon dictionary to translate between nucleotides to their corresponding amino acids. Finally, we keep note of the PANGO lineage nomenclature by referring to the corresponding GitHub repository, which is updated regularly [16]; this file allows us to completely reconstruct the hierarchy implied by the nomenclature, focusing on the alias naming scheme that represents the full hierarchical provenance. For instance, if JN is an alias for B.1.1.529.2.86.1, then JN.1 is represented by the *unaliased* lineage B.1.1.529.2.86.1.1 in our hierarchical data structure.

Data inputs

The main input of the analysis pipeline is a phylogenetic representation of designated groups of sequences. For demonstrating ConvMut, we employ and support the use of two phylogenetic tree formats, respectively provided by UCSC and Nextstrain. We detail below their semantics and characteristics.

UCSC-derived phylogeny

The first phylogeny we support is produced by the UCSC Genomics Institute team, who exploited the UShER-based comprehensive SARS-CoV-2 mutation-annotated trees [17] made available via the endpoint [18]. With the name *pango.clade-mutations.tsv*, it represents one PANGO-designated lineage [19] per line. Lines list all the nucleotide substitutions (expressed with regard to the reference genome) that are progressively acquired by one

tree-node (either representing a designated or undesignated lineage) with respect to its previous node in the hierarchy. This traces a path along the tree, starting from its root to the lineage represented in that line. Different tree-levels are marked with the symbol ">"; possibly, multiple substitutions are incorporated in the same level (meaning that these are acquired jointly by the observed node). As an example, the lineage B.1.617.2 is associated with the path "C14408T > C241N > C3037T > A23403G > G29742N > T22917G > T27638C > G28881T > G210N > C21618G > G29402T > C23604G > C25469T > C22995A,C27752T > A28461G", indicating that –starting from the root lineage designated as B– 15 levels of the phylogeny have been traversed. At each step on the tree, one mutation is acquired with respect to the previous node (first C14408T, then C241N, etc.). Finally, the two mutations C22995A, C27752T are acquired by the parent node of B.1.617.2 with respect to the grandparent node, whereas A28461G is acquired by B.1.617.2 with respect to its parent node.

As this representation is derived from the UShER tree, it should be noted that some of the tree's limitations are inherited by the downstream analysis:

1. Some genomic positions are masked in the UShER tree. Positions recommended for masking in the Problematic Sites set are masked before sequences are added to the tree, so the tree contains no mutations at those positions. Then, after sequences have been added to the tree but before the optimization, the authors perform branch-specific masking – which is a mix of completely masking some positions and masking specific reversion mutations which are usually sequencing errors (reference bias), not actual reversions. This leads to showing in some recombinant (X*) lineages reversions that are not actual reversions (whereas real reversions would have a different meaning here).
2. Some lineages (including major VoCs) have deletions and/or insertions that UShER does not annotate; these are not included in *pango.clade-mutations.tsv*.
3. The PANGO tool [20] masks the non-coding 5' and 3' untranslated regions (UTRs) at the beginning and at the end of the genome, but the UShER tree masks only the first 55 and last 100 bases, so it has some mutations that become invisible in PANGO. This creates an overrepresentation of some "unknown" mutations (e.g., C241N). Additionally, nucleotides "N" are used in other locations to compensate for the instability of some branches in the tree.

At the time of our reported analysis, this file represents 3,163 lineages. After we removed the lines with names that do not comply with the PANGO nomenclature, we retained 3,106 including 129 recombinant lineages. We note that 6 lineages are missing (when compared to the PANGO list).

Nextstrain-derived phylogeny

The other phylogeny we used was the one produced by Nextstrain [21] and made available at the link [22]. It is an Auspice JSON file, typically used to render the Nextstrain SARS-CoV-2 phylogeny. With the name *staging_nextclade_sars-cov-2.json*, it presents a hierarchical structure where each JSON element corresponds to a node in the phylogeny, either a PANGO-designated lineage or an undesignated node (i.e., group of sequences).
Overall, the JSON document presents three branches: one for lineage B, one for all (non-recombinant) descendants of B, and one for recombinant lineages.

Compared to the previous representation, it is richer with information. For each node, the list of children nodes is presented. Different from the UCSC representation, it also keeps track of intermediate nodes that have not been designated by PANGO. Moreover, both nucleotide substitutions and deletions (with respect to the previous node in the tree) are presented; their amino acid-level corresponding changes are also reported.
For our purposes, we only retain nodes that represent PANGO-designated lineages. To reconstruct the entire set of mutations for each lineage, we visit the tree using a Depth-First Search algorithm. When -along the hierarchy- a specific location presents multiple mutation occurrences, we only retain the one exhibited by the most recent lineage (with its corresponding alternative nucleotide).
At the time of our reported analysis, this file reports 4,711 lineages. After we removed PANGO-undesignated lineages, we retained 3,460 lineages including 129 recombinant lineages. We note that 5 lineages were missing (when compared to the PANGO list).

Inputs comparison

We observed a limited set of discrepancies when comparing the data contained in the two input phylogenies. Specifically, *pango.clade-mutations.tsv* presents 3,106 designated lineages (of which 61 are not present in the other one) whereas *staging_nextclade_sars-cov-2.json* presents 3,460 designated lineages (of which 415 are not present in the other one). Specifically, we noticed that many lineages in the AY group are not present in *staging_nextclade_sars-cov-2.json* (44 lineages not present) along with some other 'early' lineages. Due to the limitations of the UShER tree, it derives that *pango.clade-mutations.tsv* misses deletions, has an overrepresentation of unknown (i.e., "N") mutations, and misses several reversions (with respect to the root of the tree). Note that only two PANGO-designated lineages, i.e., B.1.415 and B.1.1.482, are absent in both files (possibly due to lack of systematic annotation at the beginning of the pandemic).
Given these results, we base our main deployment on the *staging_nextclade_sars-cov-2.json* file (which is integrated with its 61 missing lineages); however, the framework supports all functionalities also using *pango.clade-mutations.tsv*. A full list of the considered lineages (divided by input source) along with details on their differences is provided as a Supplementary File.

Data processing

From each line in *pango.clade-mutations.tsv,* we extracted the lineage (e.g., JN.1), computed its corresponding unaliased name (e.g., B.1.1.529.2.86.1.1), and retrieved the list of its nucleotide mutations – with respect to the reference genome. Here, we remove lineages with names that are illegal in the PANGO nomenclature (e.g., containing *misc* or *dropout* substrings).
From *staging_nextclade_sars-cov-2.json*, we extracted all nodes, with their lineage name, their unaliased name, and mutations acquired by that node starting from the tree root – by unifying all the mutations acquired along the root-node path. Similarly, here we only retained nodes with names that are legal in PANGO.
Once the two data structures were merged, we obtained a total of 3,521 designated lineages, with 129 recombinant ones.

Tree reconstruction and translation

We rebuilt the original parent-child and child-parent links in the hierarchical phylogeny of SARS-CoV-2, only including designated lineages.
We built *template genomes* for each lineage (by updating the reference genome with the alternative nucleotide indicated in the mutations that are assigned to the lineage). Note that reversions did not produce any change here.
Then, we translated each template genome into its amino acid-level string, by exploiting the positions of the coding intervals for the SARS-CoV-2 reference genome [23]. During this process, we translated codons according to the widely accepted translation scheme; in addition, empty codons '---' were translated as an amino acid deletion '-'. When deleted nucleotides within a codon were less than three, or unknown nucleotides "N" were featured, we translated the triplet into the symbol "?", indicating an "undefined" value.

