Journal of proteome research

pubs.acs.org/jpr

Article

CoVe-Tracker: An Interactive SARS-CoV-2 Pan Proteome Evolution Tracker

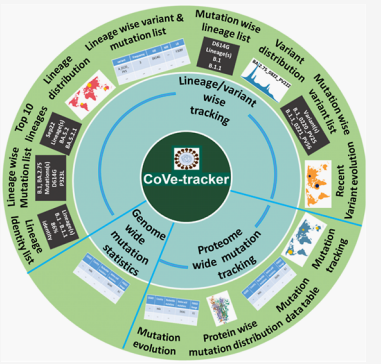
Chakkarai Sathyaseelan, Murali Aadhitya Magateshvaren Saras, L. Ponoop Prasad Patro, Patil Pranita Uttamrao, and Thenmalarchelvi Rathinavelan*

Cite This: *J. Proteome Res.* 2023, 22, 1984−1996

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: SARS-CoV-2 has significantly mutated its genome during the past 3 years, leading to the periodic emergence of several variants. Some of the variants possess enhanced fitness advantage, transmissibility, and pathogenicity and can also reduce vaccine efficacy. Thus, it is important to track the viral evolution to prevent and protect the mankind from SARS-CoV-2 infection. To this end, an interactive web-GUI platform, namely, CoVe-tracker (SARS-CoV-2 evolution tracker), is developed to track its pan proteome evolutionary dynamics (https://project.iith.ac.in/cove-tracker/). CoVe-tracker provides an opportunity for the user to fetch the country-wise and protein-wise amino acid mutations (currently, 44139) of SARS-CoV-2 and their month-wise distribution. It also provides position-wise evolution observed in the SARS-CoV-2 proteome. Importantly, CoVe-tracker provides month- and country-wise distributions of 2065 phylogenetic assignment of named global outbreak (PANGO) lineages and their 177564 variants. It further provides periodic updates on SARS-CoV-2 variant(s) evolution. CoVe-tracker provides the results in a user-friendly interactive fashion by projecting the results onto the world map (for country-wise distribution) and protein 3D structure (for protein-wise mutation). The application of CoVe-tracker in tracking the closest cousin(s) of a variant is demonstrated by considering BA.4 and BA.5 PANGO lineages as test cases. Thus, CoVe-tracker would be useful in the quick surveillance of newly emerging mutations/variants/lineages to facilitate the understanding of viral evolution, transmission, and disease epidemiology.



KEYWORDS: *SARS-CoV-2, pan proteome, viral evolution, variant tracker, lineage tracker, mutation tracker, web-GUI database*

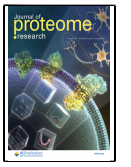
INTRODUCTION

Since its emergence in December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become a major global healthcare challenge. Among the three outbreaks caused by coronaviruses in the past 2 decades, *viz.*, severe acute respiratory syndrome coronavirus (SARS-CoV), middle east respiratory syndrome coronavirus (MERS CoV), and SARS-CoV-2,[1] the ongoing SARS-CoV-2 pandemic is the deadliest. As reported by the World Health Organization (WHO), the virus has infected 753 million people and has caused 6.8 million deaths in 224 countries and territories around the world as of January 31, 2023.

The transmission of SARS-CoV-2 from person to person leads to changes in the viral genome/proteome and may assist the virus in escaping from the host immune response. One such early example is the D614G mutation in the spike protein of SARS-CoV-2 which has replaced the wild-type and has proven to have efficient replication and transmission.[2] Thus, mutations in the SARS-CoV-2 genome have led to the emergence of several new variants,[3,4] among which some of them have become variants of concern (VOCs)[5] (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/) and have led to many successive waves of SARS-CoV-2. Notably, the SARS-CoV-2 variants emerged in the United Kingdom (known as 20I/501Y.V1, VOC 202012/01, or B.1.1.7),[6] South Africa (known as 20H/501Y.V2 or B.1.351),[7] and Brazil (known as P.1)[8] have been declared as VOCs by the WHO due to their higher viral load and enhanced transmissibility.[9,10] Alarmingly, the recently emerged fast-spreading heavily mutated omicron's subvariants such as BA.4, BA.5, BA.2.74, BA.2.75, BA.2.76, XBB, BQ.1, and so forth pose a major challenge in different parts of the world, leading to the next wave of omicron.[11,12] Such a fast evolution of SARS-CoV-2 complicates the action of existing vaccines and,[11,13−16] thus, necessitates the update on the current vaccines. Still, there are possibilities for the virus to incorporate changes in the genome which perhaps lead to the emergence of lethal variants, necessitating a systematic tracking of the viral evolution.

Luckily, the explosion of the SARS-CoV-2 genome sequence during the past 3 years has put the world in a better position to analyze, track, and understand the viral evolution. However, due

Received: February 3, 2023
Published: April 10, 2023



to the lack of a proper surveillance system, it becomes more challenging to track the spread of SARS-CoV-2 variant(s).[17] This necessitates a dedicated variant surveillance system. Considering the need to track SARS-CoV-2 proteome mutations and variants, a user-friendly, web-GUI database, namely, CoVe-tracker (SARS-CoV-2 pan proteome evolution tracker) (https://project.iith.ac.in/cove-tracker/), has been developed here by using the high-coverage SARS-CoV-2 complete genome sequences deposited in GISAID[18] as of January 2023. Silent mutations in the genome may not directly reflect the actual amino acid change in the protein. Since proteome mutations will take care of such silent mutations due to the amino acid degeneracy, proteome mutations have been focused on here. Furthermore, proteome mutations may quickly give a clue about the functional impact of a mutation. Thus, CoVe-tracker provides information about the prevalence of various SARS-CoV-2 proteome mutations and the associated variants disseminated across 200 countries in an interactive mode. Although various SARS-CoV-2 mutation databases[19−21] are currently available, CoVe-tracker is different from others as it provides a mutation-based variant classification of various SARS-CoV-2 lineages. CoVe-tracker provides a periodic update on SARS-CoV-2 evolution, and it facilitates the quick identification of variants that are circulating domestically and internationally. Thus, a variant can be identified within a short period of its emergence by CoVe-tracker before it is over-represented globally.

MATERIALS AND METHODS

Creation of SARS-CoV-2 Whole Proteome and Genome Local Repositories

5.7 million high-coverage (*viz.*, the number of undefined nucleotides "N"(s) were less than 1% ; which means that the sequences having less than or equal to 300 Ns in the SARS-CoV-2 genome of 30kb size) SARS-CoV-2 whole genome sequences of human origin which were collected on or before December 31, 2022, and deposited in GISAID on or before January 3, 2023, were downloaded. It is noteworthy that a cutoff criterion of ≤1% was used for filtering the sequences with "N" while downloading the sequence from GISAID. If the presence of one or more "N"s in a genomic sequence affects the protein translation, then the sequence was excluded from the analysis. As of December 31, 2022, a total of 2.4 million (out of 5.7 million) such SARS-CoV-2 whole genome sequences were excluded in the database. Subsequently, the downloaded sequences were verified manually and organized in a month-wise and country-wise manner. Finally, ∼3.3 million filtered genomic sequences were considered for the creation of the database. In the next step, these filtered sequences were compared with the first published SARS-CoV-2 sequence (Genbank Id: NC_045512.1)[22] using nucleotide BLAST (BLASTn) of BLAST 2.6.0+.[23] The sequences having alignment bit scores in the range of 54674 to 54975 with a statistical significance value (P) < 0.01 were alone considered for further analyses. After aligning each query genome against the reference genome, the genomic mutations corresponding to each protein-coding region were fetched from BLASTn output. The entire process was automated using Python and bash scripts. The coding regions of these sequences were further translated *in silico* into individual proteins using the in-house Python script and were subjected to comparative proteome analyses. Protein BLAST (BLASTp) was carried out individually for each protein of the SARS-CoV-2 proteome against the corresponding translated reference protein. The BLOSUM62 scoring matrix was used for the alignment. In the next step, mismatch(s), deletion(s), and insertion(s) were automatically fetched from the BLASTp output (using the in-house Python code) of the 26 proteins (coded from the canonical ORFs) of the SARS-CoV-2 genome and stored in an orderly manner based on the country of origin and month of sequence collection (Figure S1). The above process was executed in an automated fashion to carry out the alignment of each proteome sequence against the reference sequence using a bash script. Finally, the overall statistics of each mutation were calculated.

Deriving the Amino Acid and Nucleotide Mutations of SARS-CoV-2 Sequences

As discussed in our previous studies,[24−28] the variations (mutations) in the amino acid sequences such as substitutions, deletions, and insertions corresponding to a total of 26 proteins were automatically obtained in the CSV file format and used as a master file in the custom data visualization. Using this master data file, the high (percentage frequency (PF) greater than 10%), moderate (PF = 1−10%), and low (PF below 1% but occur at least three times) recurring mutations were calculated as discussed elsewhere.[24−27] Additionally, the nucleotide mutations in the entire genome of SARS-CoV-2 were collected in a master ".csv" file and utilized in data visualization. The entire analysis was carried out using an in-house Python script.

