Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time

Luca Cilibrasi [1], Pietro Pinoli [1]*, Anna Bernasconi [1], Arif Canakoglu [1], Matteo Chiara [2] and Stefano Ceri [1]

[1]Dept. of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, 20133 Milano, Italy
[2]Dept. of BioSciences, University of Milano, 20133 Milano, Italy

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The ongoing evolution of SARS-CoV-2 and the rapid emergence of variants of concern (VOCs) at distinct geographic locations have relevant implications for the implementation of strategies for controlling the COVID-19 pandemic. Combining the growing body of data and the evidence on potential functional implications of SARS-CoV-2 mutations can suggest highly effective methods for the prioritization of novel variants of potential concern, e.g., increasing in frequency locally and/or globally. However, these analyses may be complex, requiring the integration of different data and resources. We claim the need for a streamlined access to up-to-date and high-quality genome sequencing data from different geographic regions/countries, and the current lack of a robust and consistent framework for the evaluation/comparison of the results.

Results: To overcome these limitations, we developed ViruClust, a novel tool for the comparison of SARS-CoV-2 genomic sequences and lineages in space and time. ViruClust is made available through a powerful and intuitive web-based user interface. Sophisticated large scale analyses can be executed with a few clicks, even by users without any computational background. To demonstrate potential applications of our method, we applied ViruClust to conduct a thorough study of the evolution of the most prevalent lineage of the Delta SARS-CoV-2 variant, and derived relevant observations.

Conclusions: By allowing the seamless integration of different types of functional annotations and the direct comparison of viral genomes and genetic variants in space and time, ViruClust represents a highly valuable resource for monitoring the evolution of SARS-CoV-2, facilitating the identification of variants and/or mutations of potential concern.

Availability: ViruClust is openly available at `http://gmql.eu/viruclust/`.

Contact: pietro.pinoli@polimi.it

1 Introduction

With the progress of the COVID-19 pandemic, the genomic evolution of SARS-CoV-2 is followed with unprecedented interest; in particular, scientists wish to closely monitor the mutations that are acquired by the virus as it spreads and diversifies over time and across countries (Lauring and Hodcroft, 2021). Whole-genome sequencing of the virus, followed by bioinformatics analyses, can be used to rapidly identify lineages that are increasing in frequency locally and/or globally (Chiara *et al.*, 2021b). Identifying potential/prospective variants as they emerge allows for the early implementation of control measures and mitigation strategies.

Taking advantage of our experience in data integration for human genomics (Canakoglu *et al.*, 2019), we designed the Viral Conceptual Model (Bernasconi *et al.*, 2020) and used its structure to develop several databases and tools, including ViruSurf (Canakoglu *et al.*, 2021), VirusViz (Bernasconi *et al.*, 2021b), and EpiSurf (Bernasconi *et al.*, 2021a). We hereby present ViruClust, a tool specifically designed for supporting the comparative analysis of SARS-CoV-2 evolution across lineages, space and time.

ViruClust is based on ViruSurf_GISAID (Canakoglu *et al.*, 2021) (`http://gmql.eu/virusurf_gisaid/`) and enables users to freely select the viral populations of interest using GISAID-provided metadata (Pango lineages (Rambaut *et al.*, 2020), locations, collection dates) to drive their analyses. At the time of writing (Aug 26th, 2021), with over 3M SARS-CoV-2 genomic sequences isolated from different countries and at different times during the COVID-19 pandemic, the GISAID database (Shu and McCauley, 2017) represents the most complete and accurate resource for SARS-CoV-2 genomic data.

ViruClust currently supports four different modes of analysis; the first mode was designed to provide an overview of the prevalence of SARS-CoV-2 lineages in a geographic location of interest; the other modes allow the direct comparison of two distinct viral sub-populations or groups. A brief outline is presented in the following:

(1) Prevalence of lineages: analytical report of the prevalence of SARS-CoV-2 lineages deposited from a geographic place of interest. Users can specify given intervals of time, and different levels of granularity (continent, country, region). We count the number of sequences assigned to each lineage within all the subregions of a given geographic region. The distribution of the number of sequences collected in the region over time is represented by a histogram, and a time interval of interest can be graphically selected. The final output consists in a table with the distribution of lineages by subregions and is also visualized in the form of a heat map; in the visual representation lineages can be filtered and columns can be aggregated and/or renamed.
(2) Evolution in time: allows the comparison of allele frequencies, in any arbitrary selection of SARS-CoV-2 genomes, over time. A location of interest must be selected; additionally, users can specify a lineage. Comparisons are performed either by considering two arbitrary intervals of time (hereafter defined as the "target" and "background" groups), or by splitting an interval of time by weeks/months. The analysis produces a protein-level comparison of the prevalence of amino acid changes in the target/background; when several periods (weeks or months) are compared, each period is considered as the target group and the preceding period is used as the background, so as to capture increase/decrease in relative prevalence. The final output comprise tables, a heat map, and bar-charts to facilitate the data exploration.
(3) Evolution in space: compares two groups of genome sequences (target and background) associated with distinct geographic regions. The target is defined as all genome sequences collected from a location of interest. A specific lineage can optionally be selected. The background is the collection of sequences associated with an enclosing, larger region. A flexible mechanism can be used to exclude specific locations from the background. Additionally, the analysis can be restricted to a period of interest. The comparative analysis of target/background groups can be visualized in a table or visually inspected as a bar-chart of amino acid changes of a specific protein. Output formats are equivalent to those produced by analysis (2).
(4) Custom analysis: the user is free to select any two arbitrary groups/populations of viral genomic sequences from GISAID, and apply analyses (2) and (3). No restriction (of any type) is applied to the selection. Users are free to compare across different lineages, times and geographies.

