EpiSurf: metadata-driven search server for analyzing amino acid changes on epitopes of SARS-CoV-2 and other viral species

Anna Bernasconi [1,*], Luca Cilibrasi [1,*], Ruba Al Khalaf [1], Tommaso Alfonsi [1], Stefano Ceri [1], Pietro Pinoli [1], and Arif Canakoglu [1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133, Milano, Italy

ABSTRACT

EpiSurf is a Web application for selecting viral populations of interest and then analyzing how their amino acid changes are distributed along epitopes. Viral sequences are searched within ViruSurf, which stores curated metadata and amino acid changes imported from the most widely used deposition sources for viral databases (GenBank, COG-UK, and GISAID). Epitopes are searched within the open-source Immune Epitope DataBase (IEDB) or directly proposed by users by indicating their start and stop positions in the context of a given viral protein.

Amino acid changes of selected populations are joined with epitopes of interest; a result table summarizes, for each epitope, statistics about the overlapping amino acid changes and about the sequences carrying such alterations. Results may also be inspected by the VirusViz Web application; epitope regions are highlighted within the given viral protein, and changes can be comparatively inspected. For sequences mutated on the epitope, we also offer a complete view of the amino acid changes distribution, optionally grouped by location, collection date, or lineage. Thanks to these functionalities, EpiSurf supports user-friendly testing of epitope conservancy within selected populations of interest, which can be of uttermost relevance for designing vaccines, drugs, or serological assays. EpiSurf is available at two endpoints, `http://gmql.eu/episurf/` (for searching GenBank and COG-UK sequences) and `http://gmql.eu/episurf_gisaid/` (for GISAID sequences).

INTRODUCTION

With the COVID-19 pandemic outbreak, unprecedented efforts have been dedicated to the sampling and sequencing of the SARS-CoV-2 virus, with the objective of capturing and then studying SARS-CoV-2 variations and their effects. Leveraging on our previous experience on human genomics-targeted computational systems (1), we have directed our interest towards the integration, curation, search and analysis of viral sequences, yielding to several contributions: a conceptual model for describing viral sequences with their metadata and variants (2), an integrated search system for viral sequences (3), a data visualization application (4), and a knowledge base for studying variant effects (5). Capitalizing on the above experiences and resources, we developed and hereby present a Web-based search application for studying epitopes in the context of viral sequences that have been so far deposited worldwide.

Epitopes are strings of amino acid residues from a pathogen's protein that can be recognized by antibodies or B/T cell receptors, thus activating an immune response from the host; in particular, epitopes available for the Spike protein of SARS-CoV-2 are used in the design of COVID-19 vaccines. For epitope-based vaccine design, it is important to study their conservation; conversely, observing epitope variability has applications in disease monitoring, diagnostics settings, and drug design. We adopt the most basic conservancy measure for an epitope (i.e., a region on a protein), based on the number of changed amino acid residues with respect to the reference sequence of the virus species in the same position range. *Conserved* epitopes have a zero distance from the reference, whereas *modified* epitopes exhibit at least one amino acid change.

To date, the most relevant resource for epitopes employed by the research community is the Immune Epitope Database (IEDB) (6). It encompasses immune epitope data of a large number of species, including antibody, T cell, and major histocompatibility complex (MHC) binding contexts associated with several diseases. EpiSurf imports all epitopes available from IEDB; in addition, EpiSurf supports user-defined epitopes, intended as position-ranges on specific virus proteins.

*Co-first authors. AB and AC are corresponding authors. Tel: +39 02 2399 3655; Fax: +39 02 2399 3411; Email: anna.bernasconi@polimi.it, arif.canakoglu@polimi.it

Table 1. Comparison of resources for analyzing epitopes over a sequence population. For each system, we indicate the data source of SARS-CoV-2 sequences; the presence of sequences from other viruses; which hosts are considered; if sub-populations of sequences can be selected based on metadata; which epitope assays are considered; the data source for their extraction; and the presence of aggregation/visualization methods for showing mutations in the context of epitopes.

	SARS-CoV-2 seq.	**Other viruses**	**Hosts**	**Seq. filters**	**Epitope assay type**	**Epitope source**	**Vis/Agg. on epit.**
IEDB ECA (15)	User input	Virus agnostic	Host agnostic	-	B cell; T cell; MHC	User input	-
COVIDep (16)	GISAID	-	Human	Location	B cell (linear); T cell	Pred. from SARS-CoV	-
ViPR (17)	GenBank	All in GenBank	All in Genbank	Full metadata	B cell; T cell; MHC	User input; IEDB; pred. (NetCTL)	-
COG-UK-ME (18)	COG-UK	-	Human	-	T cell	Collected from exp. studies	With ggseqlogo
EpiSurf	GenBank; COG-UK GISAID	SARS; MERS; Dengue; Ebola	All in Genbank	Full metadata	B cell; T cell; MHC	User input; IEDB	With VirusViz

EpiSurf supports the search of viral sequences deposited on public platforms – released day by day on GenBank, COG-UK, and GISAID and then integrated within the ViruSurf platform; relevant sequences can be extracted thanks to a rich set of metadata information, including the sampling location, collection and deposition date, sequence's lineages and strains, submission laboratory. On top of this, EpiSurf provides several methods to intersect selected sequences and selected epitopes, thereby integrating information about amino acid changes and epitopes extracted from the largest and most popular data collections in the world using their metadata. Lastly, the integration with the VirusViz tool allows informative visualization of sequence variation in the specific epitopes locations. Building up on our previously developed resources, EpiSurf offers a novel, fully-independent, integrated environment for evaluating conservancy of epitopes against arbitrarily extracted viral populations, reflecting the spreading of viruses in time and space and their genetic evolution.

COMPARISON WITH EXISTING SYSTEMS

Several tools for epitope prediction have been studied in the past (7). The most used and well-known resource in this field is the suite of IEDB (6), comprising a set of T Cell and B Cell Epitope Prediction tools (see `http://tools.iedb.org/main/`).

For the specific case of SARS-CoV-2 (and closely related viruses) we report the following. COVIEdb (8) targets pan-coronavirus vaccine development, by describing a database of potential B/T cell epitopes for SARS-CoV-2, SARS-CoV, MERS-CoV, and RaTG13-CoV; database entries are predicted by using tools hosted by IEDB, exploiting the similarity of other viruses (as proposed by Grifoni et al. (9)). Similar databases are provided in CoronaVIR (10), DBCOVP (11), and CoronaVR (12). COVID miner (13) and COVID profiler (14) provide companion vaccine design tools, with a focus on prediction (the latter also providing light integration with IEDB data).

EpiSurf is not intended for epitope prediction. Instead, it may be labeled as a tool for conservancy and population coverage analysis. IEDB curates a collection of tools (`http://tools.iedb.org/main/analysis-tools/`), used for a variety of detailed analyses. Among these, EpiSurf is similar in spirit to the Epitope Conservancy Analysis (ECA) tool (15), however there the user must provide the amino acid sequences of (a) all the epitopes to be tested and (b) all the virus sequences whose changes should be tracked, whereas EpiSurf offers a seamless integration with all public sequences and variants currently available from GenBank, COG-UK, and GISAID.

COVIDep (16) is an integrative effort more similar in the approach to EpiSurf, as it joins IEDB epitopes with regularly updated GISAID sequences. The proposed "Population coverage analysis" is an interesting view providing quantifications of "conservation" and "population coverage" for each epitope. However, the provided epitopes are those that were predicted and experimentally derived (based on SARS-CoV data) at the time of publication (May 2020) by the authors of the work. This important exercise resulted into a total of 284 T cell epitopes and 58 B cell linear epitopes. On the contrary, EpiSurf keeps its list of epitopes updated, now reaching 3,690 T cell epitopes, 1,006 MHC Ligand epitopes, and 1,421 B cell epitopes—see Table 2. Comparatively, EpiSurf offers a scalable approach to epitope conservancy analysis that is very useful as we expect that new sequences and epitopes will be deposited for a long time. Moreover, the COVIDep resource provides much less freedom of choosing metadata for sequences and epitopes, no possibility of fixing particular amino acid changes, and no ways of analyzing in detail the sequences that mutate on the epitope. EpiSurf, on the contrary, does offer all sequences metadata in its result table and complements it with a table for understanding the breakdown of statistics over the different metadata of the population (for both EpiSurf and EpiSurf-GISAID) as well as VirusViz visualization functionalities (for EpiSurf).

The Virus Pathogen Database and Analysis Resource (ViPR) (17) is another important tool that connects (both predicted and experimentally derived) epitopes and proteins of sequences deposited on GenBank, whereas no link to GISAID is provided. EpiSurf is novel in that it offers several aggregations and simple statistics on both GenBank and GISAID data.