PANGO Lineage-Based Classification of SARS-CoV-2 Variants and Creation of a Local Repository

Using the SARS-CoV-2 sequences stored in the local repository, the phylogenetic assignment of named global outbreak (PANGO) lineage (which describes the viral evolution) corresponding to each SARS-CoV-2 sequence was obtained using the pangolin Python package.[29] Subsequently, a local repository of SARS-CoV-2 proteome variants was created which were grouped according to their PANGO lineages. Here, each SARS-CoV-2 proteome variant was assigned a protein variant (PV) name. The PV nomenclature was followed in such a way that the lineage corresponding to the variant was given in the first place, followed by the month and year of the first occurrence of the variant, which again was followed by a randomly assigned number (Figure S1). All these processes were done in an automated fashion using in-house bash and Python scripts.

Subsequently, the PANGO lineage and variant information along with the proteome and genome mutations master files (that contain the country of origin and the date of sequence collection) were used in the creation of a relational database using MySQL (https://www.mysql.com/).

IMPLEMENTATION OF COVE-TRACKER WEB-GUI

Using the above data, an interactive SARS-CoV-2 database named CoVe-tracker (SARS-CoV-2 pan proteome evolution tracker) was implemented with the help of the Apache HTTP server (https://httpd.apache.org) and D3.js (https://d3js.org/). The client-side user interface was implemented using HTML and PHP to enable access to the SARS-CoV-2 PANGO lineage, variant, and mutation information. Plotly Dash App (https://dash.plotly.com/) was employed for the user-friendly display of the country-wise distribution of PANGO lineages, variants, and amino mutations as well as for the month-wise distribution information of PANGO lineages, variants, and amino mutations. The JQuery Ajax method was employed for the dynamic user interface to visualize the lineage-wise variant and mutation list,

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

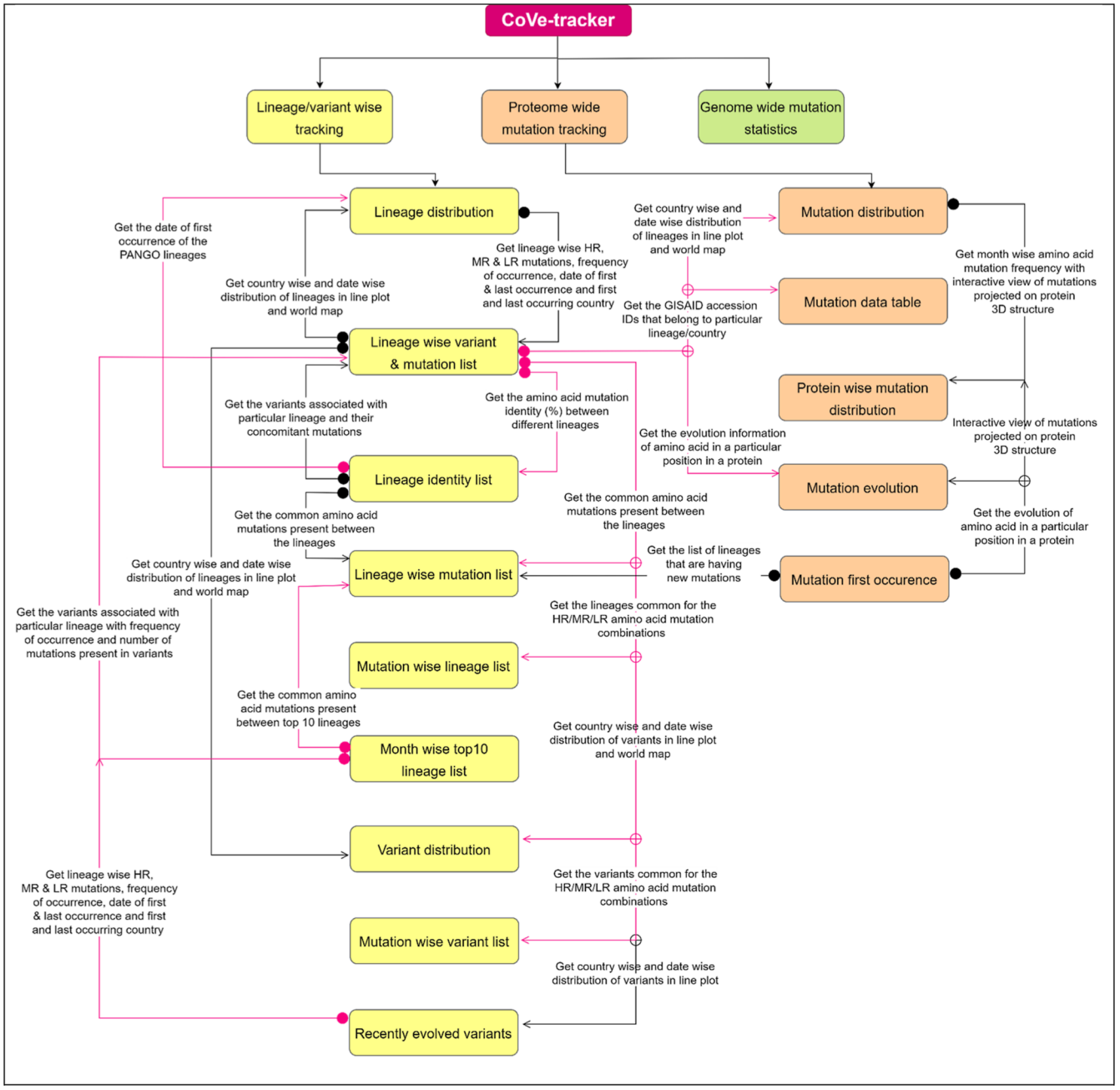


Figure 1. Overview of different modules of CoVe-tracker and their interlinks. Lineage/variant wise tracking, proteome wide mutation tracking, and genome wide mutation statistics are the three main menus available in the CoVe-tracker database. Note that there are nine and five modules under lineage/variant wise tracking and proteome wide mutation tracking menus, respectively. The arrow marks represent the interlink between different modules.

lineage identity list, lineage-wise mutation list, mutation-wise lineage list, mutation-wise variant list, and mutation evolution data. To provide structure-based insights about the protein mutations, the JSmol (http://www.jmol.org/) molecular viewer Java applet was added to visualize the location of recurrent mutations on the three-dimensional (3D) structures of SARS-CoV-2 proteins. To display the mutations on the 3D structure, I-TASSER structures were used[30−33] except in the case of the spike protein (PDB ID: 6XR8). The CoVe-tracker database is freely accessible through https://project.iith.ac.in/cove-tracker/.

Functionality of the CoVe-Tracker Web Interface

CoVe-tracker comes with the following three main menus: Lineage/variant wise tracking, Proteome wide mutation tracking, and Genome wide mutation statistics. The "Lineage/variant wise tracking" menu bar has nine submenus, using which the user can track country-wise and month-wise distribution of any lineage and variant and obtain lineage-wise variant and mutation lists, lineage identity lists, lineage-wise mutation lists, mutation-wise lineage and variant lists, month-wise top 10 lineage list, and the variant evolution trend during the past 2 months. Under the "Proteome wide mutation tracking" option, the following submenus are displayed: mutation tracking, mutation data table, protein-wise mutation distribution,