The four types of analysis are highly effective for the identification of novel amino acid changes as they emerge in the genome of the virus, and in principle can provide useful information for prioritization of novel variants of the virus of potential concern. The tool is organized so as to streamline the analysis, providing an easy access to up-to-date and high quality genome sequencing data from different geographic regions/countries, and a robust and consistent framework for the evaluation/comparison of the results.

2 Related work

Various systems and methods have recently been proposed for analysing the mutational landscape of SARS-CoV-2 sequences in an interactive way. A number of approaches have focused on computational methods for classifying genomes either along phylogenetic guidelines (see Pangolin (Rambaut *et al.*, 2020) and Nextstrain (Hadfield *et al.*, 2018)) or only employing unsupervised learning methods (see for example (Chiara *et al.*, 2021a) and (Yang *et al.*, 2020)). Such research outcomes, however, only partially allow to differentiate between similar/closely related viral genomes and lineages and to identify relevant differences or events associated with their evolution. Typically, differences are better captured by comparing specific groups of sequences associated with a different geographic origin, collected at different points in time and/or (even) assigned to different classifications. In this scenario, easy-to-use platforms for quickly visualizing distributions of lineages and variants and of their single characterizing mutations become essential.

Selected functionalities are offered by a number of already-available systems. CoV-Spectrum (Chen *et al.*, 2021b) allows to analyze the time series of the sequences annotated with specific Pangolin lineages or exhibiting one particular amino acid change. More in-depth analyses are provided by integrating models to estimate epidemiological dynamics and variant properties; note that integration with severity of infections, vaccine status, and environmental data is only available for Switzerland. The `outbreak.info` platforms (Mullen *et al.*, 2020) allows the generation of detailed reports, enriched by informative visualizations and infographics of the prevalence of distinct lineages and mutations in the viral genome. However, comparisons between genomics sequences are not supported. The Regeneron COVID-19 Dashboard (`https://covid19dashboard.regeneron.com/`) and CoronaTrend (`https://coronatrend.live/`) offer conceptually similar services, with more options to customize plots and visualization of the data, but support only the analysis of a single lineage or mutation at a time. COVID-19 CG (Chen *et al.*, 2021a) provides tools and methods for prioritizing SARS-CoV-2 mutations, which however works on just a subset of Spike mutations.

The VirusViz (Bernasconi *et al.*, 2021b) sequence data visualizer, previously developed by our group, supports a comparative analysis of SARS-CoV-2 genomics sequences. Some of the visual features of ViruClust were inspired by VirusViz; in comparison, ViruClust provides effective and orthogonal methods for organizing data analysis and new statistical instruments for prioritizing mutations, whereas VirusViz allows just a visual inspection.

3 Materials and methods

Our methods are applied to a big dataset of SARS-CoV-2 sequences (see Section 3.1) annotated with external knowledge information (see Section 3.2). The overall approach consists in comparing two sets of viral sequences, the *target* and the *background* groups, to identify amino acid changes (substitutions, insertions or deletions) that show a statistically significant difference frequency in the two datasets (see Section 3.3). Upon this mechanism, we built and deployed several services to semi-automatically analyze sets of sequences and identify relevant amino acid changes.

3.1 GISAID data

Data were downloaded as a single json file representation of the EpiCov^{TM} database from GISAID (`https://www.gisaid.org/` (Shu and McCauley, 2017)), on the basis of a specific Data Connectivity Agreement.

The following metadata were provided for every SARS-CoV-2 genome: accession ID, collection date, submission date, Pangolin lineage (Rambaut *et al.*, 2020), collection location, and the list of amino acid changes. Amino acid changes were indicated according to the notation used by GISAID. Each amino acid change was represented by the concatenation of the following elements: protein acronym, amino acid residue in the reference genome, relative position in the protein sequence and alternative amino acid residue or a dash (in case of deletions). Changes were indicated with respect to the reference genome sequence EPI_ISL_402124 (WIV04 (Okada *et al.*, 2020)).

All the records for which dates reported in the "collection date" metadata were not complete or not valid were discarded, resulting in a dataset that comprised about 96.6% of the total number of sequences (as of Aug 26th, 2021). Moreover, the location field was split into four levels (indicatively representing continent, country, region, and province); additional levels were dropped.

3.2 External knowledge data

Sequence and mutation data are complemented by a series of resources used to annotate additional information. While mutations and the reference genome annotation were directly derived from GISAID, we extract relevant information about domains, functionally characterized sites, and glycosylation sites directly from the UniProtKB resources (The UniProt Consortium, 2021) dedicated to SARS-CoV-2 proteins (see, e.g., `https://www.uniprot.org/uniprot/P0DTC2` for the Spike protein).