The COG-UK Mutation Explorer (COG-UK-ME) (18) recently released an interface dedicated to only UK data and its variation (also in the context of T cell epitopes reported by experimental studies).

Table 1 summarizes the relevant aspects of the tools that allow population conservancy and coverage analysis. Their late focus has been on SARS-CoV-2 and human hosts, however ViPR and IEDB ECA pre-existed, offering support for many kinds of viruses. Only ViPR offers the possibility to select sub-populations of sequences at the user's preference. COG-UK-ME focuses on T cell epitopes. Sources for epitopes are various: IEDB ECA only allows user input strings, whereas ViPR and EpiSurf enrich them with the IEDB corpus of epitopes. COVIDep, ViPR, and COG-UK-ME currently

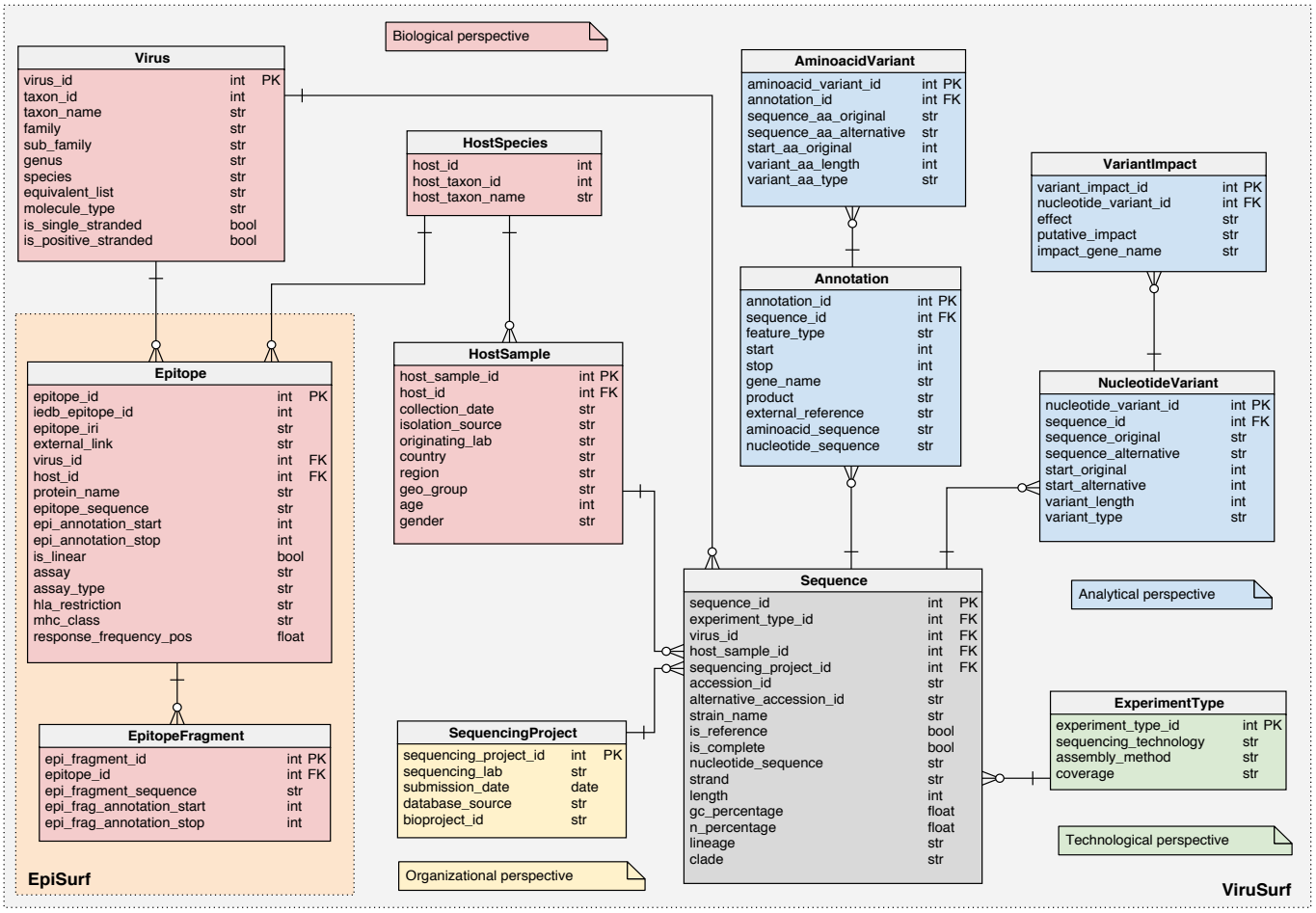

Figure 1. Logical schema of the relational database in the back-end of EpiSurf.

offer a curated list, respectively predicted from SARS-CoV, predicted with NetCTL (19), and manually extracted from experimental studies.

For data visualization, EpiSurf provides a connection to VirusViz (4) (`http://gmql.eu/virusviz/`), a Web application for visualizing and exploring the fully open source nucleotide and amino acid changes that are made available though search services or autonomously provided as input from users. When VirusViz is opened starting from an EpiSurf search, the tool visualizes a bar plot where the x-axis represents the amino acid positions of a protein, bars' heights represent the number of sequences in the selected populations that feature a change in the bar's position, and epitope position ranges are represented as blue vertical regions (see Figure 6 in Example 3).

MATERIALS AND METHODS

Database Schema

The database schema, represented in Figure 1, is partially inherited from ViruSurf (3) (`http://gmql.eu/virusurf/`), an integrated database of SARS-CoV-2 sequences (and of other similar viruses), storing all the sequences deposited to GenBank (20) and COG-UK (21). A dual database (`http://gmql.eu/virusurf_gisaid/`) stores relevant metadata and variation information of GISAID sequences (22).

For both databases, the schemas are centered on the SEQUENCE, described by biological metadata (VIRUS and HOSTSAMPLE), technological metadata (EXPERIMENTTYPE), and organizational metadata (SEQUENCINGPROJECT). The *analytical perspective* provides the ANNOTATION, AMINOACIDVARIANT, NUCLEOTIDEVARIANT and VARIANTIMPACT tables.

EpiSurf and EpiSurf-GISAID feature two novel databases, whose complete schema descriptions are available at `https://github.com/DEIB-GECO/EpiSurf/wiki/Database-and-sources`, pointing to SchemaSpy (`http://schemaspy.org/`) documents. In the following, we only detail the additions that were not present in (3):

1. The HOSTSPECIES table, a connector between the HOSTSAMPLE and the EPITOPE tables, representing the identification of the animal species both involved in the extraction of biological material to be sequenced and in the epitope design.
2. The EPITOPE table, describing the epitopes extracted from IEDB, connected both to HOSTSPECIES and to the

Table 2. Summary of EpiSurf content as of July 18th, 2021. For each taxon name (identified by a taxon ID and rank) and each source, we specify the number of distinct sequences and the number of available epitopes, with their breakdown into T cell, B cell, and MHC ligand assays.

					IEDB epitopes			
Taxon rank	Taxon ID	Taxon name	Source	#Seq.	#Total	#T cell	#B cell	#MHC lig.
No rank	2697049	Severe acute respiratory syndrome coronavirus 2	GISAID	2,390,870				
No rank	2697049	Severe acute respiratory syndrome coronavirus 2	GenBank	691,734	6,117	3,690	1,421	1,006
No rank	2697049	Severe acute respiratory syndrome coronavirus 2	COG-UK	574,061				
Species	694009	Severe acute respiratory syndrome-related coronavirus	GenBank	674	1,722	782	437	503
Species	1335626	Middle East respiratory syndrome-related coronavirus	GenBank	1,453	110	110	-	-
Species	2010960	Bombali ebolavirus	GenBank	8	-	-	-	-
Species	565995	Bundibugyo ebolavirus	GenBank	22	14	-	14	-
Species	186539	Reston ebolavirus	GenBank	58	-	-	-	-
Species	186540	Sudan ebolavirus	GenBank	39	536	240	9	287
Species	186541	Tai Forest ebolavirus	GenBank	9	-	-	-	-
Species	186538	Zaire ebolavirus	GenBank	2,932	2,113	700	487	926
Strain	11053	Dengue virus 1	GenBank	12,059	1,631	1,130	215	286
Strain	11060	Dengue virus 2	GenBank	9,646	2,024	1,396	322	306
Strain	11069	Dengue virus 3	GenBank	5,628	2,318	1,704	224	390
Strain	11070	Dengue virus 4	GenBank	2,812	1,090	782	96	212

VIRUS table. The core information is contained within the *epitope_sequence* – the amino acid sequence of the epitope, starting at position *epi_annotation_start* and finishing at position *epi_annotation_stop* of the reference sequence of a given protein, referred as *protein_name*; the *is_linear* attribute defines continuous (true) or discontinuous (false) epitopes, composed of amino acid residues that may be located on different protein regions – brought together by protein folding. We also report information on the experiments performed to retrieve the epitope. Each epitope record in EpiSurf may correspond to multiple records from IEDB (each assigned to one single experiment, i.e., assay). Therefore the following four fields can take multiple values:

- *assay* indicates the target of the experiment (allowing values ['T cell', 'B cell', 'MHC ligand']);
- *assay_type* indicates the outcome – considering possibly multiple experiments (we have 'positive', 'negative', and 'both', when positive and negative outcomes were included);
- *hla_restriction* (also referred to as 'mhc allele') indicates the list of the class (e.g., 'HLA Class I'), or lists of alleles (e.g., 'HLA-B*35:01, HLA-B*15:01') to which the epitope is restricted—this is relevant only for T cell and MHC Ligand assays;
- *mhc_class* indicates the general classes of alleles provided in the previous field (possible values are 'I', 'II', or 'I,II' if both class I and II alleles are considered).

