ViruSurf: an integrated database to investigate viral sequences

Arif Canakoglu [1,†], **Pietro Pinoli** [1,†], **Anna Bernasconi** [1], **Tommaso Alfonsi**[1],
Damianos P. Melidis [2] **and Stefano Ceri** [1,*]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano,
Italy and [2]L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany

ABSTRACT

**ViruSurf, available at http://gmql.eu/virusurf/, is a
large public database of viral sequences and inte-
grated and curated metadata from heterogeneous
sources (RefSeq, GenBank, COG-UK and NMDC); it
also exposes computed nucleotide and amino acid
variants, called from original sequences. A GISAID-
specific ViruSurf database, available at http://gmql.
eu/virusurf_gisaid/, offers a subset of these function-
alities. Given the current pandemic outbreak, SARS-
CoV-2 data are collected from the four sources; but
ViruSurf contains other virus species harmful to hu-
mans, including SARS-CoV, MERS-CoV, Ebola and
Dengue. The database is centered on sequences, de-
scribed from their biological, technological and or-
ganizational dimensions. In addition, the analytical
dimension characterizes the sequence in terms of
its annotations and variants. The web interface en-
ables expressing complex search queries in a sim-
ple way; arbitrary search queries can freely combine
conditions on attributes from the four dimensions,
extracting the resulting sequences. Several example
queries on the database confirm and possibly im-
prove results from recent research papers; results
can be recomputed over time and upon selected pop-
ulations. Effective search over large and curated se-
quence data may enable faster responses to future
threats that could arise from new viruses.**

INTRODUCTION

The pandemic outbreak of the coronavirus disease COVID-
19, caused by the virus species SARS-CoV-2, has created
unprecedented attention toward the genetic mechanisms of
viruses. The sudden outbreak has also shown that the re-
search community is generally unprepared to face pandemic
crises in a number of aspects, including well-organized
databases and search systems. We respond to such urgent
need by means of a novel integrated database and search
system collecting and curating virus sequences with their
properties. Data are captured, standardized, organized and
made accessible to the scientific community, so as to facili-
tate current and future research studies.

In our work, we are driven by the Viral Conceptual
Model (VCM) for virus sequences (1), which was recently
developed by interviewing a variety of experts of the various
aspects of virus research (including clinicians, epidemiolo-
gists, drug and vaccine developers). The conceptual model
is general and applies to any virus. The sequence of the virus
is the central information; sequences are analyzed from a *bi-
ological dimension* describing the virus species and the host
environment; a *technological dimension* describing the se-
quencing technology; an *organizational dimension* describ-
ing the project that was responsible for producing the se-
quence; and an *analytical dimension* describing properties of
the sequence, such as known annotations and variants. An-
notations include known genes, coding and untranslated re-
gions, and so on. Variants are extracted by performing data
analysis and include both nucleotide variants—with respect
to the reference sequence for the specific species—with their
impact, and amino acid variants related to the genes.

We have previously proposed another conceptual model
focused on human genomics (2), which was based on a
central entity representing files of genomic regions, sim-
ilarly described from various dimensions. We next devel-
oped and implemented an integrated database (3), search-
able through the GenoSurf (4) interface (http://gmql.eu/
genosurf/). Thanks to such previous knowledge in human
genomics, we have been able to rapidly design VCM and
then to deploy ViruSurf.

Currently, ViruSurf includes reference sequences from
RefSeq (5) and regular sequences from GenBank (6) of
SARS-CoV-2 and SARS-related coronavirus, as well as
MERS-CoV, Ebola and Dengue viruses; the pipeline is
generic and other virus species will be progressively added
next, giving precedence to those species which are most

Table 1. Summary of ViruSurf content as of 4 August 2020. For each taxon name (identified by a taxon ID and rank) and each source, we specify the number of distinct sequences and the reference genome; we also provide the average number of annotations, nucleotide variants and amino acid variants per sequence. The GISAID-only entry refers to those GISAID sequences that are not also present in the other three sources