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

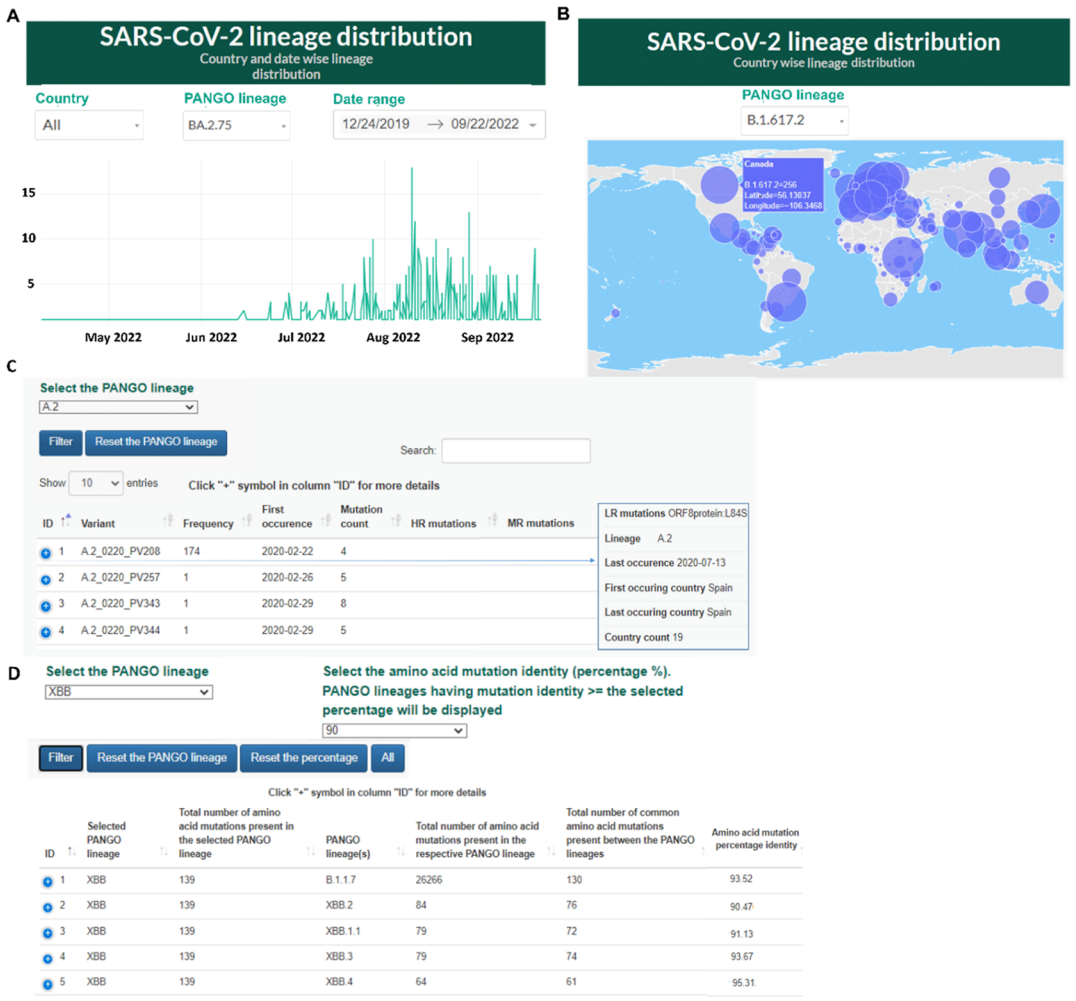


Figure 2. Examples illustrating the retrieval of PANGO lineage-associated information using CoVe-tracker. (A) Snapshot illustrating the country-wise and date-wise evolutionary dynamics of a selected PANGO lineage. (B) Global distribution of the selected PANGO lineage. (C) Table displaying the variant and mutation lists for a selected PANGO lineage. The user can click the (+) symbol in the ID column to view the details such as the country and date of first occurrence and the country and date of last occurrence corresponding to a selected variant. Refer to Figure 4A for variant ID nomenclature. (D) A snapshot demonstrates the "Lineage identity list" option of CoVe-tracker which displays the closest PANGO lineage(s) of a user-selected PANGO lineage. Note that the amino acid mutation identity (%) between the PANGO lineages is used to fetch the closest relative(s).

mutation evolution, and the first occurrence of the mutation. The "Genome wide mutation statistics" main menu provides a scatter plot of position-wise nucleotide mutation frequency and a mutation data table. The lineage, variant, amino acid mutation, and genomic mutation information of SARS-CoV-2 can be accessed through the respective menu bar. Thus, these options in CoVe-tracker provide the country-wise statistics of all the existing variants, lineages, and mutations. The month-wise statistics of the lineages and variants provide their emergence and dissemination during the selected timeline. Notably, CoVe-tracker also provides high, moderate, and low recurring (LR) amino acid mutations. CoVe-tracker further lists the month-wise top 10 lineages and the concomitant variants to track the evolution of a new lineage/variant. Thus, CoVe-tracker provides detailed information about the distribution and evolution of SARS-CoV-2 lineages, variants, and mutations. The details about the various modules available in CoVe-tracker and their interlinks are schematically represented in Figure 1. The "documentation" menu bar leads to the documentation page which provides stepwise information for the user to access the data.

RESULTS AND DISCUSSION

Monitoring the Lineage and Variant Distribution Using CoVe-Tracker

Accessing PANGO Lineage Distribution. Tracking the viral evolution through the PANGO lineage will further be useful to understand the viral evolution. In this context, the available lineages of SARS-CoV-2 were derived along with their country- and month-wise statistics using the high-coverage SARS-CoV-2 sequences (see above). As of December 31, 2022, there are 2065 SARS-CoV-2 PANGO lineages and are listed under the "Lineage distribution" submenu of the main menu "Lineage/variant wise tracking". The user can track the date-wise and country-wise dissemination of a PANGO lineage through the "Lineage distribution" submenu. Here, the user can select the country of their interest from the dropdown "Country" along with the PANGO lineage of their interest ("from PANGO Lineage" dropdown) to get the month-wise evolutionary dynamics (in terms of frequency of occurrence) of a particular lineage in the selected country in a specific time period in a line plot (Figure 2A). This submenu also gives an interactive worldwide distribution of a selected PANGO lineage (Figure

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

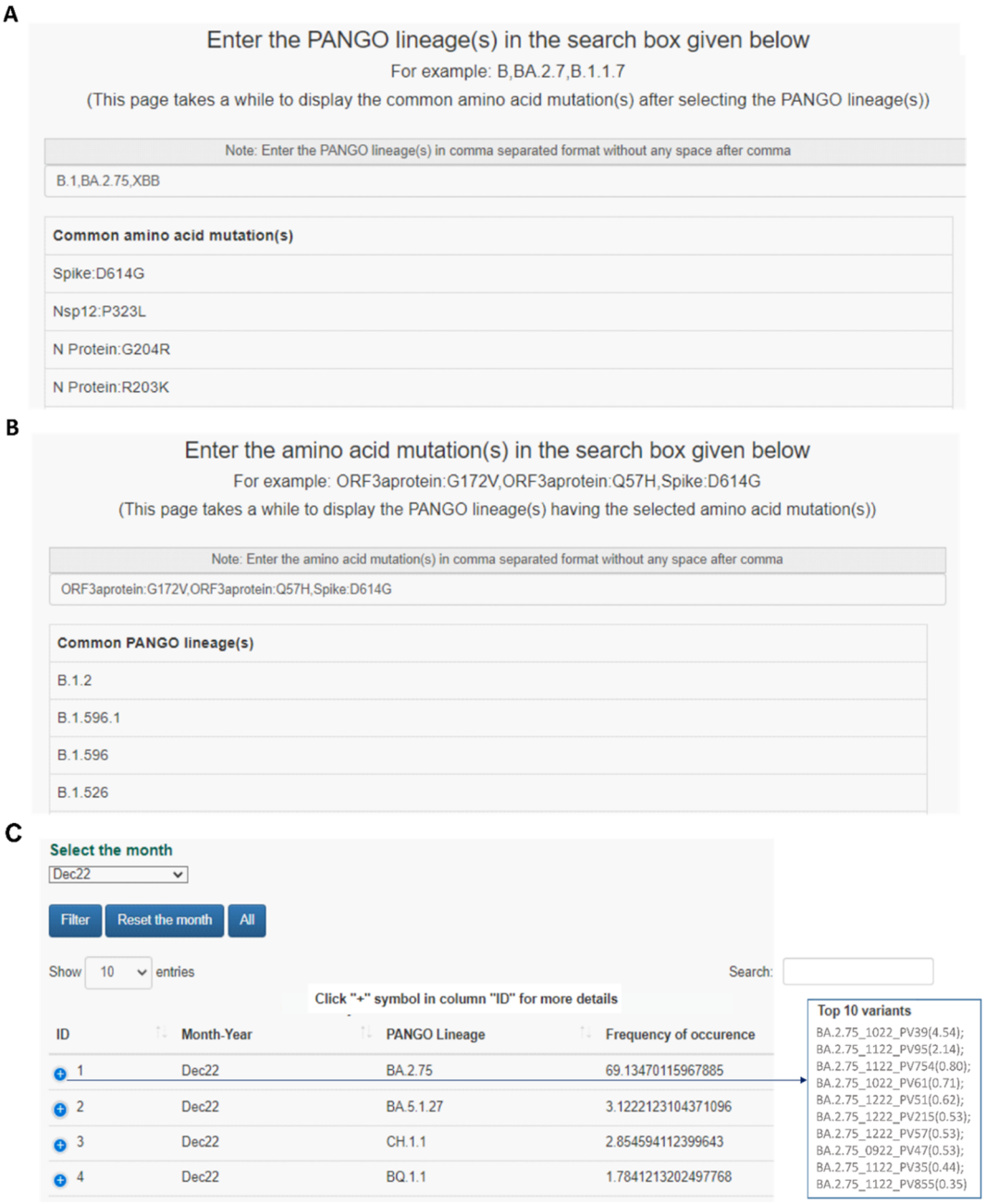


Figure 3. Examples illustrating the provisions in CoVe-tracker to (A, B) relate the mutations and PANGO lineages and (C) fetch month-wise top 10 lineages/variants. (A) Snapshot illustrating the option "Lineage wise mutation list" which displays the list of mutations that are common to user-selected PANGO lineages. (B) Snapshot illustrating the option "Mutation wise lineage list" which displays the list of PANGO lineage(s) that have the user's choice of HR and/or MR and/or LR mutations. (C) Month-wise list of top 10 lineages and the associated top 10 variants which can immediately provide information about the emergence of a new lineage/variant.

2B). To further view the mutation(s) and variant(s) associated with a lineage, one can use the submenu "lineage wise variant and mutation list" (Figure 2C). Additionally, the "Lineage identity list" submenu provides the closest lineage(s) for the user-selected lineage which can be traced with the help of commonly occurring LR (*viz.*, <1% and should appear in at least three sequences), MR (moderately recurring, *viz.*, between 1 and 10%), and HR (highly recurring, *viz.*, above 10%) mutations (Figure 2D).

Furthermore, the "Lineage wise mutation list" submenu can be used for getting the amino acid mutations that are common between any number of user-selected PANGO lineages (Figure 3A). Conversely, the "Lineage identity list" submenu and the "Mutation wise lineage list" submenu give the PANGO lineage associated with one or more mutations (Figure 3B). Furthermore, the "Month wise top 10 lineage list" submenu provides the list of month-wise occurrences of the top 10 PANGO lineages and their concomitant top 10 variants. For this, the user has to select the month from the dropdown "select the month" to get the frequency of occurrence of the top 10 PANGO lineages and the concomitant top 10 variants (Figure 3C).