Finally, we add the *response_frequency_pos*: on IEDB this measure is defined as the number of positively responded subjects (R) divided by the total number of those tested (N), summed up by mapped epitopes; however, to compensate for epitopes that are identified by a low number of assays, we employ a corrected formula (proposed in (23)) resulting as $(R-\sqrt{R})/N$, where the importance of corrections decreases as the number of assays increases.

3. The EPITOPEFRAGMENT table, which contains the segments (identified by the *epi_fragment_id*) of non linear epitopes (each of which is contained within one comprehensive epitope). In case of linear epitopes, we store a unique fragment in this table. The *epi_fragment_sequence* contains the amino acid sequence of the single fragment starting at *epi_frag_annotation_start* and finishing at *epi_frag_annotation_stop*.

Database Content

Content imported from ViruSurf. EpiSurf database is fuelled by the same automatic import pipeline that frequently extracts and processes sequences, metadata, and variant information for populating the ViruSurf database (3). Specifically, we extract sequences and their metadata from COG-UK and GenBank, whereas we compute amino acid changes according to the following steps: i) for each virus species, selection of a reference sequence and a set of annotations; ii) for each sequence, computation of the optimal global alignment to the reference by means of the Needleman-Wunsch (NW) algorithm (24); iii) identification of the sub-sequences corresponding to the reference annotations; iv) translation of coding regions into their equivalent amino acid sequences; v) alignment of translated amino acid sequences to the corresponding reference amino acid sequences (also using NW); vi) inference of amino acid changes. Thanks to a Data Connectivity Agreement with GISAID, we have access to frequently updated information downloaded from EpiCoV^{TM} database including, for each sequence, selected metadata and all amino acid changes.

Content imported from IEDB. We regularly download and process experimental epitope sequences and their metadata. The process is controlled by an automated pipeline that retrieves the DB exports of B cell, T cell, and MHC ligand in the form of CSV files from the IEDB Database Export site (`https://www.iedb.org/database_export_v3.php`) at the section "CSV Metric Export". After extraction, each file is parsed as regular tabular data, which allows for easy selection of the relevant characteristics. Indeed, the attributes available in our database are copied *as is* from the origin, with the exception of attributes regarding assays. As mentioned in the discussion of the EPITOPE table, the four attributes

assay, *assay_type*, *hla_restriction*, *mhc_class* – concerning a single assay on IEDB – are concatenated in a single epitope in EpiSurf. Similarly, the *response_frequency_pos* is calculated as an aggregation over all the positive assays that derived the epitope.

The pipeline associates three foreign keys to each imported epitope: the *virus_id*, the *protein_name*, and the *host_id*. The first two attributes are derived by directly mapping the virus name and the UniProtID respectively to the 'id' of the organism in the VIRUS table and to the product in the ANNOTATION table. The third attribute links an epitope to a host in the table HOSTSPECIES. To make sure that this foreign key can be set, before the import stage, we automatically update the hosts' table by collecting from the NCBI Taxonomy database the 'name' and 'id' of the species that are not already available in ViruSurf. Finally, we generate one row inside the EPITOPEFRAGMENT table for every epitope sequence, be it linear or non-linear, and link them through the key *epitope_id* to the EPITOPE table. In this way, it is easy to access all epitope sequences by a regular join of the two tables and selecting the *epi_fragment_sequence*.

Quantitative description. Table 2 provides a description of the current EpiSurf content; for each virus we report the rank, the NCBI Taxonomy identifier/name, and the number of sequences included from each source. In the last four columns, we provide the number of epitopes retrieved from IEDB for the indicated species. The total number is broken down into three categories: T cell, B cell and MHC ligand epitopes. The most substantial contribution in the database is provided by SARS-CoV-2 data, however the system works seamlessly also for SARS-CoV, MERS-CoV, Ebola and Dengue species. In the future, additional viruses may be added with small changes in the configuration of pipelines and no changes in the data representation and query engine.

Data Access Optimization

A number of optimizations steps were designed for allowing acceptable query performances. As the most critical part of the system involves data on SARS-CoV-2 virus and human host, the data on epitopes and matching variants regarding this fraction of the database has been precalculated into several materialized views, one for each of the 12 distinct proteins in the system (i.e., ORF1a, ORF1ab, Spike, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, ORF10) and 16 sub-proteins (from NSP1 to NSP16).

System Development and Sustainability

In terms of software architecture, EpiSurf is organized as a Web application where the back-end runs on a Flask (Python) server, and the front-end is implemented with the Javascript Vue.js framework. The underlying relational database is built with PostgreSQL (Version 10.17); continuous interactions with the database are handled with the Python sqlalchemy library. The code is available on GitHub at `https://github.com/DEIB-GECO/EpiSurf/`.

The objective of EpiSurf is to offer a concrete public endpoint for research on the interplay of epitopes with current viral sequences; its sustainability depends on the timely provision of both sequences and epitope inputs, as well as on the interplay with the ViruSurf database and VirusViz visualization application. Sequence data are currently updated on EpiSurf and EpiSurf-GISAID weekly; SARS-CoV-2 epitopes from IEDB are updated with the same frequency. We also periodically consider other species (SARS-CoV, MERS, Dengue, and Ebola).

RESULTS

The web interface of EpiSurf is composed of 4 sections, numbered in Figure 2: (1) the menu bar, for accessing services, documentation and predefined example queries; (2) the search interface over sequence metadata attributes; (3) the search interface over epitopes, available in three modes (user-defined epitopes, IEDB epitopes with – and without – the calculation of statistics on variants); (4) the results section, showing epitopes with their metadata, counters, and visualization options. Users should select exactly one host organism and one virus (pre-selected options are "homo sapiens" and "SARS-CoV-2"), as this is the default configuration for matching sequences with epitopes. Interaction over (2) and (3) is carefully designed in the three modes, as the underlying system builds complex queries that intersect the sequences resulting from (2) with the epitopes resulting from (3) by considering the amino acid changes exhibited by sequences within given epitope ranges.

Sequence Population Search

The Metadata search section is organized in four parts: *Virus* and *Host Organism* (from the *biological* perspective of the database schema), *Technology* and *Organization* (from the corresponding perspectives). It includes attributes that are present in most of the sources, described by an information tab that is opened by clicking on grey circles; values can be selected using drop-down menus. At the side of each value we report the number of sequences in the repository with that value. The user can select the desired sequence population by entering values from all the drop-down menus; the result is the set of sequences matching all the filters. For numerical fields (age, length, GC% and N%) the user must specify a range between a minimum and maximum value; in addition, the user can check the N/D (Not Defined) flag, thereby including in the result those sequences having an unknown value. Similarly, ranges of collection and submission dates can be selected using calendar-like drop-down components, supporting also the N/D flag.

Epitope Search

Interaction over epitope data can be conducted using three different modes, respectively for inputting user-defined epitopes, for extracting epitopes as they are imported from IEDB, and for associating to those epitopes statistics computed over the mutated sequences of the database. We detail the three scenarios in the following.