Taxon rank	Taxon ID	Taxon name	Source	#Seq.	Reference	Avg Annot.	Avg Nuc Var	Avg AA Var
No rank	2697049	SARS-CoV-2	GISAID all	76 664	EPI_ISL_402124	–	–	4.8
No rank	2697049	SARS-CoV-2	GISAID only	46 366	EPI_ISL_402124	–	–	4.7
No rank	2697049	SARS-CoV-2	GenBank + RefSeq	13 309	NC_045512.2	28.0	17.6	24.0
No rank	2697049	SARS-CoV-2	COG-UK	38 124	NC_045512.2	28.0	25.5	58.0
No rank	2697049	SARS-CoV-2	NMDC	295	NC_045512.2	28.1	28.8	56.5
Species	694009	SARS-CoV	GenBank + RefSeq	673	NC_004718.3	14.0	91.6	19.7
Species	1335626	MERS-CoV	GenBank + RefSeq	1381	NC_019843.3	27.0	104.0	87.4
Species	2010960	Bombali ebolavirus	GenBank + RefSeq	5	NC_039345.1	9.0	126.8	41.6
Species	565995	Bundibugyo ebolavirus	GenBank + RefSeq	22	NC_014373.1	9.0	130.5	40.9
Species	186539	Reston ebolavirus	GenBank + RefSeq	57	NC_004161.1	9.0	126.1	29.4
Species	186540	Sudan ebolavirus	GenBank + RefSeq	40	NC_006432.1	9.0	368.6	51.0
Species	186541	Tai Forest ebolavirus	GenBank + RefSeq	9	NC_014372.1	9.0	4.6	2.8
Species	186538	Zaire ebolavirus	GenBank + RefSeq	2938	NC_002549.1	9.0	503.7	66.3
Strain	11053	Dengue virus 1	GenBank + RefSeq	11 185	NC_001477.1	15.0	469.7	200.9
Strain	11060	Dengue virus 2	GenBank + RefSeq	8692	NC_001474.2	15.0	410.0	117.7
Strain	11069	Dengue virus 3	GenBank + RefSeq	5344	NC_001475.2	15.0	269.3	118.6
Strain	11070	Dengue virus 4	GenBank + RefSeq	2492	NC_002640.1	15.0	265.1	147.2

harmful to humans. For what concerns SARS-CoV-2, we also include sequences from COG-UK (7) and NMDC (http://nmdc.cn/). GenBank and COG-UK data are made publicly available and can be freely downloaded and redistributed. Special arrangements have been agreed with GISAID (8,9), resulting in a GISAID-enabled version of ViruSurf. Due to constraints imposed by GISAID, the database exposed in this version lacks the original sequences, certain metadata and nucleotide variants; moreover, GISAID requires their dataset not to be merged with other datasets. Hence, the two versions of ViruSurf should be used separately, and a certain amount of integration effort must be carried out by the user.

Although the origin sources provide well-organized data portals (NCBI Virus (10), the COVID-19 Data Portal https://www.covid19dataportal.org/ and GISAID EpiCoV™ data browser), these do not allow for an integrated search over multiple sources, nor they provide fast selection using sequence variants. A number of other integrated interfaces are being developed in alternative research contexts: UCSC SARS-CoV-2 Genome Browser (11); 2019nCoVR (12) at the Chinese National Genomics Data Center; Virus-DIP (13) at the China National GeneBank; CovSeq (14); and CARD (15). Compared to other resources, ViruSurf has much stronger query and search capabilities; we use the power of conceptual modeling to structure metadata and to organize data integration and curation; we support search queries allowing to combine filters on metadata, nucleotide and amino acid mutations in an effective and scalable way, treating all of them as first-class citizens. The closest comparison is with CoV-GLUE (16) (only containing GISAID data), which includes variants, but forces users to search one variant at a time, visualizing the list of sequences with that variant.

MATERIALS AND METHODS

Database content

For SARS-CoV-2, ViruSurf contains data from four main data sources: NCBI (including Genbank and RefSeq),

COG-UK, NMDC and GISAID. We also reviewed other available sources (GenomeWarehouse and CNGdb) but observed that they do not add substantial value to the integration effort, as most of their sequences overlap with those stored in the four cited sources.

Table 1 provides a description of the current ViruSurf content: for each virus we report the rank, ID and name from NCBI Taxonomy, the number of sequences included from each source and the reference sequence. We next provide the average number per sequence of each annotation and nucleotide/amino acid variants computed against the reference sequence. Note that, although GISAID uses a different reference sequence, provided amino acid variants are relative to protein sequences (which are the same as in other sources), hence they can be compared with other variants.

Some content of ViruSurf is extracted from the sources and used without changes; some is manually curated; and some is computed in-house (nucleotide and amino acid variants, impact, quality measures). The code repository is available on GitHub (https://github.com/DEIB-GECO/virusurf_downloader). The current content corresponds to data available at the sources on 4 August 2020. We will provide periodical updates on a monthly basis.