Accessing Variant Distribution. Although the PANGO lineage provides information about the evolution of a viral

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

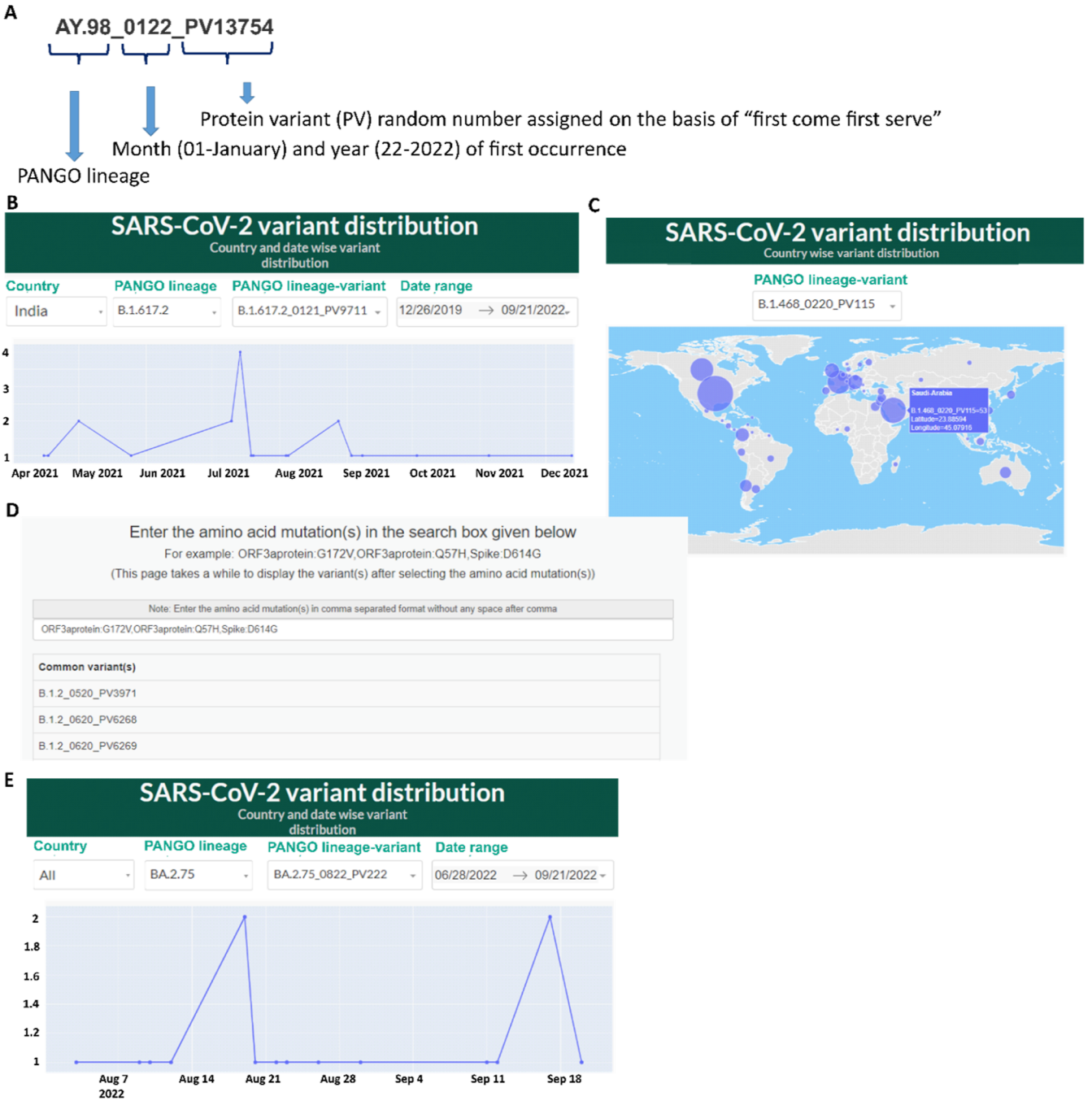


Figure 4. Examples illustrating the retrieval of variant-specific information using CoVe-tracker. (A) Example describing the variant naming scheme (see text). (B) Snapshot showing the country-wise and date-wise evolutionary dynamics of a selected variant that belongs to a PANGO lineage. (C) Global distribution of the selected variant. (D) Snapshot illustrating the display of variant(s) which have the user's choice of HR and/or MR and/or LR mutations. (E) Snapshot illustrating the country-wise and date-wise evolutionary dynamics of variants that have emerged for the first time during the past 2 months.

lineage, multiple variants come under a particular PANGO lineage due to the difference in the amino acid mutations. For example, the B.1.617.2 lineage has 8606 variants (occurs ≥3 times) due to the difference in the amino acid mutations. Thus, CoVe-tracker classifies different variants under each lineage and each variant is assigned a PV identifier. The variant name is coined based on the PANGO lineage and month and year of occurrence followed by a random number which is assigned based on "first come first serve" (Figure 4A). This PV ID is helpful in distinguishing different variants under the same PANGO lineage. For example, the B.1.617.2_0421_PV44577 and B.1.617.2_0521_PV48574 variants which come under the B.1.617.2 lineage are given different PV IDs due to the difference in their amino acid mutations. B.1.617.2_0421_PV44577 has 31 mutations (Nsp3:A488S, Nsp3:P1228L, Nsp3:P1469S, Nsp3:A1711V, Nsp4:V167L, Nsp4:T492I, Nsp6:T77A, Nsp12:P323L, Nsp12:G671S, Nsp13:P77L, Nsp14:A394V, Spike:T19R, Spike:T95I, Spike:G142D, Spike:E156G, Spike:F157, Spike:R158, Spike:L452R, Spike:T478K, Spike:D614G, Spike:P681R, Spike:D950N, ORF3aprotein:S26L,

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

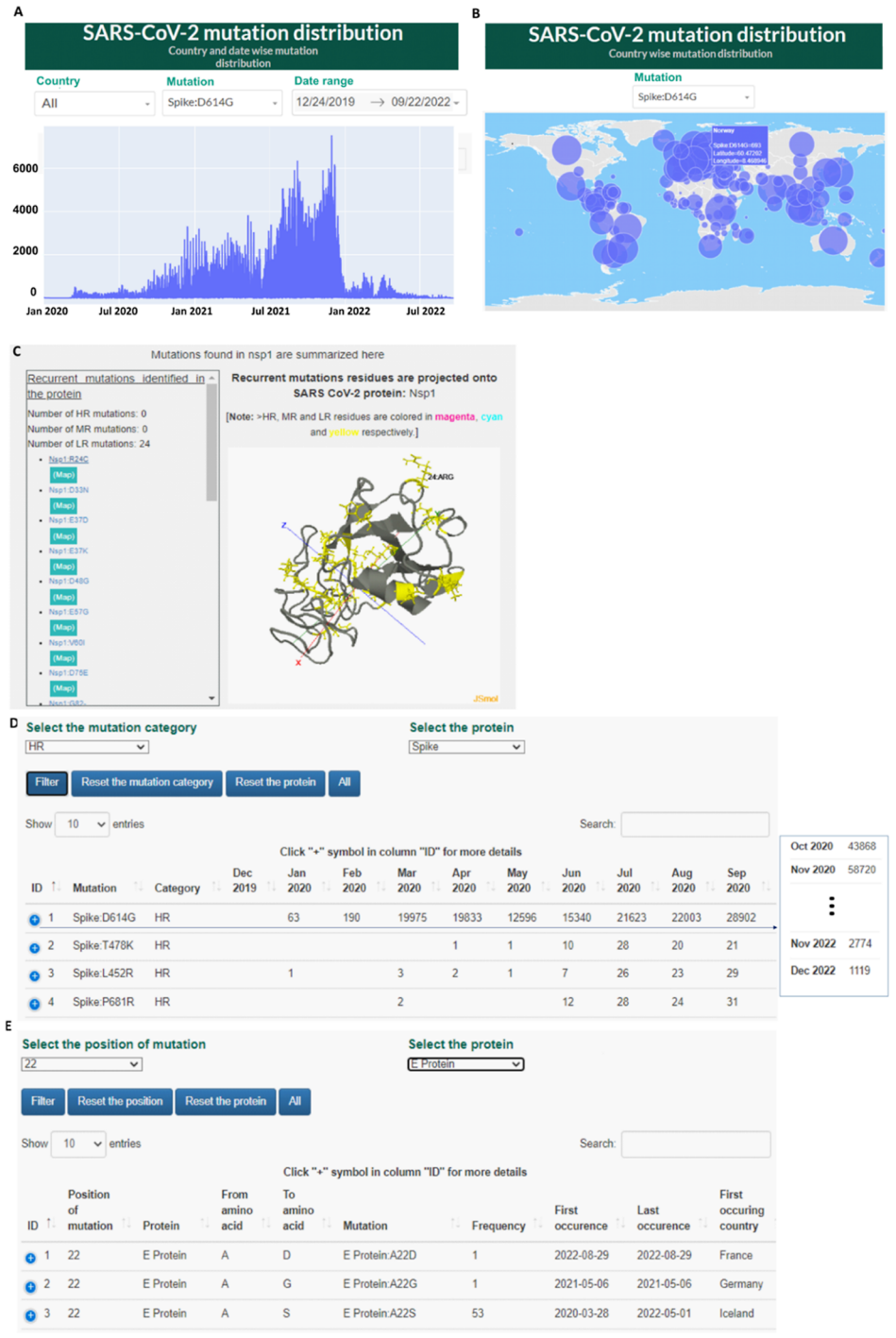
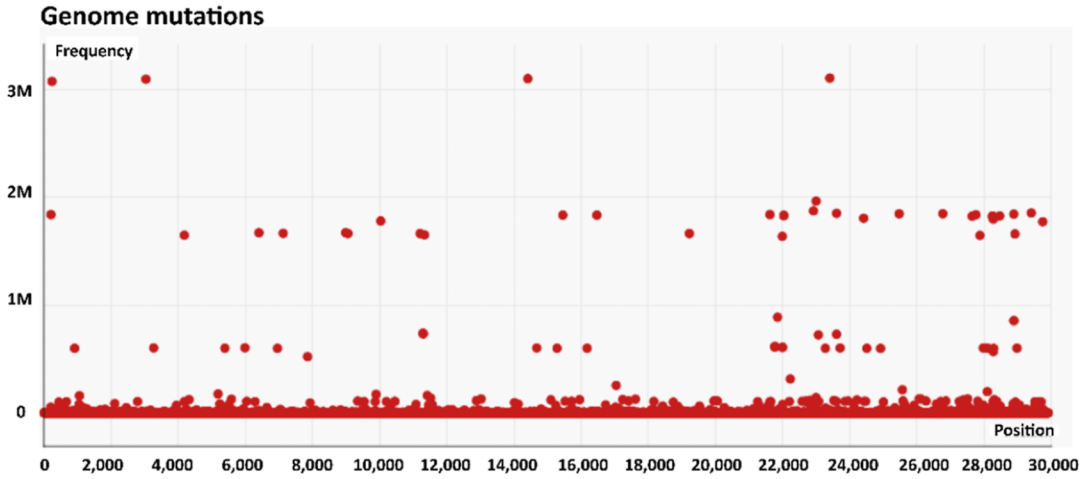