Mode 1: Custom Epitopes. This mode is particularly useful in the context of B cell epitopes, as these can be examined regardless of the HLA restriction on the targeted population. The user is provided with a panel, shown in Figure 3, for defining candidate epitopes by providing its name, a

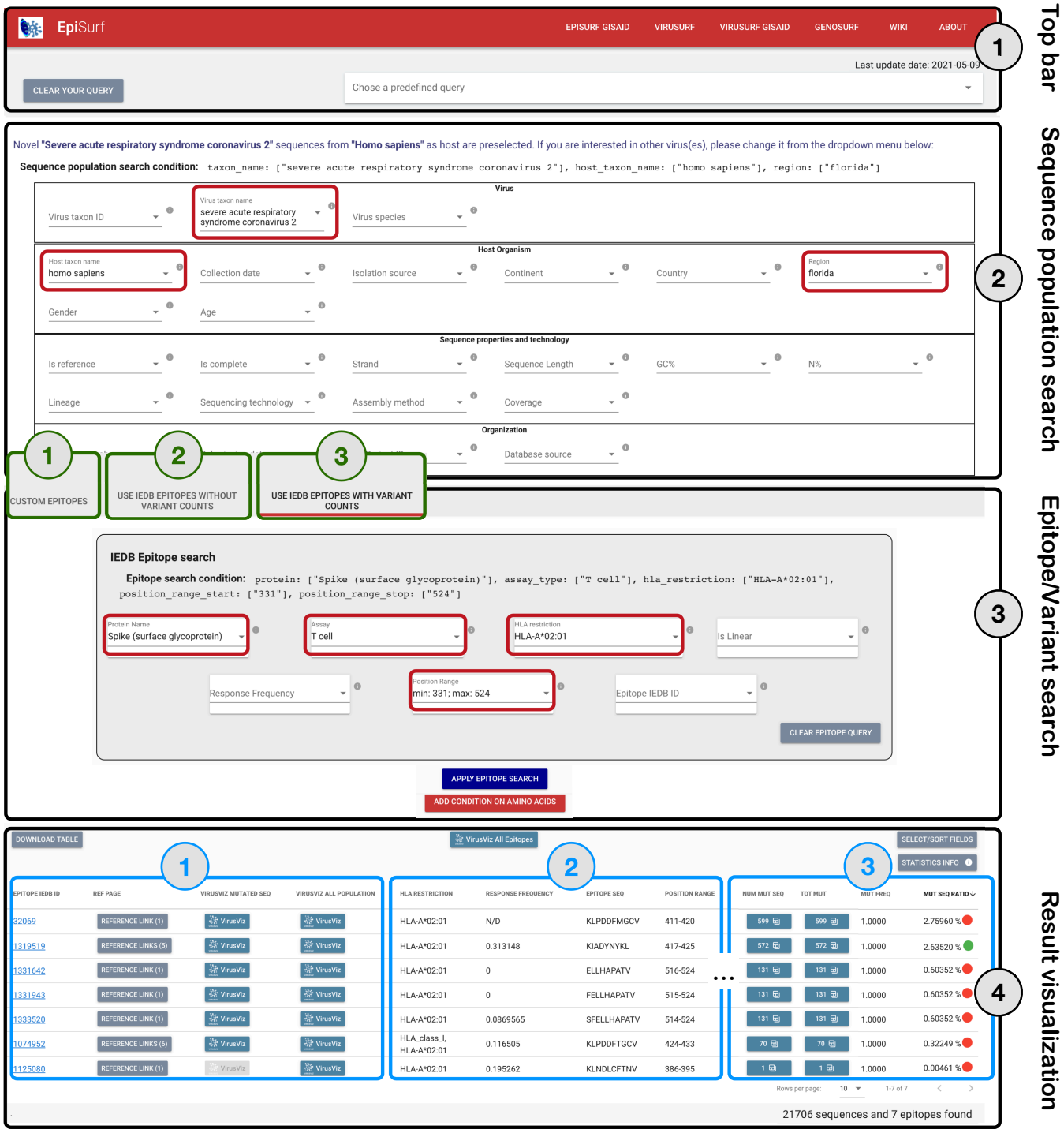

Figure 2. Overview of EpiSurf interface, divided in 4 parts (see black rectangles). *Part 1* (Top bar) allows to clear all filters selected in the interface or to select predefined example queries. *Part 2* (Sequence population search) is used for selecting sequences of interest. SARS-CoV-2 and human host are pre-selected, whereas the "Florida" location has been added as an example (see red rectangles). After metadata search, users select between three different modes (squared in green): 1) Custom epitopes (shown in detail in Figure 3); 2) Use IEDB epitopes without variant counts; 3) Use IEDB epitopes with variant counts. For parts 3 and 4 of this figure, we assume the selection of Modes 2 or 3, whereas Figure 3 shows Parts 3 and 4 after selecting Mode 1. *Part 3* (Epitope/Variant search) enables selecting Epitopes from IEDB. The Spike protein is pre-selected (but users can easily change the choice of protein), whereas other conditions allow to filter epitopes by using metadata available in IEDB. As an example, we show the selection of T cell assay, HLA-A*02:01 restriction, and a position range covering the Receptor Binding Domain (25)—see red rectangles. Finally, *Part 4* (Result visualization) provides a table describing selected epitopes, further vertically decomposed into three areas (shown in blue): *Area 1* includes a number of buttons to open the results within IEDB page, VirusViz (considering only sequences mutated on the epitope range or all the ones in the population selected in Part 2); *Area 2* includes sortable and adjustable metadata about epitopes; *Area 3* is present only in Mode 3 and includes counters that define the conservancy of the epitope in the population of interest. The columns of the table are customizable and the full table can be conveniently downloaded as a CSV file. The two most relevant counts in the search, i.e., the number of sequences and of epitopes, are provided at the bottom right corner of the web page.

specific protein on the virus and a position range (possibly discontinuous) on the protein. The epitope may be added to the list as is; in this case the statistics will be computed over the full sequence population selected with the Metadata search.

Optionally, the user may select an additional condition on one amino acid change, with the purpose of instructing the system to compute statistics over the fraction of the selected population that carries the amino acid change. The panel allows to select a specific protein, a range of coordinates, a type of variation (insertion, deletion, or substitution), an original amino acid residue and the corresponding alternative residue. The filter selection may be approved (ADD), deleted for choosing alternative ones (CLEAR), or deleted for removing the entire amino acid-related condition (CLEAR & CLOSE).

The choice of amino acid filters is supported by a practical add-on triggered by the "Analyze Substitutions" button, which allows to inspect the characteristics of a specific replacement from original into alternative amino acid. Each involved (source or target) residue is characterized by a series of structural categorical properties (such as polarity, charge and flexibility) and of numerical properties (e.g., molecular weight, hydrophobicity); the pair of residues involved in the change is associated to a measure of its impact (Grantham distance (26)); a threshold on impact maps a change into radical or conservative categories.

After addition, the new epitope is inserted in a list of user-defined epitopes, which are presented by providing summary information, including its name, creation and refresh date, protein and position range, and virus/host taxon name, number of mutated sequences and of variants. Current epitopes in the list can be downloaded as a JSON file, thereby supporting the possibility of reloading specific files representing the status of saved interaction with EpiSurf; in this way, users may organize and manage the information collected about user-defined epitopes through many sessions of EpiSurf use.

Out of the current epitope list, users may read more information on the epitope (MORE INFO); refresh its counters (REFRESH)—this option is typically used after uploading an external JSON file as discussed above; reload all the values (originally used to create that epitope) into the drop-down menus of the sequence and epitope search panels (RELOAD)—this option facilitates the creation of a new epitope with different coordinates or for testing its conservancy on a different underlying population; and finally delete the element from the list (DELETE).

The result table stores all the relevant information on the defined epitopes connecting them with statistics on the sequences mutated on each epitope's range. The table can be downloaded for subsequent data analysis as a CSV file. The last five columns of the table describe: (i) *NUM SEQ POPULATION*: the number of sequences available in the population where the epitope has been tested in EpiSurf (i.e., matching the filters in the Metadata and Amino Acid Condition columns); (ii) *NUM MUT SEQ*: the number of sequences in the selected population that have at least one amino acid change exactly matching with the epitope position range; (iii) *TOT MUT*: the number of total amino acid changes exhibited by the full population of sequences (note that any insertion counts for one); (iv) *MUTATED FREQ*: the ratio of total variants (iii) over the number of mutated sequences (ii); (v) *MUTATED SEQ RATIO*: the ratio of mutated sequences (ii) over the total of the selected population (i). When epitopes have been defined using also an amino acid condition, counters (ii) and (iii) are computed by considering the fraction of the population that exhibits the specific selected amino acid condition.

By clicking on the *NUM MUT SEQ* number, the list of mutated sequences with their metadata is shown in a table. From here, EpiSurf users may invoke VirusViz that will be opened on a variant distribution that considers all the mutated sequences and highlights the chosen epitope. By clicking on the *TOT MUT* number, the user will open a new panel called "Epitope mutation statistics", where the number of mutated sequences can be observed in a custom breakdown, grouping by several attributes concerning location, collection time, and phylogenetic classification methods. A table is generated providing, for each specific amino acid change in a row, the number of sequences exhibiting such change in each formed group.