Inter-lineage deltas

For each considered lineage (*L*), we computed a *parent-delta*, i.e., the set of reference amino acids (with their position in the reference proteins) that have mutated, happening between *L* and *L*'s parent. Exceptionally, for those few lineages for which a parent was missing, we employed the first available ancestor in the dataset.
When *L* is recombinant, we computed a *delta* with respect to each of its parents. If an amino acid with its position was present in all deltas, it was inserted in the final *L*'s delta. Otherwise, we did not retain it (as we cannot rely on known breaking point coordinates for assigning a position to specific parents). We recognize that this leads to an approximation in the following cases:

A. in the observed position (*pos*), *LP1* exhibits *a1*, *LP2* exhibits *a2* and the recombinant lineage exhibits *a3*. Here, we will encode this situation as *a1_pos_a3*
B. in the observed position (*pos*), *LP1* exhibits *a1*, *LP2* exhibits *a2* and the recombinant lineage exhibits *a1*. Even if *a1* were to be considered an actual mutation (as derived from the portion of the genome inherited from *LP2*) this would not be recognized as a mutation, given that it is present in at least one of the parental lineages (i.e., *LP1*).

Wild type-lineage deltas

For each lineage, we computed a *wildtype-delta*, by comparing the amino acid exhibited in each position of the genome with the one present in the wild type. The "delta mutations" were stored in a dictionary. We noted that, for a given lineage, its parent-delta and wildtype-delta were not necessarily the same, as the following situations may arise:

- a certain mutation was included in parent-delta but not in wildtype-delta. This corresponds to a reversion, i.e., a mutation that re-instantiates the original wild-type genome.
- a certain mutation is included in the wildtype-delta but not in the parent-delta. Here, the observed lineage has inherited a mutation from its parent.

Quantifying a mutation convergence representativeness

We group by mutations that are present in parent-deltas, thereby extracting counts of how many lineages have acquired that mutation with respect to their direct parent. These counts are used to draw a ranking among mutations; the *most convergent mutations* are those with

the highest counts, as they are the most represented ones in parent-deltas (independently from the phylogenetic tree portion where this may happen).
Note that mutations in the same position contribute equally to computed counts, e.g., considering the position *pos*, if the *L1* lineage exhibits the mutation *a1_pos_a2*, and the *L2* lineage the mutation *a3_pos_a4*, the occurrence count for *pos* is increased by 2.
The previously illustrated approach may introduce a possible bias, as it ignores intermediate nodes that correspond to undesignated lineages. Specifically, all mutations acquired by a node in the tree that represents a non-designated lineage fall into the parent-delta of its designated offspring, thus possibly overestimating the mutation count.

Web Application

ConvMut is implemented as an interactive Web data application based on the open-source Python framework Streamlit that can be deployed as a Docker-based image on any server. It presents a landing page with documentation and three tabs respectively targeted to *convergence analysis*, *codon exploration*, and *mutation exploration*.
The application leverages pre-computed data structures (prepared as described in the 'Data processing' section). The corresponding workflow can be run regularly (e.g., daily) to capture potential updates in the input files (see 'Data inputs' section). The data structures are then pre-loaded in the application when it is started and then saved as session variables so that all tabs can reuse them and make the application more responsive.

Convergence analysis

This page allows us to analyze the most frequent convergent mutations in a specific SARS-CoV-2 protein within the entire phylogenetic tree or within a portion of the phylogeny starting from a user-selected *root* lineage. The interface user experience is built as a progressive workflow where the user is asked to select values to filter the analysis (unless default values are accepted). In the following, we define each phase.

Selection of a SARS-CoV-2 protein

One among all the SARS-CoV-2 proteins must be selected. Depending on this choice, only positions/mutations on the selected protein are then considered. The default protein is 'Spike'.

Choice of root lineage

This is the starting node of the subtree considered for the downstream analysis. Note that recombinant lineages are included in the analyzed subtree when at least one of their parent lineages is included in the subtree. ConvMut provides a tabular visualization of the subtree, where each line represents a *lineage* (plus its unaliased name), *parent-delta*, and *wildtype-delta.*
Additionally, ConvMut shows a horizontal barplot featuring the count of independent acquisitions of mutations occurring in each position of the selected protein along the selected subtree. All different mutation types occurring at a given position contribute to the global count for a given position (e.g., R346T, R346S, and S346I are counted for position 346). E.g., in a sub-tree rooted in JN.1, featuring 589 sub-lineages (plus JN.1 itself), we find

58 lineages whose parent-delta contains a mutation occurring on 346. Reversions count as regular mutations. The barplot can be sorted by position or number of occurrences.

Selection of positions of interest

The set of positions to be included in the following visualization is pre-filled by default using the positions exceeding a threshold set during the previous phase. By default, this is set to the value of the third most frequent position – should there be ties, at most 10 positions are included. Users can freely add or remove positions or change the entire set with an auto-complete component, e.g., when the analyzed protein is Spike, it is possible to restrict the analysis to specific regions, such as the receptor-binding domain (RBD: amino acid positions 319-541) or receptor-binding motif (RBM: amino acid positions 438-506).

We term *SP* the set of positions selected by the user; we compute and show 1) a series of pieplots representing the variability of different amino acid residues in each position in *SP*, and 2) a table including only the lineages within the selected subtree that include at least a position within *SP*. The table features the (aliased) lineage name, the unaliased lineage name, the cluster it belongs to, the mutations acquired by lineages in the cluster, and a string explaining how the cluster composition was computed.

For meaningful well-organized visualization of the map, we *cluster* lineages. Specifically – given a lineage *L*, its parent-delta, and wildtype-delta – *L*'s corresponding cluster is computed by taking only the positions in *SP* that are included in either *L*'s parent-delta or *L*'s wildtype-delta (or both). For instance, when SP = {346, 45}, *L1*'s parent-delta = {F456L, S680I}, and *L1*'s wildtype-delta = {S31-, R346S, F456L}, *L1*'s cluster will be characterized by {346,456}, then labelled c_346_456. Instead, with, *L2*'s parent-delta = {F456L}, and *L2*'s wildtype-delta = {Q480T}, *L2*'s cluster would be characterized by {456}, then labelled simply c_456.

Note that lineages grouped in the same *cluster* do not necessarily have the same mutations with respect to the user-selected root; instead, they must exhibit the same convergent mutations (even because of a reversion process).

Convergent mutations map

ConvMut finally plots a map to represent the complete convergent mutations' landscape of the selected protein in the selected subtree, including only mutations corresponding to the positions chosen as described above. Nodes in the map are then a specific set of PANGO-designated lineages *LV* (determined as those whose parent-delta sets contain at least one mutation included in the positions of *SP*) and arrows represent a parent-child (or ancestor-child) relationship between two lineages, labeled with mutation(s) contained in the *parent-delta* of the child node. Note that the graph is developed and visualized from left to right, by representing the root and older lineages in the left-most portion of the page and the leaves (most recent lineages) in the right-most portion.

We note that two nodes that are connected by an edge in the graph are not necessarily parent/child, as -given a child- it could happen that its direct parent is not contained in *LV* (and, in turn, its grandparent, etc.). Then, edges are drawn between a node and its closest PANGO-designated ancestor included in *LV*.