Relational schema

The core schema, represented in Figure 1, is inspired to classic data marts (17), with a central fact table describing the SEQUENCE, featuring several characterizing attributes, and then four dimensions:

(1) The *Biological dimension* (including the VIRUS and HOSTSAMPLE tables) is concerned with the virus species characterization and the host organism, including the temporal/spatial information regarding the extraction of the biological sample.
(2) The *Technological dimension* (EXPERIMENTTYPE table) describes the sequencing method.
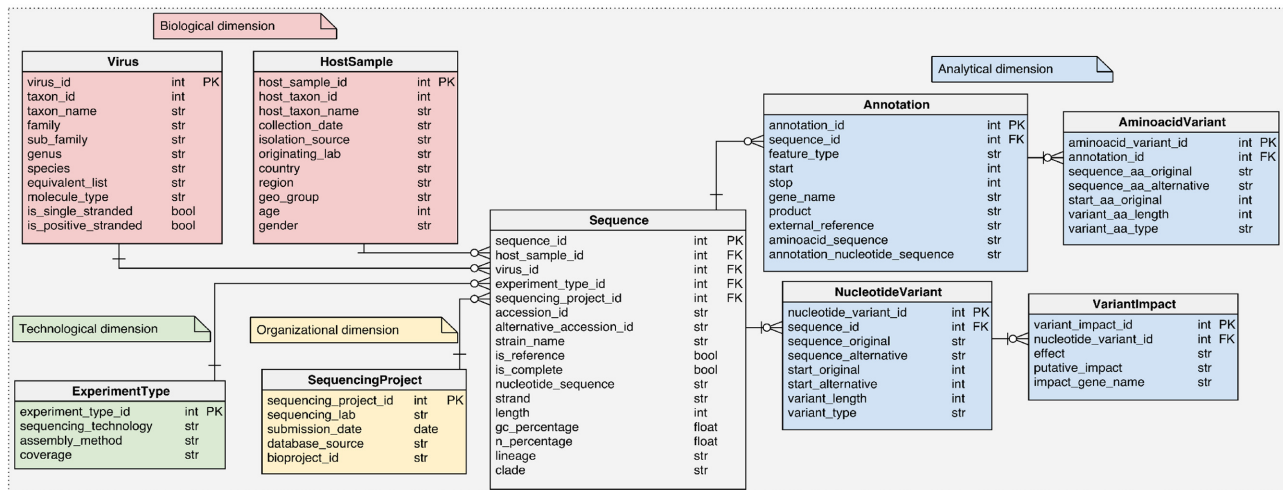(3) The *Organizational dimension* (SEQUENCINGPROJECT table) describes the project producing each sequence.

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

(4) The *Analytical dimension* provides annotations for specific sub-sequences and characterizes the variants in the nucleotide sequence and in the amino acid sequence. It includes the ANNOTATION, AMINOACIDVARIANT, NUCLEOTIDEVARIANT and VARIANTIMPACT.

All tables have a numerical sequential primary key (PK), conventionally named using the table name and the postfix '_id', and indicated as PK in Figure 1; we indicate with foreign keys (FK) the relationships from a non-key attribute to a primary key attribute of a different table. Relationships from the SEQUENCE towards VIRUS, HOSTSAMPLE, SEQUENCINGPROJECT and EXPERIMENTTYPE are functional (e.g. one SEQUENCE has one EXPERIMENTTYPE, while an EXPERIMENTTYPE may be the same for multiple SEQUENCE); instead, relationships in the analytical dimension are 1:N (e.g. one SEQUENCE has many ANNOTATIONS, and an ANNOTATION has many AMINOACIDVARIANTS). In the Supplementary Table S1, we describe the specific attributes of every table.

Data import

The pipeline used to import the content of the ViruSurf database from sources is shown in Figure 2. We use different download protocols for each source:

- For NCBI data (including GenBank and RefSeq sequences), we employ the extraction tools available in the E-utilities (18): the Python APIs allows to retrieve one complex XML file for each sequence ID available in NCBI.
- COG-UK instead provides a single MultiFASTA file on its website (https://www.cogconsortium.uk/data/); this is associated with a text file for metadata.
- NMDC exposes an FTP server with FASTA files for each sequence, while metadata are captured directly from the HTML description pages.
- GISAID provides to us an export file in JSON format, updated every 15 min. The file is produced by GISAID technical team in an *ad-hoc* agreed form for ViruSurf.

Automatic pipelines have been implemented to extract metadata and fill the SEQUENCE, VIRUS, HOSTSAMPLE, SEQUENCINGPROJECT, and EXPERIMENTTYPE tables; some attributes require data curation, as next described.