Figure 5. Examples illustrating the retrieval of proteome-wide/protein-wise amino acid mutation information from CoVe-tracker. (A) Snapshot showing the country-wise and date-wise evolutionary dynamics of a selected amino acid mutation. (B) Global distribution of the selected amino acid mutation. (C) Panel displaying the list of Nsp1 mutations and their projection onto the protein 3D structure in the JSmol interactive viewer (here, cyan- and yellow-colored spheres represent the MR and LR mutations, respectively). (D) Mutation table that lists all the mutations in a protein along with their month-wise occurrence information (see text for details). (E) Mutation evolution table that lists all the mutations in a protein along with their position(s), frequency of occurrence, country and date of first occurrence, and country and date of last occurrence (see text for details). Note that the corresponding wild-type amino acid is also displayed.

M protein:I82T, ORF7aprotein:V82A, ORF7aprotein:T120I, ORF7bprotein:T40I, N Protein:D63G, N Protein:R203M, N Protein:G215C, and N Protein:D377Y), and B.1.617.2_0521_PV48574 possesses 26 amino acid mutations (Nsp3:P822L, Nsp4:A446V, Nsp6:T10I, Nsp6:V149A, Nsp6:T181I, Nsp12:P323L, Nsp12:G671S, Nsp13:P77L, Spike:T19R, Spike:G142D, Spike:E156G, Spike:F157-, Spike:R158, Spike:A222V, Spike:L452R, Spike:T478K, Spike:D614G, Spike:P681R, Spike:D950N, ORF3aprotein:S26L, M protein:I82T, ORF7aprotein:V82A, ORF7aprotein:T120I, N Protein:D63G, N Protein:R203M, and N Protein:D377Y). As of December 31, 2022, there are 177546 SARS-CoV-2 variants whose country-wise and month-wise dissemination along with the frequency of occurrence in a specific country in a specific

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

A



B

User can filter the data based on GISAID ID, Country, Nucleotide mutations, Amino acid mutations, GISAID clade, PANGO lineage and Protein variant ID by entering the keyword in search option.

Show 10 entries Search: China

GISAID ID	Country	Nucleotide mutations	GISAID clade	PANGO lineage	Protein variant ID
EPI_ISL_1039208	China	T565C;C17825T;T21784A;C23525T;G29573A;	L	B	B_0420_PV5216
EPI_ISL_1040031	China	C8782T;G18598A;T28144C;C29095T;C29586T;	S	A	A_0220_PV176
EPI_ISL_1040034	China	C8782T;T28144C;	S	A	A_0120_PV27
EPI_ISL_1040036	China	C1077T;C8782T;G23438T;T28144C;G28878A;G29742A;	S	A	A_0220_PV147
EPI_ISL_1040037	China	C8782T;C16375T;C19185T;T21991-;T21992-;A21993-;T28144C;	S	A	A_0220_PV160
EPI_ISL_1040040	China	C8782T;G18598A;T28144C;C29095T;C29586T;	S	A	A_0220_PV176
EPI_ISL_1040042	China	G11083T;C21219T;	O	B	B_0120_PV43
EPI_ISL_1040046	China	A1015G;C8782T;G17594A;A26664G;T28144C;	S	A	A_0220_PV275
EPI_ISL_1040047	China	C6982T;C19386T;G22081A;T29483G;	L	B	B_0220_PV276
EPI_ISL_1040048	China	C8782T;T18603C;G20683T;C27925T;T28144C;C29095T;	S	A	A_0220_PV299

Showing 1 to 10 of 1,238 entries (filtered from 3,164,111 total entries) Previous 1 2 3 4 5 … 124 Next

Figure 6. Examples illustrating the retrieval of genome-wide nucleotide mutation information using CoVe-tracker. (A) Scatter plot displaying the frequency of occurrence of position-wise nucleotide mutations. (B) Table displaying all the nucleotide and amino acid mutations of SARS-CoV-2 found across different countries along with the corresponding PANGO lineage and PV ID (Refer Figure 4A). The user can also filter the data by entering a country name or GISAID accession ID or nucleotide mutations or amino acid mutation or PANGO lineage or PV ID in the search text box.

period can be accessed through the "Variant distribution" submenu (Figure 4B). The user can also obtain an interactive worldwide distribution of a variant (among 177546 proteome variants) (Figure 4C). On the other hand, the "mutation-wise variant list" submenu provides the list of variants associated with one or more mutations (Figure 4D). The submenu "Recently evolved variants" provides information about the country-wise and month-wise frequency of occurrence of new variants that have emerged for the first time during the past 2 months (from the date of the last update of the database) (Figure 4E). Figure 4 illustrates the tracking of the delta variant PANGO lineage and the associated subvariants.

Tracking the Amino Acid Mutations of the SARS-CoV-2 Proteome

To track the distribution of amino acid mutations of the SARS-CoV-2 proteome, the user has to choose the "Proteome wide mutation tracking" menu. The submenu "Mutation tracking" provides country-wise and date-wise distribution of the amino acid mutations (currently, 47901) in a similar fashion as the lineage (Figure 2A) and variant (Figure 4B) distributions (Figure 5A,B). For example, the user can select the country of choice from the dropdown "Country" along with a choice of the mutation (using the dropdown) to get its month-wise evolutionary dynamics (in terms of its frequency of occurrence) (Figure 5A). The user also has the flexibility to choose the timeline. Additionally, the user can view the projection of the frequency of occurrence of the selected amino acid mutations on the world map (Figure 5B).

The "Protein wise mutation distribution" submenu provides the individual list of mutations found in all the 26 SARS-CoV-2 proteins. This submenu lists the proteins into three categories: nonstructural proteins, accessory proteins, and structural proteins. The nonstructural proteins (Nsp) category consists of 16 proteins, namely, Nsp1, Nsp2, Nsp3, Nsp4, Nsp5, Nsp6, Nsp7, Nsp8, Nsp9, Nsp10, Nsp11, Nsp12, Nsp13, Nsp14, Nsp15, and Nsp16. The accessory proteins category consists of ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 proteins. The structural protein category consists of spike, E, M, and N proteins. When the user clicks the protein of interest (for example, Nsp1), an interactive panel lists the option to choose the LR, HR, and MR mutations present in the protein. Note that 0.12 million sequences having insertions in the range of 1 to 6 amino acid length occur only less than 1% (LR category); thus, they fall under the "LR" category. This is in contrast to the