We also note that some clusters may enclose nodes whose in-coming edges have different labels (either in the alternative amino acid residue or even in the position); this is consistent with the definition of a cluster, which gathers lineages that have the same mutations in their

parent-delta or wildtype-delta – making the stage at which a mutation was acquired irrelevant.

Codon exploration

This page allows us to explore mutations that occur in a specific codon of the selected protein. Starting from the amino acid residue, we provide information on the determinant codon. Users input a SARS-CoV-2 protein, a phylogenetic root (in the PANGO-lineage hierarchy) from which the analysis is produced, and a *position of interest* (among pre-filtered options only including those that are present in at least one *delta* in the selected sub-tree). Users can choose to visualize results based on codons or amino acid residues.

Mutation exploration

This page allows users to investigate a given position on a selected protein, by considering also its phylogenetic background. Consistently with the previous tab, we require a protein, root lineage, and position *P* to be inspected.
We identify with *LM* all the lineages in the considered subtree that include a mutation in the *P* position in their parent-delta. The user can visualize a tree-like representation of the initial lineages' sub-tree, where we pruned all branches that do not present any lineage in *LM*; we also collapse tree levels not containing any lineage (or parent of a lineage) in *LM*. Mutation labels are colored to make differences in amino acids more apparent.

Results

We were able to generate convergent evolution charts for SARS-CoV-2 proteins. In this section, we report sample results for each of the ConvMut main functions.

Figure 1 shows intermediate results for the *Convergence analysis* when the following inputs are selected: `protein = S` (Spike), `root = JN.1`, `threshold = 16` (to include positions of interest), and `selected positions` (from more represented to less represented) `= 456, 346, 31, 572, 22, 478, 59, 445`.
Panel A illustrates a table with lineages (normal and unaliased) described by their parent-delta and wildtype-delta; Panel B shows the barplot whose y-axis represents positions along the chosen protein and the x-axis represents the count of lineages acquiring a mutation in that position; Panel C shows pieplots for each of the positions extracted above the set threshold (along the barplot shown in B), showing how mutations are distributed by alternative amino acid; Panel D presents a table of lineages to be visualized in the graph, detailing their cluster, the clusters' specific mutations and how these were computed, according to the formula:

$$cluster = (parent\ delta \cup wild\ type\ delta) \cap positions\ selected.$$

As a result of the previous parameters' selection, a graph is rendered as shown in Figure 2. By selecting a specific lineage in the dropdown menu above the graph, we can highlight the specific relevant path of converging events on the subtree. For example, in Figure 2 we show the path of LP.8, a lineage that is dominant at the time of writing. Specifically, the chain to be followed from the root (left) to the selected lineage (right) can be described as follows: JN.1 acquired F456L in JN.1.11.1, which later acquired R346T in KP.1.1. This in turn acquired S31- in KP.1.1.3, which acquired a triplet <F186L, H445R, Q493E> in LP.8.

Interestingly, LP.8's parent, i.e., KP.1.1.3, is located in the same cluster (c_22_31_445_456_479) as LP.8. However, only LP.8 acquired H445R; by inspecting an older part of the tree, it is possible to verify that V445H had been formerly acquired from BA.2 to BA.2.86, so the visualization correctly captures the current state of concurrent mutations.

While Figure 2 shows the visualized tree, Figure 3 makes a zoom on the observed path, obtained by exploiting the features of the implemented interactive visualization.

Then, Figure 4 provides an overview of the possible output of the *Codon exploration* analysis, when the input `root = JN.1` and `pos = 346` are selected. Panel A shows the genetic code (correspondences between nucleotide triplets and the corresponding amino acid residues) where we underline the codon exhibited in the wild type genome in the position of interest *p* and color-highlight the codon present in that position in those lineages that feature an amino acid change in *p* and their parent-delta. Panel B shows the counts of mutations per type of codon, i.e., the number of lineages exhibiting that codon in their parent-delta, in the selected position. Panel C details the lineages contributing to those counts.

Finally, Figure 5 provides the visualization built within the *Mutation exploration* tab for `root = JN.1` and Spike protein position 346.

Discussion

The analysis of convergent evolution in a pathogen is of paramount importance to predict the changing landscape where antiviral drugs have to act. Clearly, convergent evolution is more evident for protein antigens that are targeted by the selective pressure of the immune response, either infection- or vaccine-elicited. Throughout the four years of the COVID-19 pandemic, the clearest example of convergent evolution has been within the Spike receptor-binding domain (RBD): it has happened in waves and has led to dramatic failures of recently authorized therapeutic anti-Spike monoclonal antibodies [24]. While most of the effort has historically been at studying the divergence in order to provide granular phylogenetic trees, the myriad of sublineages identified for SARS-CoV-2 has posed challenges for the classification of variants. Scientists in the community have volunteered to keep track of shared sets of Spike mutations regardless of the origin and to represent charts that could be easily interpreted by both the general public and vaccine/antibody manufacturers.

Our study has several limitations. First, ConvMut is dependent on the PANGO nomenclature, whose designation criteria are somewhat arbitrary and not yet agnostic [25]. Nevertheless, the other available phylogenies appear either too much (Nextstrain) or too little (GISAID) granular to conduct analyses. We had to merge two data sources (Nextstrain and UCSC) to include deletions in our analysis. Insertions are not accounted for yet in the current version of ConvMut. Second, for recombinant lineages, ConvMut is not able to discriminate the parental source of a given convergent mutation, but -in many instances- the exact breakpoint has not been discriminated against. Third, when estimating the commonest convergent mutations, ignoring the intermediate PANGO-undesignated nodes could have led to an overestimate of mutation counts. Finally, the delay between genome deposition in GISAID and lineage designation by the PANGO group could lead to delays in recognizing additional convergent evolutions: in this regard, our team is currently working with GISAID to embed ConvMut within their framework, in order to grant more up-to-date analyses.

The strengths of ConvMut are: 1) the uniqueness of each of its tools among the plethora of SARS-CoV-2 bioinformatics tools generated so far; 2) the automated updates to data sources, without any need to manually update the backends; 3) the reliance on open data; 4) the availability as a user-friendly graphical interface; and 5) the possibility to export the generated charts.

ConvMut can prove useful at forecasting trends in SARS-CoV-2 evolution, benefiting not just basic virological research, but also drug and vaccine manufacturers: the latter can, in fact, adjust their pipeline according to trends, minimizing failures (and their associated economic costs) and focus on drugs that have the higher likelihood of market persistence. This is increasingly important in light of the recent developments in immunobridging [26], which can shorten the time from drug design to market authorization. The architecture of ConvMut is prone to replication on viral pathogens other than SARS-CoV-2, whenever granular phylogenies have been generated.

Conclusion

We have here exposed the rationale behind and the generation of a novel Web tool to automatically explore convergent evolution in SARS-CoV-2 proteins. Our model could be readapted to different human pathogens for which granular phylogenies are available.

Declarations

Funding. The work was supported by Ministero dell'Università e della Ricerca (PRIN PNRR 2022 “SENSIBLE” project, n. P2022CNN2J), funded by the European Union, Next Generation EU, within PNRR M4.C2.1.1. Principal Investigator A.B.