Mode 2: IEDB Epitopes without variants count. Modes 2 and 3 take advantage of the epitopes publicly deposited to IEDB. In both modes, the user can select epitopes by using seven different drop-down menus, representing metadata attributes of epitopes, extracted from IEDB: the protein, the type of experiment performed to recognize the epitope, the presence of specific HLA restrictions, if it is linear or discontinuous, the allowed range of corrected response frequency, a range of coordinates (i.e., all epitopes overlapping with the coordinate range are selected), and the specific epitope identifier within IEDB. The selection condition is determined as a conjunction of the filters selected in each drop-down menu; the protein and position range accept a single value as a filter, the other attributes accommodate multiple values (intended in disjunction).

Results in the bottom panel are in a tabular format; each row of the table represents one epitope, with links that refer it back to IEDB pages and relevant metadata (columns can be sorted and selected/deselected); the full table can be downloaded as a CSV file. VirusViz may be invoked on: i) the full population of sequences and all epitopes in the results (using the button at the top of the table); ii) the full population of sequences and one specific epitope (using the button on the epitope's row). In both cases, users should invoke the visualizer after selecting populations of small size.

Mode 3: IEDB Epitopes with variants count. This mode adds to Mode 2 the computation of statistics. It is the most sophisticated use of EpiSurf, and requires a heavier computational load on the back-end (therefore, in addition to selecting small populations, users are also suggested to select a small number of epitopes). The specific feature introduced by this mode is the addition of an amino acid filter, which includes a variant position, type, original and alternative amino acid residue; its effect is to restrict the calculation of the four statistics only to sequences that exhibit matching amino acid changes. Changes may be chosen only among positions that are allowed by the previously set position range filter. As in Mode 1, we provide the "Analyze Substitutions" functionality to aid users in evaluating the characteristics of amino acid replacements.

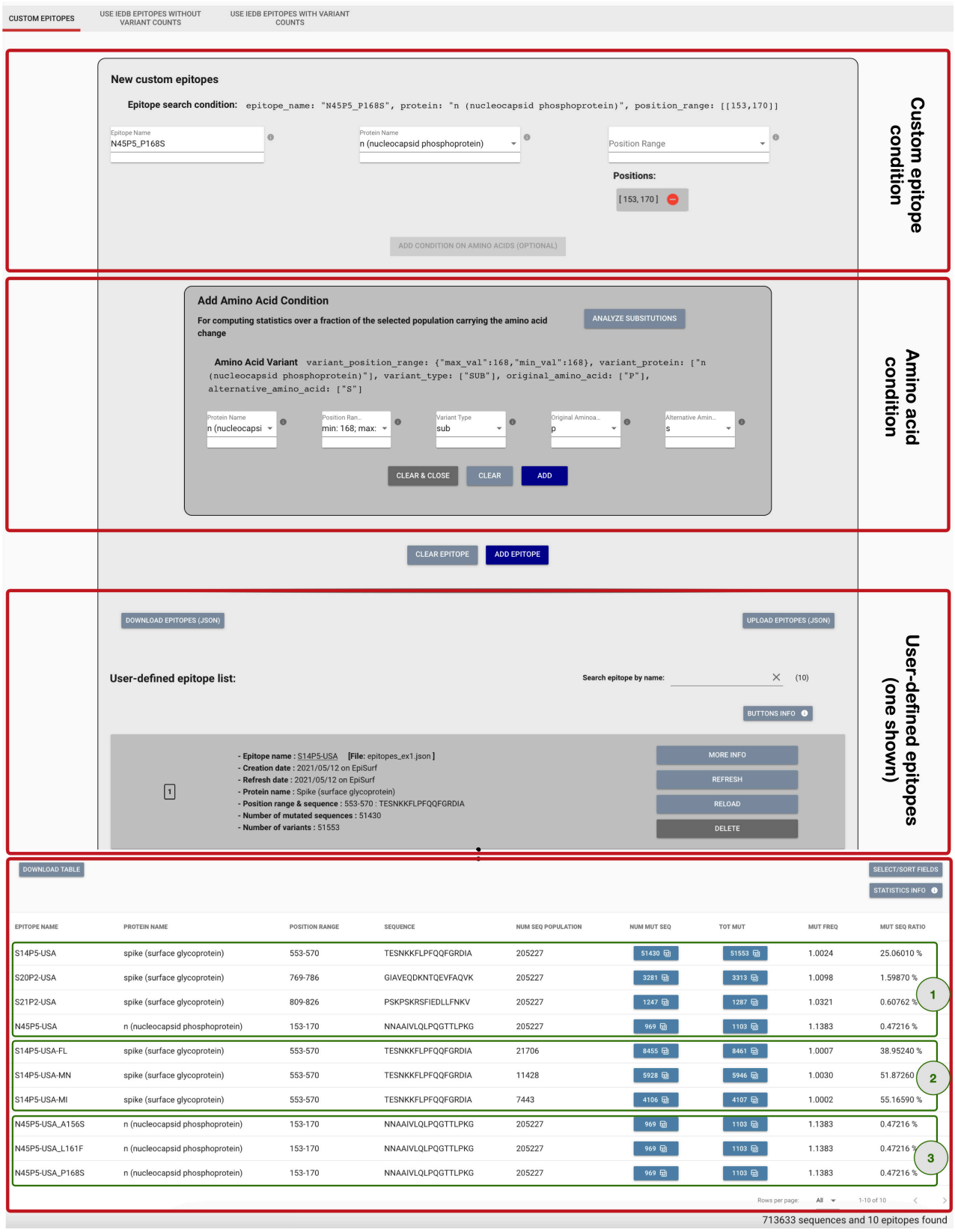

Figure 3. Epitope definition panels in the "Custom Epitope" Mode 1, where users can input their defined epitopes, inserting a name, the protein and range (or collection of ranges, when the proposed epitope is non-linear). In addition, the user can optionally add an amino acid condition for restricting the population selected in the Sequence Population Search panel (see Figure 2) to the viruses carrying a specific amino acid change. Once created, epitopes are added to a list, displayed on the page; each epitope information can be inspected by using the MORE INFO button, updated with the REFRESH button, modified with the RELOAD button, or removed with the DELETE button. Epitopes can be downloaded and then uploaded during a different session of EpiSurf use. The bottom table shows the results of the epitope design session, as further described in the Example 1 of the Use Cases section.

As in Mode 2, the result table provides links to IEDB or to invoke VirusViz, and entries describing metadata from IEDB; in addition, as in Mode 1, the result table provides in the last four entries a quantitative description of sequence changes over each epitope: (i) *NUM MUT SEQ*: the number of sequences in the selected population that exhibit at least one amino acid change within the epitope position range; (ii) *TOT MUT*: the number of total amino acid changes exhibited by the full population of sequences. (iii) *MUT FREQ*: the ratio of total variants (ii) over the number of mutated sequences (i); (iv) *MUT SEQ RATIO*: the ratio of mutated sequences (i) over the total of the selected population.

For this mode, we have designed a mechanism that ensures that users carefully select epitopes to be intersected with the population of interest. Indeed, while for B cell epitopes no attention is needed w.r.t. alleles expressed in the population, much concern should be dedicated when T cell or MHC ligand epitopes are targeted. In these cases, we recommend considering epitopes with a high response frequency, by setting a threshold – whose value was suggested by experts to be at least 0.2 – using the response frequency provided by IEDB. We opted to use this threshold with the corrected formula proposed in (23), being the threshold even more conservative in this case. Besides this, users should consider the percentages of *MUT SEQ RATIO* with care, ensuring that the HLA restriction is appropriate for the observed population (by checking suitable population alleles databases, e.g., the Allele Frequency Net Database (27)). Note that we have assigned a color code to support users in understanding how much they can rely on the observed statistics:

- `green` denotes epitopes that have been derived by B cell assays and/or by T cell/MHC ligand assays with a positive assay response frequency ≥ 0.2;
- `orange` denotes epitopes that have been derived by reliable assays (i.e., B cell assays or T cell/MHC ligand assays with response frequency ≥ 0.2) but also by less reliable assays (i.e., T cell/MHC ligand assays with response frequency < 0.2);
- `red` denotes epitopes that have been derived by T cell/MHC ligand assays with a positive assay response frequency < 0.2.

Similarly to Mode 1, the user may click on *NUM MUT SEQ* and *TOT MUT* numbers to activate further analysis features.

GISAID-specific EpiSurf.