Annotation and variant calling

In order to provide homogeneous information for sequence annotations and variants, we use a unique annotation procedure for GenBank, COG-UK and NMDC; resulting variants for amino acid sequences are consistent with those provided by GISAID. We extract: structural annotations, nucleotide and amino acid sequences for each annotated segment, nucleotide variants and their impact, amino acid variants for the proteins and other information such as percentage of specific nucleotide bases.

For each virus, we manually select a reference sequence and a set of annotations, comprising coordinates for codifying and structural regions, as well as the amino acid sequences of each protein. Usually, such data are taken from the RefSeq entry for the given virus (e.g. NC_045512 for SARS-CoV-2). For each imported sequence, the pipeline starts by computing the optimal global alignment to the reference by means of the dynamic programming Needleman-Wunsch (NW) algorithm (19). The time and space complexity of NW is quadratic in the length of the aligned sequences, which often hinders its adoption in genomics, but viral sequences are relatively short, thus we can use NW rather than faster heuristic methods. We configured the algorithm to use an affine gap penalty, so as to favor longer gaps which are very frequent at the ends of sequences.

Once the alignment is computed, all the differences from the reference sequence are collected in the form of variants (substitutions, insertions or deletions). Using the SnpEff tool (20), we annotate each variant and predict its impact on the codifying regions; indeed, a variant may, for example, be irrelevant (e.g when the mutated codon codifies for the same amino acid of the original codon), produce small changes or be deleterious. Based on the alignment result, the sub-sequences corresponding to the reference annotations are identified within the input sequence.
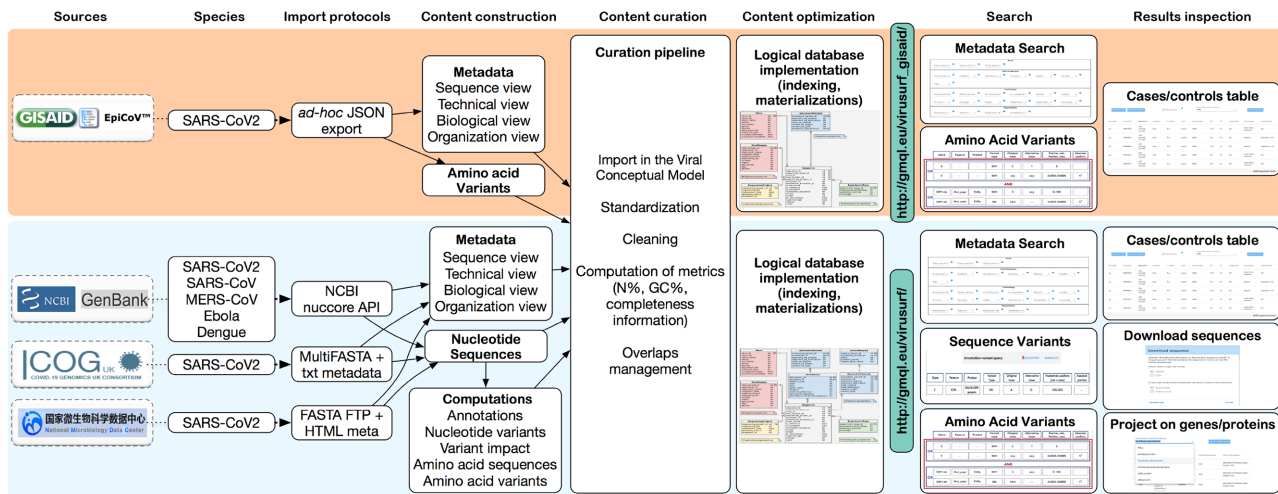
Figure 2. General pipeline of the ViruSurf platform. For given sources and species, we use download procedures to construct content, perform data curation, and load the content into two distinct databases, for GISAID and for the other sources, which are schema-compatible (the former is a subset of the latter). We then provide two Web-based interfaces supporting search and result inspection.

Coding regions are then translated into their equivalent amino acid sequences; the translation takes into consideration annotated ribosomial frameshifts events (e.g. within the ORF1ab gene of SARS-CoV-2). When translation fails (e.g. because the nucleotide sequence retrieved from the alignment is empty or its length is not a multiple of 3), we ignore the amino acid product; failures are due to incompleteness and poor quality of the input sequence, further computation of amino acid variants would produce erroneous information. Instead, when an aligned codon contains any IUPAC character ambiguously representing a set of bases (https://genome.ucsc.edu/goldenPath/help/iupac.html), it is translated into the X (unknown) amino acid, which automatically becomes a variant. Note that queries selecting known amino acids are not affected; unknown amino acids are usually not of interest. Translated amino acid sequences are then aligned with the corresponding amino acid sequences (using NW), annotated with the reference, and amino acid variants are inferred.