Institutional Review Board Statement. *Not applicable.*

Informed Consent Statement. *Not applicable.*

Data Availability Statement. All data used in the backend of the proposed application can be publicly retrieved at the endpoints [16,18,22]. Supplementary Material reports the list of the lineages considered (as of December 3, 2024) from the input files.

Acknowledgments. We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative. We gratefully acknowledge the contributors to the UCSC genome browser, with particular reference to the work of Angie Hinrichs. We gratefully acknowledge the contributors to the PANGOlin project. We gratefully acknowledge the contributors of the Nextstrain project.
We gratefully acknowledge the contribution of Dr. Fabrizio Maggi and Dr. Cesare Gruber at the National Institute for Infectious Diseases “Lazzaro Spallanzani” in Rome, Italy.

Conflicts of Interest. The authors have no conflict of interest to declare related to this manuscript.

Figure 1. *Intermediate results' visualization.* ***(A)*** *Table with all the lineages that derive from the chosen root;* ***(B)*** *Barplot that, for each position on the selected protein, counts how many lineages independently acquired a mutations in that position;* ***(C)*** *Pieplots that capture the variability of different amino acid residues in the selected positions (above threshold according to the barplot in panel B;* ***(D)*** *Table for detailing a lineage's cluster name and its mutations for each of the lineages to be plotted in the graph.*

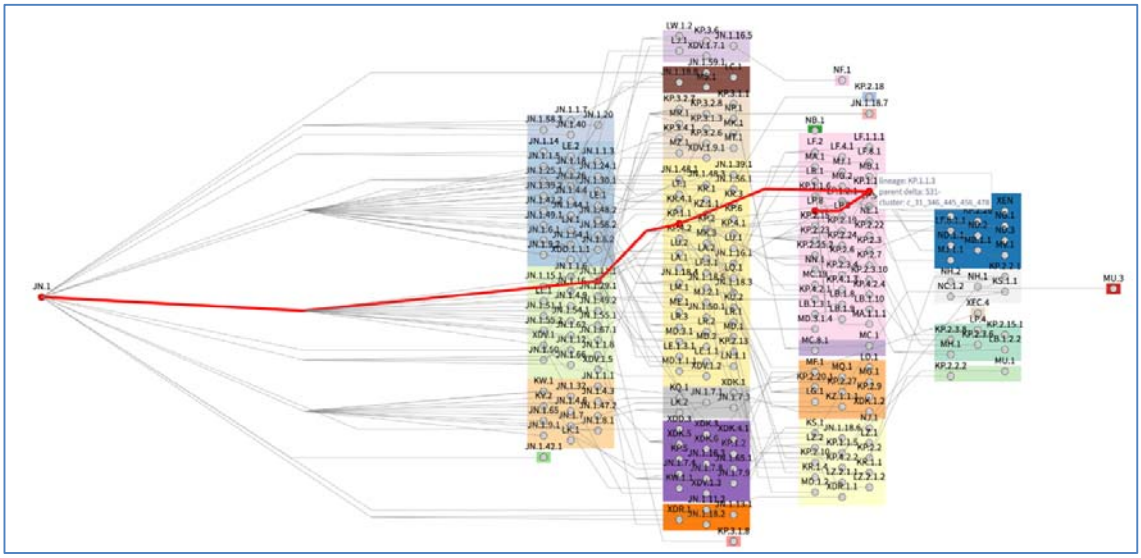


Figure 2. *The ConvMut Convergence analysis tree, for Spike's mutations* `456, 346, 31, 572, 22, 478, 59, 445` *in the JN.1 subtree.*

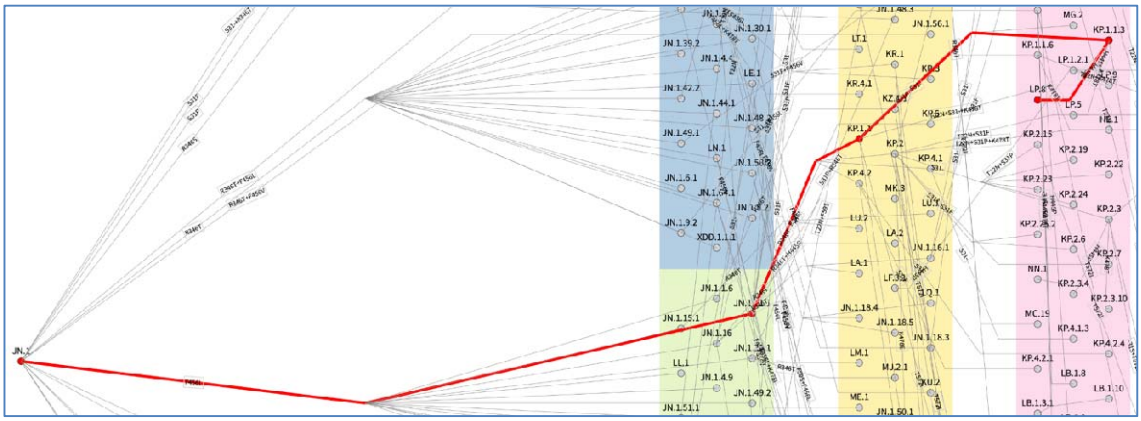


Figure 3. *Zoomed visualization of the ConvMut Convergence analysis tree, for Spike's mutations* `456, 346, 31, 572, 22, 478, 59, 445` *in the JN.1 subtree. The path leading from the root JN.1 to the recent lineage LP.8 is highlighted in red.*

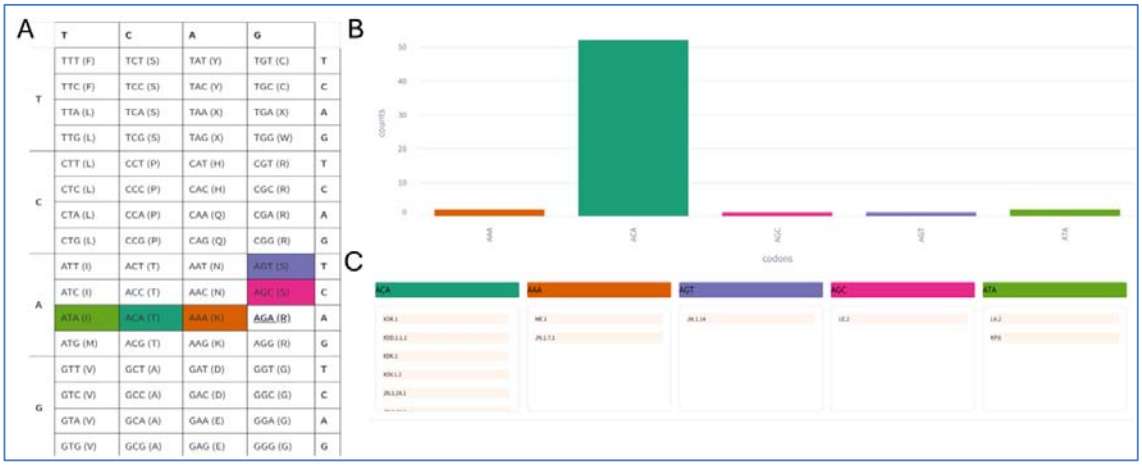


Figure 4. *The ConvMut Codon exploration view for the input selection JN.1 (root) and 346 (position).* ***(A)*** *genetic code,* ***(B)*** *counts of mutations,* ***(C)*** *lineages.*