EpiSurf presents a version that is specific for the data imported from GISAID, as the data agreement does not allow merging GISAID information with information from other sources. The GISAID version has a panel for population selection offering restricted options, otherwise Modes 1, 2 and 3 are available with no change, except that VirusViz buttons are not available. Of course, since amino acid variants are sourced from different databases, epitope mapping to amino acid variants produces different counts in the two systems.

USE CASES

Example 1. Amrun et al. (28) present four different immunodominant B cell assay epitopes, to be used as highly specific and sensitive serological diagnostic targets, i.e., to test for the presence of the virus in patients potentially exposed to SARS-CoV-2. The candidate epitopes are named S14P5, S20P2, S21P2 on the Spike protein, and N4P5 on the N protein. In (28) authors studied the conservation of these epitopes across 17k SARS-CoV-2 sequences publicly available at the time of their writing. They reported low rate of potential amino acid changes over the epitopes.

By using Mode 1 of EpiSurf (Custom Epitopes), it is possible to replicate such epitopes as ranges of positions on the proteins, respectively on [553-570], [769-786], [809-826] on Spike and [153-170] on N. These may be checked, for instance, against the EpiSurf sequence population from the USA, of about 205k sequences as of May 9th, 2021. The previously introduced Figure 3 shows a particular snapshot of the analysis session where the user has already inserted all four epitopes. See the third red rectangle framing the user-defined epitopes, where – for brevity – we only show the card produced for the first S14P5 epitope. When all four have been inserted, the user will be provided with the results framed by the green rectangle (1). It is worth noting that the first epitope has a high ratio of altered sequences, i.e., 25.1%. The user may be interested in inspecting the breakdown of such a consistent set of sequences. By clicking on the number of total amino acid changes (i.e. *TOT MUT*), we open the "Epitope mutation statistics" functionality. We may group by the country attribute, thereby obtaining a table that, for each amino acid change, reports the total of sequences exhibiting such changes, and the breakdown of such amount by 'country'. Through sorting by descending total count, we observe that the Spike A570D position is the most commonly mutated one. We also check the most impacted US states (grouping by the attribute 'region'), which are Florida, Minnesota, and Michigan. An alternative grouping can be performed on 'lineage', highlighting that the sequences with this mutation are almost always assigned to the B.1.1.7 lineage, corresponding to the Variant of Concern (first defined on a Virological.org post in December 2020 (29)). We can make our conservancy analysis more specific by adding new epitopes to our list, tested against smaller populations; in Figure 3, Result section, box (2) we have created the candidate epitopes S14P5-USA-FL, S14P5-USA-MN, and S14P5-USA-MI. In the *MUT SEQ RATIO* column, we observe that Minnesota (MN) and Michigan (MI) have a higher incidence of mutations on this epitope.

We then focus on N4P5 (reported in (28) as the most stable epitope out of the four). In our knowledge base (a corpus of variants' annotations regarding their increased/decreased effects on kinetics/epidemiology/immunology levels, obtained through a systematic search of published or preprint literature (5)), we find three amino acid changes falling within the scope of this epitope (namely, A156S, L161F, P168S); it has been claimed that they may lower the protein stability and modify the protein flexibility (30). Due to these changes there is the possibility that the specificity and sensitivity of serological tests for COVID-19 diagnosis may be impacted (leading to false negatives) (31). In Figure 3, in the first red box we show how we insert the custom epitope condition. In the second red box we input an amino acid condition, N protein, 168-168 position range for a substitution for P (Proline) to S (Serine). By adding this condition we are

EPITOPE IEDB ID	HLA RESTRICTION	RESPONSE FREQUENCY	EPITOPE SEQ	POSITION RANGE	NUM MUT SEQ	TOT MUT	MUT FREQ	MUT SEQ RATIO ↓
2802	HLA-A*02:01	0.292893	ALNTPKDHI	138-146	4868	4870	1.0004	31.5776 %
34851	HLA-A*02:01	0.310859	LALLLLDRL	219-227	3789	3789	1.0000	24.5784 %
54690	HLA-A*02:01	N/D	RLNQLESKV	226-234	566	566	1.0000	3.67150 %
1330575	HLA-A*02:01	0.292893	RLNQLESKM	226-234	566	566	1.0000	3.67150 %
1310539	HLA-C*07:02	N/D	KKADETQAL	374-382	496	496	1.0000	3.21740 %
1309136	HLA-B*27:05, HLA-C*07:02, HLA-C*07:01	0.105263	QRNAPRITF	9-17	267	267	1.0000	1.73200 %
1333386	HLA-C*07:01	0	RRGPEQTQGNF	276-286	75	75	1.0000	0.48651 %
1309134	HLA-A*02:01	0.25359	NTASWFTAL	48-56	63	64	1.0159	0.40867 %
37515	HLA-A*02:01	0.224851	LLLLDRLNQL	221-230	57	57	1.0000	0.36975 %
1330713	HLA-A*02:01	0	AADLDDFSKQL	397-407	47	47	1.0000	0.30488 %
3956	HLA-A*02:01	0.4	AQFAPSASA	305-313	40	40	1.0000	0.25947 %
1332478	HLA-C*07:01	0	KKQQTVTLL	387-395	40	40	1.0000	0.25947 %
1309129	HLA-B*07:02, HLA-A*02:01, HLA-C*07:02	0.311952	LSPRWYFYYLGTGPEAGL	104-121	37	37	1.0000	0.24001 %
1125053	HLA-A*02:01	N/D	DLDDFSKQL	399-407	36	36	1.0000	0.23352 %
1125146	HLA-A*02:01	0.195262	YLGTGPEAGL	112-121	36	36	1.0000	0.23352 %
1330485	HLA-A*02:01	0.5	IIWVATEGA	130-138	25	25	1.0000	0.16217 %
1330583	HLA-A*02:01	0.5	RTATKAYNV	262-270	21	21	1.0000	0.13622 %
38881	HLA-A*02:01	0.292893	LQLPQGTTL	159-167	19	19	1.0000	0.12325 %
69720	HLA-A*02:01	0.307104	VLQLPQGTTL	158-167	19	19	1.0000	0.12325 %
21347	HLA-A*02:01	0.0680757	GMSRIGMEV	316-324	15	15	1.0000	0.09730 %
27182	HLA_class_I, HLA-A*02:01	0.0474282	ILLNKHIDA	351-359	11	11	1.0000	0.07135 %
37473	HLA-A*02:01	0.223588	LLLDRLNQL	222-230	9	9	1.0000	0.05838 %
1125078	HLA-A*02:01	0.195262	KLDDKDPNF	338-346	7	7	1.0000	0.04541 %
1332792	HLA-C*07:01	0	MKDLSPRWY	101-109	7	7	1.0000	0.04541 %
125100	HLA-A*02:01	0.292893	ILLNKHID	351-358	3	3	1.0000	0.01946 %

Figure 4. Result table of the query performed in Example 2: list of epitopes with their statistics descriptive of mutation rate over the selected population (Campania, Italy). Color codes are use to discriminate between epitopes with a response frequency > 0.2 (green) or < 0.2 (red).

instructing the system to compute statistics only over the fraction of the selected population (all SARS-CoV-2, human host sequences in this example) that carry the specified amino acid change. In the result section, green box (3), we observe that the mutated sequences ratios for the N4P5 epitope over the three defined populations are quite low (below 0.5%). Note that, if the *MUT FREQ* is close to 1 – when one of these three changes is present – no other change is carried within the epitope scope. Overall, the impacts on the epitope are minimal but attention should be paid and further investigations needed to proceed with its use for serologic assays.

Example 2. Aiming to pave the way for designing novel vaccine candidates, Rakib et al. (32) propose to focus on epitopes located along the nucleocapsid (N) protein, especially for its high conservancy and dominant/long-lasting immune response (previously reported against SARS-CoV (33) and infectious bronchitis virus (34)). From an immunological point of view, vaccine development has historically relied mostly on B cell immunity, but recent discoveries (35) revealed that T cell epitopes are also promising, leading to a more long-lasting immune response mediated by CD8+ T cells (thus recognizing viral peptides in the MHC class I area). By using EpiSurf in its Mode 2 we may perform a search on the updated corpus of IEDB deposited epitopes and then test their conservancy on selected populations. Drop-down menus may be employed to select epitopes on the *protein = N (nucleocapsid phosphoprotein)*, with *assay type = T cell*, and *hla restriction = HLA class I*. We then recommend selecting a high response frequency for positive assays (at least 0.2). Alternatively, one could select specific alleles from the *hla restriction* menu and test epitopes only on a population of sequences from hosts that exhibits such alleles (in a statistically significant way).