Alignment, variant calling and variant impact algorithms are computationally expensive, so we decided to parallelize this part of the pipeline, taking advantage of Amazon Elastic Compute Cloud (Amazon EC2). We implemented a chunked and parametrized execution modality for distributing the analysis of the sequences associated with each virus to multiple machines so that the total execution time of the process can be divided by the number of available machines.

Data curation

Our first curation contribution is to provide a unique schema (VCM (1)) for different data sources. Metadata of each source use different terms, the mappings between VCM attribute names and those used at the original sources are in Supplementary Tables S2–S5. Specific value curation efforts have been dedicated to the location information, the collection and submission dates, the completion of virus and host taxonomy names/identifiers, the choice of the ap-
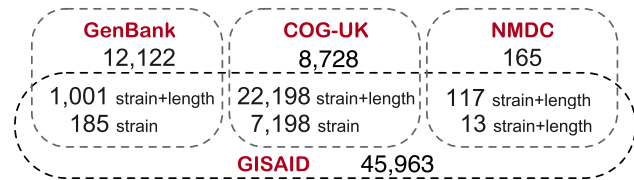
Figure 3. Counts of SARS-CoV-2 overlapping sequences from each source. Overlaps are computed by means of either the strain name, or both strain name and length.

propriate reference sequence (cross-checking with several research papers to ascertain that the typical reference sequence used for variant calling is defined) and the coverage of the sequencing assay. We also compute metrics regarding the percentage of G and C bases, of unknown bases and the information about sequence completeness.

Some sequences are deposited to multiple sources; we detect such redundancy by matching sequences based on their strain name or the pair of strain name and length. Overlaps among sources is illustrated in Figure 3. As all overlaps occur between GISAID and the other sources, we store the information about overlaps within the GISAID database and allow the possibility of performing 'GISAID only' queries, i.e. restricted to GISAID sequences that are not present in GenBank, COG-UK or NMDC.

RESULTS

Web interface

The web interface of ViruSurf is composed of four sections, numbered in Figure 4: (1) the menu bar, for accessing services, documentation and query utilities; (2) the search interface over metadata attributes; (3) the search interface over annotations and nucleotide/amino acid variants; (4) the result visualization section, showing resulting sequences with their metadata. Results produced by queries on the search interface (2) are updated to reflect each additional

Figure 4. Overview of ViruSurf interface. Part 1 (Top bar) allows to reset the previously chosen query or select predefined example queries. Queries are composed by using Part 2 (Metadata search) and Part 3 (Variants search). In our example, Part 2 includes three filters on Virus taxon name, Is complete and *N*%. Part 3 includes three panels. Panel 'A' is a query on amino acid variants, selecting sequences with RK and GR changes in gene N; Panel B' is a query on nucleotide variants, selecting sequences with a variant at position 28 881. Panels 'A' and 'B' are closed, they can be removed but not changed. Panel 'C' is another query on amino acid variants, currently open; it includes two filters selecting given positions of the Spike protein, and visualizes available values for the original amino acid involved in the change. Part 4 shows the Result Visualization. Resulting sequences already reflect the filters of Part 2 and the queries of the closed panels 'A' and 'B' of Part 3, applied in conjunction. Results can be downloaded, in csv or FASTA format; they can be selected as either cases (default) or controls (switch), and both the nucleotide and amino acid sequences can be projected on a specific protein; table columns can be omitted and reordered. On the bottom right corner, the number of sequences resulting from the search is visualized (in the Figure we show only three sequences out of 14 sequences found).

search conditions, and counts of matching sequences are dynamically displayed to help users in assessing if query results match their intents. The interface allows to choose multiple values for each attribute at the same time (these are considered in *disjunction*); it enables the interplay between the searches performed within parts (2) and (3), thereby allowing to build complex queries given as the logical conjunction—of arbitrary length—of filters set in parts (2) and (3).

Menu bar. The menu bar includes links to the GISAID-specific ViruSurf system, to the GenoSurf system, and pointers to the data curation detail page, to the wiki, to a video compilation, and to a pedagogical survey supporting

the user by documenting the aspects of search queries; on the top right of the interface we provide various 'Predefined queries'.