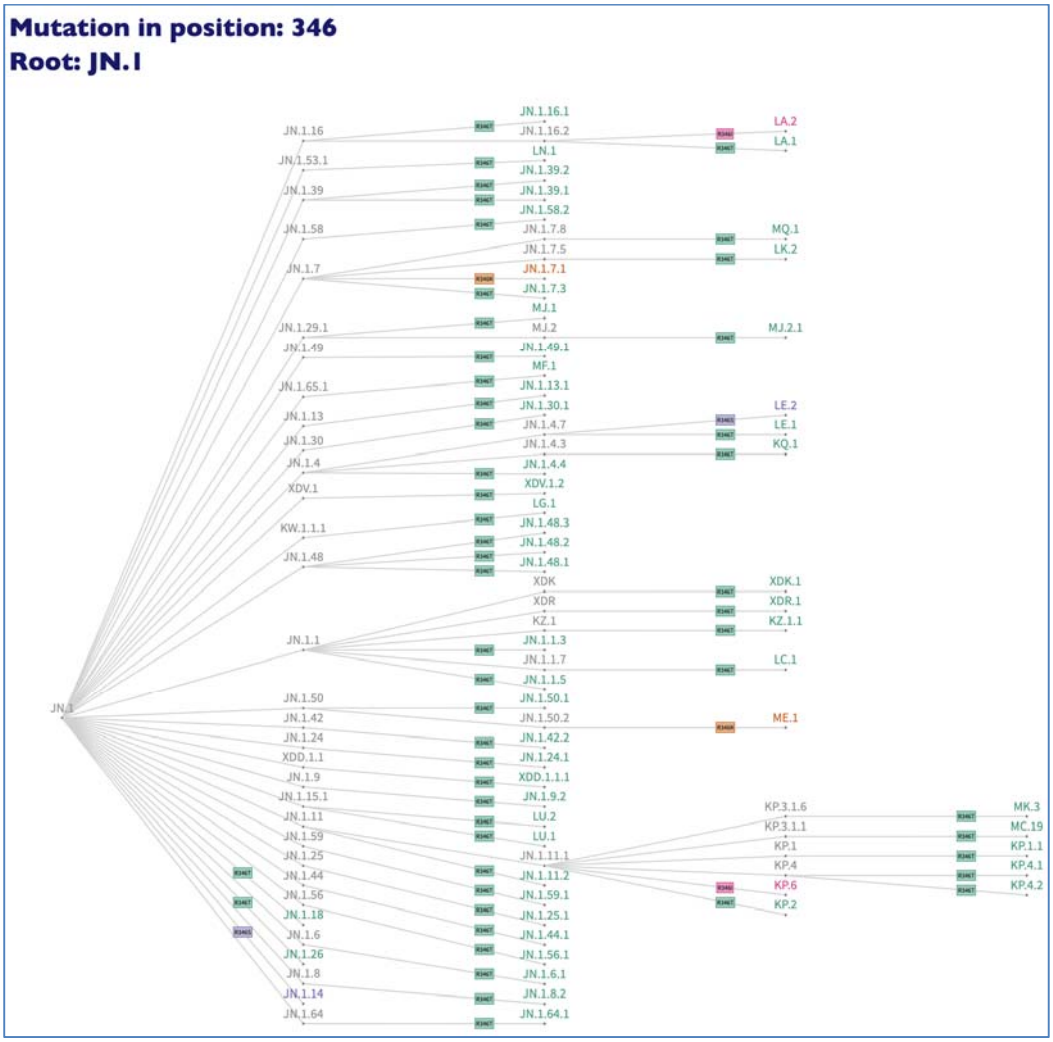


Figure 5. *The ConvMut Mutation exploration tree, for root JN.1 and position 346. Labels on edges indicate the different mutation occurred on this position and acquired by the node on the right of the edge. Note that we use different colors to mark different exhibited amino acid residues.*

Bibliography

[1] Wang, Y., Ye, M., Zhang, F., Freeman, Z.T., Yu, H., Ye, X. and He, Y., 2024. Ontology-based taxonomical analysis of experimentally verified natural and laboratory human coronavirus hosts and its implication for COVID-19 virus origination and transmission. Plos one, 19(1), p.e0295541.

[2] Oude Munnink, B.B., Sikkema, R.S., Nieuwenhuijse, D.F., Molenaar, R.J., Munger, E., Molenkamp, R., Van Der Spek, A., Tolsma, P., Rietveld, A., Brouwer, M. and Bouwmeester-Vincken, N., 2021. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. science, 371(6525), pp.172-177.

[3] Pickering, B., Lung, O., Maguire, F., Kruczkiewicz, P., Kotwa, J.D., Buchanan, T., Gagnier, M., Guthrie, J.L., Jardine, C.M., Marchand-Austin, A. and Massé, A., 2022. Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human transmission. Nature Microbiology, 7(12), pp.2011-2024.

[4] Fischhoff, I.R., Castellanos, A.A., Rodrigues, J.P., Varsani, A. and Han, B.A., 2021. Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2. Proceedings of the Royal Society B, 288(1963), p.20211651.

[5] Caserta, L.C., Martins, M., Butt, S.L., Hollingshead, N.A., Covaleda, L.M., Ahmed, S., Everts, M.R., Schuler, K.L. and Diel, D.G., 2023. White-tailed deer (Odocoileus virginianus) may serve as a wildlife reservoir for nearly extinct SARS-CoV-2 variants of concern. Proceedings of the National Academy of Sciences, 120(6), p.e2215067120.

[6] Domańska-Blicharz, K., Munnink, B.B.O., Orłowska, A., Smreczak, M., Opolska, J., Lisowska, A., Trębas, P., Socha, W., Giza, A., Bomba, A. and Iwan, E., 2023. Cryptic SARS-CoV-2 lineage identified on two mink farms as a possible result of long-term undetected circulation in an unknown animal reservoir, Poland, November 2022 to January 2023. Eurosurveillance, 28(16), p.2300188.

[7] Nextstrain, 2024. Genomic epidemiology of SARS-CoV-2 with subsampling focused globally since pandemic start. Available online at

https://nextstrain.org/ncov/gisaid/global/all-time. Last accessed 3rd Dec, 2024.

[8] Cornelius Roemer. CoV Lineages, Pango Designation. Available online at https://github.com/cov-lineages/pango-designation/blob/7651872d633c8e94437830f792b272712d4a22fc/lineage_notes.txt. Last accessed 3rd Dec, 2024.

[9] World Health Organization. Tracking SARS-CoV-2 variants. Available online at https://www.who.int/activities/tracking-SARS-CoV-2-variants. Last accessed 3rd Dec, 2024.

[10] Focosi, D., McConnell, S. and Casadevall, A., 2022. The Omicron variant of concern: Diversification and convergent evolution in spike protein, and escape from anti-Spike monoclonal antibodies. Drug Resistance Updates, 65, p.100882.

[11] Focosi, D., Maggi, F., McConnell, S. and Casadevall, A., 2022. Spike mutations in SARS-CoV-2 AY sublineages of the Delta variant of concern: implications for the future of the pandemic. Future microbiology, 17(4), pp.219-221.