In addition, when specific lineages are analyzed, we provide a simple annotation of mutations that are increasingly observed in sequences of the specific lineage. We automatically extract such information from the ECDC dedicated website (`https://www.ecdc.europa.eu/en/covid-19/variants-concern`).

3.3 Sequence set comparison

As mentioned above, the core analysis is based on the comparison of two sets of viral sequences, the *background* and the *target* (e.g., in the case of an "Evolution in space" analysis, this could correspond to all the sequences assigned to the B.1 lineage worldwide vs. the B.1 sequences collected in Italy). The aim is to identify a group of amino acid changes that show a statistically significant difference in prevalence in the target compared with the background. For the sake of reproducibility, we here present a formal description of the analysis. In our system, we represent sequences in a data cube, where the facts are the pairs $\langle id, C \rangle$, corresponding to a sequence identifier along with a set C of the sequence amino acid changes with respect to the reference sequence EPI_ISL_402124. Each change is represented by the usual quadruple comprising the protein name, the reference amino acid, the relative position and the alternative amino acid (e.g., *Spike_D614G* indicates the alteration of the 614*th* amino acid of the Spike protein from aspartic acid to glycine). The dimensions associated to the sequence are its lineage p, as assigned by the most recent version available of Pangolin, the deposition date d and the location l, further organized into continent, country, region and province. Thus, the system allows to analyze a database

$$D = \langle id, C, p, d, l \rangle$$

composed of all the valid GISAID sequences along with the relevant metadata.

The user is provided with guided procedures to slice and dice the database D composing queries on (the combination of) the dimensions lineage, location and time interval, to build the background B and the target T sets of sequences. In particular, the lineages are organized within a hierarchical drop down menu; geographical data for the selection of the location are organized as four progressive pie-charts each allowing the selection at a different level of granularity; time-intervals are defined using two disjoint bar-plot sliders. The guided selection wizards, through which the user defines the two sets, ensures that the two sets are disjoint, i.e., $B \cap T = \emptyset$.

Then, for each amino acid change c, such that:

$$c \in \bigcup_{t}^{T} t.C,$$

i.e., all the changes that appear in at least one sequence of the target set, the number of occurrences of c in the background and in the target are computed as follows:

$$c_B = |b \in B : c \in b.C|$$

and

$$c_T = |t \in T : c \in t.C|.$$

The odds ratio of c is computed as:

$$o_c = \frac{c_T |B|}{c_B |T|}.$$

The o_c provides a simple but very informative quantitative indication of the over/under presence of c in the target set with respect to the background. Additionally, the system computes a χ^2 test for each change, to identify statistically significant differences in the prevalence of amino acid changes between the background and the target. To correct for multiple testing, p-values are adjusted by the Bonferroni correction.

The combination of odds ratio and associated p-value can be used to identify the relevant amino acid changes, i.e., the ones that clearly distinguish the target set from the background set.

4 Results

ViruClust is an open system providing a powerful and intuitive Web-based interface. The tool supports four main different types of analyses, each discussed in the following sections. They share the same visualization tools and typically consist of a short pipeline in which a user can:

- Visually define the *scope of the analysis* by selecting lineage, times, and space (e.g., see the time selector in the left part of Figure 1).
- Visualize the analysis result using *heat maps* (e.g., see the right part of Figure 1 and many other examples in the use case).
- Analyze results within *summary tables* that include, for given amino acid changes, relevant prevalence statistics (e.g., see Supplementary Figure S1).
- Compare the distribution of changes of two populations, denoted as target and background, using *annotated bar plots* (e.g., see Figure 2).

Locations are described using a 4-level hierarchy of levels, as provided by GISAID EpiCov^{TM} metadata: continent/country/region/province. Local filters can be used to adapt the visualizations, e.g., by choosing a protein of interest or setting thresholds on statistical tests.

4.1 Prevalence of lineages

The "Prevalence of lineages" analysis allows users to visually explore the prevalence of different lineages within a set of arbitrarily selected genomic sequences. Selection is performed by specifying an interval of time for the collection date and a geographic location of interest, at a given level of granularity (continent/country/region), thereby enabling a comparison of

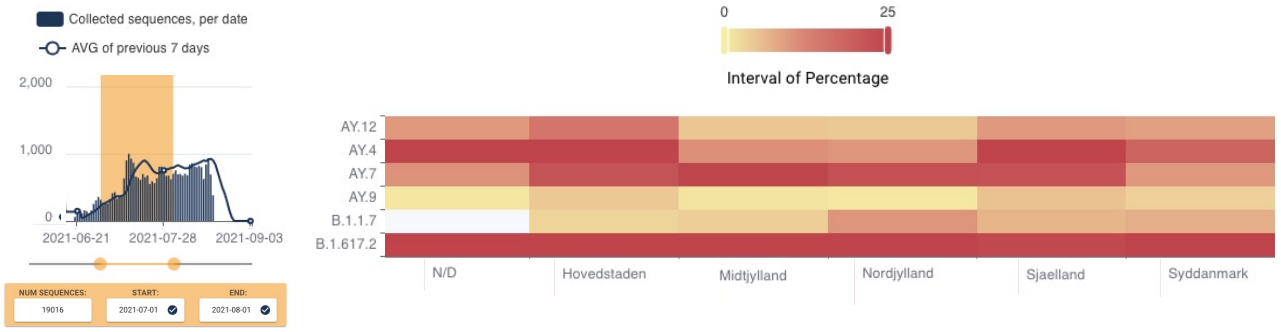

Fig. 1. Heat map showing the prevalence of the most common lineages across Denmark regions during July 2021; distinct lineages (AY) of the Delta variant appear in several regions.