Metadata search. The Metadata search section is organized in four parts: *Virus* and *Host Organism* (from the *biological* dimension), *Technology* and *Organization* (from the respective dimensions). It includes attributes which are present in most of the sources, described by an information tab that is opened by clicking on blue circles; values can be selected using drop-down menus. At the side of each value we report the number of items in the repository with that value.

The user can compose desired queries by entering values from all the drop-down menus; the result is the set of sequences matching all the filters. Note that the special value N/D (Not Defined) indicates the null value, that can also be used for selecting items. 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 flag, thereby including in the result those sequences having the value set to N/D. Similarly, collection date and submission date have a calendar-like drop-down components, supporting a range of dates and the N/D flag.

Variant search. The Variant search section allows searching sequences based on their nucleotide variants (with their impact) and the amino acid variants. When the user selects 'ADD CONDITION ON AMINO ACIDS' or 'ADD CONDITION ON NUCLEOTIDES' buttons, a dedicated panel is opened, with a series of drop-down menus for building search conditions. A user can add multiple search conditions within the same panel; these are considered in disjunction. Once the panel is completed, it is registered; registered panels can be then deleted from a query, if needed. Variants selected in different panels are intended in conjuction. Heterogeneous variant searches (i.e. on amino acid and nucleotide ones) can only be combined in panels, thus in conjunction.

In the example shown in Figure 4 (which represents the construction of the 'Predefined query 8', from Pachetti et al. (21)) the user is choosing all SARS-CoV-2 sequences that are complete, have a maximum percentage of unknown bases of 0.5%, and have R to K or G to K amino acid changes in gene N and a nucleotide variant at position 28 881. The filters set up to this point have selected a set of 14 sequences (as indicated at the bottom right of the page) In the represented snapshot, a third variant panel is in the process of being compiled with an amino acid condition that could be added to the two existing ones by pressing 'APPLY'. This holds a filter on the spike protein and on the position of the variant on the protein (>1000); the filter on original amino acid allows to select P (3 sequences available) or APHG (1 sequence).

Result visualization and download. The result table describes the sequences resulting from the selections of the user. The columns of the table can be ordered/included/excluded from the visualization; the resulting table can be downloaded for further processing. Whenever the user either adds or removes a value in the

Metadata search, by clicking on a drop-down menu, the results table is updated; instead, it is updated only when a panel of the Variant search section is complete.

The 'Show control' switch allows to visualize the sequences of the control group, defined by those sequences selected by the Metadata search filters for which there exist some variant and the variant filters are not satisfied. This option, suggested to us by virologists, is the most sensible for describing the effects of variant analysis.

A user can select from a drop-down menu which subpart of a nucleotide or amino acid sequence should be visualized; the default returns a 'FULL' nucleotide sequence (leaving the amino acid field empty), but with this menu option it is possible to return in the result the specific segment of interest. The whole result table—as it is visualized, inclusive of selected metadata and nucleotide or amino acid sequences—can be downloaded for further analysis as a CSV. Alternatively, the user may download either full or selected sequences by using their accession ID, either as CSV or FASTA files.

GISAID-specific virusurf. ViruSurf presents a version that is specific for data imported from GISAID, as requested by a specific Data Agreement. This interface presents limited functionalities but is nevertheless powerful and allows for combining its results with the ViruSurf main interface. Notable differences are here summarized: (i) after selecting filters, a user must explicitly apply her search by pressing an execution button. (ii) Searches may be performed on the full dataset from GISAID or on the specific subset of sequences that are only present on GISAID (button 'Apply GISAID specific')—this may result particularly useful when the user wishes to compare or sum up results from the two interfaces (see Q3 in the following for an example). (iii) Both drop-down menus and the result table's columns hold the original GISAID attribute name, when available—when this differs from ViruSurf's, the second one is provided in second position inside parentheses.

Example queries

By means of complex search queries over our database it is possible to help virus research, according to the requirements provided by several domain experts; this is not currently supported by existing systems, which typically offer very nice visual interfaces reporting results of data analysis but limited search capabilities. We cite some examples inspired by recent research works.

Q1. To support SARS-CoV-2 vaccine design efforts, it is useful to track antigenic diversity. Typically, pathogen genetic diversity is categorized into distinct *clades* (i.e. a monophyletic group on a phylogenetic tree). In Gudbjartsson *et al.* (22), specific sequence variants are used to define clades/haplogroups (e.g. the 'A3 group' is characterized by the 11083 and 29742 nucleotides G mutated to T, by the 1397 nucleotide G mutated to A, and by the 28 688 nucleotide T mutated to C). ViruSurf supports all the information required to replicate the definition of SARS-CoV-2 clades proposed in the study.