[12] Focosi, D., McConnell, S. and Casadevall, A., 2022. The Omicron variant of concern: Diversification and convergent evolution in spike protein, and escape from anti-Spike monoclonal antibodies. Drug Resistance Updates, 65, p.100882.

[13] Focosi, D., Spezia, P.G. and Maggi, F., 2024. Subsequent Waves of Convergent Evolution in SARS-CoV-2 Genes and Proteins. Vaccines, 12(8), p.887.

[14] Focosi, D., Franchini, M., Casadevall, A. and Maggi, F., 2024. An update on the anti-Spike monoclonal antibody pipeline for SARS-CoV-2. Clinical Microbiology and Infection.

[15] Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y. and Yuan, M.L., 2020. A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), pp.265-269.

[16] Cornelius Roemer. CoV Lineages, Pango Designation. Available online at https://github.com/cov-lineages/pango-designation/blob/master/pango_designation/alias_key.json. Last accessed 3rd Dec, 2024.

[17] McBroome, J., Thornlow, B., Hinrichs, A.S., Kramer, A., De Maio, N., Goldman, N., Haussler, D., Corbett-Detig, R. and Turakhia, Y., 2021. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Molecular biology and evolution, 38(12), pp.5819-5824.

[18] Hinrichs, A., 2024. Pango Clade mutations file. Available online at https://hgwdev.gi.ucsc.edu/~angie/pango.clade-mutations.tsv. Last accessed 3rd Dec, 2024.

[19] Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., Du Plessis, L. and Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology, 5(11), pp.1403-1407.

[20] O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B. and Yeats, C., 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus evolution, 7(2), p.veab064.

[21] Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. and Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), pp.4121-4123.

[22] Nextstrain, 2024. SARS-CoV-2 phylogeny. Available online at https://nextstrain.org/staging/nextclade/sars-cov-2. Last accessed 3rd Dec, 2024.

[23] NCBI Reference Sequence: NC_045512.2. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. Available online at https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=genbank. Last accessed 3rd Dec, 2024.

[24] Focosi, D., Franchini, M., Casadevall, A. and Maggi, F., 2024. An update on the anti-Spike monoclonal antibody pipeline for SARS-CoV-2. Clinical Microbiology and Infection.

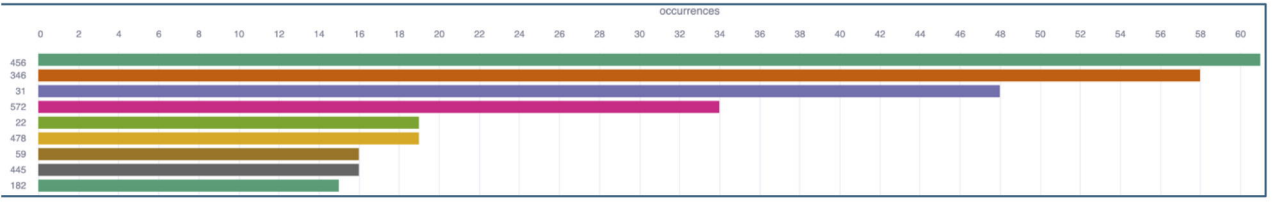
[25] McBroome, J., de Bernardi Schneider, A., Roemer, C., Wolfinger, M.T., Hinrichs, A.S., O'Toole, A.N., Ruis, C., Turakhia, Y., Rambaut, A. and Corbett-Detig, R., 2024. A framework for automated scalable designation of viral pathogen lineages from genomic data. Nature Microbiology, 9(2), pp.550-560.

[26] Schmidt, P., Li, Y. and Popejoy, M., 2024. Immunobridging for Pemivibart, a Monoclonal Antibody for Prevention of Covid-19. New England Journal of Medicine, 391(19), pp.1860-1862.

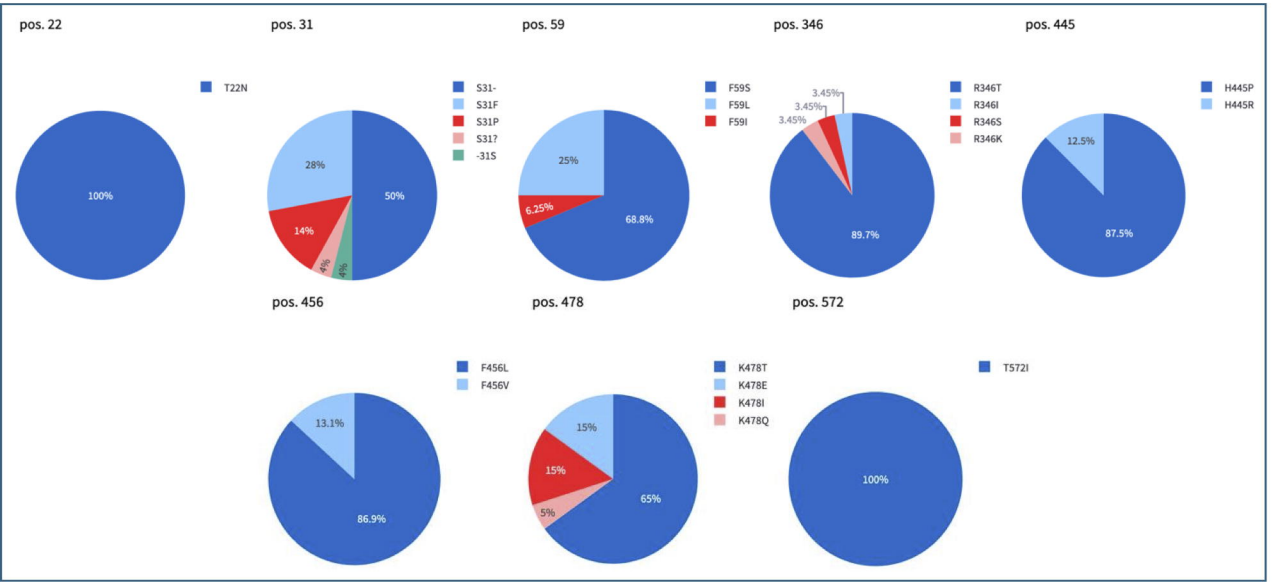
A

	lineage	unaliased lineage	parent delta	wildtype delta
0	JN.1	B.1.1.529.2.86.1.1	L455S	T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G N211? L212? V213G L
1	JN.1.1	B.1.1.529.2.86.1.1.1		T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G N211? L212? V213G L
2	JN.1.1.1	B.1.1.529.2.86.1.1.1.1	T572I	T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G N211? L212? V213G L
3	JN.1.1.10	B.1.1.529.2.86.1.1.1.10		T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G N211? L212? V213G L
4	JN.1.1.2	B.1.1.529.2.86.1.1.1.2	G181E	T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G G181E N211? L212? V
5	JN.1.1.3	B.1.1.529.2.86.1.1.1.3	R346T	T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G N211? L212? V213G L
6	LT.1	B.1.1.529.2.86.1.1.1.3.1	L176F F456V	T19I R21T L24? P25- P26- A27? S50L I68? H69- V70? V127F G142D Y144- F157S R158G L176F N211? L212? V