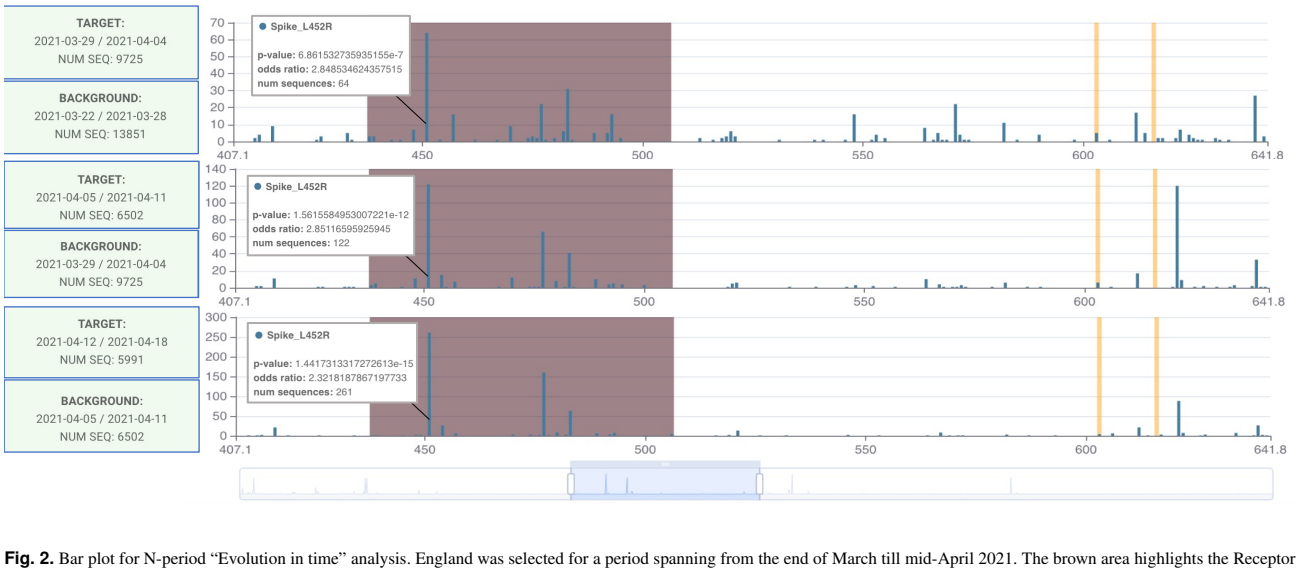

Fig. 2. Bar plot for N-period "Evolution in time" analysis. England was selected for a period spanning from the end of March till mid-April 2021. The brown area highlights the Receptor Binding Motif, positions in yellow correspond to glycosylation sites. The prevalence increase of characteristic changes of the Delta variant is clear, see positions 452 and 484.

the regions at a lower level of granularity. For example, in Figure 1, by choosing Denmark as country, the analysis is performed by considering the different geographic regions of that country. A histogram provides the distribution of the number of sequences collected from the selected locations over time; an interactive slider supports a graphical selection of the interval of interest in an informed manner. Results are provided in both a tabular and heat map format; filters can be used to include only lineages whose prevalence is above a user-defined threshold.

4.2 Evolution in time

The analysis starts by selecting one or more locations of interest, and possibly a lineage. Two types of analysis are available:

- Two-period comparison, respectively the target and the background period.
- N-period analysis, selected by setting fixed periods (weeks, months, given number of days) within a time interval.

In the two-period comparison mode, users select two distinct intervals respectively as target and background, by using sliders upon the histogram of the number of sequences collected from the selected locations. Prevalence of amino acid changes between the target and background are compared and mutations showing a statistically significant under/over representation are identified. Advanced filters can be applied to select amino acid changes associated with a specific protein, and/or satisfying user-defined prevalence constraints in the target and/or background. Additionally, p-values and odd ratio filters can be applied to select mutations associated with a desired level of significance.

The main output is presented in tabular form, and includes p-values, odd ratios and prevalence in the target and in the background. Columns can be reordered and additional annotations (e.g., mutations observed in the Spike glycoprotein of VOCS as defined by ECDC; or mutations with a prevalence >75% in a lineage) can be loaded to highlight important changes/residues – for an example of such table, see Supplementary Figure S1. The target-to-background comparison is also presented using bar plots.

In the N-period analysis, an interval of time is segmented into smaller intervals, and pairwise comparisons between consecutive intervals are performed. Users can select a start and end date, and subsequently, the span in time of the intervals that should be compared (weeks, months, or user-defined). The output consists of a collection of tables, one for every comparison, each in the same format described in the previous paragraph. A heat map representation of the data, comparing the prevalence of amino acid changes at the different intervals, is also produced to show trends in relative prevalence.