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

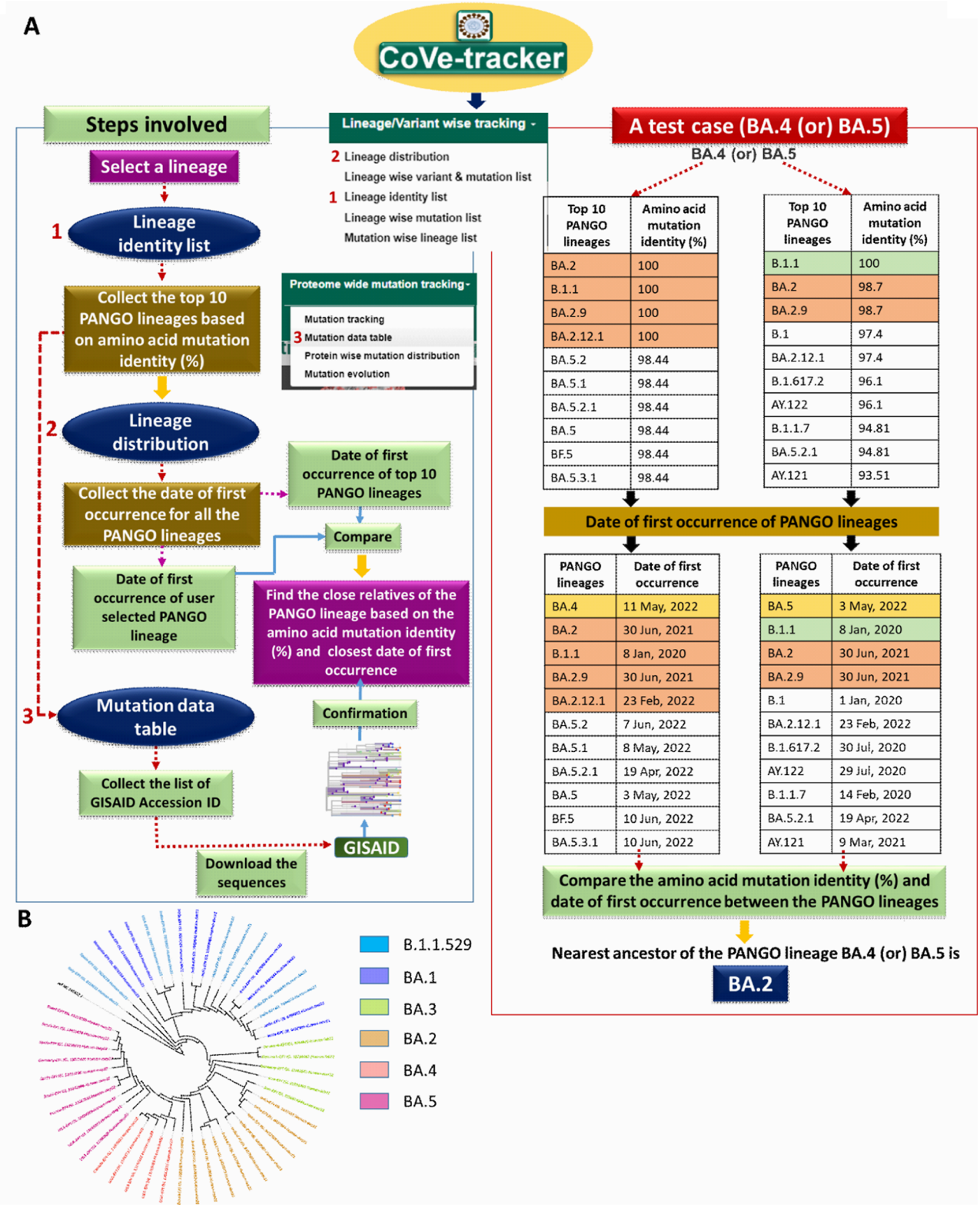


Figure 7. Application of CoVe-tracker is shown by considering BA.4 and BA.5 PANGO lineages as test cases. (A, Left) Flowchart describing the various steps involved in tracking the closest relatives of a variant. (A, Right) Tracking the parent lineage of BA.4 and BA.5 using CoVe-tracker shows that BA.2 is the parent lineage. (B) Phyloproteome analyses of SARS-CoV-2 sequences from six PANGO lineages suggest that BA.2 as the closet ancestor of the lineage BA.4 (or) BA.5 in conformity with the CoVe-tracker results. See text for details.

deletion which falls under HR or MR or LR categories. Soon after the user chooses an amino acid mutation listed under LR/HR/MR, the mutation is projected onto the known 3D structure of the protein in a JSmol window (Figure 5C). In the 3D structure, the mutations are shown as spheres which are color-coded according to their recurrence frequency [high (magenta), moderate (cyan), and low (yellow)]. The user can see the zoomed-in view of the protein 3D structure that has the

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

mutation of interest which can be selected from the panel (Figure 5C).

The "Proteome wide mutation tracking" page also provides a table that contains information about the month-wise frequency of occurrence of HR, MR, and LR mutations. Here, when a user looks for "high recurring mutations" using the key "HR" in the search bar, CoVe-tracker lists all the HR amino acid mutations of SARS-CoV-2 (Figure 5D). Similarly, all the "moderate recurring mutations" and "low recurring mutations" can be listed using the "MR" and "LR" keys, respectively. The user is further given an option to fetch the category-wise and/or protein-wise amino acid mutation by choosing the HR or MR or LR in the "Select the mutation category" dropdown and the name of the protein (see above) listed in the "Select protein" dropdown (Figure 5D).

Since SARS-CoV-2 keeps evolving, it is important to see how the individual positions of the proteome undergo changes. Thus, the "Mutation evolution" submenu provides the amino acid position-wise evolution. For this, the user has to choose the protein and the amino acid position in the protein to get the variations along with the frequency of occurrence and the date and country of the first and last occurrences (Figure 5E).

The submenu "Mutation data table" lists the details such as GISAID ID, country, nucleotide substitutions, amino acid substitutions, GISAID clade, PANGO lineage, and CoVe-tracker PV ID. Using the GISAID ID, the user can readily fetch the metadata information associated with a variant or a mutation of their interest from GISAID.

Genome Mutation Statistics

The "Genome wide tracking" menu provides information about the position-wise nucleotide mutation frequency observed in the SARS-CoV-2 genome. The user can view the nucleotide mutations and their frequency of occurrence by hovering the scatter points in the interactive scatter plot (Figure 6A). The page also provides a table that lists the nucleotide and amino acid substitution(s) corresponding to each SARS-CoV-2 sequence used in the analysis along with its GISAID accession ID, the country of occurrence, GISAID clade, PANGO lineage, and CoVe-tracker PV ID (Figure 6B). The user can filter the data by entering the information such as mutation or country of choice, and so forth in the search tab which is given at the right-hand-side corner of the table (Figure 6B).

Application of CoVe-Tracker in Tracking SARS-CoV-2 Variant Evolution

Besides providing the month- and country-wise dissemination of SARS-CoV-2 mutation/lineage/variant information, CoVe-tracker facilitates the understanding of the variant evolution. This has been demonstrated by considering the omicron sublineages BA.4 and BA.5 as test cases to fetch their parent lineage (Figure 7), which emerged in South Africa during December 2021 and January 2022, respectively.[34] These variants have been the major cause of the hike in the number of infections throughout the globe from April 2022 to June 2022.[13] These variants are further found to affect the vaccine efficacy as discussed below. The efficacy of BNT162b2 (Pfizer-BioNTech), which is one of the most common vaccines used against COVID-19, is seen to wane against BA.4 and BA.5 over time in South Africa (the vaccine's protective efficacy declined to 47 and 26% after 3−4 and 5−6 months, respectively, after full vaccination).[13] Similarly, these lineages resist antibody neutralizing activity by the serum isolated from the individuals vaccinated with triple doses of AstraZeneca-Oxford or Pfizer vaccine.[16] A similar reduction of neutralizing antibodies in the serum isolated from individuals who have received the Moderna booster dose is also observed.[14] BA.4 and BA.5 sublineages with their immunity-evading mutations (Spike:F486V and Spike:L452R) significantly affect the antibody neutralization and, thus, the vaccine effectiveness.[16]

To fetch the parent lineage of BA.4 and BA.5, the user can make use of the "Lineage identity list" option [step 1, Figure 7 (left)]. Upon choosing the lineage (for example, BA.4) of interest under "Lineage identity list" and applying the filtering option HR and MR and/or LR mutations along with the amino acid mutation percentage identity, the module lists the lineages that have the amino acid mutation percentage identity greater than the selected value with respect to the lineage of the user's choice (*viz.*, BA.4). When HR and MR mutations are alone considered (which is preferred) for tracking the parent lineage of BA.4/BA.5 lineage in CoVe-tracker, 64 and 77 HR and MR mutations, respectively, from BA.4 from BA.5 were used in the comparison with the other lineages. The top 10 lineages among the lineages having an amino acid mutation identity greater than 90% were further considered to individually identify the parent lineage of BA.4 and BA.5. For BA.4, lineages such as BA.2 (100%), B.1.1 (100%), BA.2.9 (100%), BA.2.12.1 (100%), BA.5.2 (98.44%), BA.5.1 (98.44%), BA.5.2.1 (98.44%), BA.5 (98.44%), BF.5 (98.44%), and BA.5.3.1 (98.44%) were reported as the top 10 lineages based on amino acid mutation identity [Figure 7 (right)]. Similarly, lineages such as B.1.1 (100%), BA.2 (98.7%), BA.2.29 (98.7%), B.1 (97.4%), BA.2.12.1 (97.4%), B.1.617.2 (96.1%), AY.122 (96.1%), B.1.1.7 (94.81%), BA.5.2.1 (94.81%), and AY.121 (93.51%) were reported as the top 10 lineages for BA.5 [Figure 7 (right)]. Subsequently, the first date of occurrence of the top 10 lineages under BA.4 and BA.5 was collected [Figure 7 (right)] from the "Lineage distribution" submenu under the "Lineage/variant wise tracking" menu [step 2, Figure 7 (left)]. Next, the date of the first occurrence of BA.4 and BA.5 was parallelly compared with the date of the first occurrence of the top 10 lineages. Based on the closest date of the first occurrence, BA.2 was tracked as the parent lineage for both BA.4 and BA.5, respectively [Figure 7 (right)]. To further get the list of variants corresponding to the BA.4 lineage, one has to use the "Lineage wise variant & mutation list" dropdown of the "Lineage/variant wise tracking" menu that provides the date and country of the first and last occurrences. If the user needs to get the sequences corresponding to these variants for further analysis, the user has to initially obtain the appropriate GISAID ID from the "Mutation data table" dropdown of the "Proteome wide mutation tracking" menu and download the sequence from the GISAID (https://www.gisaid.org/).[18] These steps are detailed in Figure 7. Another example illustrating the application of CoVe-tracker indicates the possibility of omicron evolution through the recombination between alpha and delta lineages (Supporting Information Table S1) as suggested earlier based on the spike protein mutations.[35]