B

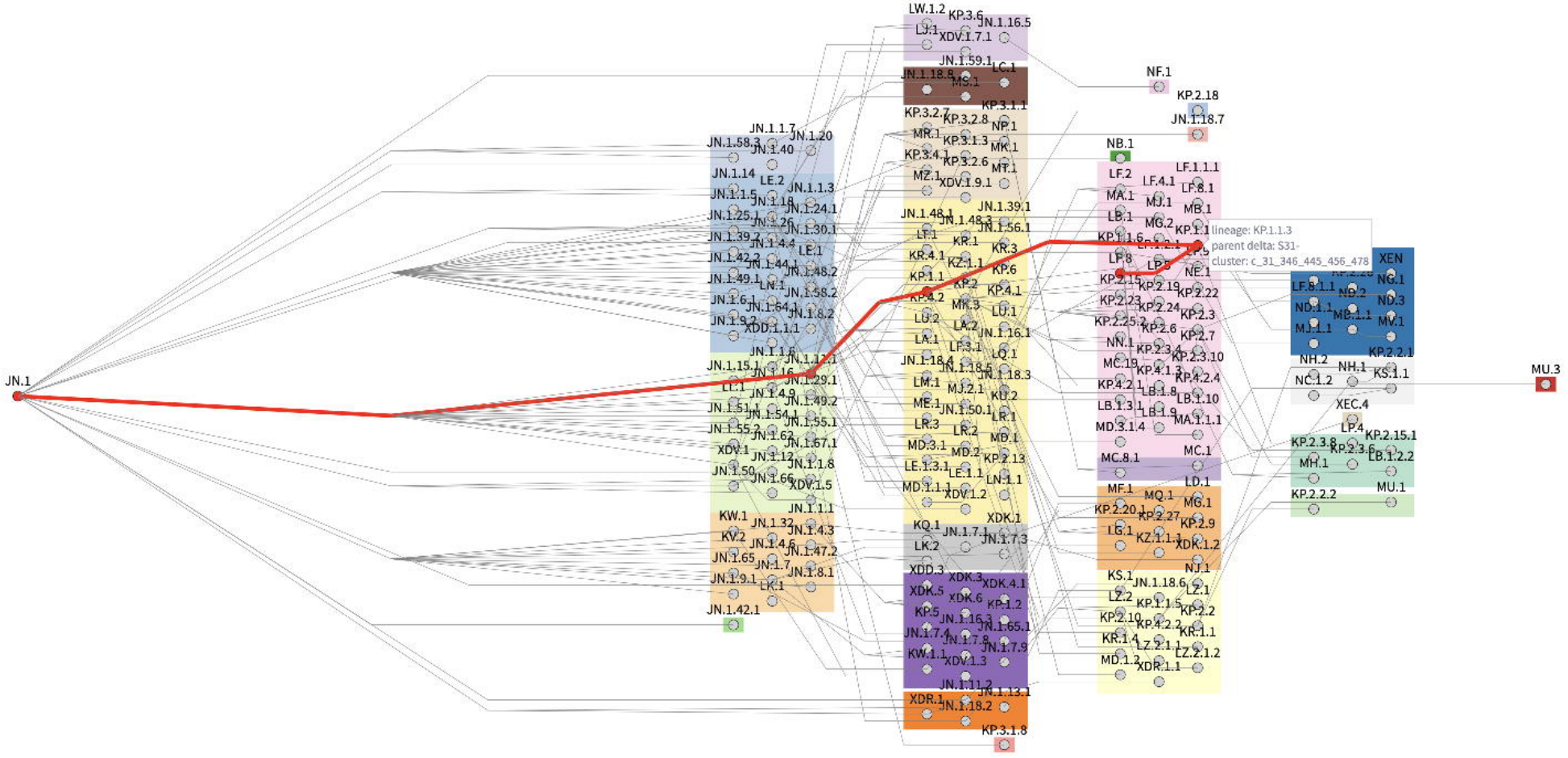


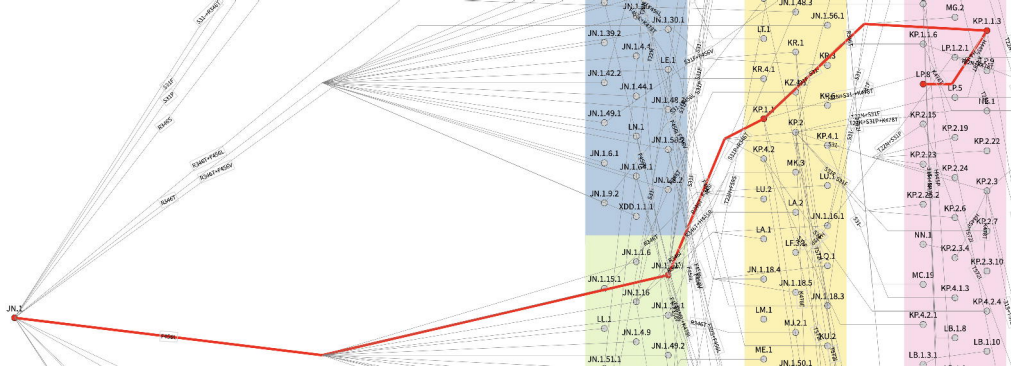
C



D

	lineage	unaliased lineage	parent delta	cluster	cluster mutations	cluster mutations formula
0	JN.1.1.1	B.1.1.529.2.86.1.1.1.1	T572I	c_445_478_572	V445H T478K T572I	([T572I] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157S, R15
1	JN.1.1.3	B.1.1.529.2.86.1.1.1.3	R346T	c_346_445_478	R346T V445H T478K	([R346T] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157S, R1
2	JN.1.1.5	B.1.1.529.2.86.1.1.1.5	R346T	c_346_445_478	R346T V445H T478K	([R346T] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157S, R1
3	JN.1.1.6	B.1.1.529.2.86.1.1.1.6	F456L	c_445_456_478	V445H F456L T478K	([F456L] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157S, R1
4	JN.1.1.7	B.1.1.529.2.86.1.1.1.7	S31F	c_31_445_478	S31F V445H T478K	([S31F] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S31F, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157
5	JN.1.1.8	B.1.1.529.2.86.1.1.1.8	H445P F456L	c_445_456_478	H445P V445P F456L T478K	([H445P, F456L] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F1
6	JN.1.11.1	B.1.1.529.2.86.1.1.11.1	F456L	c_445_456_478	V445H F456L T478K	([F456L] ∪ [T19I, R21T, L24?, P25-, P26-, A27?, S50L, I68?, H69-, V70?, V127F, G142D, Y144-, F157S, R1

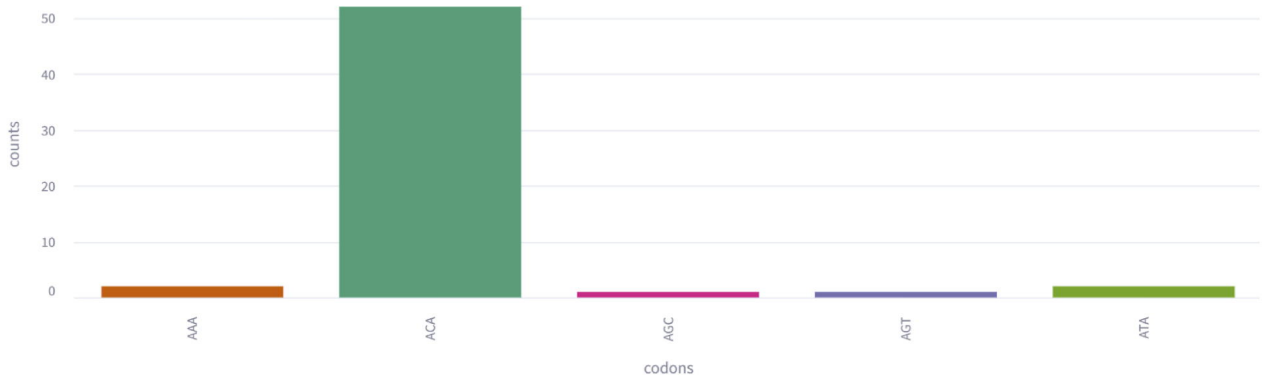

JN.1
lineage: KP.1.1.3
parent delta: S31-
cluster: c_31_346_445_456_478
NF.1
KP.2.18
JN.1.18.7
NB.1
XEN
MU.3
MU.1
JN.1.42.1
KP.3.1.8