Q2. In SARS-CoV-2, the G-T transversion at 26 144, which caused an amino acid change in ORF3 protein

(G251V), is investigated in Chaw *et al*. (23). The paper claims that this mutation showed up on 22 January 2020 and rapidly increased its frequency. We can use ViruSurf to find out that GenBank currently provides three complete sequences with such mutation collected before 22 January 2020, while GISAID provides other 13 sequences, non-overlapping with GenBank ones.

Q3. A study from Scripps Research, Florida, found that the mutation D614G stabilized the SARS-CoV-2 virus's spike proteins, which emerge from the viral surface. As a result, the viruses with D614G seem to infect a cell more likely than viruses without that mutation; the G genotype was not present in February and was found with low frequency in March; instead, it increased rapidly from April onward. The scientific manuscript by Zhang *et al.* (24), cited by mass media (https://www.nytimes.com/2020/06/12/science/coronavirus-mutation-genetics-spike.html), has not been peer reviewed yet, but others on the same matter are (25)).

ViruSurf can be used to illustrate this trend. Let us consider two queries on complete sequences. Sequences with the D614G mutation collected before 30 March are 6592, against 4664 without the mutation; sequences with the D614G mutation collected after 1 April are 23 649, against 3331 without the mutation. In both queries, case/control checks are obtained by using the 'Show control' switch, which retrieves—for the population specified by metadata filters—sequences that either have or do not have the chosen variants.

The same queries can be repeated on the GISAID-specific version of ViruSurf. Sequences with the D614G mutation collected before 30 March are 15 034, against 8821 without the mutation; sequences with the D614G mutation collected after 1 April are 18 421, against 3369 without the mutation. By summing up the query results from the two non-overlapping databases, we obtain that the sequences with the D614G mutation are 61.6% of those collected before 30 March and 86.3% of those collected after 1April.

Comparison. Note that queries Q1-Q3 could not be performed directly on the native sources. While these provide fairly advanced filters concerning metadata describing sequences, they do not provide in-house computed variants (neither on nucleotides nor on amino acids). Moreover, ViruSurf provides an integration between four sources, with duplicate elimination, thus it provides a good estimate of the sizes of results matching a given query.

DISCUSSION

ViruSurf provides a single point of access to curated and integrated data resources about several Virus species. Example queries show that ViruSurf is able to replicate research results and to monitor how such results are confirmed over time and within different segments of available viral sequences, in a simple and effective way. The relevance of ViruSurf as a tool for assisting the research community will progressively increase with the growth of available sequences and of the knowledge about viruses.

While today's efforts are concentrated on SARS-CoV-2, ViruSurf can similarly be useful for studying other virus species, such as other Coronavirus species, the Ebola Virus, the Dengue Virus and MERS-Cov, epidemics which are a current threat to mankind; ViruSurf will also enable faster responses to future threats that could arise from new viruses, informed by the knowledge extracted from existing virus sequences available worldwide.

In our future work, we plan to add epitopes (amino acid subsequences that can used for designing vaccines, with their lineage, host, evidence and type of response), retrieved from IEDB and stored within new ViruSurf tables, and then searched using suitable Web interface extensions. We also plan to design data analysis services that can provide sophisticated use of our big data collection.

REFERENCES

1. Bernasconi,A., Canakoglu,A., Pinoli,P. and Ceri,S. (2020) Empowering Virus Sequences Research through Conceptual Modeling. In: *39th InternationalConference on Conceptual Model, Nov. 2020*.
2. Bernasconi,A., Ceri,S., Campi,A. and Masseroli,M. (2017) Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data. In: Mayr,H.C., Guizzardi,G., Ma,H. and Pastor,O. (eds). *Conceptual Modeling*, Springer International Publishing, Cham, pp. 325–339.
3. Bernasconi,A., Canakoglu,A., Masseroli,M. and Ceri,S. (2020) META-BASE: a Novel Architecture for Large-Scale Genomic Metadata Integration. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, https://ieeexplore.ieee.org/document/9104916.
4. Canakoglu,A., Bernasconi,A., Colombo,A., Masseroli,M. and Ceri,S. (2019) GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database*, **2019**, baz132.
5. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D.

et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

6. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.

7. The COVID-19 Genomics UK (COG-UK) consortium (2020) An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe.*, **1**, E99E100.