Figure S2 illustrates how CoVe-tracker can help in catching the newly emerging PANGO lineages/variants/mutations by considering the month of December 2022 as an example. Among the top 10 lineages, BA.2.75 occurs above the PF of 60 and BA.5.1.27 (2.92%), CH.1.1 (2.67%), XBB (2.00%), BQ.1.1 (1.83%), and CM.2 (1.33%) occur at the top 5 positions (Figure S2A). Eventually, one can track the prevalence of these lineages across different countries by using the "Lineage distribution" submenu bar. One can further get information about HR, MR, and LR mutations associated with the variants under these lineages along with the date and country of first occurrence by

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

using the "Lineage wise variant and mutation list" submenu bar. This has been illustrated in Figure S2B by considering BQ.1.1 (found to be resistant to antibody neutralization[36]) as a case in point which shows a steady increase since its first emergence in September 2022. A total of 92 variants are seen for BQ.1.1, among which the variant with 5 HR and 13 MR mutations is more prevalent. Notably, BQ.1.1 which was not among the top 10 lineages in the month of September 2022, now falls under the top 5, suggestive of the highly transmissible nature of the BQ.1.1 variant and warns of the need for immediate countermeasures.

Limitations of CoVe-Tracker

Although CoVe-tracker can provide detailed information about the variations in the amino acids of the SARS-CoV-2 proteome, it cannot directly provide concomitant metadata information. The user has to obtain the metadata information from the GISAID database by using the GISAID ID provided in the CoVe-tracker. Second, the database is currently being updated every month (the details about the update can be accessed through the link: https://project.iith.ac.in/cove-tracker/status.html). This may delay the capturing of a newly emerging lineage/variant/mutation. To circumvent this issue, CoVe-tracker will be updated frequently in the near future.

CONCLUSIONS

The fast-evolving SARS-CoV-2 genome poses a major public healthcare challenge as the existing vaccines are poorly matched with the newly emerging variants. This mandates the upgradation of variant-proof vaccines which requires an understanding of viral evolution. To this end, a user-friendly and interactive web-GUI database, CoVe-tracker, has been developed to track the viral evolution and spread. The highlight of CoVe-tracker is that it provides month-wise and country-wise statistics of SARS-CoV-2 amino acid mutations, PANGO lineages, and variants. This aids in the tracking of region-specific mutations, lineages, and variants. Such user-friendly information provided by CoVe-tracker can be a good source for artificial intelligence-based prediction of future SARS-CoV-2 variants. CoVe-tracker allows the user to obtain the PANGO lineages/variants having one or more mutations, thus helping in the understanding of the variant evolutionary dynamics. Thus, CoVe-tracker would act as an important tool to understand SARS-CoV-2 evolution and the concomitant epidemiological dynamics as well as for theragnostic development.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00068.

Occurrence of HR and MR mutations in different VOCs (XLSX)

Flowchart describing the methodological development of CoVe-tracker and example illustrating the application of CoVe-tracker to track the evolution of newly emerging top 10 PANGO lineages and the associated top 10 variants by considering the month of December 2022 as an example (PDF)

AUTHOR INFORMATION

Corresponding Author

Thenmalarchelvi Rathinavelan − *Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana State 502285, India;* orcid.org/0000-0002-1142-0583; Email: tr@bt.iith.ac.in

Authors

Chakkarai Sathyaseelan − *Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana State 502285, India*

Murali Aadhitya Magateshvaren Saras − *Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana State 502285, India*

L. Ponoop Prasad Patro − *Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana State 502285, India*

Patil Pranita Uttamrao − *Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana State 502285, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.3c00068

Author Contributions

C.S., M.A.M.S., and L.P.P.P. contributed equally. M.A.M.S. and P.P.U. collected and organized the whole genome sequence data of SARS-CoV-2. C.S. developed the database and implemented the web-GUI interface. L.P.P.P. wrote the code in the bash script to derive the mutation statistics and also implemented the interactive 3D view of the proteins. M.A.M.S. converted the bash scripts into Python scripts that are being used in maintaining the database. C.S. and M.A.M.S. implemented the interactive display of the sequence data. C.S., L.P.P.P., P.P.U., and M.A.M.S. analyzed the data to derive the variant and lineage classifications. C.S. and T.R. wrote the manuscript. T.R. designed and supervised the entire project.

Notes

The authors declare no competing financial interest.

Bash and Python scripts used in this analysis along with the sample dataset have been deposited in the GitHub repository that can be accessed through the link: https://github.com/MBL-lab/CoVe-tracker.

ACKNOWLEDGMENTS

The authors thank the GISAID database (https://www.gisaid.org) for sharing SARS-CoV-2 sequence data. The Indian Institute of Technology Hyderabad and CDAC are acknowledged for their computational resources. BIRAC-SRISTI GYTI [grant (PMU_2019_007)] is acknowledged for the financial support.

REFERENCES

(1) Civljak, R.; Markotic, A.; Kuzman, I. The third coronavirus epidemic in the third millennium: what's next? *Croat. Med. J.* **2020**, *61*, 1−4.

(2) Hou, Y. J.; Chiba, S.; Halfmann, P.; Ehre, C.; Kuroda, M.; Dinnon, K. H., 3rd; Leist, S. R.; Schafer, A.; Nakajima, N.; Takahashi, K.; Lee, R. E.; Mascenik, T. M.; Graham, R.; Edwards, C. E.; Tse, L. V.; Okuda, K.; Markmann, A. J.; Bartelt, L.; de Silva, A.; Margolis, D. M.; Boucher, R. C.; Randell, S. H.; Suzuki, T.; Gralinski, L. E.; Kawaoka, Y.; Baric, R. S. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **2020**, *370*, 1464−1468.

(3) Callaway, E. How months-long COVID infections could seed dangerous new variants. *Nature* **2022**, *606*, 452−455.

(4) Aleem, A.; Akbar Samad, A. B.; Slenker, A. K. Emerging Variants of SARS-CoV-2 And Novel Therapeutics Against Coronavirus (COVID-