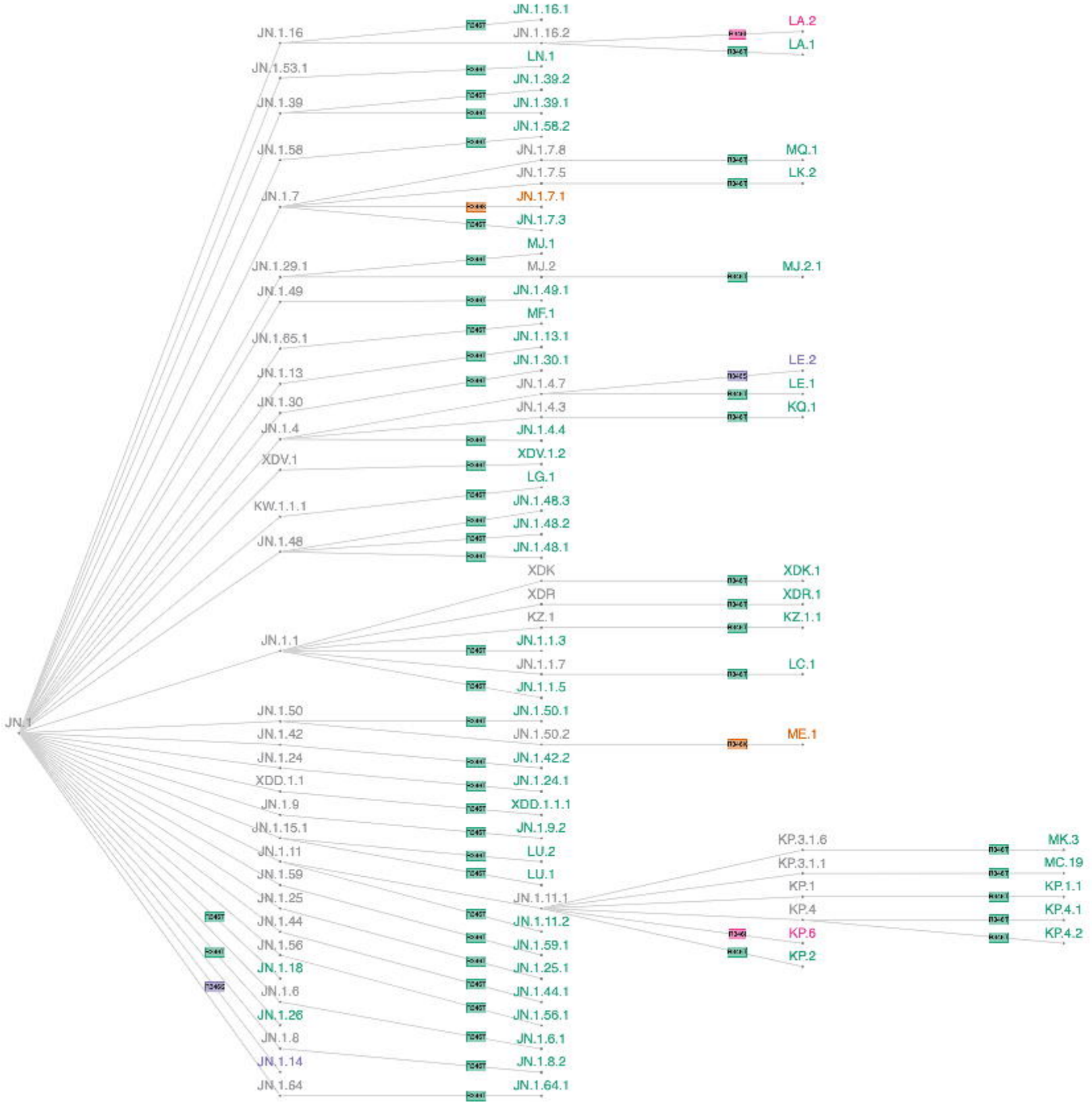

JN.1
JN.1.48.3
JN.1.56.1
JN.1.30.1
JN.1.39.2
JN.1.4.1
JN.1.42.2
JN.1.44.1
JN.1.49.1
JN.1.6.1
JN.1.9.2
XDD.1.1.1
JN.1.1.6
JN.1.15.1
JN.1.16
JN.1.4.9
JN.1.49.2
JN.1.51.1
LE.1
KR.1
KR.4.1
KZ.1
KP.1.1
KP.2
KP.4.1
KP.4.2
MK.3
LU.2
LA.2
JN.1.16.1
LA.1
LF.3.2
LQ.1
JN.1.18.4
JN.1.18.5
JN.1.18.3
LM.1
MJ.2.1
LKU.2
ME.1
JN.1.50.1
MG.2
KP.1.1.3
KP.1.1.6
LP.1.2.1
LP.8
LP.5
HE.1
KP.2.15
KP.2.19
KP.2.22
KP.2.23
KP.2.24
KP.2.3
KP.2.28.2
KP.2.6
KP.2.7
NN.1
KP.2.3.4
KP.2.3.10
MC.19
KP.4.1.3
KP.4.2.4
KP.4.2.1
LB.1.8
LB.1.10
LB.1.3.1

A

	T	C	A	G	
T	TTT (F)	TCT (S)	TAT (Y)	TGT (C)	T
	TTC (F)	TCC (S)	TAC (Y)	TGC (C)	C
	TTA (L)	TCA (S)	TAA (X)	TGA (X)	A
	TTG (L)	TCG (S)	TAG (X)	TGG (W)	G
C	CTT (L)	CCT (P)	CAT (H)	CGT (R)	T
	CTC (L)	CCC (P)	CAC (H)	CGC (R)	C
	CTA (L)	CCA (P)	CAA (Q)	CGA (R)	A
	CTG (L)	CCG (P)	CAG (Q)	CGG (R)	G
A	ATT (I)	ACT (T)	AAT (N)	AGT (S)	T
	ATC (I)	ACC (T)	AAC (N)	AGC (S)	C
	ATA (I)	ACA (T)	AAA (K)	AGA (R)	A
	ATG (M)	ACG (T)	AAG (K)	AGG (R)	G
G	GTT (V)	GCT (A)	GAT (D)	GGT (G)	T
	GTC (V)	GCC (A)	GAC (D)	GGC (G)	C
	GTA (V)	GCA (A)	GAA (E)	GGA (G)	A
	GTG (V)	GCG (A)	GAG (E)	GGG (G)	G

B



C

Mutation in position: 346
Root: JN.1