Suppose we are observing the Italian population of sequences from the Campania region (to date, most Italian

MUTATION	↓ TOTAL [5148]	B_1_596 (2020-11) [1]	B_1_596 (2020-12) [2]	B_1_596 (2021-01) [63]	B_1_596 (2021-02) [191]	B_1_596 (2021-03) [105]	B_1_596 (2021-04) [14]
P681H	4082	0	0	1	1	0	0
Q677P	369	1	2	59	189	103	14
Q677H	321	0	0	1	1	0	0
A688V	103	0	0	2	0	0	0
H655Y	98	0	0	0	0	0	0

Rows per page: 5 1-5 of 30

Figure 5. Epitope mutation statistics result of Example 3.

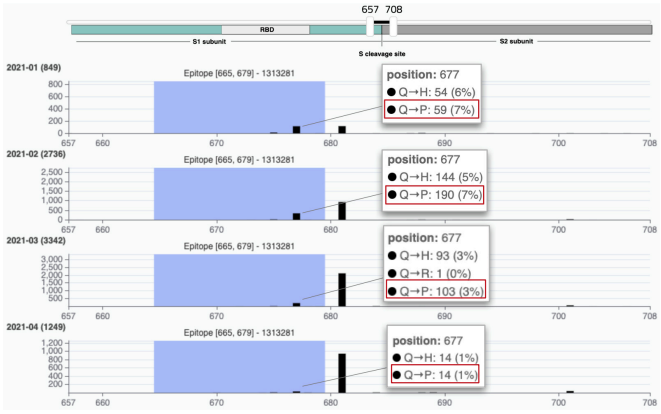

Figure 6. VirusViz compare functionality was applied to four groups of sequences collected in Texas in the first months of 2021 (see Example 3). We highlight a particular epitope that does not include the highly mutated 681 position (belonging to B.1.1.7 lineage, known as UK variant of concern) but does include the position 677 and thus the Q766P mutation.

sequences were sequenced here and GISAID depositions are about 15k, as of May 9th, 2021). From the Allele Frequency Net Database web portal (27)) we gather that the most present (above a 40% threshold) alleles in 'Italy South Campania Region' are DRB1*11 (49.6%), A*02 (43.0%), and C*07 (41.7%). In EpiSurf we find no occurrence of T cell assay epitopes restricted to HLA-DBR1*11. However, we do find 20 ones restricted to HLA-A*02 and 7 ones restricted to HLA-C*07 (with two different subtypes). For observing the distribution of mutations over such epitopes we switch to Mode 3, which preserves our previous selections. By applying the epitope search, we obtain the table shown in Figure 4, where we appreciate the total of 25 epitopes (note that two of these, IEDB 1309129 and 1309136 match multiple HLA restriction filter conditions). Out of these, 13 of them exhibit a green mark, as their response frequency (for positive assays) is above the 0.2 threshold, whereas the other 12 show a red mark, meaning that users should carefully consider the statistics, as they may not be meaningful in the selected population. We also note that the most mutated epitope (ID 2802) has almost 4900 amino acid changes in the Campania population, most of which are of type L139F (also reported to modify protein flexibility and stability (30)).

Example 3. In the last months of 2020 there has been interest on studying seven independent lineages circulating in the USA all having a change of the Glutamine (Q) amino acid at position 677 of Spike protein. These all seemed to have originated and spread in the last few months (36). Q at 677 is more commonly mutated into Histidine (H) – leading to a non-radical change – however, in a considerable number of cases (predominantly in Texas) the change to Proline (P) has been observed, being this a radical change (Grantham distance = 76).

According to Hodcroft et al. (36), this change should be monitored as 677 is nearby to – though outside of – the furin binding pocket (polybasic site), important for S1/S2 cleavage; therefore hypothetically the presence of Proline in this precise spot could influence cleavage of S1/S2. As this change is radical, with potentially interesting effects, and geographically quite delimited, it serves as a good candidate to be monitored within EpiSurf. From our system, we choose the full sequence population of Texas (about 12k sequences, as of May 9th, 2021), then – using Mode 3 – we select epitopes located on the Spike protein that overlap the 677 position (obtaining 24 results), finally we add the condition on the presence on the substitution at 677 into P residue. As a result, we observe that all 24 epitopes exhibit 369 sequences where such change occurs. This set of sequences may be further inspected by clicking on any number in the *TOT MUT* column. The shown "Epitope mutation statistics" functionality can be employed to group by lineage and collection month to observe – as shown in Figure 5 – that all such sequences belong to lineage B.1.596 and that the Q677P change has been observed mostly between January and March 2021. A user may further investigate the selected Texas population by, for example, clicking on the "VirusViz All Epitopes" button. The tool opens directly on the distribution of amino acid variants of the Spike protein. From the left menu "Highlight region", users may select epitopes one at a time, thereby highlighting a specific position range in the bar plot. Note that immediately from the first visualization it is evident that 34% of the population exhibits the P681H amino acid change, feared for impacting antibody recognition of linear SARS-CoV-2 epitopes, reducing class 3 antibody recognition (even if this was only suggested in non-peer reviewed literature (37) so far). The user may then want to remove all epitopes that include this critical position. This may be achieved in the 'Regions' page, where epitopes

are presented in the form of a list and can be dropped. Only two remain: IEDB ID 1313281 (position range 655-679) and 1310485 (position range 666-680). In the Population page the user may observe that there has been a considerable amount of depositions of sequences collected since 2021. It may be interesting to observe how the variant distributions behave (in terms of percentages) in the initial months of this year. By building different groups for the first four months of 2021, we produce a comparative visualization shown in Figure 6. Here we observe that the mutation Q677P decreases its occurrences in percentage, therefore diminishing the concerns on its possible effects on epitopes designed in this position range.

Example 4. The focus of EpiSurf is on SARS-CoV-2, however variation over epitopes of other viral species may be analyzed. Chen et al. (38) generated 12 monoclonal antibodies as experimental candidates to develop antibodies neutralizing Dengue Virus serotype 1 (DENV-1). They define an epitope on the domain III of the DENV-1 E protein, spanning from residues 346 and 360 with the sequence TQNGRLITANPIVTD, deemed as a highly conserved region among different genotypes of DENV-1.

The conservation of the epitope can be checked against EpiSurf sequences. We restrict our search to Dengue Virus 1 sequences deposited on Genbank that were collected from human hosts before October 2017 (matching the publication date of the work from Chen et al.). We also select only complete sequences to ensure the accuracy of the variation calling algorithm, thus retrieving a total of 1,665 sequences. By using Mode 1 of EpiSurf (Custom Epitopes), we then build the target epitope on the E protein with the range 346-360. As a result, we obtain that 64 of the sequences exhibit at least one mutation (*MUT FREQ* of 1.06), representing the 3.84% of the total set (*MUT SEQ RATIO*).

Chen et al. perform multiple sequence alignment of residues corresponding to the proposed epitope, thereby showing that it contains three conserved residues, namely G349, R350, P356. By using the "Epitope mutation statistics" panel, we can check if any mutation occurs at these specific positions; specifically, we find only one sequence with the G349D substitution and one with the P356H substitution. We can also confirm the presence of L351V, an additional amino acid substitution mentioned by Chen et al., appearing in two sequences. Incidentally, we notice that the mutation T346I, not mentioned in their work, is the most present in the dataset (29 sequences) and could thus be further investigated.

DISCUSSION

During 2020 and the beginning of 2021, fueled by the outbreak of the COVID-19 pandemics, huge interest has been focused on studying epitopes, parts of the SARS-CoV-2 sequence that can be recognized by vaccines, drugs, and serological tests. For epitopes, IEDB is recognized as the most important, fully public repository, as of today collecting about five thousand epitopes of SARS-CoV-2 (along with many other viruses), well-described by means of attributes and search panels.

Several computational tools are used for supporting the prediction of epitopes. Our EpiSurf system covers a different need, as it provides a flexible interface for testing their conservancy, measured as the presence/absence of amino acid changes over epitope sequences. The unique aspect of EpiSurf is the ability to perform such conservancy testing by intersecting epitopes of interest, extracted by means of queries on IEDB, against the amino acid changes that are present on arbitrarily selected viral sequences, e.g., by lineage, location, or time of sequence collection. Such queries upon viral sequences can also be used against custom epitopes, freely entered by EpiSurf users.

Extensive analytical and visual support is offered to the users, including aggregations by metadata, statistics of mutated sequences, and distribution plots (through our connection to VirusViz). EpiSurf aims to be a solid companion tool for researchers designing epitopes that need to rely on the big corpus of sequences available online.

DATA AVAILABILITY

The code of EpiSurf is available on GitHub at `https://github.com/DEIB-GECO/EpiSurf/` and on Zenodo at `http://doi.org/10.5281/zenodo.5121287`. The system is documented in the related WIKI at `https://github.com/DEIB-GECO/EpiSurf/wiki/`.