8. Shu,Y. and McCauley,J. (2017) GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**, doi:10.2807/1560-7917.ES.2017.22.13.30494.

9. Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, **1**, 33–46.

10. Hatcher,E.L., Zhdanov,S.A., Bao,Y., Blinkova,O., Nawrocki,E.P., Ostapchuck,Y., Schäffer,A.A. and Brister,J.R. (2017) Virus Variation Resource–improved response to emergent viral outbreaks. *Nucleic Acids Res.*, **45**, D482–D490.

11. Fernandes,J.D., Hinrichs,A.S., Clawson,H., Navarro Gonzalez,J., Lee,B.T., Nassar,L.R., Raney,B.J., Rosenbloom,K.R., Nerli,S., Rao,A. *et al.* (2020) The UCSC SARS-CoV-2 Genome Browser. *Nat. Genet.*, doi:10.1101/2020.05.04.075945.

12. Zhao,W.-M., Song,S.-H., Chen,M.-L., Zou,D., Ma,L.-N., Ma,Y.-K., Li,R.-J., Hao,L.-L., Li,C.-P., Tian,D.-M. *et al.* (2020) The 2019 novel coronavirus resource. *Yi Chuan*, **42**, 212–221.

13. Wang,L., Chen,F., Guo,X., You,L., Yang,X., Yang,F., Yang,T., Gao,F., Hua,C., Ding,Y. *et al.* (2020) VirusDIP: Virus Data Integration Platform. bioRxiv doi: https://doi.org/10.1101/2020.06.08.139451, 09 June 2020, preprint: not peer reviewed.

14. Liu,B., Liu,K., Zhang,H., Zhang,L., Bian,Y. and Huang,L. (2020) CoV-Seq: SARS-CoV-2 Genome Analysis and Visualization. bioRxiv doi: https://doi.org/10.1101/2020.05.01.071050, 02 May 2020, preprint: not peer reviewed.

15. Shen,L., Maglinte,D., Ostrow,D., Pandey,U., Bootwalla,M., Ryutov,A., Govindarajan,A., Ruble,D., Han,J., Triche,T.J. *et al.* (2020) Children's Hospital Los Angeles COVID-19 Analysis Research Database (CARD)-A Resource for Rapid SARS-CoV-2 Genome Identification Using Interactive Online Phylogenetic Tools. bioRxiv

doi: https://doi.org/10.1101/2020.05.11.089763, 23 May 2020, preprint: not peer reviewed.

16. Singer,J., Gifford,R., Cotten,M. and Robertson,D. (2020) CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. doi:10.20944/preprints202006.0225.v1.

17. Bonifati,A., Cattaneo,F., Ceri,S., Fuggetta,A. and Paraboschi,S. (2001) Designing Data Marts for Data Warehouses. *ACM T. Softw. Eng. Meth.*, **10**, 452–483.

18. Sayers,E. (2009) The E-utilities In-Depth: Parameters, Syntax and More. https://www.ncbi.nlm.nih.gov/books/NBK25499/.

19. 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. *J. Mol. Biol.*, **48**, 443–453.

20. Cingolani,P., Platts,A., Coon,M., Nguyen,T., Wang,L., Land,S., Lu,X. and Ruden,D. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

21. Pachetti,M., Marini,B., Benedetti,F., Giudici,F., Mauro,E., Storici,P., Masciovecchio,C., Angeletti,S., Ciccozzi,M., Gallo,R.C. *et al.* (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.*, **18**, 179.

22. Gudbjartsson,D.F., Helgason,A., Jonsson,H., Magnusson,O.T., Melsted,P., Norddahl,G.L., Saemundsdottir,J., Sigurdsson,A., Sulem,P., Agustsdottir,A.B. *et al.* (2020) Spread of SARS-CoV-2 in the Icelandic Population. *New Engl. J. Med.***382**,23022315.

23. Chaw,S.-M., Tai,J.-H., Chen,S.-L., Hsieh,C.-H., Chang,S.-Y., Yeh,S.-H., Yang,W.-S., Chen,P.-J. and Wang,H.-Y. (2020) The origin and underlying driving forces of the SARS-CoV-2 outbreak. *J. Biomed. Sci.*, **27**, 73.

24. Zhang,L., Jackson,C.B., Mou,H., Ojha,A., Rangarajan,E.S., Izard,T., Farzan,M. and Choe,H. (2020) The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv doi: https://doi.org/10.1101/2020.06.12.148726, 12 June 2020, preprint: not peer reviewed.

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