In addition, amino acid changes of each target period can be visualized and compared in the form of bar-chart plots. Figure 2 shows the "Evolution in time" analysis for England during three weeks, from the end of March 2021 till mid April 2021, when the Delta variant started emerging. Each week is used first as target against the previous week, then as a background against the following week. Diagrams are presented for given a protein (in this case, the Spike protein); a slider dynamically adjusts the scope of the visualizer. Absolute counts/prevalence of amino acid changes in the target group are displayed along with external annotations (i.e., protein domains, or functional annotations of single amino acid residues) to facilitate the

Table 1. Prevalence of selected amino acid changes in Spike and N

Mut	% India	% Asia (no India)	% World (no Asia)	% Kappa
Spike_D950N	72.66%	94.85%	97.55%	0.19%
Spike_E156G	29.43%	91.91%	94.17%	0.15%
Spike_F157-	29.54%	91.89%	93.55%	0.15%
Spike_R158-	29.65%	91.81%	94.20%	0.15%
Spike_G142D	24.41%	76.90%	89.33%	44.71%
N_G215C	29.91%	49.87%	67.33%	0.06%

Table 2. Prevalence of selected Spike changes in distinct regions of India

	Maharasthra	Telengana	West Bengal	Gujarat	Tamil Nadu	Andhra Pradesh	Karnataka
Spike_D950N	60.35%	80.18%	99.76%	71.60%	93.31%	88.27%	90.89%
Spike_E156G	21.42%	23.15%	90.17%	49.21%	<1%	<1%	51.30%
Spike_F157-	21.42%	23.15%	90.17%	49.21%	<1%	<1%	51.30%
Spike_R158-	21.42%	23.15%	90.17%	49.21%	<1%	<1%	51.30%
Spike_G142D	48.27%	5.17%	24.14%	82.72%	12.00%	4.82%	5.50%
N_G215C	31.13%	33.55%	12.13%	28.14%	45.20%	45.00%	44.70%

interpretation of the results and the identification of potentially important mutations/residues.

4.3 Evolution in space

The "Evolution in space" analysis allows a direct comparison between genomic sequences collected from a specific geographic location and matched sequences associated with a geographic entity of higher order. Selection of multiple locations is also supported. Users can optionally restrict the comparison to a specific lineage and interval of time. To define the target, a location of interest must be selected using the continent/country/region/province levels menu. The background is formed by considering geographic entities of higher level (World for Continent, Continent for Country, and so on). When multiple locations are selected (e.g., California, Nevada, Washington, and Utah), all pairwise comparisons are performed (e.g., California vs USA, Nevada vs USA, Washington vs USA, and Utah vs USA).

Final results are presented as one or more summary tables and, when multiple locations are selected, a heat map displaying the prevalence of different amino acid changes at each locations; results can also be visually inspected as a bar-chart on a given protein. In the heat map, all selected targets appear as columns and color maps may represent alternatively:

- the percentage of sequences in the target;
- the difference between target and background percentages;
- the odds ratio.

One example of such heat map is illustrated in Figure 4, which shows the spatial distribution of amino acid changes of the B.1.617.2 lineage in different geographic regions of India.

4.4 Custom analysis

The "Custom" analysis allows users to compare any 2 sets of genomic sequences, without any restriction. The target and background sets can be specified either by applying the set of filters and tools described in the previous sections or by providing directly GISAID accession ids for both the target and the background. Overlapping sequences present in both the target and background may be removed from both sets or from any one of them. When selected GISAID accession ids are used, at least 50 non-overlapping ids must be provided both for the target and background, to ensure the statistical soundness of the comparison.

For example, using the "Custom" analysis, it is possible to separately define the target as Europe/Italy/Lombardy and the background as Europe/Switzerland/Ticino, thereby comparing two regions that are not hierarchically related.