FUNDING

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

ACKNOWLEDGEMENTS

The authors would like to thank Matteo Chiara, Alba Grifoni, Carla Mavian, Brittany Rife Magalis, Marco Salemi, Anna Sandionigi, for their contribution to requirements elicitation and for inspiring future developments of this research. They would also like to thank Andrea Gulino for the development of VirusViz and Damianos P. Melidis for the first prototype of IEDB data import pipeline. The authors are grateful to the GISAID Initiative for the data sharing agreement that allowed the development of the GISAID-specific version of EpiSurf. 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.

Conflict of interest statement. None declared.

REFERENCES

1. Canakoglu, A., Bernasconi, A., Colombo, A., Masseroli, M., and Ceri, S. (2019) GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database,* **2019**.
2. Bernasconi, A., Canakoglu, A., Pinoli, P., and Ceri, S. (2020) Empowering Virus Sequence Research Through Conceptual Modeling. In Dobbie, G., Frank, U., Kappel, G., Liddle, S. W., and Mayr, H. C., (eds.), *Conceptual Modeling*, Cham: Springer International Publishing pp. 388–402.
3. Canakoglu, A., Pinoli, P., Bernasconi, A., Alfonsi, T., Melidis, D. P., and Ceri, S. (2020) ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Research,* **49**(D1), D817–D824.
4. Bernasconi, A., Gulino, A., Alfonsi, T., Canakoglu, A., Pinoli, P., Sandionigi, A., and Ceri, S. (2021) VirusViz: Comparative analysis and

effective visualization of viral nucleotide and amino acid variants. *Nucleic Acids Research,* p. gkab478.
5. Al Khalaf, R., Alfonsi, T., Ceri, S., and Bernasconi, A. (2021) CoV2K: a Knowledge Base of SARS-CoV-2 Variant Impacts. In *print on LNCS Proceedings of the International Conference on Research Challenges in Information Science 2021*.
6. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2019) The immune epitope database (IEDB): 2018 update. *Nucleic acids research,* **47**(D1), D339–D343.
7. Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., and Rosales-Mendoza, S. (2015) An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *Journal of biomedical informatics,* **53**, 405–414.
8. Wu, J., Chen, W., Zhou, J., Zhao, W., Sun, Y., Zhu, H., Yao, P., Chen, S., Jiang, J., and Zhou, Z. (2020) COVIEdb: A database for potential immune epitopes of coronaviruses. *Frontiers in Pharmacology,* **11**, 1401.
9. Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., and Sette, A. (2020) A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell host & microbe,* **27**(4), 671–680.
10. Patiyal, S., Kaur, D., Kaur, H., Sharma, N., Dhall, A., Sahai, S., Agrawal, P., Maryam, L., Arora, C., and Raghava, G. P. (2020) A Web-Based Platform on Coronavirus Disease-19 to Maintain Predicted Diagnostic, Drug, and Vaccine Candidates. *Monoclonal Antibodies in Immunodiagnosis and Immunotherapy,* **39**(6), 204–216.
11. Sahoo, S., Mahapatra, S. R., Parida, B. K., Rath, S., Dehury, B., Raina, V., Mohakud, N. K., Misra, N., and Suar, M. (2021) DBCOVP: A database of coronavirus virulent glycoproteins. *Computers in biology and medicine,* **129**, 104131.
12. Gupta, A. K., Khan, M., Choudhury, S., Mukhopadhyay, A., Rastogi, A., Thakur, A., Kumari, P., Kaur, M., Saini, C., Sapehia, V., et al. (2020) CoronaVR: A Computational Resource and Analysis of Epitopes and Therapeutics for Severe Acute Respiratory Syndrome Coronavirus-2. *Frontiers in microbiology,* **11**, 1858.
13. Massacci, A., Sperandio, E., D'Ambrosio, L., Maffei, M., Palombo, F., Aurisicchio, L., Ciliberto, G., and Pallocca, M. (2020) Design of a companion bioinformatic tool to detect the emergence and geographical distribution of SARS-CoV-2 Spike protein genetic variants. *Journal of translational medicine,* **18**(1), 1–7.
14. Ward, D., Higgins, M., Phelan, J. E., Hibberd, M. L., Campino, S., and Clark, T. G. (2021) An integrated in silico immuno-genetic analytical platform provides insights into COVID-19 serological and vaccine targets. *Genome medicine,* **13**(1), 1–12.
15. Bui, H.-H., Sidney, J., Li, W., Fusseder, N., and Sette, A. (2007) Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC bioinformatics,* **8**(1), 1–6.
16. Ahmed, S. F., Quadeer, A. A., and McKay, M. R. (2020) COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2. *Nature protocols,* **15**(7), 2141–2142.
17. Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., et al. (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research,* **40**(D1), D593–D598.
18. de Silva, T. I., Liu, G., Lindsey, B. B., Dong, D., Shah, D., Mentzer, A. J., Angyal, A., Brown, R., Parker, M. D., Yin, Z., et al. (2021) The impact of viral mutations on recognition by SARS-CoV-2 specific T-cells. *bioRxiv,* https://doi.org/10.1101/2021.04.08.438904.
19. Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC bioinformatics,* **8**(1), 1–12.
20. Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2019) GenBank. *Nucleic acids research,* **47**(D1), D94–D99.
21. The COVID-19 Genomics UK (COG-UK) consortium (2020) An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe,*.
22. Elbe, S. and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges,* **1**(1), 33–46.
23. Carrasco Pro, S., Sidney, J., Paul, S., Lindestam Arlehamn, C., Weiskopf, D., Peters, B., and Sette, A. (2015) Automatic generation of validated specific epitope sets. *Journal of immunology research,* **2015**.
24. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* **48**(3), 443–453.
25. Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S., Zhou, Y., and Du, L. (2020) Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cellular & molecular immunology,* **17**(6), 613–620.
26. Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science,* **185**(4154), 862–864.
27. Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. d., Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020) Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic acids research,* **48**(D1), D783–D788.
28. Amrun, S. N., Lee, C. Y.-P., Lee, B., Fong, S.-W., Young, B. E., Chee, R. S.-L., Yeo, N. K.-W., Torres-Ruesta, A., Carissimo, G., Poh, C. M., et al. (2020) Linear B-cell epitopes in the spike and nucleocapsid proteins as markers of SARS-CoV-2 exposure and disease severity. *EBioMedicine,* **58**, 102911.
29. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. `https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563` Accessed: July 13th, 2021.
30. Rahman, M. S., Islam, M. R., Alam, A. R. U., Islam, I., Hoque, M. N., Akter, S., Rahaman, M. M., Sultana, M., and Hossain, M. A. (2021) Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *Journal of medical virology,* **93**(4), 2177–2195.
31. Petherick, A. (2020) Developing antibody tests for SARS-CoV-2. *The Lancet,* **395**(10230), 1101–1102.
32. Rakib, A., Sami, S. A., Islam, M., Ahmed, S., Faiz, F. B., Khanam, B. H., Marma, K. K. S., Rahman, M., Uddin, M. M. N., Nainu, F., et al. (2020) Epitope-Based Immunoinformatics Approach on Nucleocapsid Protein of Severe Acute Respiratory Syndrome-Coronavirus-2. *Molecules,* **25**(21), 5088.
33. Peng, H., Yang, L.-t., Wang, L.-y., Li, J., Huang, J., Lu, Z.-q., Koup, R. A., Bailer, R. T., and Wu, C.-y. (2006) Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein in SARS-recovered patients. *Virology,* **351**(2), 466–475.
34. Seah, J., Yu, L., and Kwang, J. (2000) Localization of linear B-cell epitopes on infectious bronchitis virus nucleocapsid protein. *Veterinary microbiology,* **75**(1), 11–16.
35. Gilbert, S. C. (2012) T-cell-inducing vaccines–what's the future. *Immunology,* **135**(1), 19–26.
36. Hodcroft, E. B., Domman, D. B., Snyder, D. J., Oguntuyo, K., Van Diest, M., Densmore, K. H., Schwalm, K. C., Femling, J., Carroll, J. L., Scott, R. S., et al. (2021) Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677. *medRxiv,*.
37. Haynes, W. A., Kamath, K., Lucas, C., Shon, J., and Iwasaki, A. (2021) Impact of B. 1.1. 7 variant mutations on antibody recognition of linear SARS-CoV-2 epitopes. *medRxiv,*.
38. Chen, W.-H., Chou, F.-P., Wang, Y.-K., Huang, S.-C., Cheng, C.-H., and Wu, T.-K. (2017) Characterization and epitope mapping of Dengue virus type 1 specific monoclonal antibodies. *Virology journal,* **14**(1), 1–11.