19). In *StatPearls*; StatPearls Publishing, 2022. Treasure Island (FL); (Bookshelf ID: NBK570580).
(5) Dennehy, J. J.; Gupta, R. K.; Hanage, W. P.; Johnson, M. C.; Peacock, T. P. Where is the next SARS-CoV-2 variant of concern? *Lancet* **2022**, *399*, 1938−1939.
(6) Volz, E.; Mishra, S.; Chand, M.; Barrett, J. C.; Johnson, R.; Geidelberg, L.; Hinsley, W. R.; Laydon, D. J.; Dabrera, G.; O'Toole, A.; Amato, R.; Ragonnet-Cronin, M.; Harrison, I.; Jackson, B.; Ariani, C. V.; Boyd, O.; Loman, N. J.; McCrone, J. T.; Goncalves, S.; Jorgensen, D.; Myers, R.; Hill, V.; Jackson, D. K.; Gaythorpe, K.; Groves, N.; Sillitoe, J.; Kwiatkowski, D. P.; Flaxman, S.; Ratmann, O.; Bhatt, S.; Hopkins, S.; Gandy, A.; Rambaut, A.; Ferguson, N. M.; Ferguson, N. M. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **2021**, *593*, 266−269.
(7) Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E. J.; Msomi, N.; Mlisana, K.; von Gottberg, A.; Walaza, S.; Allam, M.; Ismail, A.; Mohale, T.; Glass, A. J.; Engelbrecht, S.; Van Zyl, G.; Preiser, W.; Petruccione, F.; Sigal, A.; Hardie, D.; Marais, G.; Hsiao, N. Y.; Korsman, S.; Davies, M. A.; Tyers, L.; Mudau, I.; York, D.; Maslo, C.; Goedhals, D.; Abrahams, S.; Laguda-Akingba, O.; Alisoltani-Dehkordi, A.; Godzik, A.; Wibmer, C. K.; Sewell, B. T.; Lourenco, J.; Alcantara, L. C. J.; Kosakovsky Pond, S. L.; Weaver, S.; Martin, D.; Lessells, R. J.; Bhiman, J. N.; Williamson, C.; de Oliveira, T. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **2021**, *592*, 438−443.
(8) Voloch, C. M.; da Silva Francisco, R., Jr.; de Almeida, L. G. P.; Cardoso, C. C.; Brustolini, O. J.; Gerber, A. L.; Guimarães, A. P. d. C.; Mariani, D.; da Costa, R. M.; Ferreira, O. C., Jr.; Cavalcanti, A. C.; Frauches, T. S.; de Mello, C. M. B.; Leitão, I. d. C.; Galliez, R. M.; Faffe, D. S.; Castiñeiras, T. M. P. P.; Tanuri, A.; de Vasconcelos, A. T. R. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **2021**, *95*, 00119.
(9) Puhach, O.; Adea, K.; Hulo, N.; Sattonnet, P.; Genecand, C.; Iten, A.; Jacquerioz, F.; Kaiser, L.; Vetter, P.; Eckerle, I.; Meyer, B. Infectious viral load in unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2. *Nat. Med.* **2022**, *28*, 1491−1500.
(10) Lyngse, F. P.; Molbak, K.; Skov, R. L.; Christiansen, L. E.; Mortensen, L. H.; Albertsen, M.; Moller, C. H.; Krause, T. G.; Rasmussen, M.; Michaelsen, T. Y.; Voldstedlund, M.; Fonager, J.; Steenhard, N.; Brandt, J.; Knuttson, S.; et al. Increased transmissibility of SARS-CoV-2 lineage B.1.1.7 by age and viral load. *Nat. Commun.* **2021**, *12*, 7251.
(11) Callaway, E. Fast-evolving COVID variants complicate vaccine updates. *Nature* **2022**, *607*, 18−19.
(12) Callaway, E. What Omicron's BA.4 and BA.5 variants mean for the pandemic. *Nature* **2022**, *606*, 848−849.
(13) Collie, S.; Nayager, J.; Bamford, L.; Bekker, L. G.; Zylstra, M.; Gray, G. Effectiveness and Durability of the BNT162b2 Vaccine against Omicron Sublineages in South Africa. *N. Engl. J. Med.* **2022**, *387*, 1332−1333.
(14) Qu, P.; Faraone, J. N.; Evans, J. P.; Zheng, Y. M.; Yu, L.; Ma, Q.; Carlin, C.; Lozanski, G.; Saif, L. J.; Oltz, E. M.; Gumina, R. J.; Liu, S. L. Durability of Booster mRNA Vaccine against SARS-CoV-2 BA.2.12.1, BA.4, and BA.5 Subvariants. *N. Engl. J. Med.* **2022**, *387*, 1329−1331.
(15) Shen, X.; Chalkias, S.; Feng, J.; Chen, X.; Zhou, H.; Marshall, J. C.; Girard, B.; Tomassini, J. E.; Aunins, A.; Das, R.; Montefiori, D. C. Neutralization of SARS-CoV-2 Omicron BA.2.75 after mRNA-1273 Vaccination. *N. Engl. J. Med.* **2022**, *387*, 1234−1236.
(16) Tuekprakhon, A.; Nutalai, R.; Dijokaite-Guraliuc, A.; Zhou, D.; Ginn, H. M.; Selvaraj, M.; Liu, C.; Mentzer, A. J.; Supasa, P.; Duyvesteyn, H. M. E.; Das, R.; Skelly, D.; Ritter, T. G.; Amini, A.; Bibi, S.; Adele, S.; Johnson, S. A.; Constantinides, B.; Webster, H.; Temperton, N.; Klenerman, P.; Barnes, E.; Dunachie, S. J.; Crook, D.; Pollard, A. J.; Lambe, T.; Goulder, P.; Paterson, N. G.; Williams, M. A.; Hall, D. R.; Fry, E. E.; Huo, J.; Mongkolsapaya, J.; Ren, J.; Stuart, D. I.; Screaton, G. R.; Conlon, C.; Deeks, A.; et al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. *Cell* **2022**, *185*, 2422−2433.
(17) Maxmen, A. Omicron blindspots: why it's hard to track coronavirus variants. *Nature* **2021**, *600*, 579.
(18) Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **2017**, *22*, 30494.
(19) Fang, S.; Li, K.; Shen, J.; Liu, S.; Liu, J.; Yang, L.; Hu, C. D.; Wan, J. GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. *Nucleic Acids Res.* **2021**, *49*, D706−D714.
(20) Alam, I.; Radovanovic, A.; Incitti, R.; Kamau, A. A.; Alarawi, M.; Azhar, E. I.; Gojobori, T. CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants. *Lancet Infect. Dis.* **2021**, *21*, 602.
(21) Canakoglu, A.; Pinoli, P.; Bernasconi, A.; Alfonsi, T.; Melidis, D. P.; Ceri, S. ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res.* **2021**, *49*, D817−D824.
(22) 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.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265−269.
(23) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(24) Patro, L. P. P.; Sathyaseelan, C.; Uttamrao, P. P.; Rathinavelan, T. The evolving proteome of SARS-CoV-2 predominantly uses mutation combination strategy for survival. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3864−3875.
(25) Magateshvaren Saras, M. A.; Patro, L. P. P.; Uttamrao, P. P.; Rathinavelan, T. SARS-CoV-2 whole-proteome sequences from environment as an indicator of community viral distribution, evolution and epidemiological dynamics: A cohort analysis of Austria. *Environ. Microbiol. Rep.* **2022**, *14*, 850−855.
(26) Magateshvaren Saras, M. A.; Patro, L. P. P.; Uttamrao, P. P.; Rathinavelan, T. Geographical distribution of SARS-CoV-2 amino acids mutations and the concomitant evolution of seven distinct clades in non-human hosts. *Zoonoses Public Health* **2022**, *69*, 816−825.
(27) Patro, L. P. P.; Sathyaseelan, C.; Uttamrao, P. P.; Rathinavelan, T. Global variation in SARS-CoV-2 proteome and its implication in pre-lockdown emergence and dissemination of 5 dominant SARS-CoV-2 clades. *Infect., Genet. Evol.* **2021**, *93*, 104973.
(28) Uttamrao, P. P.; Sathyaseelan, C.; Patro, L. P. P.; Rathinavelan, T. Revelation of Potent Epitopes Present in Unannotated ORF Antigens of SARS-CoV-2 for Epitope-Based Polyvalent Vaccine Design Using Immunoinformatics Approach. *Front. Immunol.* **2021**, *12*, 692937.
(29) O'Toole, A.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J. T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; Yeats, C.; du Plessis, L.; Maloney, D.; Medd, N.; Attwood, S. W.; Aanensen, D. M.; Holmes, E. C.; Pybus, O. G.; Rambaut, A. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **2021**, *7*, veab064.
(30) Yang, J.; Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174−W181.
(31) Huang, X.; Pearce, R.; Zhang, Y. De novo design of protein peptides to block association of the SARS-CoV-2 spike protein with human ACE2. *Aging* **2020**, *12*, 11263−11276.
(32) Zhang, C.; Zheng, W.; Huang, X.; Bell, E. W.; Zhou, X.; Zhang, Y. Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1. *J. Proteome Res.* **2020**, *19*, 1351−1360.
(33) Huang, X.; Zhang, C.; Pearce, R.; Omenn, G. S.; Zhang, Y. Identifying the Zoonotic Origin of SARS-CoV-2 by Modeling the Binding Affinity between the Spike Receptor-Binding Domain and Host ACE2. *J. Proteome Res.* **2020**, *19*, 4844−4856.
(34) Tegally, H.; Moir, M.; Everatt, J.; Giovanetti, M.; Scheepers, C.; Wilkinson, E.; Subramoney, K.; Makatini, Z.; Moyo, S.; Amoako, D. G.; Baxter, C.; Althaus, C. L.; Anyaneji, U. J.; Kekana, D.; Viana, R.; Giandhari, J.; Lessells, R. J.; Maponga, T.; Maruapula, D.; Choga, W.;

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996

Matshaba, M.; Mbulawa, M. B.; Msomi, N.; Bester, A. P.; Claassen, M.; Doolabh, D.; Mudau, I.; Mbhele, N.; Engelbrecht, S.; Goedhals, D.; Hardie, D.; Hsiao, N. Y.; Iranzadeh, A.; Ismail, A.; Joseph, R.; Maharaj, A.; Mahlangu, B.; Mahlakwane, K.; Davis, A.; Marais, G.; Mlisana, K.; Mnguni, A.; Mohale, T.; Motsatsi, G.; Mwangi, P.; Ntuli, N.; Nyaga, M.; Olubayo, L.; Radibe, B.; Ramphal, Y.; Ramphal, U.; Strasheim, W.; Tebeila, N.; van Wyk, S.; Wilson, S.; Lucaci, A. G.; Weaver, S.; et al. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* **2022**, *28*, 1785−1790.

(35) Ou, J.; Lan, W.; Wu, X.; Zhao, T.; Duan, B.; Yang, P.; Ren, Y.; Quan, L.; Zhao, W.; Seto, D.; Chodosh, J.; Luo, Z.; Wu, J.; Zhang, Q. Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduction Targeted Ther.* **2022**, *7*, 138.

(36) Jiang, X. L.; Zhu, K. L.; Wang, X. J.; Wang, G. L.; Li, Y. K.; He, X. J.; Sun, W. K.; Huang, P. X.; Zhang, J. Z.; Gao, H. X.; Dai, E. H.; Ma, M. J. Omicron BQ.1 and BQ.1.1 escape neutralisation by omicron subvariant breakthrough infection. *Lancet* **2023**, *23*, 28−30.

https://doi.org/10.1021/acs.jproteome.3c00068
J. Proteome Res. 2023, 22, 1984−1996