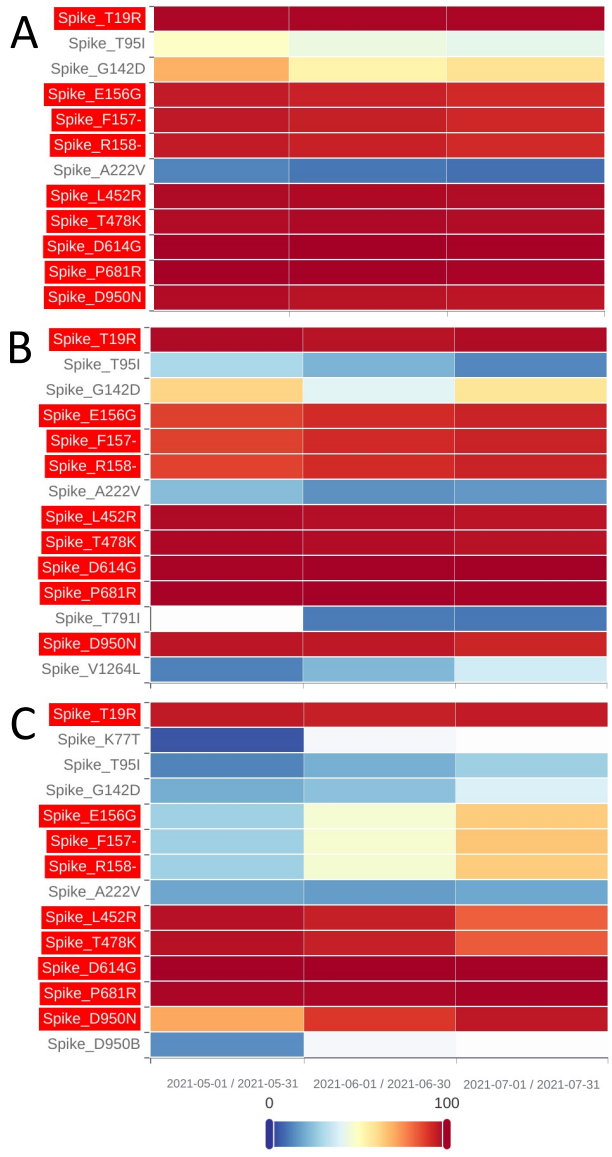

Fig. 3. Heat maps of allele frequencies of the Spike glycoprotein in the B.1.617.2 lineage between April and July 2021. Heat maps illustrate the prevalence in time of different amino acid changes in the Spike glycoptrotein in time A) Worldwide, B) in Asia C) in India. Changes are reported on the rows, intervals of time on the columns

5 Use case: study of evolution of the B.1.617.2 lineage

To demonstrate the effectiveness of ViruClust for studying the evolution of SARS-CoV-2, our tool was applied to identify potentially interesting patterns of mutation/evolution in B.1.617.2, the most prevalent, and first emerged lineage of the SARS-CoV-2 Delta variant. Delta is currently the most widespread variant of SARS-CoV-2 worldwide (Otto *et al.*, 2021). First identified in the state of Maharashtra (India) in late 2020 (Cherian *et al.*, 2021), this SARS-CoV-2 lineage spread throughout India, and in several countries worldwide, displacing the previously dominant B.1.1.7 (Otto *et al.*, 2021) (Alpha) variant and other pre-existing lineages.

Several independent lines of evidence suggest that not only Delta is about 60% (Li *et al.*, 2021) more efficient in infecting human cells than

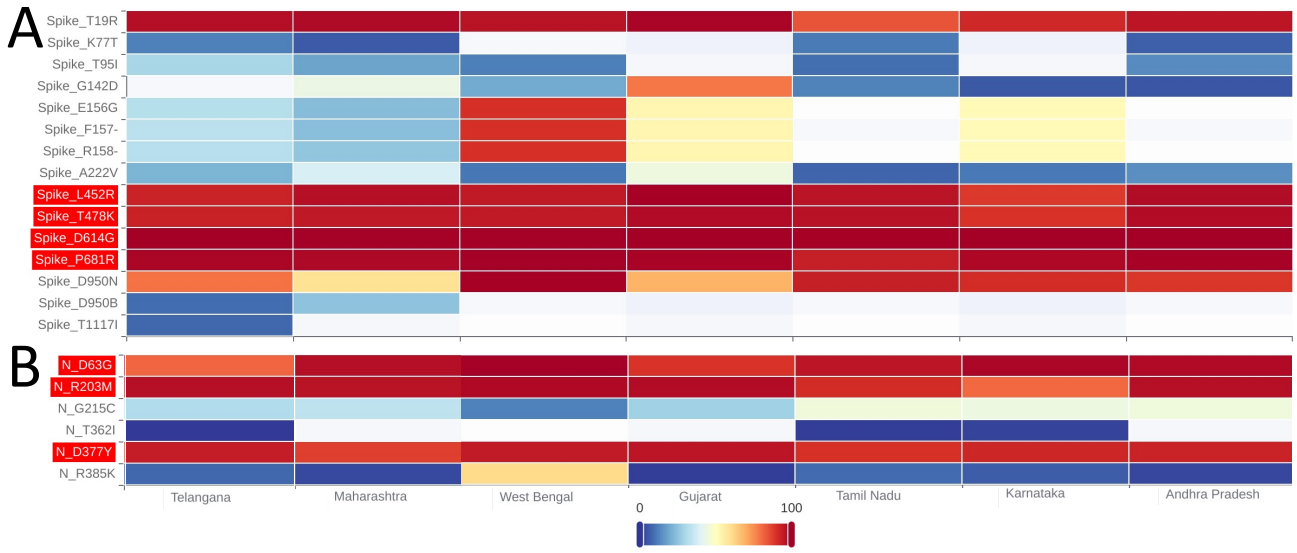

Fig. 4. Heat map of allele frequencies of the Spike glycoprotein (Panel A) and in the N nucleocapsid protein (Panel B) in the B.1.617.2 SARS-CoV-2 lineage in different Geographic Regions of India. The heat map reports the frequencies of different amino acid changes from distinct regions in India for which 1000 or more genome sequences are available. Changes are reported on the rows, geographic regions on the columns. Only changes with a frequency >= 5% at at least one interval of time are displayed.

any other variant of the virus, but also an enhanced ability to escape immune response (Planas *et al.*, 2021), thus explaining (at least in part) why this novel variant has been so successful in outcompeting other variants of SARS-CoV-2. Although mechanistic models are not currently available, characteristic mutations in the Spike glycoprotein and in the N nucleocapsid protein of the Delta variant have been recently reported to facilitate immune escape and/or enhance viral replication (Scudellari, 2021; Syed *et al.*, 2021; Liu *et al.*, 2021).

Direct comparison, by "Evolution in space" analysis, of SARS-CoV-2 B.1.617.2 lineage genomic sequences isolated from India and outside of India (Supplementary Tables S1A–D) highlights some interesting patterns of variation of the Spike glycoprotein in this VOC. While, as it should be expected, we observe a biased distribution of frequencies, with some amino acid changes being preferentially associated with distinct geographic areas (India, Asia outside of India or outside of Asia), we also notice that five amino acid changes in Spike (Spike_D950N; Spike_E156G, Spike_F157-, Spike_R158-; and Spike_G142D) – cumulatively observed in more than 75% of the SARS-CoV-2 genomes assigned to the B.1.617.2 lineage – have a significantly lower prevalence in India (Table 1). Several patterns/dynamics of evolution are compatible with this observation, including for example: i) founder effects in the propagation of the lineage outside of India; ii) later emergence and subsequent fixation of these allelic variants (not necessarily in India); and/or iii) convergent evolution and independent fixation. Notably, mutations at Spike residues 142 and 158 have been previously reported to potentially (McCallum *et al.*, 2021) contribute evasion of the innate immune response, by reducing the affinity to potent neutralising antibodies derived from convalescent donors sera, suggesting potentially important functional implications. Equivalent analyses on the N nucleocapsid protein (Supplementary Tables S2A–D) reveal a somewhat limited pattern of amino acid changes, with N_R203M, the single amino acid replacement recently reported to increase viral replication by approximately a 50 fold (Syed *et al.*, 2021), consistently showing a very high prevalence at every level of geographic granularity (India, Asia or worldwide). Interestingly, we observe an increase outside of India and worldwide in the prevalence of N_G215C. Similar to N_R203M and other amino acid substitutions in the N protein associated with improved viral replication, N_G215C is localized in the "linker" region of the protein and thus could bear potential functional implications.

"Evolution in time" analyses of allele frequencies of B.1.617.2 lineage genomes indicate that, while the prevalence of the five Spike changes reported above does not change significantly and is consistently above 80% when genomes isolated outside of India are considered (Figure 3A–B, Supplementary Figure S2A–B), each has a distinct and characteristic frequency in India, indicating independent patterns of fixation. Additionally, while our data support a steady and constant increase in prevalence, none of the changes seem to reach complete fixation when only genomes isolated from India are considered (Figure 3C, Supplementary Figure S2C). Interestingly, equivalent analyses on the N nucleocapsid protein (Supplementary Figure S3A–C) highlight a general and constant increase in the prevalence of N_G215C in time, suggesting a progressive fixation of this amido acid change in B.1.617.2.

Consistent with the phylogeographic origin of the lineage, these observations potentially indicate a higher genomic diversity in India. Interestingly (Table 2, Figure 4), spatial analysis of distinct geographic regions in India indicate an even more variable pattern of allele frequency of the Spike_D950N; Spike_E156G, Spike_F157-, Spike_R158-, Spike_G142D and N_G215C changes in that country, with some regions such as West Bengal, where some of the changes seem to have reached complete fixation, and others, such as Odisha and Chhattisgarh where these changes are barely observed. While these observations are consistent with possible founder effects in the propagation of the lineage outside of India, and potentially indicate a distinct pattern of fixation for the six changes, detailed analyses of the first (by deposition date) 1,500 B.1.617.2 genomes, as retrieved from GISAID (Supplementary Table S3) indicate that the first genomes of the B.1.617.2 lineage with this combination of changes were initially isolated outside of India, both in Asian and non Asian countries. This potentially suggests fixation out of India and a subsequent re-introduction in that country.

Notably, by executing a "Custom" analysis for the comparison of two different lineages, we observe that while the mutation at Spike G142, was already present and is also observed in the closely related Kappa (B.1.617.1) SARS-CoV-2 lineage, mutations at Spike residues 156-158 and 950 – along with two additional mutations at residue 478 and 19 –

form a haplotype of the Spike glycoprotein that is observed exclusively in Delta and never in Kappa (Supplementary Figure S1, Table 1), suggesting potential implications for the acquisition of novel/improved epidemiologically relevant features. Interestingly, we also observe that, while the N_R203M amino acid change – recently reported to improve viral replication by a > 50 fold (Syed *et al.*, 2021) – is shared between Delta and Kappa, the same consideration does not apply to the N_G215C mutation that is rapidly increasing in prevalence only in Delta.

While more detailed and specific evolutionary analyses (and potentially a more capillary sampling of early Delta variants genomes) will be required to disclose the exact chain of events that lead to the evolution of this highly pathogenic variant of SARS-CoV-2, we remark that the results herein described are consistent with recent reports of unexplored genetic diversity in the Delta SARS-CoV-2 lineage (Stern *et al.*, 2021), and uncover potentially important events in the evolution of the main lineage of this VOC.

Comparative analyses of SARS-CoV-2 B.1.617.2 genomic sequences were based on ViruClust dataset updated on August 15th 2021. Heat maps for the study of the prevalence of different amino acid changes of the Spike glycoprotein in different geographic regions were obtained by considering sequences deposited between 04-01-2021 and 07-31-2021, an interval of time for which a large number of sequences (>1000) was available for all the locations included in the analyses. Prevalence was compared at different months, so as to guarantee the inclusion of a sufficient number of sequences. Equivalent analyses were also repeated without imposing any constraints on time intervals, in order to derive the complete distribution. Only amino acid changes with a prevalence of at least 5% and at least one time point were included in the graphical representation. "Evolution in space" and "Custom" analyses were executed without imposing any temporal constraint and by considering the complete set of available sequences.

6 Conclusion

To the best of our knowledge, no currently available method enables the dynamic comparisons of user-selected groups of SARS-CoV-2 genomic sequences, across lineages, space and time. ViruClust offers the first complete platform with such a wide spectrum of interaction possibilities. Users are not only allowed to explore space and time distributions of already established lineages, but also to identify specific sequences extracted from the GISAID database that are characterized by specific mutational signatures in specific periods of time or given locations.

By enabling the rapid and systematic comparison of distinct groups of viral genomic sequences, ViruClust can greatly facilitate the discovery and identification of potentially important events in the evolution of SARS-CoV-2, the generation of "testable" evolutionary hypotheses, and ultimately the identification of important and relevant mutations of potential concern.

System availability

A documented docker version of ViruClust is available at `https://github.com/DEIB-GECO/Docker_ViruClust/`. The code of the Web application is available at `https://github.com/DEIB-GECO/ViruClust/` and on Zenodo at `https://doi.org/10.5281/zenodo.5524656/`. The WIKI documentation is at `https://github.com/DEIB-GECO/ViruClust/wiki/`.

Acknowledgements

The authors would like to thank Alba Grifoni, Carla Mavian, Brittany Rife Magalis, Marco Salemi and Shay Fleishon for useful discussions inspiring this research. The authors are grateful to the GISAID Initiative for the data sharing agreement that allowed the development of ViruClust. They also 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 shared via the GISAID Initiative.

Funding

ViruClust is supported by the ERC Advanced Grant 693174 "Data-Driven Genomic Computing (GeCo)".

References

Bernasconi, A. *et al.* (2020). Empowering Virus Sequence Research Through Conceptual Modeling. In G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, and H. C. Mayr, editors, *Conceptual Modeling*, pages 388–402, Cham. Springer International Publishing.

Bernasconi, A. *et al.* (2021a). EpiSurf: metadata-driven search server for analyzing amino acid changes on epitopes of SARS-CoV-2 and other viral species. *Database*, **2021**.

Bernasconi, A. *et al.* (2021b). VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. *Nucleic Acids Research*, **49**(15), e90.

Canakoglu, A. *et al.* (2019). GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database*, **2019**.

Canakoglu, A. *et al.* (2021). ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Research*, **49**(D1), D817–D824.

Chen, A. T. *et al.* (2021a). COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife*, **10**, e63409.

Chen, C. *et al.* (2021b). CoV-Spectrum: Analysis of globally shared SARS-CoV-2 data to Identify and Characterize New Variants. *arXiv preprint arXiv:2106.08106*.

Cherian, S. *et al.* (2021). SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms*, **9**(7).

Chiara, M. *et al.* (2021a). Comparative Genomics Reveals Early Emergence and Biased Spatiotemporal Distribution of SARS-CoV-2. *Molecular Biology and Evolution*, **38**(6), 2547–2565.

Chiara, M. *et al.* (2021b). Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics*, **22**(2), 616–630.

Hadfield, J. *et al.* (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**(23), 4121–4123.

Lauring, A. S. and Hodcroft, E. B. (2021). Genetic variants of SARS-CoV-2–what do they mean? *Jama*, **325**(6), 529–531.

Li, B. *et al.* (2021). Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *medRxiv*.

Liu, Y. *et al.* (2021). Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *bioRxiv*.

McCallum, M. *et al.* (2021). N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell*, **184**(9), 2332–2347.

Mullen, J. L. *et al.* (2020). Outbreak.info. `https://outbreak.info/`. Last accessed: Aug 26th, 2021.

Okada, P. *et al.* (2020). Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Eurosurveillance*, **25**(8), 2000097.

Otto, S. P. *et al.* (2021). The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr Biol*, **31**(14), R918–R929.

Planas, D. *et al.* (2021). Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature*, **596**(7871), 276–280.

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), 1403–1407.

Scudellari, M. (2021). How the coronavirus infects cells - and why Delta is so dangerous. *Nature*, **595**(7869), 640–644.

Shu, Y. and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**(13).

Stern, A. *et al.* (2021). The unique evolutionary dynamics of the SARS-CoV-2 Delta variant. *medRxiv*.

Syed, A. M. *et al.* (2021). Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles. *Science*, page eabl6184.

The UniProt Consortium (2021). Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**(D1), D480–D489.

Yang, H.-C. *et al.* (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**(48), 30679–30686.